Tesis para obtener el título de doctor en el programa de Tecnologías Multimedia

Reconocimiento de señales sismo-volcánicas mediante canales específicos basados en modelos ocultos de Markov

Guillermo Cortés Moreno

Directores

Mª del Carmen Benítez Ortúzar

Jesús M. Ibáñez Godoy

Noviembre de 2015

Instituto Andaluz de Geofísica

Universidad de Granada

Dpto. Teoría de la Señal, Telemática y Comunicaciones

Editor: Universidad de Granada. Tesis Doctorales
Autor: Guillermo Cortés Moreno
ISBN: 978-84-9125-449-2
URI: http://hdl.handle.net/10481/42050

La estadística propone mejores modelos a partir de la observación mientras la naturaleza se encarga de desacreditarlos contrastándolos con la realidad aún por descubrir. En este juego interminable todos ganamos...

El doctorando **D. Guillermo Cortés Moreno** y los directores de la tesis, **Dra. Mª del Carmen Benítez Ortúzar** y **Dr. Jesús M. Ibáñez Godoy**, garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección de los directores de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Granada, 13 de noviembre de 2015

Director/es de la Tesis: Doctorando:

Fdo.: **Mª del Carmen Benítez Ortúzar** Fdo.: **Guillermo Cortés Moreno**

Fdo.: **Jesús M. Ibáñez Godoy**

A nuestra Sarita, y a su mamá y a todos los que la han hecho posible y van a cuidar de ella!

Y a los que siempre nos quieren. A pesar de como somos.

A todas las cosas que nos hacen soñar al ir a la cama....

... y a todas las cosas que nos hacen querer levantarnos al día siguiente.

Índice general

Agradecimientos

Este trabajo ha sido parcialmente financiado por varios proyectos y grupos de investigación, a los que formalmente agradezco su apoyo económico e institucional:

- Grupo de investigación en geofísica y sismología RNM104 de la Junta de Andalucía
- Proyecto APASVO (*Algoritmos avanzados de procesamiento de la señal PAra reconocimiento y caracterización de las señales Sismo-VOlcánicas*), TEC2012-315511 del Grupo de investigación en Señales, Telemática y Comunicaciones (GSTC), TIC-123 de la Junta de Andalucía
- Proyectos españoles de investigación BES-2012-051822, CGL2012-31472/BTE y subprograma de ayudas FPI-MICINN
- Proyecto MED-SUV (*MEDiterranean SUpersite Volcanoes*) de la Unión Europea, que ha sido subvencionado por el 7º Programa de la UE, EC-FP7, para la investigación, desarrollo tecnológico y aplicación bajo la concesión nº 308665

Con el mismo carácter de formalidad sigo agradeciendo (y seguiré) la enorme paciencia y dedicación de mis directores de tesis, *Jesús Ibáñez* y *Carmen Benítez*, por darme la oportunidad de desarrollarme como investigador y como persona. También me gustaría dar las gracias institucionales al *Instituto Andaluz de Geofísica (IAG)* y al *Dpto. de Teoría de la Señal, Telemática y Comunicaciones (TSTC)*. Bellos lugares para trabajar y conocer gente interesante. Y emocionarse con ell@s. Y aprender. Mucho de tod@s. Lo que me recuerda otras dos personas con las que he tenido la suerte de ~~trabajar~~ aprender, *José Carlos Segura* y *Ángel de la Torre*. Y otras dos con las que me gustaría seguir aprendiendo, *Gerardo Alguacil* y *Antonio Rubio*.

Y en un tono más absolutamente informal... gracias *jefes,* de nuevo. Gracias *Jesús* por decirme no solo las cosas que hice bien, sino, y sobre todo, por mostrarme las que hago mal. Y por llevarme a comer a sitios tan bonitos en Canarias. Y por animarme en el CoV.6, y por ser tan cercano, que es desde donde se aprende más. Gracias *Carmen* por la paciencia, paciencia, paciencia... Ya sé que te preocupas por mí, y que soy demasiado cabezón, pero al final, las cosas no nos han ido tan mal ¿verdad?.

Recuerdo la época de las clases de doctorado y el dpto. TSTC. Me gustaba trabajar allí. Se aprendía mucho. Me acuerdo de la gente que conocí, y de las fiestas de disfraces de *Juanjo*. Fue una bonita época. Cuando *Ligdamis*, el *Chilango* y yo estábamos en el zulo del dpto. de Física. Y de aquella chica rara de la habitación de enfrente. Creo que era italiana. No saludaba. Le caí mal desde el principio. Ahora

la echo de menos. Me acuerdo de regresar por la noche después de cantar con *Raúl* y *Rolando*, y caerme al suelo luego. Esa acera se movía muy rápido.

Y me pregunto... ¿qué tiene que ver la *música* en todo esto? ¿eh *Juanjo*? ¿*Javi*? ¿*Ángel*? ¿*Antonio*? ¿*Luz*? ¿*José Carlos*?. Supongo que a todos nos gustan las señales, no solo estudiarlas, también vivirlas.

Y luego la época del IAG, y los viajes a Tenerife. Lo tranquilo que me sentía cuando *Justo* venía. Lo sereno que estaba siempre. Igual que *Beni*. Y todo lo que aprendí del geo-showman *David Calvo*. Que rápido iba. No solo conduciendo. Me acuerdo perfectamente como *Luz* se quitó su abrigo para dármelo a mí mientras yo me tiraba al suelo a terminar de estropear una Guralp. Y de las 20000 horas que *Isaac* se metía entre pecho y espalda conduciendo. Y de *Enrique*, de lo ilusionado que estaba al conducir el 4x4. El IAG perdió a uno grande cuando te fuiste. Y del otro *Enrique*, de sus fotos y de su felicidad contagiosa. Y del *Carmona*, preguntándome cómo me va. Y del *Carrión* y sus tiempos de paracaidista.

Es una alegría ir a trabajar al IAG. Desde que entras por la puerta y te encuentras las viñetas de *Paco Vidal*. Y luego encontrar a mis compis de despacho *Vane.1*, *J.Ángel*, *Antonio*. Y en el banco a *Caro*, *Evelyn* e *Iván*. Y por los pasillos a *Inma*, *Merche*, *J.Manuel*, *Pepe*, *Flor*, *Vane.2*, *Rosa* (¿cuando te vendrás a comer?), *Javi.M*, *Antonio*,... incluso encontraros a los que ahora no estáis *Araceli*, *Rafa*, *Cintia*, *Luciana*, *Ianire*. La gente además de lista es simpática. Ya no te dejo más ganar al Catán *Jose*. Ya no volveremos a verte al teatro *Antonio*. Ya no tardaré tanto en comer *Dani*. Ya no te preguntaré más cuestiones absurdas *Gerardo*. En fin... todos sabemos que lo volveré a hacer.

Especialmente recuerdo los *amigos* que me llevo. A mi *Codini*, a *Pablo*, *Juanjo*, *Ligdamis*, *Rolando*,... Y con los que vine. Los sempiternos perros verdes.

Me siento especialmente afortunado por tener siempre a mi gran *familia*, mis *padres*, *Álex*, *Pedrito* , mis *tit@s* y mis *prim@s*. Y la nueva, para siempre, mis 2 niñas, *Sarita* y *Mª Angelitas!*. Y mi *Bee* y todas sus amigas con las que tanto me he reído. Y mis *sobrinos guapos*! y sus *papás* y la *Chi* y sus *papás*. Siempre los padres. Y las madres. Sobre todo las madres.

Y recuerdo a aquellos *que no he nombrado aún*. Ya sabéis que la memoria no es lo mío.

Y viendo todo esto, me doy cuenta que he tenido mucha suerte de haberos conocido. Y que fueron tiempos bonitos. Ahora espero el futuro. A ver con quién nos encontramos. Hay mucho que aprender.

Gracias por estar ahí y allí conmigo!

Prólogo

Resumen de la tesis

La actividad volcánica en nuestro planeta genera un gran impacto económico y social. Actualmente la monitorización de volcanes se fundamenta principalmente en el análisis de la actividad sísmica de los eventos considerados precursores de erupciones. Un sistema automático que sea capaz de detectar y clasificar eventos sismo-volcánicos en tiempo real permitiría una gestión más eficaz al evaluar del riesgo volcánico sobre todo cuando previo a una erupción el incremento de la actividad es tal que compromete la fiabilidad de la clasificación supervisada llevada a cabo por los técnicos de los observatorios. Un análisis detallado de la situación es crucial a la hora de tomar decisiones que pueden ser críticas como la necesidad de evacuación de la población.

Los sistemas automáticos de reconocimiento de señales sismo-volcánicas (*Volcano-Seismic Recognition - VSR*) en una etapa de aprendizaje construyen modelos probabilísticos para cada tipo de evento o *clases* a partir del análisis de datos previamente clasificados por técnicos expertos. Dichos modelos permiten posteriormente una clasificación sobre registros continuos de forma automática y no supervisada. El funcionamiento en tiempo real de estos sistemas ha sido tímidamente explorado por la comunidad científica lo que se une al problema complejo del modelado dada la naturaleza y variabilidad de las señales sismo-volcánicas sometidas a solapamiento entre eventos, efectos de sitio, ruidos, etc. Tomando inspiración en los últimos avances en las áreas de inteligencia artificial, reconocimiento de patrones y aprendizaje automático, se abre un mundo de lineas de investigación muy interesantes que están atrayendo la atención de los geofísicos y los observatorios, no solo por la posibilidad de monitorizar el grado de actividad sísmica en tiempo real, sino también por la ventaja de contar con una herramienta robusta y fiable de clasificación automática no susceptible de sufrir errores inevitablemente asociados a la condición humana como la falta de un criterio unificado, el cansancio y la variabilidad en la toma de decisiones debido a factores subjetivos o psicológicos. Por ello, son cada vez más los observatorios vulcanológicos que incorporan sistemas expertos automatizados de monitarización y métodos de predicción de erupciones (Carniel et al., 2006; Ham et al., 2012; Boué et al., 2015), lo que explica el auge que los sistemas de reconocimiento automático de eventos sismo-volcánicos están teniendo en los últimos 10 años (Orozco-Alzate et al., 2012).

Complementariamente, la capacidad del cerebro humano de describir y analizar una

situación a distintos niveles conceptuales es un reto (o *el* reto) apasionante en el que se centra gran parte de los últimos trabajos de inteligencia artificial: el aprendizaje profundo (*deep learning*) o cómo enseñar a las máquinas a describir y aprender lo verdaderamente importante. Aspecto que también hay que tener en cuenta en los sistemas VSR: enseñar al sistema qué características son importantes para describir los eventos y cómo evaluar correctamente los resultados de clasificación.

Los Modelos Ocultos de Markov (*Hidden Markov Models - HMMs*), dada su naturaleza estructurada y su capacidad para modelar datos doblemente estocásticos en el espacio secuencial (el tiempo en nuestro caso) y en el espacio de descripción de los datos, se han convertido en una de las técnicas más utilizadas en el área VSR (Ohrnberger, 2001; Alasonati et al., 2006; Benítez et al., 2007; Ibáñez et al., 2009; Beyreuther et al., 2012). En esta tesis, proponemos una evolución de los sistemas VSR clásicos basados en HMMs a un sistema estructurado en paralelo (*Parallel System Architecture - PSA*) compuesto por distintos canales de reconocimiento cada uno de ellos especializado en un tipo de evento volcánico o *clases* concretas (Cortés et al., 2014). Esto permite el análisis por independiente de clases de eventos especialmente relevantes, así como el estudio de la mejor configuración y el mejor conjunto de características para describir cada tipo de canal (evento), contribuyendo así a incrementar la eficacia de reconocimiento y la capacidad de análisis así como la flexibilidad y funcionalidad del sistema. El objetivo último es la construcción de un sistema automático no supervisado de carácter general que sea fácilmente integrable en los centros de monitorización de volcanes activos.

Dicha universalidad y la demanda de un reconocimiento sobre flujos continuos de datos en tiempo cuasi-real obliga a crear modelos capaces de describir de forma general y unívoca los eventos que aún observados en distintos volcanes se asocien a la misma clase. La selección de características que describen los eventos de un mismo tipo, juega por tanto un papel clave en el diseño de los sistemas VSR que aspiran tener un funcionamiento no supervisado, influyendo además en la rapidez de ejecución, el coste computacional, la eficacia y la fiabilidad de los resultados de reconocimiento. En este escenario, esta tesis contribuye en una doble vertiente:

1. Se realiza un exhaustivo análisis las principales técnicas de descripción de datos, con especial énfasis en las características diseñadas para modelar eventos sismo-volcánicos. A partir de dicho estudio se proponen varias parametrizaciones de distinta naturaleza, construyendo un vector de características híbrido que consigue mejores resultados que otros esquemas homogéneos (Álvarez et al., 2009; Cortés et al., 2014).

2. Con el objetivo de modelar solo la información realmente relevante de los eventos se examinan distintas técnicas de reducción de dimensionalidad del vector de características. Proponemos el algoritmo DFS generalizado como una mejora al DFS (*Discriminative Feature Selection*) de Álvarez et al. (2011) que obtuvo unos resultados notables al ser aplicado sobre las señales VSR permitiendo una interpretación geofísica más directa de los eventos y un modelado

más eficaz.

Los resultados del sistema VSR-PSA, configurado como un conjunto de detectores específicos, obtienen el mejor promedio en la tasa de reconocimiento respecto a la opción clásica de sistemas VSR-SSA en serie. Asimismo, facilita la evaluación de los resultados gracias a la especialización de los canales para discriminar eventos bajo circunstancias ruidosas o en presencia de eventos solapados siendo una valiosa herramienta para el etiquetado semi-supervisado al ofrecer al técnico experto distintas opciones de clasificación incluyendo las tasas de fiabilidad de cada una de ellas. Respecto a la reducción de dimensionalidad, el algoritmo DFS generalizado se evalúa y certifica como la mejor opción entre varios métodos actuales (Cortés et al., 2015) en cuanto a la eficacia medida como la relación entre la tasa de reconocimiento y el coste computacional. La técnica DFS al ser una selección de características manteniendo el significado geofísico del vector de descripción facilitando la interpretación y el análisis posterior.

Tanto el sistema propuesto VSR-PSA de canales específicos en paralelo como la generalización del DFS proporcionan una herramienta fundamental para la consecución del objetivo marcado, la construcción de un sistema VSR no supervisado, abriendo una interesante linea de investigación para el futuro del reconocimiento automático y la monitarización de volcanes activos.

Estructura

Esta tesis se divide en 4 grandes bloques. La Parte I presenta un introducción al reconocimiento automático en el campo de las señales sismo-volcánicas (VSR). El Capítulo 1 introduce los fundamentos de la sismología volcánica y de la monitorización de volcanes activos. El Capítulo 2 se adentra en los sistemas de clasificación automática y su aplicación al reconocimiento de eventos sismo-volcánicos describiendo las particularidades del problema y repasando los principales métodos empleados. Se revisan los trabajos más populares en este área haciendo especial énfasis en sus ventajas e inconvenientes contextualizando los logros obtenidos hasta ahora y extrayendo los principales retos y problemas actuales que son necesarios superar para alcanzar el objetivo común: construir un sistema VSR eficaz y que funcione en tiempo real de manera no supervisada.

Como contribución a la solución, en la Parte II se desarrolla el marco teórico del sistema VSR-PSA con una arquitectura en paralelo (*Parallel System Architecture - PSA*) que constituirá nuestra propuesta. El Capítulo 3 estudia el origen y propiedades de los datos que usaremos en este trabajo así como la descripción que hacemos de ellos en vectores de características. Se construye el sistema de clasificación VSR-SSA (*Serial System Architecture - SSA*) basado en HMMs, el cual constituirá nuestro punto de referencia. La reducción de la dimensionalidad del vector de descripción, necesaria para mejorar la efectividad de los modelos, se aborda en el Capítulo 4

donde comparamos varios esquemas clásicos de extracción y selección de características con otros que desarrollamos específicamente para esta tesis. En el Capítulo 5 proponemos una evolución de los sistemas clásicos VSR-SSA a una arquitectura en paralelo VSR-PSA que se compone de canales especializados en el reconocimiento de un tipo concreto de eventos, la *clase propia* del canal, eligiendo las configuraciones y el conjunto de características que maximicen su eficacia para discriminar los eventos pertenecientes a dicha clase propia.

La aplicación práctica de los sistemas VSR desarrollados teóricamente se expone en la Parte III. En el Capítulo 6 se construye un sistema VSR-PSA sobre dos corpus de datos muy distintos, el volcán de la isla Decepción y el volcán de Fuego de Colima, seleccionándose las características que mejor describen la clase propia de cada canal mediante el algoritmo discriminativo DFS generalizado detallado en el Capítulo 4 y analizándose en profundidad los resultados obtenidos. Las conclusiones y lineas de investigación futuras son esquematizadas en la Parte IV. Complementariamente, la Parte V presenta algunas cuestiones prácticas que nos aparecen en la implementación experimental de los sistemas VSR.

Parte I.

INTRODUCCIÓN AL RECONOCIMIENTO AUTOMÁTICO DE EVENTOS SISMO-VOLCÁNICOS (VSR)

1. Señales sismo-volcánicas

En este capítulo haremos una breve introducción a la sismología volcánica haciendo hincapié en los conceptos que nos serán más útiles para diseñar un sistema automático de reconocimiento de señales sismo-volcánicas. Empezaremos en la Sección 1.1 por caracterizar el proceso eruptivo que tiene lugar en volcanes y el tipo de eventos asociados a cada etapa. Prestaremos especial atención al describir dichos eventos, tanto los que se dan internamente como los que se manifiestan en el exterior debido a la erupción. La clasificación automática de estas señales será el objetivo de nuestro sistema de reconocimiento propuesto en la Parte II de esta tesis.

En la Sección 1.2 nos enfrentaremos al problema que sobre la población conlleva el riesgo de convivir junto a volcanes activos. Estudiaremos los métodos que la sismología tiene para monitorizar el estado dentro del ciclo eruptivo en el que se encuentra cada volcán y el papel que juegan los sistemas de reconocimiento automático.

1.1. Sismología volcánica

Mediante el estudio de las sismos registrados en los volcanes la sismología volcánica pretende describir las fuentes y la dinámica de los procesos físicos del magma y fluidos hidrotermales que generan estos terremotos para caracterizar su evolución y localización. Toda esta información constituye una herramienta fundamental de la geofísica para comprender la estructura interna, analizar los episodios eruptivos y evaluar el riesgo asociado a un volcán (Chouet, 1996a). Omori (1911) y Sassa (1936) empezaron a estudiar los eventos que acompañaban a las erupciones sentando las bases de la sismología volcánica. Con la llegada de los primeros sensores sísmicos portables en los 60's y 70's se empezaron a desarrollar las primeras teorías relacionando la actividad sísmica con los procesos físicos que la genera.

El desarrollo de las técnicas de tomografía por Aki et al. (1977) invirtiendo los tiempos de viaje del frente de ondas de los sismos representa un antes y un después en la historia de la sismología volcánica que al permitir detectar cámaras de fluido. La aplicación de sísmica activa y pasiva para hacer un mapa de la estructura interna del volcán Kilauea por Aki y su equipo fue el punto de partida para asociar la información sísmica a un modelado analítico de la dinámica magmática (Aki et al., 1978; Helz, 1993), llevando a un grupo de científicos pioneros a relacionar los eventos de baja frecuencia registrados en los volcanes con el modelado de su fuente mediante el estudio de la generación y resonancia de ondas acústicas (Chouet, 1981; Fehler and Chouet, 1982). Los sismómetros de banda ancha y las antenas sísmicas en los 80's y 90's suponen una gran evolución hacia la sismología volcánica actual posibilitando las tomografías en 3D de alta resolución (García-Yeguas et al., 2012; Prudencio et al., 2015) y las antenas sísmicas (Almendros et al., 1999) que permiten localizar sismos de baja frecuencia relacionados con las propiedades dinámicas, acústicas y mecánicas del magma (Chouet, 2003).

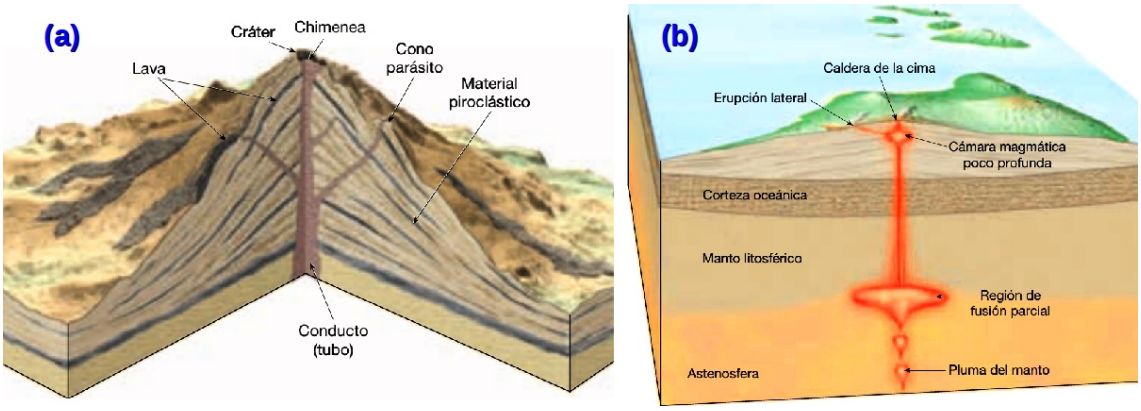


Figura 1.1.1.: Estructura de un volcán. **(a)** Esquema interno de un estratovolcán **(b)** Ascensión de magma directamente desde la astenosfera en un volcán escudo**.** Figuras originales en Tarbuck et al., 2012.

1.1.1. El proceso eruptivo

Al igual que ocurre con las fuerzas acumuladas entre placas tectónicas que se liberan en forma de terremotos, como observamos en la Figura 1.1.2 los volcanes activos están sometidos a un ciclo eruptivo de una duración variable (desde menos de una decena a miles de años) que comienza por el ascenso de material magmático desde la astenosfera hacia la corteza. La energía se va transmitiendo a las capas más superficiales, llegando a interaccionar con los sistemas hidrotermales. La presión de los fluidos va rompiendo las rocas produciendo grietas y cavidades que va rellenando. Finalmente, cuando la energía acumulada es suficiente se libera en un corto periodo de tiempo provocando la salida del magma, explosiones con liberación de gases y fluidos y terremotos en la etapa propiamente denominada como erupción (McNutt, 1996, 2005).

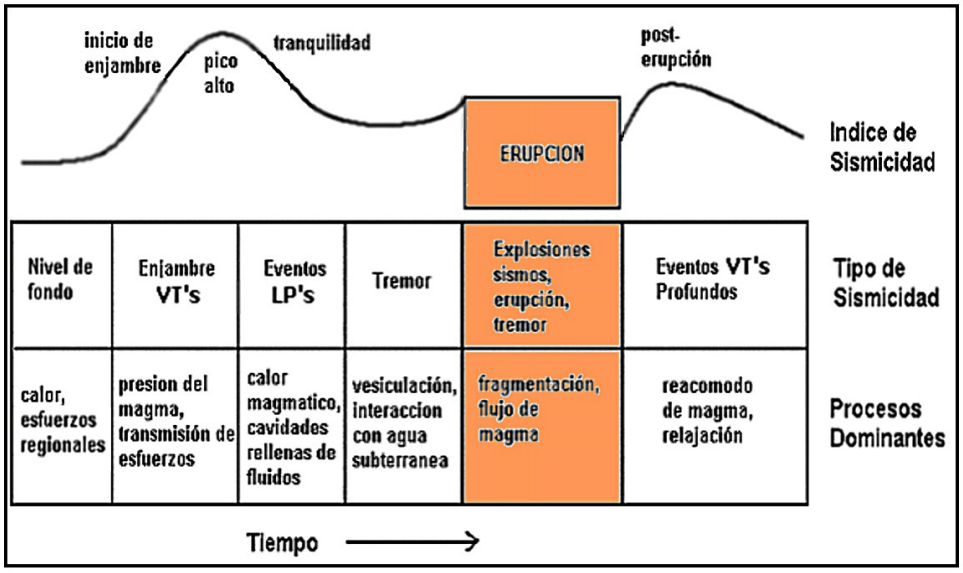


Figura 1.1.2.: Ciclo eruptivo en volcanes activos. Esquema general de la relación existente entre el tipo de eventos sísmicos detectados (*VTs*, *LPs*, *tremores*,...), los procesos de las fuentes que los genera y la etapa correspondiente en el ciclo eruptivo. Figura original de McNutt (1996).

El riesgo asociado a una erupción se relaciona fundamentalmente con la viscosidad y composición del magma y la capacidad que tenga el gas disuelto en él para liberarse. Básicamente podemos distinguir entre:

- *Erupciones explosivas con magma alta viscosidad.* La salida suele estar taponada por lava que se ha enfriado y solidificado en anteriores episodios aumentando la presión de los gases volátiles en el conducto que al liberarse lo hacen con una explosión violenta fracturando el edificio volcánico. Este efecto se multiplica si el magma ha interaccionado con sistemas hidrotermales, incorporando vapor de agua a muy alta presión.

- *Erupciones efusivas de magma fluido*, por lo que el gas escapa fácilmente de su interior. En el caso de que la lava se enfríe y tapone parcialmente la salida pueden ocurrir pequeñas explosiones de baja energía que fragmenten progresivamente el tapón, sin grandes riesgos en las áreas cercanas. Un magma fluido es la consecuencia que las coladas de lava discurran a poca velocidad.

Según este criterio y recurriendo a los episodios históricos en el vulcanismo, las erupciones se clasifican en orden creciente de explosividad como hawaianas, estrombolianas o mixtas, vulcanianas, plineanas o vesubianas y peleanas.

1.1.2. Señales sísmicas registradas en los volcanes

Los eventos registrados en los volcanes pueden ser clasificados desde 2 perspectivas principales:

1. *La representación temporal* en el sismograma y espectrograma. Tiene como inconveniente la enorme variabilidad que puede tener un mismo tipo de evento registrado en un volcán u otro, o, incluso, en diferentes estaciones de un mismo volcán, por lo que es susceptible de aplicar criterios pocos generales y subjetivos a la hora de interpretar la información dada por el sismograma
2. *Los mecanismos de fuente* que genera las señales. Basado en un sólido criterio físico, requiere una información (localización, magnitud, mapas de velocidad,...) que no siempre está disponible.

El criterio más aceptado es diferenciar los eventos en función de su forma de onda y propiedades espectrales (intervalo $[f_L, f_H]$ de concentración de energía, envolvente, evolución temporal) que presentan a partir del sismograma (Chouet, 1996b; McNutt, 1996). Sin embargo, debe ser tenido en cuenta que en ambos dominios, en la medida de lo posible, la representación del evento debe distinguirse entre la parte debida a su generación en la fuente y la que se ve afectada por los efectos de sitio y los efectos de propagación en el medio (Sección 1.1.2). En este trabajo dividiremos las señales registradas en los volcanes en 4 tipos atendiendo a su origen: tectónico, volcánico interno, volcánicas externas observables a simple vista desde fuera del volcán y ruidos.

Efectos de sitio y de propagación. Es necesario hacer una consideración de extrema importancia antes de clasificar las señales, sobre todo, si dicha clasificación se realiza solamente teniendo en cuenta la información a partir del sismograma de una única estación: *la forma del sismograma de un evento varía enormemente en función del emplazamiento del sensor*: un mismo evento puede parecer otro completamente distinto si se registra en otra estación (McNutt, 2005). Esto es debido a los efectos de sitio y propagación que son especialmente relevantes en los volcanes dada su compleja estructura interna a la que los volcanes están localizados en zonas tectónicas

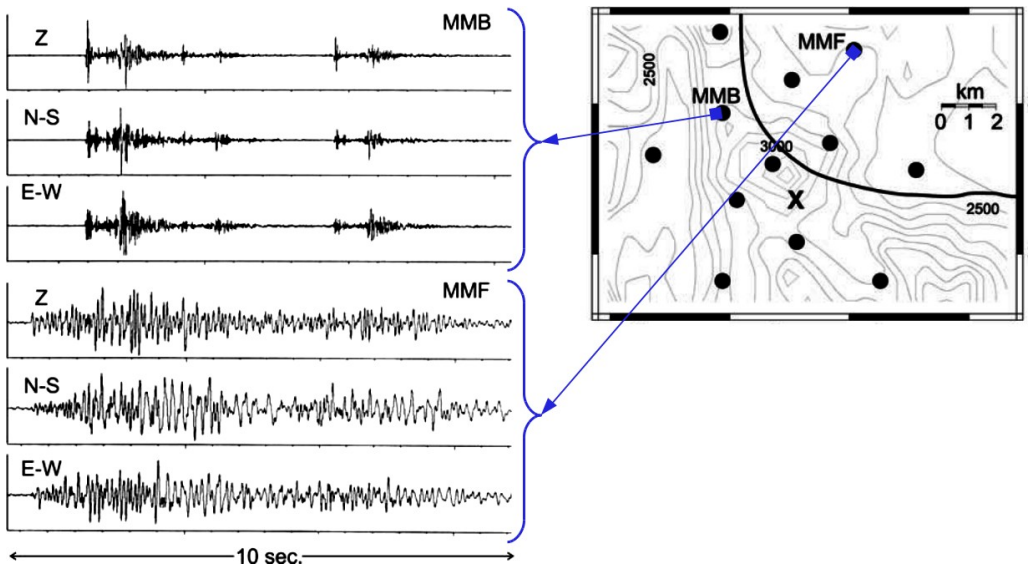


Figura 1.1.3.: Influencia de los efectos de sitio y propagación. Se constata como las mismas señales registradas en distintas estaciones (MMB y MMF) pueden ser interpretadas y clasificadas de forma distinta: 3 pequeños terremotos con llegadas P y S distinguibles en la estación MMB pierden energía en las altas frecuencias apareciendo como un evento de baja frecuencia en la estación MMF. Figuras de McNutt (2005).

ya de por sí activas. Los efectos de propagación (anisotropía, distribución de esfuerzos, permeabilidad,...) y de sitio (amplificación y atenuación, debido a la estructura geológica local) se solapan con los de dispersión, radiación, reflexión y refracción de las ondas sísmicas de origen exclusivamente tectónico, resultando en un aumento de la distorsión registrada en las estaciones volcánicas. Afectan especialmente a la banda de frecuencias medias y altas [5, 25] Hz.

La Figura 1.1.3 ilustra la importancia de los efectos de sitio y propagación en zonas volcánicas y la conveniencia de observar simultáneamente más de una estación a la hora de clasificar señales. Mora et al. (2001) estudiaron los efectos de sitio en el volcán Arenal mediante un array, evidenciando que, incluso para sensores separados por menos de 500 m., la influencia de sedimentos superficiales (tefra) puede ser muy alta en fenómenos de amplificación para la banda espectral entre [2, 4] Hz.

1.1.2.1. Eventos producidos estrictamente por la sismicidad volcánica

- **Terremotos volcano-tectónicos (VT)**. Son sismos tectónicos producidos dentro del marco local al volcán, debido a tensiones en el edificio volcánico. Se asocian a inyecciones de fluidos que causan una acumulación de energía y variaciones de presión capaces de reactivar fallas locales o generar fracturas y deformaciones en un medio rocoso frágil y que pueden solaparse con la actividad tectónica local. En general tienen una duración corta (en torno al minuto)

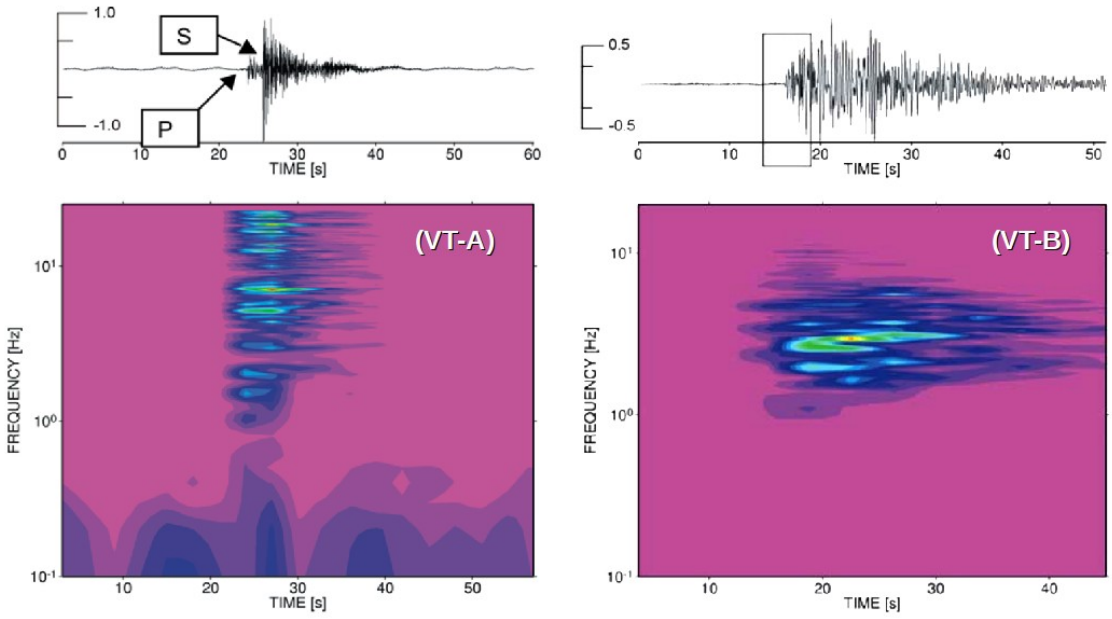


Figura 1.1.4.: Terremotos volcano-tectónicos. Se representan el sismograma y espectrograma de dos subclases de sismos volcano-tectónicos: profundos ($VT-A$) y superficiales ($VT-B$). Figuras de Wassermann (2012).

con un inicio rico en altas frecuencias seguido de una coda decae exponencialmente respecto al tiempo. Su energía espectral logarítmica representada en función de la frecuencia se mantiene constante hasta llegar a una frecuencia de corte f_c, que depende de la magnitud del sismo, a partir de la cual decae linealmente. Solo se diferencian de los sismos tectónicos locales y regionales en que los VTs se suelen presentar en *enjambres;* conjunto de sismos del mismo tipo parecidos en intensidad y duración que ocurren intermitentemente con un intervalo corto de tiempo entre eventos. Se subdividen en 2 tipos (Figura 1.1.4):

- *VTs profundos (**VT-A**)*: localizados a una profundidad p superior a 2 km., tienen un alto contenido espectral, $f_H > 10$ Hz (por lo que también se denominan eventos de alta frecuencia o *High Frecuency - HF*), y llegadas del frente de ondas P y S diferenciables y una alta impulsividad. Se originan por fuerzas de cizalla que causan un deslizamiento sobre el plano de falla (mecanismos de *doble par o* par de fuerzas acopladas). La llegada impulsiva y el alto contenido en frecuencias rápidamente se dispersan al llegar a regiones con baja atenuación y alta dispersión.
- *VTs superficiales (**VT-B**)*: con un profundidad entre $[1, 2]$ km., presentan una llegada emergente (poco impulsiva) del frente P con lo que no siempre es fácil determinar la llegada de las ondas S. Concentran la energía espectral en la banda de $[1, 10]$ Hz y tienen una frecuencia dominante

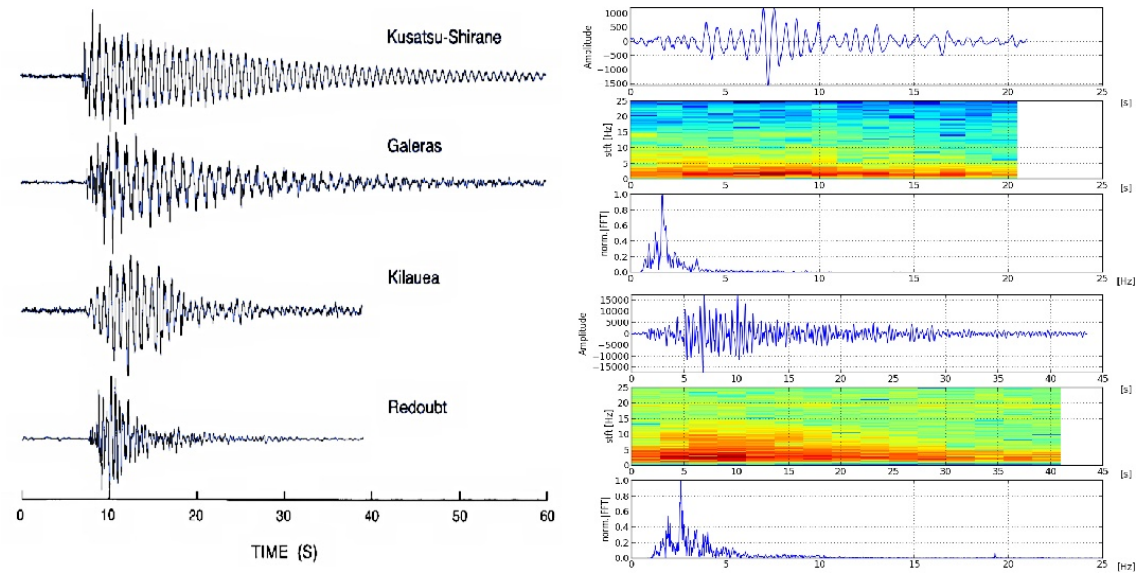


Figura 1.1.5.: Terremotos de largo periodo (*LP*). *Izq*: sismos LP registrados en distintos volcanes (Chouet, 2003). *Der.*: Se representan el sismograma, espectrograma y magnitud espectral de un LP registrado en el volcán de Decepción (*sup*) y de otro en el volcán de Colima (*inf*).

$f_0 \approx 5$Hz. Al tener el hipocentro muy superficial el contenido de las altas frecuencias se dispersa muy rápidamente pero la energía del sismo tarda más en distribuirse por lo que la coda es más larga.

- **Eventos de largo periodo (*LP*)** o baja frecuencia (*Low Frequency* - ***LF***). Señales muy superficiales ($d \lesssim 2$ km.) con una envolvente de amplitud en forma de huso con una duración parecida a los VTs y de contenido espectral en un intervalo de bajas frecuencias [0,1, 10] Hz. Se presentan en trenes de ondas donde es difícil distinguir las llegadas S . Presentan un carácter monocromático centrado en una frecuencia dominante aunque a veces tienen más de una frecuencia resonante que incluso genera armónicos observables. Suelen tener una llegada emergente con componentes espectrales más dispersas ($f_H \lesssim 10$ Hz.) pero de menos energía comparada con la coda. Los cambios de amplitud en el sismograma no implican necesariamente una variación en el espectrograma, lo que demuestra que la frecuencia es independiente de la cantidad de energía liberada en la fuente.

  En la Figura 1.1.5 se representan diferentes LPs registrados en varios volcanes. Podemos observar la variabilidad (tanto en el dominio temporal como en el espectral) que caracteriza a la mayoría de los eventos sismo-volcánicos. Su modelo de fuente se ha asociado a cavidades abiertas que resuenan cuando el magma asciende (Chouet, 1996b) y a una respuesta a variaciones transitorias de presión que resuenan debido a la interacción de la mezcla poco viscosa de fluidos y gases con el magma (Seidl et al., 1981), por lo que varios autores

los correlacionan con una sismicidad precursora de erupciones, tanto como en su forma aislada (Chouet, 1996b) como cuando se presentan con una frecuencia dominante de muy largo periodo (*Very Large Period* - ***VLP***, Arciniega-Ceballos et al., 1999) o cuando están agrupados en enjambres sísmicos al igual que los VTs (Chouet et al., 1994).

- **Tremor volcánico (*T*)** o ruido volcánico. Señal principalmente de baja frecuencia ([0,5, 10]Hz., Konstantinou and Schlindwein, 2003) con una alta variabilidad en su duración (desde minutos a días o semanas) que se caracteriza por tener un sismograma y espectrograma casi constantes respecto al tiempo, por lo que también se denomina *ruido volcánico*. De hecho, apenas varía sus características dentro de un mismo episodio eruptivo (Almendros, 1999). No hay una teoría comúnmente aceptado en cuanto a su fuente, sugiriendo un origen debido a desgasificaciones, resonancia de cavidades y variaciones de presión y temperatura en sistemas hidrotermales. Otros autores incluyen los tremores en el grupo de eventos de baja frecuencia (*LF*) a los que también pertenecen los LP y VLP asumiendo el mismo modelo de fuente pero sometida a una excitación continua. Almendros et al. (1997) propone que se debe a un solapamiento de eventos más simples. Los tremores han sido estudiados como precursores (Barberi et al., 1992; Langer et al., 2011). Se distinguen tipos principales de tremor (Arámbula, 2011) que podemos ver representados en la figura Figura 1.1.6 junto a una secuencia de pulsos de baja frecuencia o pulsos LP:

  - *Tremor armónico* o *resonante* (***TR***): trenes de ondas muy monocromáticas que se generan por una (o varias) frecuencias dominantes (*resonantes*) en el intervalo [0,1, 5] Hz. y en menor grado por sus respectivos armónicos. El número de frecuencias resonantes y sus armónicos pueden variar dentro de un mismo evento, fenómeno que Sturton and Neuberg (2003) explica mediante un proceso de compresión y descompresión de un tubo resonante con gas y fluidos. Leet (1988) proponen su origen a sistemas hidrotermales que fluctúan debido a variaciones en la presión y temperatura. Se considera que el TR es el resultado de la superposición de eventos LP.
  - *Tremor espasmódico* (***TS***): menos monocromático que el TR, su energía espectral se encuentra de forma más o menos dispersa en la banda de [1, 10] Hz percibiéndose una única frecuencia dominante. Se ha asociado al encadenamiento de sismos VTs, lo que explica su contenido en bandas intermedias del espectro y su relativa variación de amplitud y frecuencia respecto a otros tremores.
  - *Tremor pulsante* (***TP***): se considera un evento mixto formado por un TS sobre el que se solapan grupos de ondas que forman *pulsos* temporales secuencialmente tanto en la forma de onda como en el espectrograma. Los pulsos están espaciados por intervalos variables de unos 10 s. y su

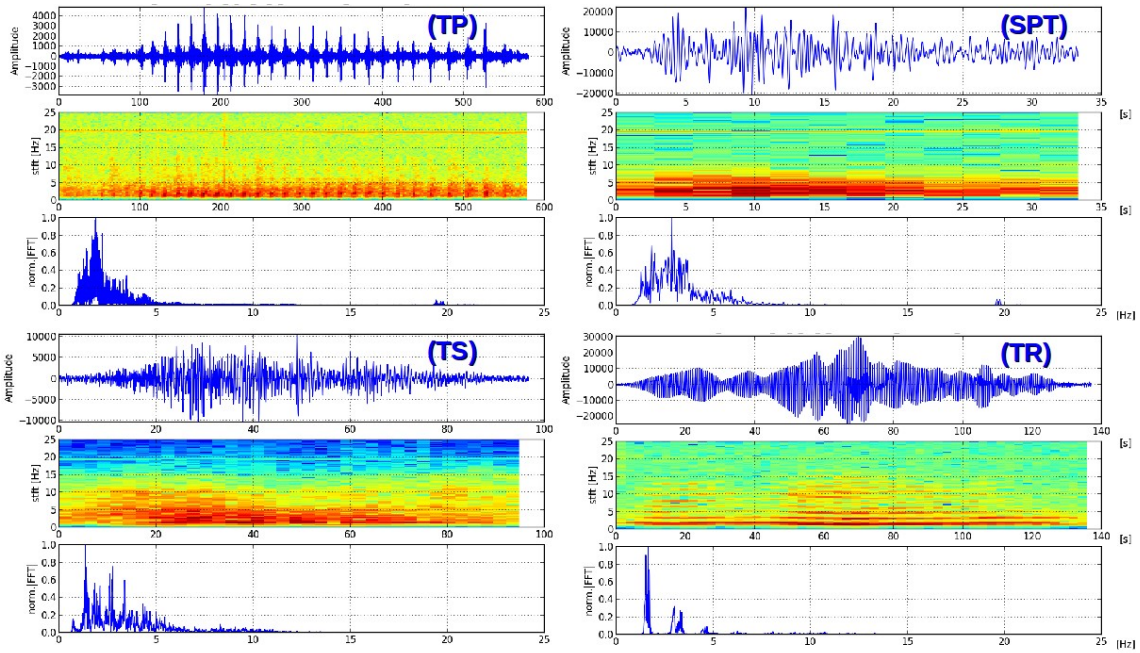


Figura 1.1.6.: Tremores en el volcán de Colima. Tremor pulsante (TP), pulsos de baja frecuencia (SPT), tremor espasmódico (TS) y tremor armónico o resonante (TR).

espectro parece centrarse en una frecuencia dominante.

Nótese no todos los autores están de acuerdo en esta subclasificación de tremores. En concreto, en el modelo de fuente asociada para cada tipo. Wassermann (2012) distingue entre un *tremores viscosos* relacionado con sistemas resonantes en volcanes con lava viscosa en el que agrupa los TR y TPs con pulsos debidos a eventos híbridos y multifase (Figura 1.1.2.1) y un *tremores poco viscosos* típico de volcanes basálticos con baja viscosidad en los que observa tremores eruptivos que cambian su forma según hayan columnas de ceniza o efusiones de vapor, tremores espasmódicos sobre los que se superponen explosiones y otros tipos de tremores rítmicos (derivados de los TPs). En la Figura 1.1.7 encontramos un ejemplo típico del grado de interrelación existente entre estas señales que llevan a otros trabajos a incluirlas bajo un único grupo (eventos LF): un enjambre de LPs que progresivamente va reduciendo progresivamente la separación temporal entre ellos hasta convertirse en un tremor pulsante (TP) y posteriormente en un TS. Al cabo de unos minutos los pulsos se van separando temporalmente para dar paso de nuevo a un TP, luego a un TS y finalmente a un enjambre de LPs.

- **Eventos híbridos (*HY*)** y **eventos multifase** (*Multi-Phase* - ***MP***). Son señales mixtas que engloban propiedades de eventos VTs y señales de baja frecuencia (LF): llegada impulsiva de alta energía distribuida en una amplia

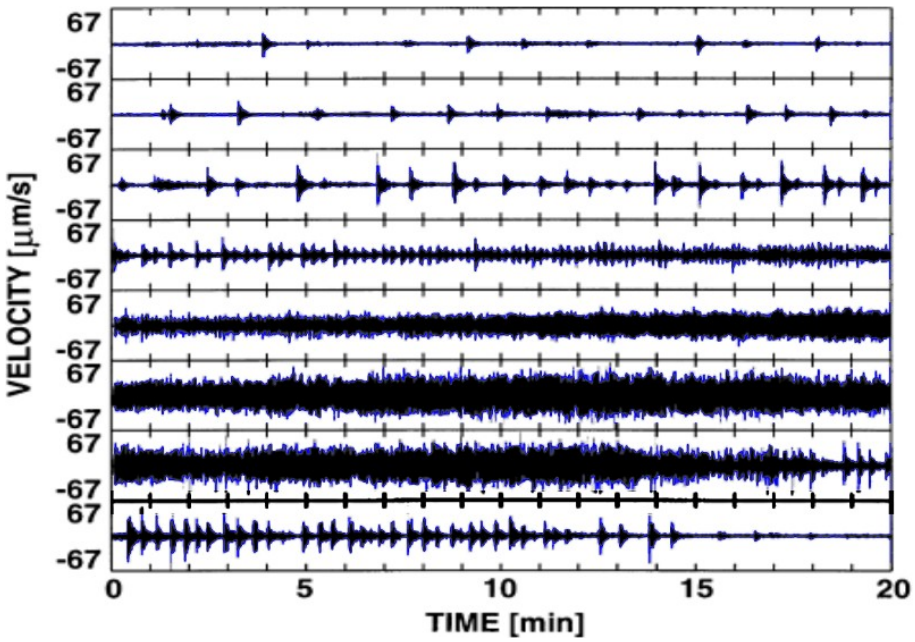


Figura 1.1.7.: Relación entre LPs y tremores. Un enjambre de LPs incrementa su tasa de ocurrencia hasta convertirse en un tremor pulsante y posteriormente en un tremor espasmódico hasta que vuelven a separase los pulsos para de nuevo presentarse en forma de secuencia de LPs . Figura original de Neuberg and O'Gorman (2002).

banda del espectro ([1, 25] Hz.) que se asocia a una fractura en un medio frágil y una coda de paquetes de ondas poco dispersas en torno a una frecuencia dominante en la banda inferior como las de los LPs o tremores, interpretada como la resonancia producida en una fractura al interaccionar con fluidos (Lahr et al., 1994). Nótese que la fractura generada en la llegada no tiene que ser necesariamente la que resuena: un microterremoto podría disparar un evento LP en una grieta cercana (Wassermann, 2012). A veces se presentan en enjambres superficiales que se han asociado a una actividad muy superficial relacionada con el crecimiento del domo (Miller et al., 1998).

La Figura 1.1.8 representa varios ejemplos de eventos mixtos evidenciando sus similaridades en la forma de onda. En ella observamos el parecido entre señales HYs y VT-B y como los eventos multifase (***MP***) comparten características espectrales con los HYs aunque con una banda espectral más amplia en su coda (Figura 1.1.8). Los MP responden a múltiples llegadas de paquetes de ondas que acaban superponiéndose. Al igual que los enjambres de HYs, las señales multifase se asocian a sismos muy superficiales que ocurren en el proceso de crecimiento del domo formado por lava viscosa al mismo tiempo que comparten con los LPs su modelo de fuente (Wassermann, 2012).

- **Explosiones (*EXP*)** y **eventos de muy bajo periodo (*VLP, ULP*)**. Liberaciones energéticas de energía que suceden en un intervalo corto de tiempo

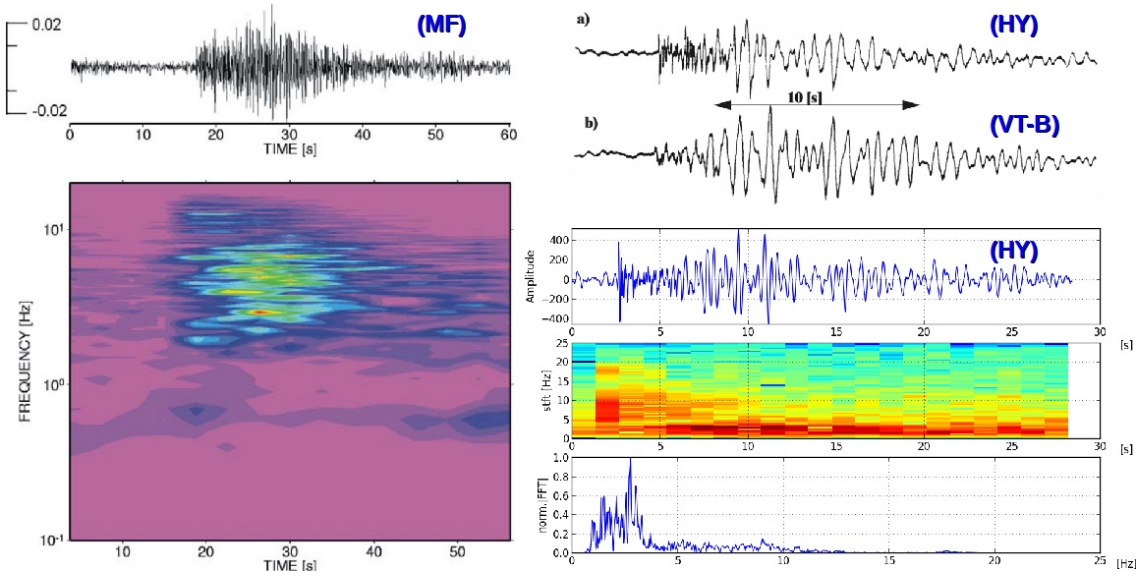


Figura 1.1.8.: Terremotos híbridos. *Izq*: evento multifase (MF). *Der-sup*: Comparación entre un evento híbrido (HY) y volcano-tectónico superficial ($VT-B$). *Der-inf*: sismograma, espectrograma y magnitud espectral de una señal híbrida registrada en la isla Decepción. Figuras de Wassermann (2012).

(del orden de 30 s.) debido a variaciones rápidas de presión y temperatura o de cambio de fase de materiales que, tras explotar, salen despedidos al exterior. Es el evento que más se asocia con las erupciones, probablemente porque puede escucharse debido a la expansión rápida del gas en el conducto que sale a la superficie y suele generar otros fenómenos visibles tales como expulsión de material, columnas eruptivas y derrumbes. Se da principalmente en episodios eruptivos estrombolianos (burbujas de gas que ascienden y al llegar a la superficie explotan, Ripepe et al., 2001) y vulcanianos (domo de lava que se enfría y tapona la salida de gases, hasta que la presión de estos lo fragmentan violentamente). Su sismograma presenta cierta similaridad con eventos LF profundos (sin incluir la fase debida a las ondas audibles transmitidas por el aire) al tener forma de huso, normalmente poco impulsivas. Como observamos en la Figura 1.1.9 su espectrograma también recuerda a los eventos LP que tras un inicio debido al fenómeno de rápida descompresión del material explosivo, la energía tiende a agruparse en torno a una frecuencia dominante relativamente baja ($f_0 \in [1, 5]$ Hz.) asociada a la onda acústica registrada por los sismómetros (McNutt, 1986).

La Figura 1.1.9 muestra la relación entre las explosiones superficiales y eventos LF como VLPs y *Ultra Long Period* - ***ULP*** (con una $f_0 < 0{,}01$ Hz. y $d < 3$ km.) en volcanes con actividad estromboliana y freática y con lava de baja a media viscosidad (Kawakatsu et al., 2000; Rowe et al., 1998; Wassermann, 2012). Las señales VLP y ULP se asocian al movimiento de magma y gases en

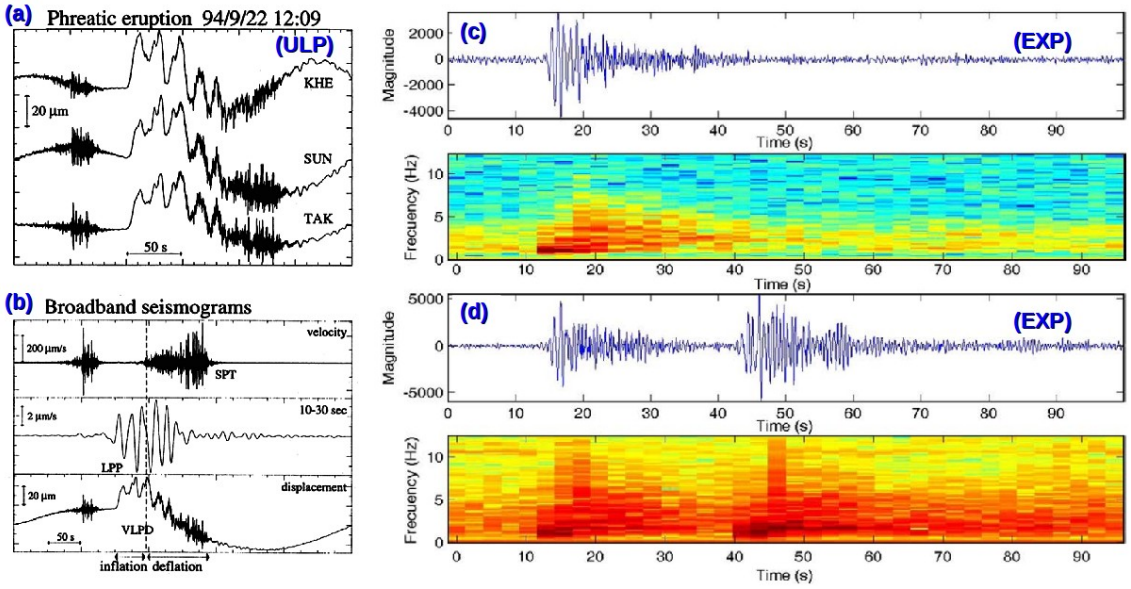


Figura 1.1.9.: Sismos debido a explosiones (*EXP*) y eventos de muy bajo periodo (*ULP*). (a) sismo ULP registrado en 3 estaciones solapado al **(b)** inicio de una EXP freática en el volcán de Aso: velocidad, BPF{velocidad} y desplazamiento registrados en una de las estaciones (Kawakatsu et al., 2000). **(c)** y **(d)** Sismograma y espectrograma de explosiones observadas en Stromboli (Ibáñez et al., 2009).

los conductos (Chouet, 1996a; Rowe et al., 1998).

1.1.2.2. Eventos externos del volcán

Incluye señales relacionadas con la actividad volcánica que pueden observarse a simple vista y que pueden tener un origen no exclusivamente sismo-volcánico, tales como flujos de lava, flujos piroclásticos, lahares, lluvia de cenizas, fumarolas, etc. (Figura 1.1.10). Este tipo de eventos también genera señales sísmicas que, por lo general, tienen un contenido espectral más alto y disperso que los eventos internos del volcán.

Los eventos que implican derrumbes de material desde la cresta del volcán; lahares, flujos piroclásticos y colapsos se caracterizan por tener una forma ahuesada (de puro) y un espectrograma en las bandas altas ($[5, 25]$ Hz.) no centrados en ninguna frecuencia dominante. Al igual que las explosiones, también generan ondas acústicas audibles. Junto a la lluvia de cenizas, su monitorización es especialmente importante si hay poblaciones cercanas al volcán. Los más importantes ordenados en orden decreciente de peligrosidad son:

- **Flujos piroclásticos (*PYF*)** o *nubes ardientes.* Las erupciones explosivas expulsan fragmentos de lava, rocas de varios tamaños y gases a más de 200 C°. Parte de la columna eruptiva asciende hasta varias decenas de km. mientras

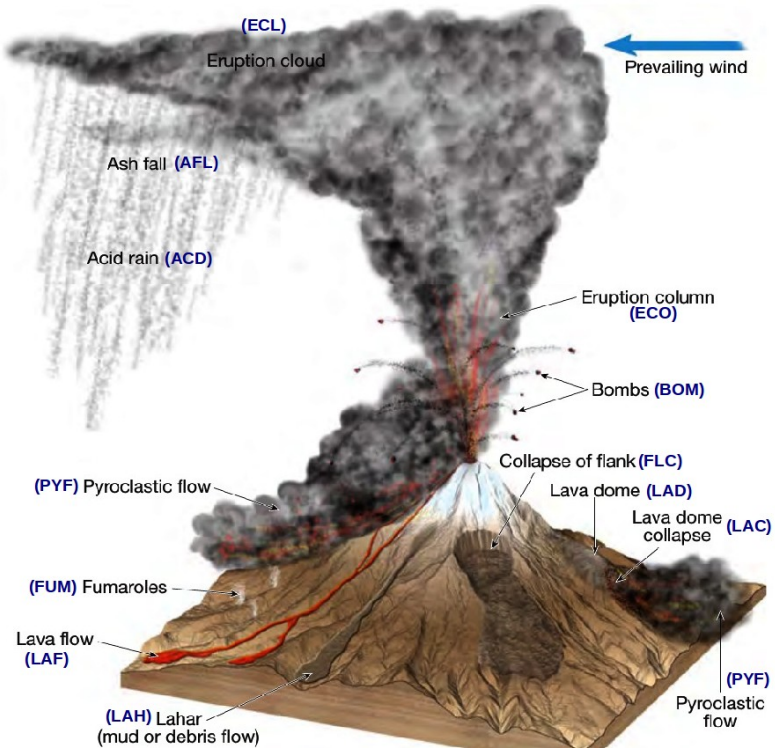


Figura 1.1.10.: Eventos eruptivos. Tipos de fenómenos derivados de una erupción volcánica: nube eruptiva (ECL), lluvia de cenizas (AFL), lluvia ácida (ACD), flujo piroclástico (PYF), fumarolas (FUM), flujos de lava (LAF), lahares (LAH), colapso del domo (LAC), domo (LAD), derrumbe lateral (FLC), *bombas* o fragmentos balísticos (BOM) y columna eruptiva (ECO). Figura original en Tarbuck et al. (2012).

que los materiales con más densidad por efecto de la gravedad bajan por las laderas del volcán alcanzando velocidades de cientos de km/h por lo que son un riesgo potencial para núcleos habitados próximos en un radio de hasta 50 km. Son especialmente peligrosas en el caso de explosiones laterales en volcanes peleanos.

- **Lahares (*LAH*)** (flujos de lodo). Material volcánico (rocas, cenizas y sedimentos) que al mezclarse con agua se convierte en una avalancha de barro volcánico que va arrasando con otros objetos que encuentra a su paso, incluso enterrando pueblos enteros. Si en el volcán hay nieve o glaciares el riesgo de lahares es especialmente alto en un episodio eruptivo. Igual ocurre en el caso de lluvias torrenciales. Pueden alcanzar considerables distancias de hasta 100 km., lo que unido a la velocidad que alcanzan (del orden de 50 km/h.) y que pueden aparecer asociados o no a una erupción, los convierte en unos fenómenos muy peligrosos e impredecibles. Takahashi and Satofuka (2002) proponen una separación en 2 capas para este flujo, una inferior responsable de las ondas sísmicas registradas con fragmentos que colisionan entre sí y arrastran mate-

rial de la superficie por la que discurres y otra superior formada por partículas en suspensión. En 1985 el volcán Nevado del Ruiz (Colombia) fundió el hielo de su cima al entrar en erupción provocando un peligroso flujo de lodo que sumergió a la ciudad de Armero causando más de 20.000 muertos.

- **Colapsos (*COL*).** Caídas de material volcánico como parte del flujo de lava, fragmentos del domo y rocas, similar al lahar pero sin involucrar agua que cambia sus propiedades al enfriarse y se convierte en un derrumbe de material fluido (Takahashi and Tsujimoto, 2000). Dependiendo de la forma del volcán, pueden llegar a recorrer decenas de km aunque su duración extrañamente supera las decenas de minutos. En general los derrumbes de materiales (lahares, colapsos y flujos piroclásticos) comparten similitudes en su forma de onda y evolución de la energía espectral, aunque en los lahares predomina respecto a los colapsos la forma en punta de flecha en su sismograma (Figura 1.1.11). Se observa en todos ellos una distribución dispersa de la energía espectral en casi toda banda de frecuencias registradas y una ligera acumulación de energía en el intervalo central de duración del evento.

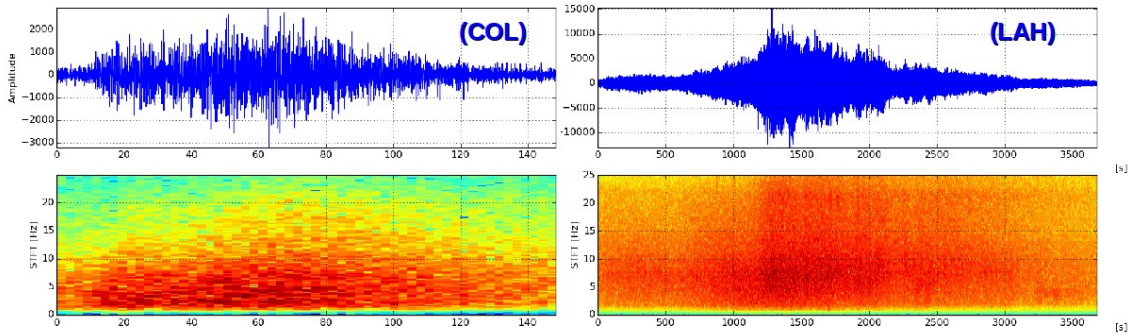


Figura 1.1.11.: Derrumbes de material. Se representa el sismograma y evolución de la energá espectral de un colapso (*COL*) y un lahar (*LAH*) registrados en la erupción del volcán de Colima en 2004.

- **Lluvia de cenizas (*AFL*).** Gases y ceniza que una vez expulsados en la columna eruptiva se enfrían al llegar al límite entre estratosfera y troposfera, condensándose y cayendo sobre la superficie cubriendo terrenos y ciudades. Provocan incendios, derrumbes de tejados y problemas respiratorios. Las nubes de ceniza pueden ser arrastradas por el viento hasta cientos de kilómetros precipitándose en poblaciones muy alejadas y colapsando el tráfico aéreo. La lluvia mezclada con ceniza y gases tóxicos (SH_2, CO, CO_2 y SO) origina la lluvia ácida, perjudicial para toda la vida animal (***ACD***). Gases como el CO_2 pueden acumularse en los llamados *valles de la muerte* incluso provocando víctimas mortales. En erupciones severas, el material expulsado bloquea la luz del sol provocando bajadas de temperatura, incluso a nivel mundial. Se ha considerado que pudieron ser el origen de alguna glaciación.

- **Ríos de lava (*LAF*).** Formados por el magma que sale al exterior, descienden lentamente desde el cráter. Engullen todo lo que se encuentran en su curso y causan incendios a su alrededor debido a la alta temperatura de la lava (más de 1000 C°). Sin embargo, no suelen causar víctimas mortales pues gracias a su baja velocidad y bajo nivel de viscosidad su trayectoria es fácil de predecir.

1.1.2.3. Señales ruidosas

Entendiendo *ruido* como cualquier señal que no pertenezca a ninguna de las clases que queremos reconocer, independientemente de cual sea su origen, la detección de este evento también tiene su importancia en sismología volcánica, aunque solo sea porque caracterizar el ruido nos servirá para discriminar y segmentar los eventos de otras clases que queremos clasificar (la Figura 1.1.12 muestra ejemplos de estas señales). Varios tipos de ruido pueden registrarse en los sensores sísmicos:

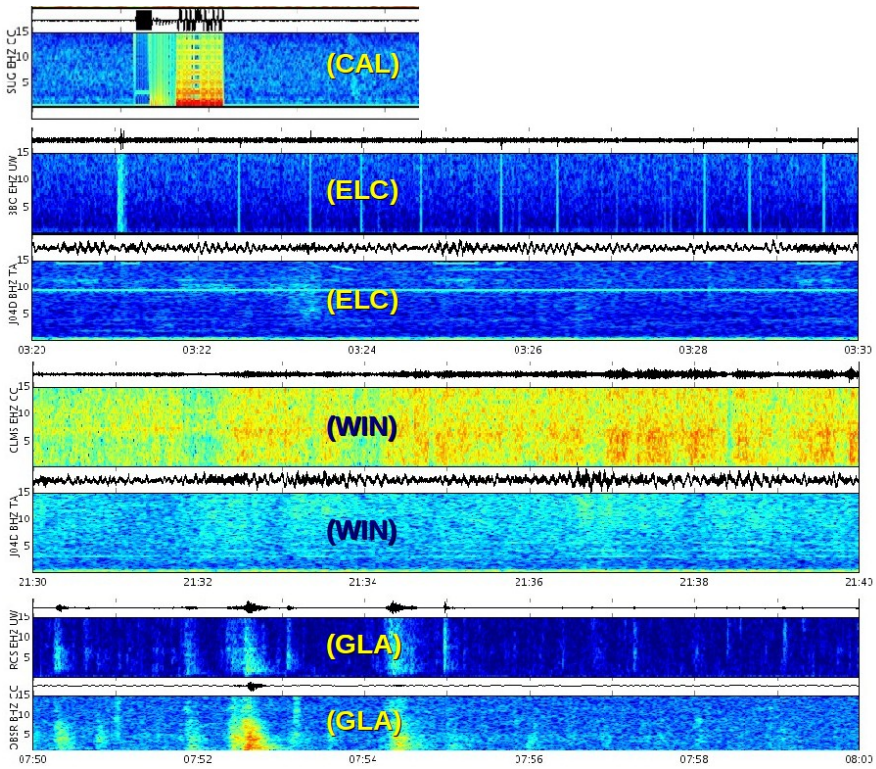


Figura 1.1.12.: Ruido registrado en los sismogramas. Sismogramas y espectrogramas registrando durante 10 minutos la presencia de diversos ruidos: señal de calibración de sensores (*CAL*), ruidos electrónicos varios (*ELC*), ráfagas de viento (*WIN*) y agrietamiento de glaciares (*GLA*). Figura original de la red sísmica del pacífico noroeste (`http://pnsn.org/spectrograms`).

- Ruido debido al **viento**. Aunque en general produce una alteración en un margen amplio de la banda de frecuencia (en torno a $[1, 15]\,Hz$) y comúnmente

en forma de ráfagas, a veces genera energía espectral en un un intervalo que puede confundirse con señales de tremor volcánico.

- **Glaciares**. En la Figura 1.1.12 se representan ruidos procedentes de los glaciares del Mount Rainier registrados durante 10 minutos en 2 estaciones.
- Ruidos **electrónicos**, de diversas fuentes.
- Señales de **calibración** de los propios sensores sísmicos.
- Ruido **antropogénico**. La presencia cercana de actividad humana y poblaciones genera todo un abanico de ruidos, desde motores de automóviles, aviones, ruidos de ciudades... La Figura 1.1.13 muestra el paso de varios autobuses cercanos al sensor, en un rango de $[5, 40]$ Hz, pudiendo interferir en la banda de interés de los sismos.

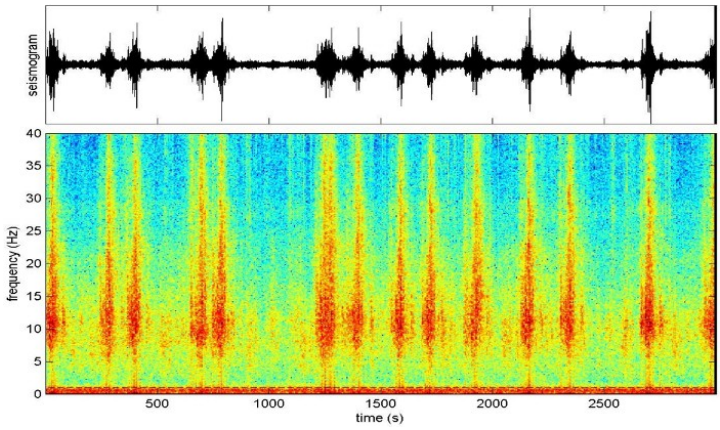


Figura 1.1.13.: Ruido debido al tráfico rodado. Registro del paso de autobuses a 50 m. de una estación de corto periodo en el parque del Timanfaya (cortesía de Javier Almendros).

Muchas veces la clave para aislar estos ruidos de las señales de interés es la *simultaneidad* en los registros de distintas estaciones cercanas; a pesar de los efectos de sitio y la dispersión, los eventos de origen sismo-volcánico deben registrarse en un intervalo temporal relativamente corto en todas las estaciones. Algunos ruidos como el calibrado de instrumentos y viento extrañamente aparecen simultáneamente. La *periodicidad* de las señales también es importante; por ejemplo, sabemos que el ruido antropogénico debe tener un ciclo de 24 horas y diferenciarse actividad diurna de la nocturna.

1.2. Monitorización de volcanes activos

Las erupciones volcánicas son junto a los terremotos y tornados los fenómenos naturales que más duramente azotan a la población. Existen unos 500 volcanes en la

Tierra de los que se estima que están activos (que han erupcionado en los últimos 20.000 años). Teniendo en cuenta que una erupción media puede desencadenar una energía equivalente a 20.000 bombas atómicas, un estudio profundo del vulcanismo, sus causas y su estadio en el proceso eruptivo (Figura 1.1.2) es fundamental de cara a monitorizar la actividad sísmica para evaluar un posible riesgo y prevenir desastres en la sociedad (Yueqing et al., 1996; McNutt, 1996). Actualmente se tiende a integrar en un solo sistema de predicción de erupciones la información que junto a la sismicidad proporcionan otras herramientas de monitorización disponibles (Giacinto et al., 1997; Aspinall et al., 2006; Carniel et al., 2006; Wassermann et al., 2007).

En esta sección detallaremos cual es la manera más efectiva de registrar las señales sismo-volcánicas en función de la información que se pretenda obtener de ellas y estudiaremos que las soluciones que ofrece la sismología volcánica para vigilar los volcanes activos. Por último listaremos otras técnicas complementarias de monitarización.

1.2.1. Registro de sismos

La variedad de los eventos sismo-volcánicos requiere usar distinta metodología a la hora de su adquisición y almacenamiento digital, tanto a nivel de instrumentos a utilizar como su distribución topológica entre distintas estaciones. Aunque un estudio detallado queda fuera de los objetivos de este capítulo, sí que señalaremos algunas nociones de especial relevancia para el reconocimiento automático de sismos. Havskov and Alguacil (2010) ofrecen una amplia visión sobre el tema.

Nuestras señales se registran mediante sensores sísmicos que se colocan en lugares estratégicos en las proximidades del cráter del volcán. El emplazamiento de las estaciones es sumamente importante para intentar minimizar los efectos de propagación y de sitio (Sección 1.1.2) y los ruidos (Subsubsección 1.1.2.3). Normalmente se utilizan redes de monitorización de al menos, 5 estaciones, aunque según qué casos se instalan arrays de sensores o redes de arrays.Existe una enorme variedad y calidad de sismómetros que repercute directamente en las propiedades del *sismograma* registrado que puede representar el desplazamiento, la velocidad o la aceleración del suelo. Es muy importante tener en cuenta que un sismograma no cuantifica directamente las ondas generadas en la fuente, sino que es la convolución temporal de la fuente, el camino (como vimos en la Sección 1.1.2) y la respuesta al impulso del instrumento que digitaliza las ondas sísmicas. Básicamente nos interesa distinguir entre dos tipos de sensores en función de su respuesta en frecuencia, determinada por la banda espectral de interés $[f_L, f_H]$ donde el sismómetro debe responder con una ganancia constante y de forma lineal para entradas con un rango de frecuencias dentro de dicho intervalo:

1. *Sensores de corto periodo* o *banda estrecha*: con una banda espectral útil a partir de $f_L \geq 1\,[Hz]$, normalmente con una frecuencia F_s de muestreo de hasta

50 $[Hz]$. Suelen ser aparatos sencillos que digitalizan las señales en muestras de pocos bits (12 o 16 a lo sumo) en forma de números enteros.

2. *Sensores de largo periodo* o *banda ancha:* cuya respuesta en frecuencia se extiende por debajo del Hz $f_L \leq 1$ $[Hz]$, incluso hasta 0.01 o más (se habla de *periodo de muestreo,* T_s en segundos en lugar de frecuencia de muestreo) y una Fs que puede llegar hasta 200 $[Hz]$. Utilizan desde 20 hasta 32 bits para describir una muestra, en formato de números enteros o reales, por lo que tienen un margen dinámico mucho mayor y también una mayor relación señal ruido (*Signal to Noise - SNR*).

Otra posible clasificación de los sensores puede hacerse en función de si proporcionan datos solo en la componente vertical (Z) o también si registran el movimiento del suelo en las 2 componentes superficiales (NS, EW). En cuanto a la distribución y localización de los sensores cabe resaltar las siguientes cuestiones (Wassermann, 2012):

- *Monitorización con un único sensor.* Esta era la forma más común de registrar señales en los 60's y 70's, e incluso hoy en día cuando no hay suficientes recursos. Aún así, no hay que menospreciar la información que se puede obtener con una sola estación: estudios de polarización y movimiento de partículas, análisis de energía espectral de la señal (*Seismic Spectral Amplitude Measurement - SSAM,* Rogers and Stephens, 1991) y de la energía media acumulada del sismograma (*Real-Time Seismic Amplitude Measurement - SSAM,* Endo and Murray, 1991). Estas medidas son la base de los sistemas rudimentarios de monitorización y de los sistemas de alerta temprana basados en el método de fallo de material (*Failure Forecast Method - FFM*, *Voight, 1989*). También la mayoría de los sistemas actuales VSR de clasificación automática de señales sísmicas (*Volcano-Seismic signal Recognition - VSR*, Orozco-Alzate et al., 2012) usan solo una estación de componente vertical como veremos en el Capítulo 2.
- *Topología de la red*: área cubierta, localización y número de las estaciones y tipos de sensores. Las redes amplias (de diámetro $\Delta > 20$km.) son ideales para distinguir entre sismicidad tectónica y volcánica y para localizar eventos y movimientos de fluidos profundos. Las redes pequeñas, concentradas en los flancos del volcán son menos propensas a sufrir efectos de sitio, dispersión y atenuación, siendo más adecuadas para registrar eventos sismo-volcánicos superficiales de baja amplitud y alto contenido espectral. Las señales de largo periodo (LPs y VLPs) se captan más eficazmente con sensores de banda ancha localizados lo más cerca posible de la actividad sísmica.
- *Arrays sísmicos.* Son conjuntos de unos 6-12 sensores instalados en una pequeña área de diámetro $\Delta < 100$m. con el objetivo de construir mapas de radiación de los frentes de onda que permitan adquirir información de la estructura superficial de velocidades y localización de fuentes de sismicidad de largo periodo (Almendros et al., 1997; Almendros, 1999).
- *Redes de arrays* o *arrays de pequeña apertura,* en los que cada estación es

un pequeño array, usualmente compuesto por un sensor de banda ancha de 3 componentes rodeado por sensores verticales de corto periodo. Estas distribuciones mejoran la resolución de los arrays convencionales al localizar eventos de distinto tipo y al representar las propiedades del campo de ondas.

1.2.2. Sismología volcánica como herramienta de monitorización

Los sistemas de predicción de erupciones basados en monitorizar el incremento de actividad sísmica de los eventos de bajo periodo (LPs, VLPs y tremores) han sido aplicados satisfactoriamente a numerosos volcanes: Redoubt (Chouet et al., 1994), Galeras (Fischer et al., 1994), Popocatépetl (Arciniega-Ceballos et al., 1999),... La compleja estructura heterogénea que presentan las zonas volcánicas, unida a los efectos de sitio y la dificultad de modelar los fenómenos magmáticos e hidrotermales involucrados en los episodios eruptivos hacen de la sismología volcánica una ciencia que necesita abarcar múltiples lineas de investigación (Chouet, 2003; Chouet and Matoza, 2013):

- *Tomografías 3D de alta resolución*, que permiten detallar la estructura interna del volcán y la existencia de fluidos en cámaras y grietas. Se realizan mapas del cociente de velocidades v_P/v_S de los frentes de onda P y S. El hecho de que las ondas de cizalla S no se propagen a través de líquidos indica la posible existencia de cámaras magmáticas.
- *Características espacio-temporales de los sismos de largo periodo.* Las antenas sísmicas densas (arrays de apertura pequeña) son necesarias para construir los mapas de potencia espectral de lentitud aparente con los que localizar estos terremotos, debido a que LPs y tremores no tienen una llegada impulsiva de su frente de ondas. El estudio de la sismicidad asociada a estas señales es el método más fiable y usado en la monitorización de volcanes activos (Chouet, 1996a). Estos eventos normalmente preceden y se simultanean con los episodios eruptivos, por lo que son considerados como precursores.
- *Recorrido, localización y volumen del magma,* extraído mediante la inversión del tensor momento en eventos de muy largo periodo (*Very Large Period - VLP*). Las señales VLP están correlacionadas con los mecanismos de fuente y el transporte de material involucrados en las erupciones volcánicas (Rowe et al., 1998).
- *Propiedades acústicas del magma y fluidos hidrotermales.* Chouet (1994; 1996b) identifica los eventos LP como precursores de las erupciones. Su análisis espectral revela que tras una llegada con más o menos componentes en alta frecuencia los LPs se comportan como un oscilador armónico sometido a una excitación transitoria cuyas frecuencias dominantes se asocian con las propiedades acústicas que tienen los tubos de resonancia, los cuales identifica con

grietas rellenas de fluido: el valor de la frecuencia dominante y su ancho de banda asociado está correlacionado con la morfología de la cavidad y con el tipo de fluido que contiene (Ibáñez et al., 2000). Los tremores se relacionan también con cavidades resonantes cuya excitación de mantiene de forma continuada. El mayor problema en este área es que una cavidad puede contener distintos tipos de fluido y definir una correspondencia entre el tipo de fluido y la sismicidad de largo periodo es una dificultad añadida. Tampoco se ha establecido una relación directa entre LPs profundos (a más de 30 km) y episodios eruptivos.

- *Conteo de eventos precursores*. Aún siendo un método sencillo, una gran parte de sistemas de alerta temprana utilizan esta técnica. Particularmente popular es la técnica de monitorización por fallo de material (*Failure Forecast Method - FFM*, *Voight, 1989*) que basándose en las variaciones del grado de actividad asociada a eventos precursores estima la fecha de una erupción invirtiendo la curva de aceleración en la frecuencia de dicha actividad. Kilburn (2003) lograron predecir en tiempo real la erupción del volcán Soufriere Hills en Montserrat. Estos métodos han evolucionado combinándose con sistemas VSR de reconocimiento automático (Orozco-Alzate et al., 2012) y técnicas probabilísticas (Boué et al., 2015).

Los retos a lograr en un futuro cercano pasan por estudiar el comportamiento del magma a partir de los fenómenos de desgasificación, de ebullición y de sus desplazamientos por los conductos y cámaras así como la descripción de las propiedades físico-químicas de los fluidos multifase (Chouet and Matoza, 2013). Para conseguirlo no solo hay que hacer un esfuerzo al interpretar las observaciones sísmicas para convertirlas en información en torno a dinámica de fluidos, si no que hay que estudiar la física que gobierna los procesos de interacción y transformación entre las burbujas de gas y el resto de fluido volcánico. Esto requiere de:

- *Investigación multidisciplinar* de todos los fenómenos físicos, químicos y geológicos involucrados, que son primordiales en la predicción a corto plazo de erupciones.
- *Métodos complementarios de monitorización* que ayuden a identificar precursores mediante la deformación del terreno (Bamler et al., 2009), la gaseometría (Vergniolle and Jaupart, 1990), la microgravimetría o la grabación de infrasonidos y observación con cámaras de infrarrojos.

2. Reconocimiento de señales volcano-tectónicas

En este capítulo nos adentraremos en los fundamentos de la clasificación automática de señales y más concretamente en el reconocimiento de eventos sismo-volcánicos.

Empezaremos en la Sección 2.1 exponiendo conceptos generales en el ámbito del aprendizaje automático y reconocimiento de patrones: objetivos, tipos de algoritmos, estadística gausiana, detección y clasificación de eventos. El lector familiarizado con estos temas puede seguir avanzando, si lo desea, por la sección siguiente. Continuaremos estudiando las particularidades y requerimientos que nos encontramos al intentar reconocer señales sísmicas registradas en los volcanes (Sección 2.2). En la Sección 2.3 presentaremos las principales estrategias de clasificación automática, estudiando detalladamente los métodos más usados para las señales de origen sísmico y más concretamente para aquellas relacionadas con la actividad sismo-volcánica. Finalmente repasaremos los trabajos más populares de la comunidad científica en este área, haciendo especial hincapié en las ventajas y desventajas de las técnicas actuales lo que nos dará el punto de partida (Sección 2.4) para detallar nuestra propuesta: un sistema de análisis y reconocimiento en paralelo presentado en posteriores capítulos.

2.1. Desde el aprendizaje automático hasta el reconocimiento de eventos.

A menudo, en el ámbito de la computación nos encontramos con áreas íntimamente relacionadas lo que nos lleva a errores conceptuales. El término *aprendizaje automático* se refiere a una disciplina dentro de la inteligencia artificial encargada de diseñar algoritmos que permitan a las máquinas aprender mediante la *inducción del conocimiento*: generalizando comportamientos a partir del análisis de datos. La forma usual de conseguir este objetivo es con la construcción de *modelos* o *reglas* que permiten a un sistema realizar acciones como tomar *decisiones* o hacer *predicciones* determinadas ante unas entradas concretas. Formalmente:

> (Mitchell, 1997)*: "Dada una tarea T cuyo rendimiento puede medirse mediante una función R, un programa es capaz de aprender de unos datos experimentales E si al desempeñar la tarea T mejora su rendimiento tras analizar E."*

Muchas tareas pueden ser englobadas bajo esta definición, desde la extracción de características hasta la separación ciega de señales pasando por la clasificación de eventos. Atendiendo a la manera de interactuar del sistema con los datos disponibles y a la información a priori que estos proporcionen, el aprendizaje automático se suele categorizar como (Theodoridis and Koutroumbas, 2009; Sutton and Barto, 1998):

- *Aprendizaje supervisado:* Se intenta construir una función o modelo que generalice la relación entre la entrada al sistema y la salida deseada a partir de datos de entrenamiento con muestras *etiquetadas* (ejemplos de entradas a las que corresponden salidas conocidas).
- *Aprendizaje no supervisado,* en el que se intenta extraer propiedades y generalizar estructuras de datos *no etiquetados.*
- *Aprendizaje semi-supervisado*: Usa métodos no supervisados para extraer información de muestras no etiquetadas e información a priori de datos etiquetados para construir el modelo de predicción de salida.
- *Aprendizaje por refuerzo* o *aprendizaje por ensayo y error*: Inspirado en el conductismo, el sistema aprende observando las respuestas del entorno a una serie de acciones que genera sobre él. Se busca maximizar una función recompensa a las secuencias de acciones óptimas. El ejemplo típico es una máquina que automáticamente va aprendiendo a conducir un coche maximizando una función de acierto.

Otras extensiones de estas categorías incluyen el aprendizaje por *transducción* (Gammerman et al., 1998) que apuesta por la *inferencia transductiva* frente a la inducción para resolver problemas concretos de clasificación sin generalizar un modelo predictivo. En el *aprendizaje multitarea (Caruana, 1996)* se aprende a solucionar conjuntamente problemas relacionados entre sí a partir de reglas y modelos que el sistema mismo infiere analizando datos con múltiples representaciones.

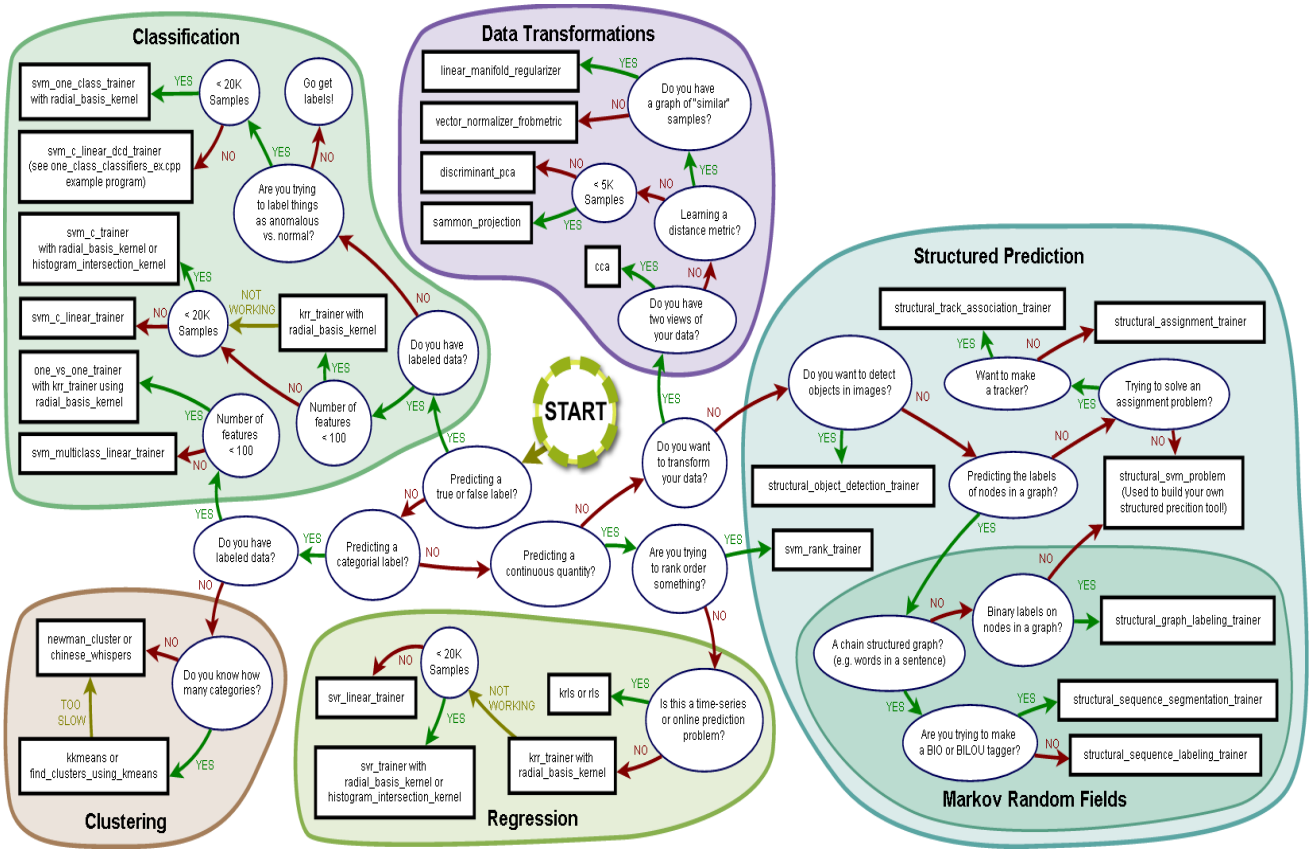


Figura 2.1.1.: Algoritmos de aprendizaje automático agrupados según la tarea que desempeñan y el tipo de datos sobre los que se aplican. (King, 2009).

Dependiendo del cometido a desempeñar por nuestra máquina y de la naturaleza de su salida, Bishop (2007) agrupa las tareas del aprendizaje automático en (Figura 2.1.1):

- *Reducción de dimensionalidad*: Se pretende disminuir la dimensión del espacio generado por las variables de entrada extrayendo la información que sea de más utilidad y eliminando la que sea redundante. El paso previo a la reducción de dimensionalidad es la *parametrización de los datos*: la extracción de características que mejor describan a los datos de entrada. Esquemas clásicos de parametrización y reducción de dimensionalidad incluyen el análisis de las componentes principales (*Principal Component Analysis - PCA*) y el análisis discriminante de Fisher (*Fisher's Discriminant Analysis - FDA*). El Capítulo 4 está dedicado en su totalidad a este tema.

- *Estimación de la densidad* de probabilidad en el espacio de descripción de las entradas (de características de los datos). Los métodos más conocidos son los histogramas y las estimaciones mediante núcleos (*Kernel Density Estimation - KDE*), también llamados ventanas de Parzen.

- *Regresión* o mejor *ajuste*: Se construye una función analítica que se ajuste a los datos minimizando el error entre los valores predichos por esta función y los realmente observados en los datos. La salida del sistema es del mismo tipo que las entrada (normalmente continua). El ejemplo típico es el ajuste de puntos a una curva definida por un polinomio. Tareas como la *interpolación* o *extrapolación* de datos están muy relacionadas con la regresión. En ellas, los datos de entrenamiento deben necesariamente ser idénticos a las predicciones del modelo.
- *Clasificación*: el sistema debe aprender a asociar cada entrada a un grupo o *clase* de muestras con propiedades parecidas. La salida es por tanto siempre un valor o símbolo (que representa a la clase) en el dominio discreto. Las redes neuronales, las máquinas de vectores soporte o los modelos ocultos de Markov son los clasificadores más comunes.
- *Agrupamiento* o *clustering*: De manera no supervisada se agrupan datos en categorías *similares* conforme un criterio o medida. Nótese que el número de grupos no tiene necesariamente que estar determinado a priori y que una misma muestra puede pertenecer a uno o más grupos. A menudo el término clustering se usa para designar a la clasificación no supervisada. El algoritmo *k-medias* es el más popular de estos métodos.

Otros conceptos que se relacionan o solapan con el aprendizaje automático es la *minería de datos* y el *reconocimiento de patrones*. La minería de datos (*data mining*) intenta descubrir propiedades existentes en un conjunto de datos sin tener un conocimiento previo de la estructura subyacente de las muestras. Vapnik (2000) presenta el reconocimiento de patrones como uno de los *tres problemas básicos del aprendizaje*

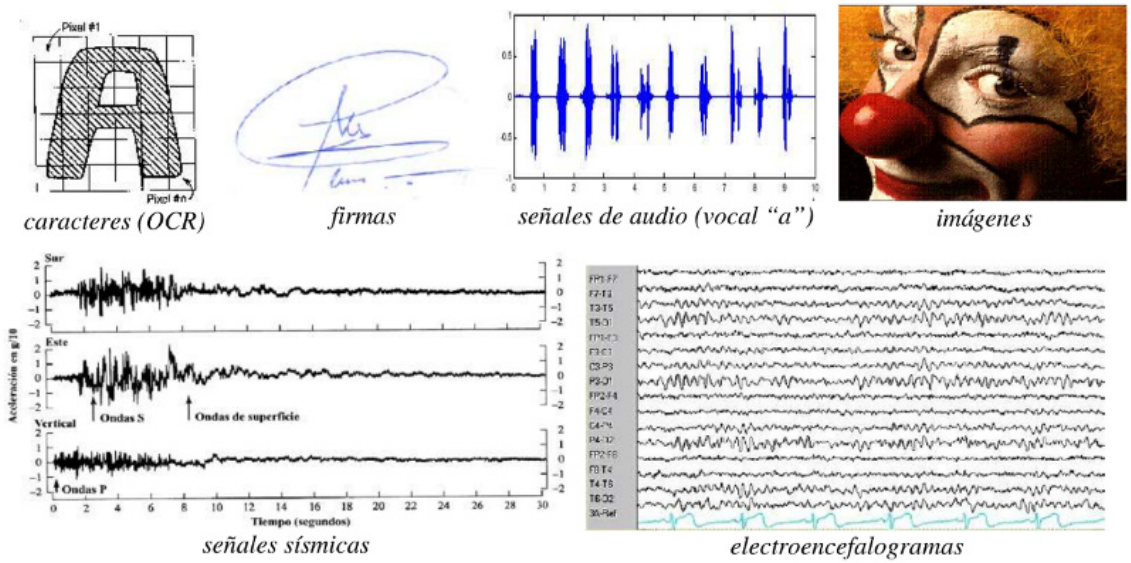


Figura 2.1.2.: Ejemplos de distintos tipos de *patrones*.

automático junto a la regresión y la estimación de densidades de probabilidad en el espacio de descripción de los datos. Definiendo de manera general un *patrón* como un conjunto de valores, reglas o estructuras determinadas que caracterizan a un grupo de muestras (Figura 2.1.2), se considera que un sistema de reconocimiento de patrones engloba la extracción y el *etiquetado* de esos patrones vía la adquisición de datos mediante un sensor a partir de una señal física o fuente de información y su adecuada descripción mediante el proceso de extracción de características (Figura 2.1.3). En esta ocasión, el concepto de etiquetado es más amplio que en el caso de clasificación. De nuevo, en función de dichas etiquetas o salidas, podemos distinguir entre:

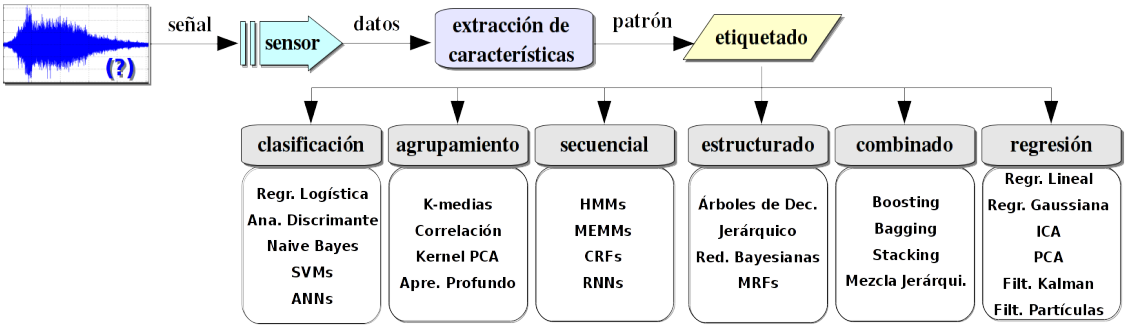


Figura 2.1.3.: Tipos de sistemas de reconocimiento de patrones conforme al tipo de etiquetado.

- Etiquetas de *categorías* o *clases* usadas en las tareas de clasificación y agrupamiento. En este caso, el patrón de cada clase se entiende como los valores modelo de las variables aleatorias asociadas al vector de características.
- *Combinación de algoritmos* supervisados con el objetivo de conseguir unas predicciones más fiables y una mayor flexibilidad que la obtenida con solo un método de aprendizaje a costa de una mayor complejidad computacional.
- Etiquetas en el dominio de características que generan nuevas muestras, o secuencias de ellas, como salida de los métodos de *regresión.*
- Patrones *estructurados* para caracterizar a objetos relacionados entre sí de forma jerárquica o atendiendo a reglas sintácticas o morfológicas. La segmentación y seguimiento de objetos en escenas de vídeo y realidad virtual o el análisis sintáctico de textos son ejemplos en los que usar este enfoque es más adecuado. Un caso particular de estructura es una *secuencia* de símbolos, que aparecen en áreas como el reconocimiento del habla, modelado de lenguaje, reconocimiento óptico de caracteres y firmas digitales.

Nótese que esta diversificación tiene una gran similitud con la realizada por Bishop (2007) para el aprendizaje automático. No en vano, muchos métodos y algoritmos se usan indistintamente en varias tareas del aprendizaje automático y reconocimiento

de patrones e, incluso, unos forman parte de otros, ya sea en su fase de inicialización o de manera complementaria. Existen excelentes libros clásicos para ampliar el conocimiento sobre esta área (Fukunaga, 1990; Bishop, 2007; Theodoridis and Koutroumbas, 2009; Hastie et al., 2009; James et al., 2013). Sea como fuere, en nuestro caso solo estamos interesados en las técnicas que nos permitan:

- Reconocer eventos de origen sismo-volcánico
- Reducir la dimensionalidad del vector de características usado al representar dichos eventos

Debido a que contamos con bases de datos etiquetadas, prestaremos más atención a los algoritmos supervisados, pues suelen ser más eficientes de cara a la clasificación y selección de características y requerir menos recursos computacionales que los no supervisados.

2.1.1. Clasificación supervisada de eventos

Se pueden escoger diversas estrategias para conseguir que una máquina clasifique de manera automática. No todas ellas son aplicables o adecuadas para resolver un determinado problema, por lo que en cada caso concreto conviene prestar atención a la hora de escoger la técnica de aprendizaje. El problema de la clasificación supervisada de patrones puede plantearse como:

Algoritmo 2.1 Clasificación supervisada de patrones.

Sean:

$W = \{w_1, ..., w_C\}$ un conjunto disjunto de C clases o etiquetas en las que pueden agruparse los datos, representados en el espacio de descripción de patrones o características, $\Omega_{\mathbf{X}}$, generado por vectores $\mathbf{x} = (x_1, ..., x_K)$ con K características.

$c : \Omega_{\mathbf{X}} \to W$ la función de clasificación experta o etiquetado, que asocia a cada vector $\mathbf{x}_i$ una clase que lo representa etiquetada como w_i, y con ello, su correspondiente región Γ_i dentro de $\Omega_{\mathbf{x}}$ tal que $c(\mathbf{x}_i) = w_i \Leftrightarrow \mathbf{x}_i \to w_i$.

DB_{tr}, DB_{ev} bases de datos de eventos etiquetados de entrenamiento $DB_{tr} = \{\mathbf{x}_{tr} \to w_{tr}\}$ y evaluación $DB_{ev} = \{\mathbf{x}_{ev} \to w_{ev}\}$, con $\mathbf{X}_{tr} = \{\mathbf{x}_{tr}\}$ y $\mathbf{X}_{ev} = \{\mathbf{x}_{ev}\}$ subconjuntos del espacio de características $\Omega_{\mathbf{X}}$.

La clasificación supervisada de patrones pretende construir a partir de reglas de decisión, inferidas mediante la aplicación $c(\mathbf{x} = \mathbf{x}_{tr})$, una función $d : \Omega_{X_{ev}} \to W$ de *decisión* o *decodificación* que sea *la mejor estimación posible* de la función de clasificación experta $c(\mathbf{x} = \mathbf{x}_{ev})$:

$$d(\mathbf{x}_{ev}) \equiv \hat{c}(\mathbf{x}_{ev}) \approx c(\mathbf{x}_{ev})$$

El diseño de un sistema de clasificación automática supervisada se estructura en varias etapas, esquematizadas en la Figura 2.1.4:

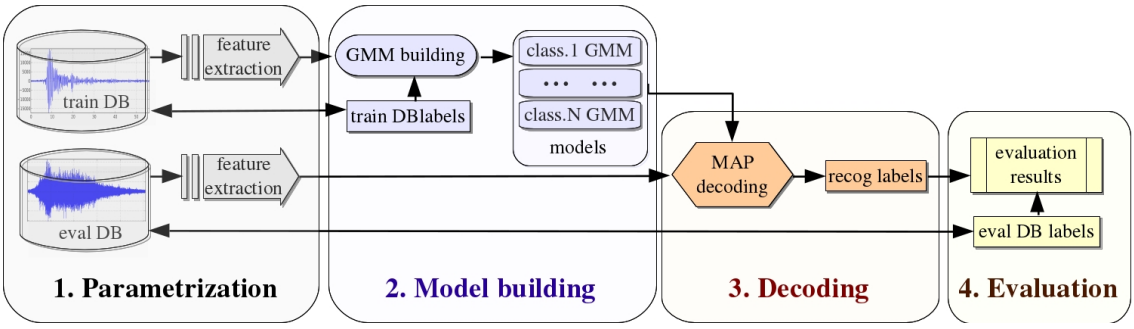


Figura 2.1.4.: Sistema de clasificación supervisada (basado en modelos estadísticos bayesianos)

1. **Adquisición** o **inicialización de datos.** Comprende la adquisición de señales, la construcción de las bases de datos y su parametrización y acondicionamiento para pasar a eventos o secuencias de eventos representables matemáticamente en un espacio de representación de patrones o espacio de características $\Omega_{\mathbf{x}}$. Los eventos se *etiquetan* o agrupan manualmente en clases por personal experto en el área por lo que la función $c(\mathbf{x})$ de *clasificación experta*, o *etiquetado*, no suele ser fácil de describir de forma analítica.

2. **Entrenamiento** o **aprendizaje.** Se crea la función $d(\mathbf{x})$ de *clasificación automática* o *decisión* en base a unas *reglas de decisión* inferidas al aplicar $c(\mathbf{x})$ a los datos de la base de entrenamiento DB_{tr}. Existen dos formas diferenciadas de construir $d(\mathbf{x})$:

   - Los métodos *paramétricos* incluyen en $d(\mathbf{x})$ un modelado analítico explícito de los datos de entrenamiento dado por un conjunto de parámetros $\boldsymbol{\theta}_p = \{\theta_p\} = \{\theta_1, .., \theta_P\}$ tal que $d(\mathbf{x}) = d(\mathbf{x}; \boldsymbol{\theta}_p)$. En la etapa de aprendizaje o de *estimación de parámetros* se hallan los valores de $\boldsymbol{\theta}_p$ que mejor se adaptan a la forma analítica de los modelos propuestos. De una manera relajada, a $d(\mathbf{x}; \boldsymbol{\theta}_p)$ también se le conoce como el *modelo* o *hipótesis* de clasificación, o, simplemente el *clasificador*, si bien es más exacto pensar en $d(\mathbf{x})$ como una función automática de clasificación que opera según unas reglas de decisión y, opcionalmente, también incorpora un modelado de datos. La mayoría de los clasificadores se encuentran en este grupo.

   - Los métodos *no paramétricos* construyen sus reglas de decisión directamente del análisis de los datos, sin proponer ningún modelo analítico previo. Los clasificadores basados en estimación de densidad mediante Parzen o los basados en vecindades (*k-Nearest Neighbors* - *kNNs*) son los ejemplos más representativos.

3. **Clasificación**. Una vez hallada $d(\mathbf{x})$, su aplicación sobre todos los vectores $\{\mathbf{x}_c\}$ asignados a la clase w_c genera un subconjunto Γ_c del espacio $\Omega_{\mathbf{X}}$ denominado *región de decisión* de w_c que está delimitado por su *frontera de decisión*, dibujada en la Figura 2.1.5. Usualmente, las regiones son subconjuntos disjun-

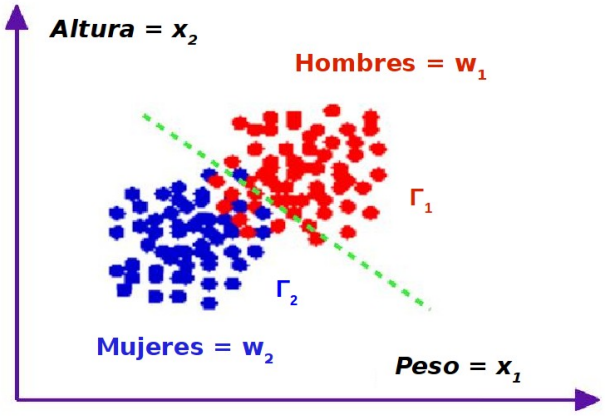


Figura 2.1.5.: Espacio $\Omega_{\mathbf{X}}$ de *descripción de patrones* o de *características* para datos agrupados en las clases $Hombres = w_1$ y $Mujeres = w_2$ asociadas respectivamente a las regiones de decisión Γ_1 y Γ_2. Cada dato se describe mediante el vector de características $\mathbf{x} = (x_1 = Peso, x_2 = Altura)$. Una vez construida la frontera entre Γ_1 y Γ_2, la función de decisión d etiqueta a un nuevo vector $\mathbf{x}$ con una de las dos clases posibles.

tos de $\Omega_{\mathbf{X}}$ cumpliéndose que $\Omega_{\mathbf{X}} = \Gamma_1 \cup \Gamma_2 \cup ... \cup \Gamma_C$. Delimitadas las regiones de cada clase junto con las reglas de decisión inferidas y, opcionalmente con los modelos de los datos, el clasificador $d(\mathbf{x})$ puede usarse para asignar de manera automática un evento $\mathbf{x}$ a una de las clases $\{w_c\}$ tal que $d(\mathbf{x}) = w_{\mathbf{x}}$. Como detallaremos en la Subsección 2.1.4, dependiendo de la estructura de la base de datos, podemos distinguir entre:

- *Clasificación en aislado*: cada registro o fichero f de la base de datos se corresponde solamente con un evento $\mathbf{x}$ a clasificar.
- *Clasificación en continuo*: cada fichero f contiene un número indeterminado f_i de eventos $\{\mathbf{x}_f\} \equiv \{\mathbf{x}_{f_1}, \mathbf{x}_{f_2}, ..., \mathbf{x}_{f_i}\}$ que se presentan como un flujo continuo de señal, correspondiendo al sistema delimitar temporalmente o *segmentar* dicho flujo en eventos para poder ser etiquetados uno a uno por la función de decisión $d(\mathbf{x})$. Un caso particular es la *clasificación en tiempo real* o en *vivo*, donde las secuencias de datos provienen directamente de una fuente que está continuamente emitiendo señal, y que requiere la detección y clasificación de eventos al mismo tiempo que se está emitiendo o, con un retraso mínimo.

Lógicamente, la clasificación con una base de datos DB, en continuo implica una complicación adicional al tener que delimitar temporalmente los eventos

a clasificar, lo que revierte en una eficiencia de reconocimiento generalmente menor si usamos el mismo corpus DB con los eventos ya aislados.

4. **Evaluación** del sistema. La bondad de la estimación $d(\mathbf{x}) \approx c(\mathbf{x})$ se mide sobre los datos de evaluación DB_{ev}. Aunque existen diversas técnicas para evaluar modelos, como veremos en la Sección 3.4, básicamente todas ellas se basan en contar los errores de clasificación cometidos cuando la clase $w_{\mathbf{x}}$ asociada a una muestra $\mathbf{x}_{ev}$ de evaluación por la función d no coincide con la etiqueta dada de forma supervisada, $d(\mathbf{x} = \mathbf{x}_{ev}) = w_{\mathbf{x}} \neq w_{ev} = c(\mathbf{x}_{ev})$. Definido matemáticamente lo que entendemos por error mediante una *función de coste* o *de error*, decimos que una regla de decisión es *óptima* si minimiza el error en la clasificación. Según la base de datos usada en la evaluación se habla de:

   - *Test cerrado*: en el caso de que $DB_{ev} = DB_{tr}$, esto es, los datos usados en la etapa de aprendizaje para construir la función de decisión $d(\mathbf{x})$ y, opcionalmente, sus modelos analíticos, son los mismos con los que se evalúa el clasificador.
   - *Test abierto* o *ciego*: el sistema se evalúa con datos distintos a los usados en el aprendizaje; $DB_{ev} \neq DB_{tr}$. Se pretende cuantificar la *capacidad de generalización* de los resultados obtenidos en el test cerrado.

   Ambos tipos de evaluaciones tienen su utilidad, si bien, en la práctica, lo más interesante son los sistemas con una alta capacidad de generalización, para poder usarlos de manera no supervisada sobre datos continuos sin etiquetar. Opcionalmente se usa una tercera base de datos o una partición de DB_{tr} que llamamos *datos de sintonización* o de *validación*, DB_{val}, utilizados en un proceso de ensayo-error con el objetivo de ajustar los mejores valores de configuración o de los parámetros de los modelos y comprobar que el sistema se comporta conforme se espera.

Como estudiaremos en la Sección 2.3, hay numerosas opciones para escoger una técnica adecuada a cada problema de la clasificación supervisada que podemos agrupar en dos perspectivas fundamentales:

1. **Clasificadores no estadísticos**, que no requieren métodos estadísticos al construir sus reglas de decisión o al modelar sus datos. Ejemplos de estos clasificadores sonlos clasificadores basados en reglas lógicas o heurísticas como los árboles de decisión (*Decision Trees - DT*, Murthy, 1998), los sistemas expertos (*Expert Systems - ES*) que Vogel et al. (1992); Yueqing et al. (1996) usan para evaluar el riesgo volcánico, la comparación de plantillas (*Template Matching*), los basados en medidas de similitud como los clasificadores por vecindades (*k-Nearest Neighbors - kNNs* o los *Nearest Mean Classifiers - NMC*).
2. **Aproximación estadística y clasificadores probabilísticos**. Usan inferencia estadística en la definición de $d(\mathbf{x}) = d(\mathbf{x}; \boldsymbol{\theta}_p)$, ya sea en el modelado

de datos, en el proceso de decisión o en ambos casos. El término *clasificadores probabilísticos* se aplica cuando para cada clase representada por la etiqueta w_c se proporciona la distribución de probabilidad $P(w_c|\mathbf{x})$. La ventaja inmediata de este enfoque es que dado un evento $\mathbf{x}_0$, el valor $P(w_c|\mathbf{x} = \mathbf{x}_0)$ sirve como base para medir fiabilidad de la decisión $\mathbf{x}_0 \rightarrow w_c$. En este sentido, no todos los clasificadores estadísticos son además probabilísticos.

Al igual que en el caso de los algoritmos de aprendizaje automático, las técnicas de clasificación pueden ser categorizadas de múltiples maneras, a menudo solapadas. Dentro de la aproximación estadística, es común asociar $P(w_c|\mathbf{x} = \mathbf{x}_0) = \sum_i W_{ci}\mathbf{x}_0^i$, siendo W_{ci} la matriz de coeficientes o pesos de la clase w_c aplicada al evento $\mathbf{x}_0$ elevado a su i-ésima potencia. Así se habla de clasificadores lineales (si $i = 1$), cuadráticos ($i = 2$), etc... Nótese que esta asignación solo es posible si todos los datos $\{\mathbf{x}\}$ de DB_{tr} tienen la misma longitud. A pesar de su sencillez, entre los *clasificadores lineales* encontramos algunos muy populares como los de máxima entropía, de máximo margen (SVMs), discriminadores lineales (LDA, FDA) o algunas redes neuronales. Otro caso interesante son los clasificadores difusos (Bezdek and Pal, 1992; Pal and Mitra, 1999), donde no se cumple necesariamente que $\Omega_{\mathbf{X}} = \Gamma_1 \cup \Gamma_2 \cup ... \cup \Gamma_C$ y cuya equivalencia con los clasificadores estadísticos queda demostrada en Kuncheva (1996).

2.1.2. Clasificadores estadísticos: inferencia estadística y aproximación bayesiana

La *inferencia estadística* pretende inducir un conocimiento sobre una población analizando una parte de ella. Para ello asignar un grado de certeza a una *hipótesis* o modelo H acerca de la población observando muestras o *evidencias* E de ella. En nuestro caso, podemos interpretar los datos $\{\mathbf{x}\}$ y sus clases dadas por las etiquetas $\{w\}$ como variables aleatorias que pueden ser descritas mediante funciones de probabilidad, asociando las hipótesis a las clases $\{w\}$ y las evidencias a los datos u observaciones $\{\mathbf{x}\}$, y estamos interesados en estimar $P(w = w_c|\{\mathbf{x}_c\})$, el modelo de la clase w_c a partir de los datos $\{\mathbf{x}_c\}$ de entrenamiento que manualmente asociamos a w_c. La regla de Bayes relaciona estos conceptos:

$$P(w|\mathbf{x}) = \frac{p(\mathbf{x}|w)}{p(\mathbf{x})}P(w) = P(H|E) = \frac{P(E|H)}{P(E)}P(H) \equiv \Delta(E)P(H) \qquad (2.1.1)$$

La interpretación del concepto de probabilidad en la Ecuación 2.1.1 origina dos ramas principales de inferir información estadísticamente (Bolstad, 2004):

1. **Escuela clásica** o **frecuentista**: considera la probabilidad como una medida de resultados obtenidos en la experimentación. $P(E)$ o $P(H)$ dan la proporción de que una población cumpla una propiedad o resultado E o H, $P(E|H)$ la de obtener E tras haber obtenido H y $P(H|E)$ viceversa. Asume que el vector de parámetros $\boldsymbol{\theta}_p = \{\theta_p\}$ y la probabilidad $P(w;\boldsymbol{\theta}_p)$ son desconocidos pero

objetivos, estimables a partir de los datos de entrenamiento. No considera, en todo caso, los parámetros $\{\theta_p\}$ como variables aleatorias, y por tanto, no se les puede asociar a distribuciones de probabilidad . El primer clasificador estadístico, el discriminante lineal de Fisher (1936) es el ejemplo clásico de inferencia frecuentista.

2. **Enfoque bayesiano** o **epistemológico**: que interpreta la probabilidad como un grado de certeza y explícitamente distingue entre un conocimiento *a priori*, subjetivo o experto, antes de cualquier observación, o *a posteriori*, conocimiento experimental extraído en base de observaciones. tal que:

   $P(H)$ es la *probabilidad a priori*, antes de realizar ninguna observación, de que la hipótesis H es cierta.

   $P(E)$ es la *probabilidad marginal* de observar la evidencia E.

   $P(E|H)$ es la *probabilidad condicional* de observar E asumiendo H como cierta. Se conoce como la *función de verosimilitud* (o acuerdo entre la hipótesis H y los observables E) cuando la describimos como función de parámetros aleatorios $\boldsymbol{\theta}_p = \{\theta_p\}$ del modelo y la evaluamos dados unos datos E concretos: $L_H(E|\boldsymbol{\theta}_p) = P(E|H;\boldsymbol{\theta}_p)$ (Subsubsección 2.1.3.2).

   $P(H|E)$ es la *probabilidad a posteriori,* actualizada, de que H sea cierta una vez observada E. Representa el grado de certeza sobre H modificado respecto a su incertidumbre medida por la función de verosimilitud $L_H(E|\boldsymbol{\theta}_p)$.

   En este paradigma, $P(w)$ puede formar parte de la información a priori, pudiéndose definirse directamente en la Ecuación 2.1.1. Usa la regla de Bayes para actualizar de forma continua la hipótesis $P(H)$ mediante las evidencias representados por el término $\Delta(E) \equiv \alpha P(E|H) = P(E|H)/P(E)$ siendo $\alpha = 1/Z$ y Z conocida como la constante de normalización.

Nótese que el uso o no de la regla de Bayes no implica necesariamente la adopción de la propuesta bayesiana. Los bayesianos relacionan de forma intuitiva una *causa* dada por un conocimiento experto a priori y un *efecto* representado por las observaciones experimentales. Asumen que tras observar un número suficiente de evidencias estas compensarán la información a priori en caso de no ser correcta y, finalmente, el proceso de estimación del modelo convergerá. Los frecuentistas no comparten que la convergencia tenga que darse y prefieren evitar cualquier información subjetiva estimando las probabilidades a priori.

Modelado discriminativo versus generativo. La manera escogida para hallar la probabilidad $P(w|\mathbf{x})$ (o, al menos, una estimación de ella) define las dos alternativas principales del enfoque estadístico:

- *Modelado discriminativo* o *condicional*: $P(w|\mathbf{x})$ se estima directamente. Incluye clasificadores bien conocidos como: SVMs, ANNs, CRFs, los basados en boosting, análisis discriminativo y clasificadores de máxima entropía (regresión logística).
- *Modelado generativo*: Como los GMMs, HMMs, naive Bayes y las máquinas restringidas de Boltzmann (*Restricted Boltzmann Machines* - *RBMs*); $P(w|\mathbf{x})$ se estima indirectamente, aprendiendo la probabilidad $p(\mathbf{x}, w)$ del suceso conjunto $(\mathbf{x}, w)$ o el modelo de datos $p(\mathbf{x}|w)$ de la clase w y haciendo uso de la Regla de Bayes.

La discusión entre clasificaciones discriminativas frente a generativas es una de las más eternas y prolíficas en el mundo del aprendizaje automático: el compromiso entre *desviación frente a variabilidad* (detallado en el siguiente apartado) en el modelado. Se asume que, las técnicas discriminativas obtienen mejores resultados promedio (con menor *desviación* respecto a un hipotético modelo perfecto) cuando existen suficientes y fiables datos de entrenamiento. En contraposición, la aproximación generativa es capaz de modelar de manera más robusta (con menor *variabilidad*, entendida como sensibilidad a la partición de entrenamiento) estructuras complejas y bases de datos pequeñas (Ng and Jordan, 2001; Pernkopf and Bilmes, 2005; Xue and Titterington, 2008). Otros autores apuestan por estrategias híbridas que reduzcan (usando algoritmos discriminativos) el error asintótico inherente a una estimación inexacta de la probabilidad conjunta realizada por los modelos generativos (Raina et al., 2003; Bouchard et al., 2004; Lasserre and Bishop, 2007; Bicego et al., 2013).

Probabilidad de error de clasificación: clasificador óptimo. Dados unos datos $\{\mathbf{x}\}$ de la partición de evaluación DB_{ev} y el conjunto de clases $\{w\}$ al que pueden asignarse, una forma de evaluar a un clasificador $d(\mathbf{x})$ es definir una función de coste o error sobre él. En el dominio de la estadística, el *error de clasificación* se asocia a una variable aleatoria e con una probabilidad $p(e)$:

$$p(e;d) = \sum_{w} p(e,w) = \sum_{c=\{1:C\}} p(e|w_c)P(w_c) \tag{2.1.2}$$

$$= \sum_{\mathbf{x}} p(e,\mathbf{x}) = \sum_{\mathbf{x}\in DB_{ev}} p(e|\mathbf{x})p(\mathbf{x}) \tag{2.1.3}$$

Nótese que dada una clase concreta w_A, la probabilidad condicional de error $p(e|w_A;d)$ puede expresarse como la suma de las probabilidades condicionales de los datos $\{\mathbf{x}_A\}$ que perteneciendo a w_A, tal que $c(\mathbf{x}_A) = w_A$, son erróneamente asignados a cualquier otra clase w_c, $d(\mathbf{x}_A) = w_c \neq w_A$:

$$p(e|w_A) = \sum_{\mathbf{x}_A \nrightarrow w_A} p(\mathbf{x}_A|w_A;d) \tag{2.1.4}$$

Equivalentemente, dado un evento concreto $\mathbf{x}_A$ correspondiente a la clase w_A, el error que se respecto a él se computa como la la probabilidad que existe de clasificarlo

erróneamente:

$$p(e|\mathbf{x}_A) = \sum_{w_c \neq w_A} P(w_c|\mathbf{x}_A; d) = 1 - P(w_A|\mathbf{x}_A; d) \tag{2.1.5}$$

Una forma intuitiva de entender $p(e; d)$ es verlo simplemente como un operador que calcula el error debido a las zonas de solapamiento de los modelos de clase $p(x|w_c)$ en las regiones de decisión. Una vez definido analíticamente $p(e)$, se habla de un *clasificador estadísticamente óptimo* $d(\mathbf{x}) \equiv d^*(\mathbf{x})$ si sus reglas de decisión son tales que minimizan la probabilidad de error $p(e)$:

$$d^*(x) \equiv \arg\min_d \{p(e; d)\} \tag{2.1.6}$$

Nótese que la operación de minimizar sobre d simboliza la *mejor* elección posible de sus reglas de decisión. El mínimo valor posible de la probabilidad de error que se puede cometer dados un evento $\mathbf{x}$ y un clasificador d se conoce como *error de Bayes*, $p_B(e)$, que sustituyendo la Ecuación 2.1.6 en la Ecuación 2.1.5 se alcanza si d es un clasificador óptimo (Fukunaga, 1990):

$$p_B(e|\mathbf{x}; d) \equiv \min_d \{p(e|\mathbf{x}; d)\} = 1 - \max_w \{P(w|\mathbf{x}; d)\} = p(e|\mathbf{x}; d = d^*) \tag{2.1.7}$$

Clasificador bayesiano y reglas de decisión. Un clasificador bayesiano es un clasificador probabilístico que adopta el enfoque bayesiano junto al teorema de Bayes para modelar la clase w_c a partir de los datos de entrenamiento $\{\mathbf{x}_c\}$ asociados a ella. La función $d_{Bayes}(\mathbf{x}) \equiv d_B(\mathbf{x})$ se termina de definir escogiendo una regla de decisión concreta. Lo más común es que dado un vector de características $\mathbf{x}$ se le asigne aquella clase $\hat{w}_x$ que maximice la probabilidad a posteriori de ser observada (regla *MAP - Maximum A Posteriori*):

$$\hat{w}_{\mathbf{x}} = d_{B:MAP}(\mathbf{x}) \equiv MAP[\mathbf{x}] = \arg\max_w \{P(w|\mathbf{x})\} = \arg\max_w \left\{ \frac{p(\mathbf{x}|w)}{p(\mathbf{x})} P(w) \right\} \tag{2.1.8}$$

y, teniendo en cuenta que el argumento del máximo no depende de $p(\mathbf{x})$:

$$d(\mathbf{x}) \equiv d_{B:MAP}(x) = MAP[\mathbf{x}] = \arg\max_w \{p(\mathbf{x}|w)P(w)\} \tag{2.1.9}$$

La regla MAP es una regla de decisión óptima para el clasificador bayesiano $d_B(\mathbf{x})$ por que minimiza la probabilidad de error de clasificación $p(e|\mathbf{x}; d)$ definida en la Ecuación 2.1.7. Una forma simplificada de la regla MAP es asumir antes de cualquier evidencia que la probabilidad a priori de observar una clase u otra es la misma, tal que $P(w) = Cte$ en la Ecuación 2.1.9. Esta hipótesis se conoce como la regla de máxima verosimilitud (regla *ML - Maximum Likelihood*) por maximizar la probabilidad condicional o función de verosimilitud $p(\mathbf{x}|w)$:

$$d(x) \equiv d_{B:ML}(\mathbf{x}) = ML[\mathbf{x}] = \arg\max_w \{p(\mathbf{x}|w)\} \tag{2.1.10}$$

Otra aproximación muy usada en la Ecuación 2.1.9 para reducir la complejidad de los modelos y, con ello, la necesidad de tener grandes bases de datos de entrenamiento y el tiempo requerido en el proceso de aprendizaje, es suponer que las características usadas para describir un evento son estadísticamente independientes entre sí: $p(\mathbf{x}) = p(x_1, x_2, ..., x_K) \equiv \prod_k p(x_k)$. Esta simplificación define al modelo $d_{NB}(\mathbf{x})$ *simple* de Bayes (o *naive* Bayes):

$$d(\mathbf{x}) \equiv d_{NB:MAP}(x) = \arg\max_w \left\{ P(w) \prod_k p(x_k|w) \right\} \tag{2.1.11}$$

$$d(\mathbf{x}) \equiv d_{NB:ML}(x) = \arg\max_w \left\{ \prod_k p(x_k|w) \right\} \tag{2.1.12}$$

2.1.3. Entrenamiento de modelos

Algunos clasificadores como los basados en vecindades o los que usan medidas de (di)similitud no necesitan modelar explícitamente los eventos de cada clase para generar las reglas y regiones de decisión. Otros, sin embargo, necesitan definir analíticamente un modelo $d(\mathbf{x}) = d(\mathbf{x}; \boldsymbol{\theta}_p)$ como función del vector de parámetros $\boldsymbol{\theta}_p$. El proceso de hallar los valores óptimos de cada elemento θ_p se conoce como el *ajuste* o *entrenamiento del modelo* o la *estimación de parámetros*. El objetivo de este proceso es obtener una alta eficiencia de clasificación tanto con los datos de entrenamiento (que el sistema no tenga *desviación* respecto a un hipotético modelado perfecto de DB_{tr}) como con eventos que no pertenezcan a la DB_{tr} (que sea lo suficientemente flexible o *variable* para garantizar una *capacidad de generalización*). Esto se consigue al reducir al mismo tiempo los errores que provienen tanto de una desviación como de una varianza excesivas en el modelado. Lamentablemente, como esbozaremos matemáticamente en la Ecuación 2.1.13 , una vez elegido el tipo de modelo, existe un *compromiso entre la desviación y varianza* que dificulta esta tarea.

En general, cada tipo de clasificador tiene sus propios algoritmos de aprendizaje de modelos, si bien cabe examinar estas técnicas básicas de estimación:

1. Entrenamiento mediante optimización de una función de coste
2. Aprendizaje mediante estimadores estadísticos

2.1.3.1. Ajuste del modelo como un problema de optimización

Paralelamente a la metodología seguida en la Sección 2.1.2 para definir un clasificador estadísticamente óptimo, una estrategia clásica para hallar los mejores valores del vector de parámetros $\{\theta_p\}$ es planteando el ajuste de modelos como un *problema*

general de optimización: se minimiza una *función de coste* $J(c,d) = J(c(\mathbf{x}), d(\mathbf{x})) = J(w_x, \hat{w}_x)$ que represente el error cometido al aproximar $c(\mathbf{x}) \approx d(\mathbf{x}) \Leftrightarrow w_x \approx \hat{w}_x$ sobre la partición de entrenamiento. Dado un evento $\mathbf{x}$, $J(w_x, \hat{w}_x)$ debe ser siempre positiva, excepto cuando la clasificación sea correcta ($c(\mathbf{x}) = d(\mathbf{x})$). Equivalentemente podemos construir una *función objetivo* a maximizar que represente la bondad de la estimación. La función de coste más frecuente computa el error cuadrático medio (*Mean Square Error - MSE*) de la estimación. Asumiendo $E[\Delta]$ como el valor esperado de Δ y d^* como la mejor estimación posible de $c(\mathbf{x})$, que coincidiría su forma analítica $c_A(\mathbf{x})$, en caso de existir, para el evento $\mathbf{x}$ etiquetado como w_x definimos:

$$\begin{aligned} J_{MSE}(c,d) &= E[|c(\mathbf{x}) - d(\mathbf{x};\boldsymbol{\theta}_p)|^2] = E[(w_x - d)^2] \\ &= E[(w_x - d^*)^2] + (d^* - E[d])^2 + E[(d - E[d])^2] \\ &\equiv \sigma_n^2 + bias^2(d) + var(d) \end{aligned} \tag{2.1.13}$$

donde:

σ_n^2 representa el *ruido* inherente a la aproximación $d^*(\mathbf{x}) \approx w_x$. Sería nulo si $\mathbf{x} \in DB_{tr}$, pero puede ser positivo en el caso contrario, pues tanto d como d^* son estimados a partir de datos de entrenamiento.

$bias^2(d)$ es la *desviación* (o *desajuste*) del modelo construido d respecto al modelo ideal d^*. Es causada por una excesiva simplificación o sub-entrenamiento (*subfitting*) en el modelado o por asumir unas hipótesis incorrectas en este.

$var(d)$ es la *varianza* (o *variabilidad*) de $d(\mathbf{x})$ y representa cuanto varían sus predicciones entre sí o, equivalentemente, cuanto varía el modelo respecto su (predicción) media $\mu_d \equiv E[d]$. Da una idea de la sensibilidad del modelo sobre una partición $DB(i)$ concreta de datos de entrenamiento; un *modelado robusto* implica que se obtendrán modelos $d_i(\mathbf{x})$ muy parecidos sea cual sea la partición i usada para crearlos. Para reducir la desviación se puede aumentar la complejidad del modelo, incrementando así su $var(d)$, pero si la DB_{tr} no es lo suficientemente extensa, se aumenta también el riesgo de sobre-entrenamiento (*overfitting*, ver la Figura 2.1.6 para un ejemplo gráfico y la Sección A.4 para un ejemplo más cuantitativo).

Compromiso entre el ajuste y variabilidad del modelo. En el proceso de construcción de $d(\mathbf{x})$ no se puede reducir el ruido σ_n^2, delimitando el mínimo error de clasificación posible que se cometerá al evaluar el modelo sobre particiones de datos distintas a la base DB_{tr} utilizada en el entrenamiento. Sin embargo, tanto la desviación como la varianza pueden compensarse parcialmente, pero siempre acorde al llamado compromiso de la *desviación frente a la variabilidad* (Friedman, 1997): un modelo tiene poca desviación si acierta en la mayoría de sus predicciones sobre la base DB_{tr} lo que implica una alta variabilidad entre ellas (a no ser que solo exista

una única clase que predecir). Matemáticamente; en una DB_{tr} compleja, la variabilidad entre muestras $\{\mathbf{x}_c\}$ de una misma clase w_c tiende a ser alta, por lo que la región de decisión Γ_c de esa clase será extensa, y para que el modelo $d_c(\mathbf{x}_c)$ pueda abarcarla y capturar los eventos más alejados próximos a las fronteras de Γ_c deberá alejarse mucho respecto a su centro (representado por la media μ_d del modelo) lo que implica según la Ecuación 2.1.13 incrementar la cantidad $|d_c(\mathbf{x}_c) - \mu_c|$ y, por tanto, su varianza. Conforme al nivel de ajuste que tengan, los clasificadores se agrupan en:

- *Fuertes* o *inestables* cuando tienen baja desviación. Suelen ser modelos complicados y de alta variabilidad, como SVMs y ANNs, que permite representar bien la estructura estadística subyacente de la DB_{tr} compleja. Tienen el riesgo de modelar también datos atípicos u *outliers*: posibles errores de etiquetado que pueden existir incluso en el mismo corpus de entrenamiento asociados al ruido σ_n^2.
- *Débiles* o *estables* o con mucha desviación, típico de un modelado simple, como el Naive Bayes, con predicciones menos variables, por lo que quizás no sean capaces de modelar correctamente una DB_{tr} extensa.

Como se esquematiza en Figura 2.1.6, un modelado con demasiada desviación o demasiada varianza tiende a decrementar su capacidad de generalización. No hay por

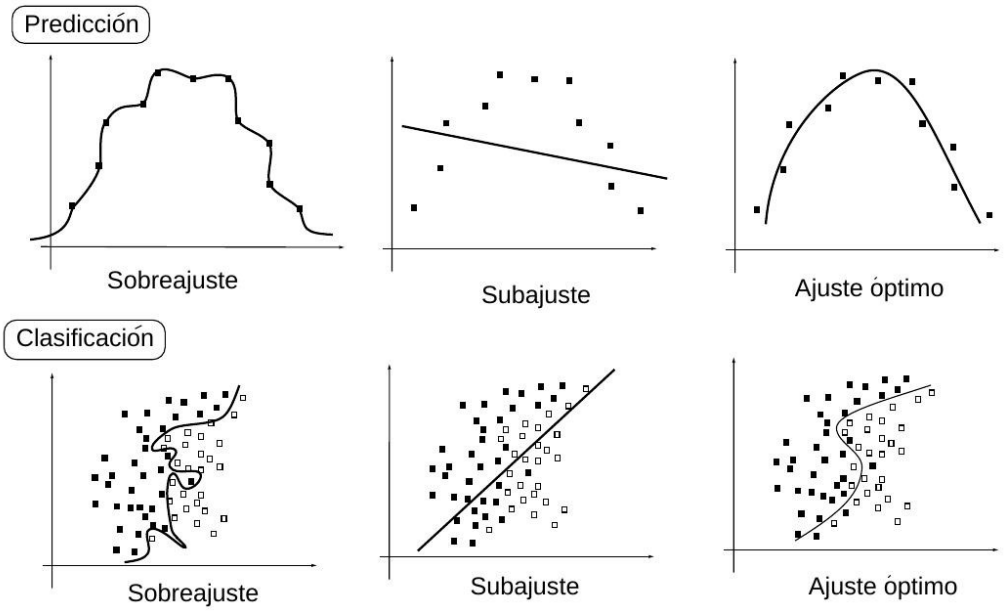


Figura 2.1.6.: Compromiso entre desviación y variabilidad. *Sobreajuste:* Un ajuste excesivo a los datos de entrenamiento genera fronteras de decisión demasiado específicas para estos que pueden reducir la capacidad de clasificación en otro corpus ligeramente distinto. *Subajuste*: modelos muy simples no discriminan bien distribuciones no lineales de datos. (Figura original en Jiménez, 2003)

tanto una correlación directa entre la complejidad de un modelo y su eficacia de clasificación y la elección de unos modelos u otros depende de cada caso en particular. Existen *técnicas de estabilización* para bajar la varianza o complejidad de los modelos, entre las que destacan los métodos de y la reducción de la dimensionalidad. Otros enfoques permiten explícitamente controlar la relación entre desviación y complejidad como los modelos híbridos, la combinación de modelos (*boosting* o *bagging*) o los modelos estadísticos de componentes (GMMs o HMMs).

Minimización de la función de coste. La minimización de $J(c,d)$ es, a menudo, una tarea no inmediata que depende tanto de la descripción analítica y complejidad del modelo $d(\mathbf{x};\boldsymbol{\theta}_p)$ como del tamaño de la base de entrenamiento DB_{tr}. En contadas ocasiones puede solucionarse directamente mediante cálculo algebraico. La mayoría de las veces hay que recurrir a algoritmos iterativos como el descenso en gradiente, métodos heurísticos u otras técnicas avanzadas de optimización local que no siempre aseguran la solución exacta. En este sentido, nótese que $J(c,d)$ puede tener varios mínimos locales que pueden ser dados como posibles soluciones en el problema de optimización. Es deseable que la función de coste tenga una forma sencilla para minimizar estos efectos, en concreto, una $J(c,d)$ convexa nos asegura un único óptimo global, objetivo que se consigue en modelos $d(\mathbf{x};\boldsymbol{\theta}_p)$ como la regresión logística (clasificadores de máxima entropía) o los SVM.

2.1.3.2. Aprendizaje estadístico: estimadores bayesianos, de máxima probabilidad (MAP) y de máxima verosimilitud (MLE).

La versión estadística del entrenamiento de modelos planteado como una optimización se reduce a definir una función de coste basada en probabilidades. En estadística bayesiana, dado un modelo $d(\mathbf{x};\boldsymbol{\theta}_p) = d_{\boldsymbol{\theta}}(\mathbf{x}) = \hat{w}$ en su forma paramétrica, se asume que cada componente del vector de parámetros $\boldsymbol{\theta}_p = \{\theta_1, .., \theta_P\}$ es una variable aleatoria independiente que lleva asociada una función de probabilidad marginal $p(\theta_p)$ y se actualiza su probabilidad a posteriori mediante una función de verosimilitud. El enfoque frecuentista asume que el valor óptimo de cada θ_p viene dado por una función analítica concreta, desconocida pero estimable, al menos puntualmente.

Estimación frecuentista. Dadas las duplas $\{(\mathbf{x}, c(\mathbf{x}) = w)\}$ de la partición de entrenamiento DB_{tr} y una función de coste $J(c, d_{\boldsymbol{\theta}})$ al aproximar $c \approx d_{\boldsymbol{\theta}}$, o, equivalentemente $J(w, \hat{w})$ que representa el error de estimar la clase correcta w como $\hat{w}$, se define la *función de riesgo* asociado al modelo d como el valor esperado de $J(c, d_{\boldsymbol{\theta}})$ como (Vapnik, 2000):

$$R(c, d_{\boldsymbol{\theta}}) \equiv E[J(c, d_{\boldsymbol{\theta}})] = \sum_{\mathbf{x}} \sum_{w} J(c, d_{\boldsymbol{\theta}}) p(\mathbf{x}, w) = \tag{2.1.14}$$

$$= \sum_{\mathbf{x}} \sum_{w} J(w, \hat{w}) p(\mathbf{x}|w) P(w) = \sum_{\mathbf{x}} \sum_{w} J(w, \hat{w}) p(\mathbf{x}|w) P(w) \tag{2.1.15}$$

Dentro del marco frecuentista, podemos abordar la estimación de parámetros como un problema general de optimización tal y como se hizo en la Subsubsección 2.1.3.1 simplemente encontrando el mejor valor del vector $\boldsymbol{\theta}_p$ que minimice el riesgo $R(c, d_\theta)$. Nótese que en la Ecuación 2.1.14 el valor de $\boldsymbol{\theta}_p$ es fijo.

Regla MAP de clasificación como minimización del riesgo. Basándonos en el riesgo $R(c, d_{\boldsymbol{\theta}}) = E[J(c, d_{\boldsymbol{\theta}})]$, dado un evento $\mathbf{x}$ se define un *riesgo condicional*, $R(\hat{w}|\mathbf{x}) = E[J(w, \hat{w})|\mathbf{x}]$ que representa el coste de asociar $\mathbf{x} \to \hat{w}$, computando la esperanza en la Ecuación 2.1.14 sobre la probabilidad condicional $P(w|\mathbf{x})$ en lugar de usar la probabilidad conjunta (Jain et al., 2000):

$$R(\hat{w}|\mathbf{x}) \equiv E[J(w, \hat{w})|\mathbf{x}] = \sum_w J(w, \hat{w}) P(w|\mathbf{x}) \tag{2.1.16}$$

La regla de máxima probabilidad a priori, que en el clasificador bayesiano (Sección 2.1.2) asigna $\mathbf{x} \to \hat{w} = \arg\max_w \{P(w|\mathbf{x})\}$, puede ser inferida minimizando el riesgo condicional bayesiano $R(w_i|\mathbf{x})$. Si en la Ecuación 2.1.19 definimos $J(w_i, w_j) \equiv |1 - \delta_{ij}|$ siendo δ_{ij} la delta de Kronecker nos queda:

$$R(w_i|\mathbf{x}) \equiv P_{error}(w_i|\mathbf{x}) = 1 - P(w_i|\mathbf{x}) \tag{2.1.17}$$

El riesgo condicional se puede asociar de esta forma a la probabilidad condicional de error $P_{error}(w_i|\mathbf{x})$ de clasificar $\mathbf{x} \to w_i$ dada en la Ecuación 2.1.5. La clase $\hat{w}$ que *optimiza* este riesgo minimizando P_{error} es exactamente la que propone la regla MAP. Nótese que partiendo de la Ecuación 2.1.14 llegamos a la misma conclusión si definimos el coste como $R(c, d) \equiv \sum_{\mathbf{x}} R_{\mathbf{x}}(c, d; \mathbf{x})$. Un posible inconveniente de asignar $J(w_i, w_j) = |1 - \delta_{ij}|$ es que en el proceso de optimización se tratan por igual los falsos negativos que los falsos positivos, lo que en ciertos problemas de clasificación puede suponer un gran problema (no es lo mismo el coste de errar al afirmar que un paciente tiene cáncer que errar al afirmar que está sano).

Estimadores bayesianos. Dado un conocimiento a priori o hipótesis sobre el parámetro aleatorio θ representado por $p(\theta)$ y una función $L(\{\mathbf{x}\}|\theta)$ de verosimilitud de θ a partir de los datos de entrenamiento $\{\mathbf{x}\}$, el enfoque bayesiano usa el teorema de Bayes para actualizar la hipótesis a partir de los resultados de experimentación de esta forma:

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}, \theta)}{p(\mathbf{x})} = \frac{L(\mathbf{x}|\theta) p(\theta)}{\int L(\mathbf{x}|\theta) p(\theta) d\theta} \tag{2.1.18}$$

Los estimadores bayesianos usan la probabilidad a posteriori $p(\theta|\mathbf{x})$ y una función de coste $J(\theta, \hat{\theta})$ para evaluar el riesgo de aproximar la variable aleatoria $\theta \approx \hat{\theta}$, siendo $\hat{\theta} = \hat{\theta}(\{\mathbf{x}\})$ una estimación a partir de los datos $\{\mathbf{x}\}$. La función de riesgo, conocida

ahora como el *riesgo condicional de Bayes,* $R_B(\hat{\theta}|\mathbf{x})$, es la esperanza de un coste $J(\theta, \hat{\theta})$ tomada sobre la probabilidad a posteriori $p(\theta|\mathbf{x})$:

$$R_B(\hat{\theta}|\mathbf{x}) \equiv E[J(\theta, \hat{\theta})|\mathbf{x}] = \int J(\theta, \hat{\theta}) p(\theta|\mathbf{x}) d\theta \tag{2.1.19}$$

Equivalentemente al papel que juega el error de Bayes en la definición de los clasificadores óptimos (Sección 2.1.2), un *estimador de Bayes* de θ es aquel que minimiza el riesgo condicional de Bayes $R_B(\hat{\theta}|\mathbf{x})$ para todos los datos $\{\mathbf{x}\}$ de la base de entrenamiento DB_{tr}:

$$\hat{\Delta}_{Bayes}[\theta, J] \equiv \hat{\theta}_B[J] \equiv \arg\min_{\hat{\theta}} \{R_B(\hat{\theta}|\mathbf{x})\} \tag{2.1.20}$$

Estimación MMSE bayesiana. Un estimador de Bayes queda construido una vez dados los datos $\{\mathbf{x}\}$, la función de probabilidad $p(\theta|\mathbf{x})$ y el coste $J(\theta, \hat{\theta})$ de aproximar $\theta \rightarrow \hat{\theta}(\mathbf{x})$. Paralelamente a lo que ocurre con las reglas de decisión en el clasificador bayesiano(Sección 2.1.2), nos encontramos con diferentes tipos de estimadores bayesianos dependiendo de como definamos la función de coste $J(\theta, \hat{\theta})$. Por ejemplo, asignando $J(\theta, \hat{\theta}) \equiv MSE(\theta, \hat{\theta})$ nos lleva al estimador de Bayes más popular, la estimación de mínimo error cuadrático medio (*Minimum Mean Square Error - MMSE*). En ella, la estimación de θ queda como la media de su probabilidad a posteriori, $\hat{\theta}_{B:MMSE}[MSE(\theta, \hat{\theta})] = mean\{p(\theta|\mathbf{x})\} = E[\theta|\mathbf{x}] = \int \theta p(\theta|\mathbf{x}) d\theta$.

Influencia del tamaño de las bases de datos en las estimaciones. Conforme la base de datos DB_{tr} se va agrandando, $p(\theta)$ va pareciéndose más a una distribución normal, por lo que su influencia sobre $p(\theta|\mathbf{x})$ se va reduciendo y los estimadores de Bayes van convergiendo a funciones más simples. De hecho, puede comprobarse que el estimador $\hat{\theta}_{B:MMSE}[J = MSE(\theta, \hat{\theta})]$ tiende a una distribución normal si el número de observaciones tiende a infinito. Por el contrario, a medida que la base de datos es más pequeña la influencia de $p(\theta)$ se incrementa en la estimación final. El estudio cuando el número T de eventos en la DB_{tr} es muy grande, $T \rightarrow \infty$, se denomina el *comportamiento asintótico*.

Estimadores puntuales de máxima verosimilitud (ML) y de máxima probabilidad a posteriori (MAP). Una forma más simple de ajustar el modelo paramétrico dado por $\boldsymbol{\theta}_p = \{\theta_1, .., \theta_p\}$ a partir de eventos $\{\mathbf{x}\}$ es usar estimadores estadísticos puntuales. A modo de equivalencia con el clasificador bayesiano (Sección 2.1.2) veremos dos casos concretos en los que se ignora la función de coste:

- Un ejemplo dentro del enfoque *frecuentista* es la *estimación de máxima verosimilitud* (*Maximum Likelihood Estimate - MLE*). Dados los observables $\{\mathbf{x}\}$, se define una *función de verosimilitud* del vector de parámetros $\boldsymbol{\theta}_p$ como una función proporcional a la probabilidad condicional de observar $\mathbf{x}$ para un valor concreto $\boldsymbol{\theta}$ de $\boldsymbol{\theta}_p$, tal que $L(\{\mathbf{x}\}|\boldsymbol{\theta}|) \equiv \alpha p(\mathbf{x}|\boldsymbol{\theta}_p = \boldsymbol{\theta})$, con α cualquier constante positiva(Fisher, 1922). $L(\{\mathbf{x}\}|\boldsymbol{\theta})$ mide el grado de concordancia entre los datos $\{\mathbf{x}\}$ medidos y el modelo $\boldsymbol{\theta}$ propuesto para ellos. Con $\alpha \equiv 1$, un estimador de máxima verosimilitud de un parámetro θ adopta esta forma:

$$\hat{\theta}_{MLE}[\{\mathbf{x}\}] \equiv \arg\max_{\theta}\{L(\{\mathbf{x}\}|\theta)\} = \arg\max_{\theta}\{p(\mathbf{x}|\theta)\} \tag{2.1.21}$$

  Puede demostrarse que el comportamiento asintótico de un estimador bayesiano con un coste $J = MSE(\theta, \hat{\theta})$ es el estimador MLE: $\hat{\theta}_{B:MMSE} \underset{T\gg 1}{\rightarrow} \hat{\theta}_{MLE}$.

- La estimación MAP *(Maximum A Posterior - MAP)* es una alternativa bayesiana a MLE que asume θ como una variable aleatoria sobre la que se maximiza su probabilidad a posteriori $p(\theta|\mathbf{x})$ vía la verosimilitud, o, equivalentemente, $\hat{\theta}_{MAP} = moda\{p(\theta|\mathbf{x})\}$:

$$\hat{\theta}_{MAP}[\{\mathbf{x}\}] \equiv \arg\max_{\theta}\{p(\theta|\mathbf{x})\} = \arg\max_{\theta}\left\{\frac{p(\mathbf{x}|\theta)p(\theta)}{\int p(\mathbf{x}|\theta)p(\theta)d\theta}\right\} \tag{2.1.22}$$

$$= \arg\max_{\theta}\{p(\mathbf{x}|\theta)p(\theta)\} \tag{2.1.23}$$

  donde $p(\mathbf{x}) = \int p(\mathbf{x}|\theta)p(\theta)d\theta$ es una constante de normalización que puede ser ignorada al no influir en el resultado de la optimización respecto a θ. El estimador MAP puede ser visto como una versión regularizada (Figura 2.1.3.1) al MLE siendo el término $p(\theta)$ el que controla la complejidad del modelo. Definiendo el coste $J = 1\delta 0(\theta, \hat{\theta})$ como nulo si $|\theta - \hat{\theta}|n < 1$ y 1 en cualquier otro caso, un estimador de Bayes tiende a la estimación MAP si $p(\delta)$ es unimodal: $\hat{\theta}_{B:1\delta 0} \underset{n\gg 1}{\rightarrow} \hat{\theta}_{MAP}$.

2.1.4. Clasificación y detección de señales

El reconocimiento sobre una señal continua implica discriminar los eventos que nos interesan delimitándolos temporalmente en la etapa de *segmentación* (o *detección*) para posteriormente asignarle una etiqueta en la etapa de *clasificación*. Normalmente esa discriminación se lleva a cabo separando los eventos del resto de señal, que llamamos genéricamente *ruido,* definiendo el ruido como otro tipo de evento en sí mismo. En la mayoría de los sistemas esta separación se realiza caracterizando al ruido como la señal base y segmentando los eventos cuando se detecta que la energía de la señal es superior a la del ruido. En los próximos apartados detallaremos el problema.

2.1.4.1. Patrones secuenciales y reconocimiento en continuo.

Vamos a suponer una fuente de información que emite un mensaje S formado por una secuencia de E símbolos $S = \{S_e\} = \{S_1, .., S_E\}$. Por ejemplo, en la Figura 2.1.7 se muestra como en el reconocimiento de voz tenemos una secuencia de fonemas que forman el sonido correspondiente a una palabra. Equivalentemente, en reconocimiento de caracteres, la palabra se compone de una secuencia letras, y una secuencia de llegadas de distintos tipos de ondas en el sismograma definen un terremoto. Para reconocer el mensaje emitido lo segmentamos en partes que describimos como una secuencia O de vectores de características u observables $O = \{O_1, .., O_6\}$. Un oyente (o sistema de reconocimiento) intentará agrupar estos observables para asociarlos de la mejor manera posible una secuencia de R símbolos que él pueda entender (patrones de los que ya tiene un modelo) como un mensaje recibido $\hat{S} = \{\hat{S}_r\} = \{\hat{S}_1, .., \hat{S}_R\}$. Idealmente, la secuencia emitida debería ser exactamente igual a la reconocida, símbolo a símbolo: $\{S_e\} = \{\hat{S}_r\}$. En este proceso, el oyente se está enfrentando a dos problemas:

1. *Detección o segmentación* de eventos en una señal continua: ¿qué reglas usar para determinar cuando termina un símbolo y empieza otro? (particionar los observables de O en R subconjuntos disjuntos, como $O = \{O_1, O_2, O_3\}\{O_4\}\{O_5, O_6\}$)

2. *Clasificación*: Una vez agrupados los vectores, ¿qué símbolo o etiqueta asociamos a cada partición de O?($\{O_1, O_2, O_3\} \rightarrow \hat{S}_2$; $\{O_4\} \rightarrow \hat{S}_1$; $\{O_5, O_6\} \rightarrow \hat{S}_3$ o bien $\{O_1, O_2, O_3\} \rightarrow \hat{S}_1$; $\{O_4\} \rightarrow \hat{S}_2$; $\{O_5, O_6\} \rightarrow \hat{S}_3$)

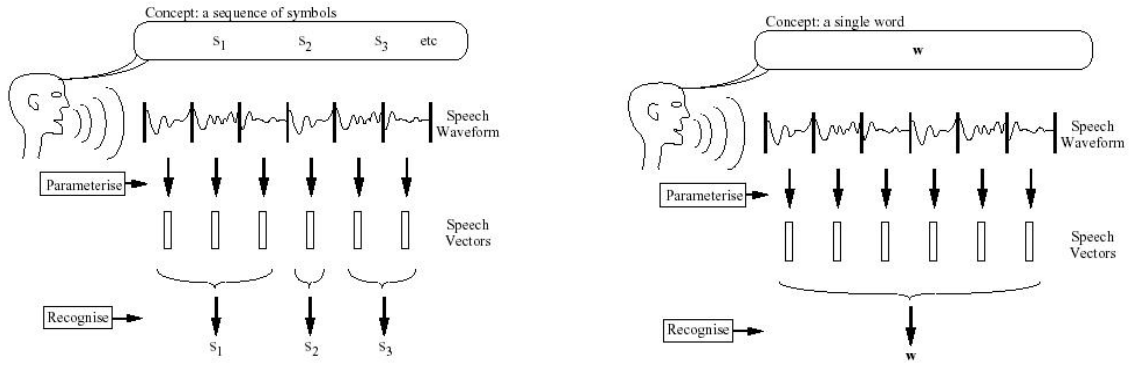


Figura 2.1.7.: Reconocimiento *continuo vs. aislado* de eventos: una fuente de información (locutor) emite continuamente una secuencia de símbolos (palabras $\{S_1, S_2, S_3\}$) transformados en otra secuencia de vectores de características que un sistema (oyente) reconoce como una secuencia estructurada de palabras (frase $\{\hat{S}_1, \hat{S}_2, \hat{S}_3\}$). Hablamos de reconocimiento aislado si la secuencia de observables emitidos solo se corresponden con un símbolo (por ejemplo, palabra w) reconocible por el oyente. Figuras originales de Young et al. (2006).

Si, como se observa en la Figura 2.1.7 a priori sabemos que el mensaje emitido y el recibido sólo pueden contener un único símbolo, el primer problema es fácilmente resoluble. Esta suposición es la que define al *reconocimiento de palabras (o eventos) aislados*, o, simplemente, *clasificación de eventos*, en contraposición con el *reconocimiento continuo* de eventos, que implica resolver ambas cuestiones, la segmentación en eventos y posteriormente la clasificación de dichos eventos. Atendiendo a la naturaleza de los mensajes emitidos por la fuente y al tipo del sistema de reconocimiento, se distingue entre:

- **Reconocimiento en continuo**: asociar una secuencia de T observables, $\{\mathbf{O}_{t=1:T}\}$, con una secuencia de R símbolos recibidos (eventos sismo-volcánicos en nuestro caso) $\hat{S} = \{\hat{S}_{r=1:R}\}$.
- **Reconocimiento de eventos aislados**: en el caso de que los eventos se encuentren ya segmentados, y se sepa de antemano cuantos observables corresponden a cada símbolo: $\{\mathbf{O}_1, ..., \mathbf{O}_{T(r)}\} \rightarrow \hat{S}_r$.
- **Reconocimiento de *frames* u observables**: cuando el clasificador etiqueta con un símbolo cada observable: $\mathbf{O}_t \rightarrow \hat{S}_t$.

2.1.4.2. Extensión a patrones secuenciales.

Cómo veremos en la Sección 2.3, solo los sistemas de reconocimiento de patrones estructurados son capaces de clasificar secuencias de observables. En la mayoría de clasificadores (SVMs, ANNs, GMMs,...) aún estando cada símbolo descrito por varios observables, el clasificador solo puede etiquetar frame a frame y se requieren técnicas e hipótesis complementarias para poder *extender* el reconocimiento a eventos aislados o en continuo.

En los clasificadores probabilísticos esta extensión se plantea de la siguiente manera; sea $\mathbf{x_t}$ el vector de características u observable en el instante t, tal que representamos cada símbolo (o evento sismo-volcánico en nuestro caso) por la secuencia de observables $\mathbf{x} = \{\mathbf{x}_t\}_{t=1:T}$. Según vimos en la Subsección 2.1.2, mediante la estimación MAP (de máxima probabilidad a posteriori) y la regla de Bayes, la estadística bayesiana resuelve el problema de clasificación a nivel de frame asignando al observable $\mathbf{x}_t$ la clase $\hat{w}_t$ entre las posibles clases $\{w_c\}$ que maximice la probabilidad a posteriori $P(w_c|\mathbf{x}_t)$:

$$\mathbf{x} \rightarrow \hat{w}_t \Leftrightarrow \hat{w}_t = \arg\max_{w_c}\{P(w_c|\mathbf{x}_t)\} = \arg\max_{w_c}\left\{\frac{p(\mathbf{x}_t|w_c)P(w_c)}{p(\mathbf{x}_t)}\right\} \tag{2.1.24}$$

A partir de Ecuación 2.1.24 podemos extender a 2 niveles:

1. *Extensión a clasificación de eventos aislados:* Entrenados los modelos de frame $p(\mathbf{x}_t|w_c)$, para un evento descrito por $\mathbf{x} = \{\mathbf{x}_t\}_{t=1:T}$ podemos evaluar cada observable $\mathbf{x}_t$ obteniendo las secuencias $\{p(\mathbf{x}_t|w_c)\}$ de probabilidades para cada modelo de cada clase w_c. El reto consiste en asociar una única etiqueta a todo

el evento: $\mathbf{x} = \{\mathbf{x}_t\}_{t=1:T} \rightarrow \hat{w}_{\mathbf{x}}$. Esto lo implementaremos al construir nuestro sistema base en la Sección 3.3.1.2.

2. *Extensión a reconocimiento en continuo:* A partir de los modelos $p(\mathbf{x}_t|w_c)$ pretendemos asociar a los observables $\mathbf{x} = \{\mathbf{x}_t\}$ la secuencia de eventos etiquetados como $\hat{\mathbf{w}} = \{w_1, w_2, ..., w_R\}$, tal que $\hat{\mathbf{w}}$ maximice la probabilidad a posteriori $P(\mathbf{w}|\mathbf{x})$. Este es un caso más complejo, que resolveremos usando métodos semi-heurísticos en el decodificador conjunto de nuestra propuesta del sistema en paralelo (Subsección 5.1.2).

En el caso de los reconocedores de patrones secuenciales, tales como los HMM, se aplica directamente la regla MAP a un evento aislado $\mathbf{x} = \{\mathbf{x}_t\}$ evaluando los modelos $p(\mathbf{x}|w_c)$. Como veremos en la Subsubsección 3.3.2.2, es posible una extensión desde la clasificación de eventos aislados a flujo continuo simplemente construyendo una macro-HMM que a partir de unos observables $\{\mathbf{x}_t\}$ correspondientes a una secuencia de eventos sin segmentar halle la secuencia de modelos etiquetados con $\hat{\mathbf{w}} = \{w_1, w_2, ..., w_R\}$ que maximicen $P(\mathbf{w}|\{\mathbf{x}_t\})$.

Pre-segmentación de la señal en el reconocimiento continuo. En los centros de monitorización que usan sistemas de reconocimiento de *tiempo (cuasi)real* una señal continua descrita por las etiquetas $\mathbf{w} = \{w_1, w_2, ..., w_R\}$ se suele analizar ininterrumpidamente (*online*) en intervalos temporales predefinidos o trazas de análisis (que pueden variar desde minutos a horas). En el caso del reconocimiento del habla el proceso de análisis y reconocimiento tiene lugar en intervalos delimitados por un detector de silencios (Rabiner and Juang, 1993; Young et al., 2006). Cuando el reconocimiento tiene lugar de forma no simultánea a la señal (*offline:* una vez adquirida la base de datos y almacenada para su posterior análisis), la duración de la secuencia $\mathbf{w}$ y el número de R eventos que contiene viene determinada por el tamaño del fichero o traza donde se ha grabado (normalmente en ficheros de una hora o 24 horas). Como veremos más adelante (Sección 2.3) la duración de las trazas de análisis, el ratio $n^o eventos/traza$ así como el número de clases a identificar, son parámetros de gran importancia a la hora de evaluar correctamente la complejidad y eficiencia de la tarea de reconocimiento.

2.1.4.3. Modelado del lenguaje y reglas gramaticales.

Junto al modelado a nivel de frame del espacio de características que describen a los eventos, es común utilizar en los reconocedores estructurados un modelado que describan cómo se interrelacionan los eventos entre sí, el llamado modelado a *nivel de eventos* o *modelado del lenguaje* (en el área del reconocimiento del habla). Las reglas gramaticales son extraídas a partir de un análisis de la *información de contexto* entre eventos y del papel o jerarquía que un evento o grupo de eventos tiene respecto a otros.

Por ejemplo, en el caso del reconocimiento del habla, los eventos son las palabras que suelen agruparse secuencialmente en una oración como $\{sujeto\} + \{verbo\} + \{predicado\}$. Podrían definirse estructuras más complejas agrupando los eventos en $\{artículos\}$, $\{verbos\}$, $\{adjetivos\}$, $\{nombres\}$, $\{preposiciones\}$, etc. y definiendo las interrelaciones entre estos grupos. Estas reglas y agrupaciones forman parte del denominado *modelado gramatical* del lenguaje. Existen otros tipos de modelado de lenguaje, en concreto el *modelado estadístico* o *estocástico*, del que ya hablamos en la Subsección 2.1.2 y pretende evaluar la probabilidad condicional $P(w_t|\mathbf{w}_{t-\tau})$ de obtener un evento w_t conociendo la secuencia $\mathbf{w}_{t-\tau} = \{w_{t-1}, w_{t-2}, ..., w_{t-\tau}\}$ de los últimos eventos observados.

2.2. Reconocimiento de patrones aplicado a señales sismo-volcánicas (VSR)

Las propiedades de las señales sísmicas, su adquisición y los métodos de evaluación generan algunas particularidades en los sistemas de clasificación de sismos (*Volcano-Seismic Recognition - VSR*). La mayor parte surge de la variabilidad temporal y espectral de los eventos que complica el proceso de construcción de bases de datos. Esta variabilidad en las señales, junto con la necesidad de un monitoreo continuo de la actividad sísmica también dificulta la etapa de clasificación e impone sobre los sistemas de reconocimiento unos requerimientos concretos como detallamos a continuación.

2.2.1. Problemas relacionados con las propiedades de los eventos sísmicos

- **Problemas de variabilidad**. La enorme variabilidad de algunos tipos de eventos sísmicos registrados en los volcanes es el principal inconveniente de cara a construir modelos eficaces. Esta variabilidad es generada de varias maneras:

  - *Variabilidad temporal y espectral inherente a ciertas clases.* Los sismos pueden ser generados con distinta intensidad y en diferente localización. Las características de la fuente generadora y del medio donde se propagan las ondas es el origen de la variabilidad que caracteriza a todas las clases sismo-volcánicas. Como se muestra en la Figura 2.2.1, algunos tipos de eventos son más propensos a sufrir este efecto: tremores que pueden ser de distinto tipo y pueden durar desde decenas de segundos a decenas de semanas. El problema con la variabilidad es que pasa a formar parte también del espacio de características que pretenden describir los modelos de reconocimiento.

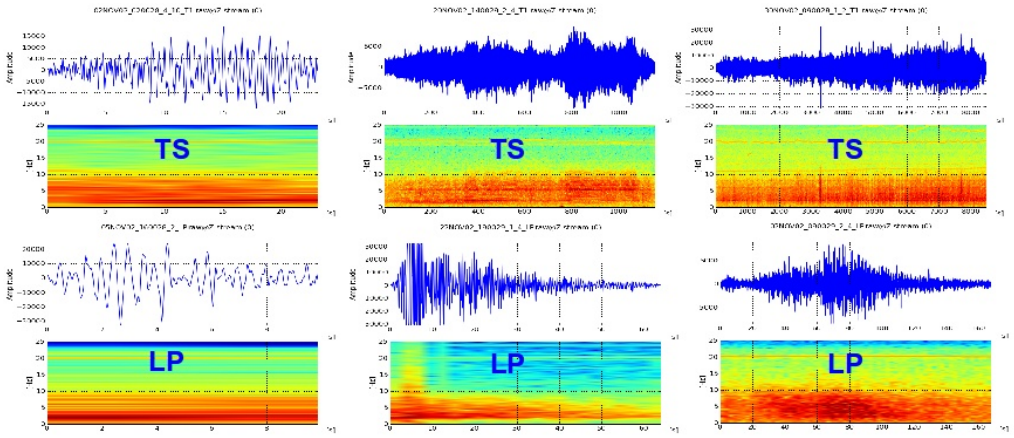


Figura 2.2.1.: Variabilidad de eventos sísmicos. Se representa la variabilidad temporal y espectral de eventos (*LP*) y (*TS*) registrados por la misma estación en el volcán de Colima.

  - *Diferencia entre eventos del mismo tipo registrados en diferentes periodos de actividad sísmica.* El origen de los sismos, su intensidad y localización así como el tipo de eventos registrados puede cambiar según el estado del volcán dentro del mismo periodo de actividad y, por supuesto, según que periodos.
  - *Diferencia entre eventos del mismo tipo pero de distintos volcanes.* Un evento de bajo periodo originado en el volcán de Colima puede ser bastante diferente a otro generado en la isla de Decepción.

- **Solapamiento de eventos.** En algunos volcanes unos eventos son capaces de generar otros apareciendo secuencialmente e, incluso, llegando a solaparse temporalmente: un tremor continuado puede desestabilizar el sistema y provocar pequeños sismos volcano-tectónicos solapados sobre él. Esto supone un problema grave para la mayoría de los sistemas de reconocimiento que solo son capaces de detectar un evento en cada instante de tiempo.
- **Adquisición de señales.** Podemos distinguir dos dificultades que encontramos relacionadas con la adquisición de los eventos:
  - *Registrado de datos por diferente tipos de estaciones,* que repercute en el espectro de la señal captada debido al uso de distintos filtros realizados sobre los registros o a una distinta respuesta en frecuencia, o ganacia de cada estación. Incluso en el caso de usar los mismos equipos, estos pueden no estar en las mismas condiciones.
  - *Efectos de sitio y propagación.* La heterogeneidad en la estructura interna de los volcanes hace que los registros de datos tiendan a sufrir efectos de sitio (atenuación y amplificación) y propagación (anisotropía, dispersión, radiación, refracción,...), que se perciben, incluso, estando las estaciones

muy próximas entre sí (a menos de 500m.). Como vimos en el Capítulo 1 (Sección 1.1.2), estos fenómenos son tan importantes como para clasificar de forma distinta al mismo evento según sea digitalizado en una estación u otra.

2.2.2. Problemas relacionados con la fiabilidad de las bases de datos

El proceso de construcción de una base de datos robusta y fiable es probablemente una de las etapas más importantes a la hora de diseñar un sistema de reconocimiento VSR basado en la probabilidad. Se requiere:

- Gran cantidad de eventos de una misma clase
- Un mínimo de calidad en la adquisición de las señales: una relación señal/ruido SNR adecuada y eventos que no estén saturados.
- Un criterio unificado y claro de etiquetado
- Técnicos que sean expertos en señales de origen sismo-volcánico

Cuando alguno de estos puntos falla lleva a problemas de diversa índole:

- **Problemas de subjetividad**. que se dan en 2 ámbitos:

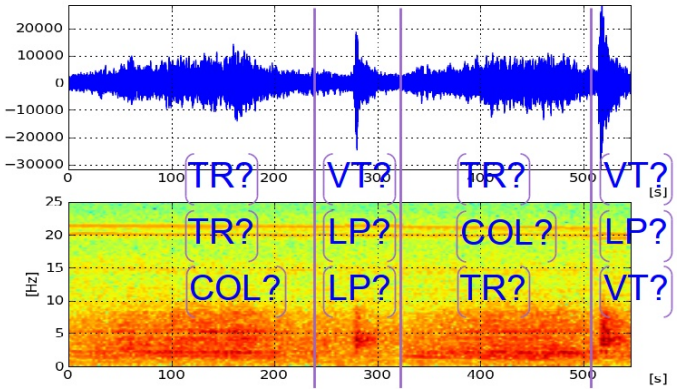


Figura 2.2.2.: Subjetividad en el etiquetado supervisado. La falta de criterios objetivos y de experiencia lleva a situaciones que dificultan la clasificación supervisada por los técnicos al construir una base de datos: ¿cómo etiquetar correctamente la secuencia de eventos arriba representados?

- *Subjetividad en la definición de eventos.* Una definición laxa de las clases, con falta de rigurosidad científica o una vaga descripción matemática de los eventos puede llevar a cierta subjetividad a la hora de que un operario aprenda las características de una clase para poder reconocerla posteriormente. Como vimos en el apartado anterior (Subsección 2.2.1),

la mayoría de las veces la definición de eventos es muy complicada debido a la variabilidad de estos (ver la Figura 2.2.2).

- *Subjetividad en el etiquetado manual*. Ya sea por falta de conocimiento del técnico, o como consecuencia de un criterio claro en la definición de eventos, se estima que la la coincidencia entre distintos expertos al etiquetar una misma base de datos rondaría el 80 % (lo que fija una barrera para la tasa de efectividad en el reconocimiento automático realizado por máquinas). A esto se le une una característica ineludible de la condición humana: el cansancio acumulado, o simplemente el estado de ánimo, pueden influir en el etiquetado supervisado.

- **Criterios de evaluación sesgados**. Existen distintas formas de evaluar los resultados de reconocimiento de un sistema. Las métricas más comunes se basan en contabilizar el número de eventos que el sistema no ha detectado o *borrado*, los que ha *insertado* (y no están en el etiquetado supervisado) y los que ha confundido con otro tipo de evento (*sustituciones*). En el caso de evaluaciones con señales sísmicas estas cuentas no son siempre eficaces debido a:

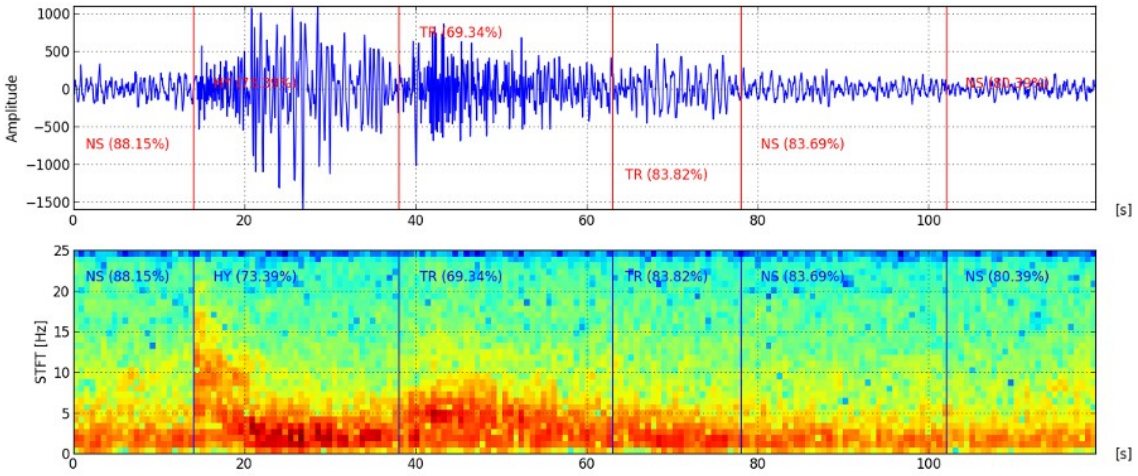


Figura 2.2.3.: Inserciones de etiquetas no geofísicamente relevantes en el reconocimiento. El sistema divide en dos el evento tremor y el ruido, asignando las etiquetas $\{TR + TR \,|\, NS + NS\}$ en lugar de $\{TR \,|\, NS\}$. sin embargo, geofísicamente estos errores de etiquetado no tienen importancia desde un punto de vista geofísico.

- *Inserciones y borrados no geofísicamente relevantes*. Como observamos en la Figura 2.2.3, no todas las clases tienen la misma relevancia desde un punto de vista geofísico: no es lo mismo insertar un ruido tras otro ruido tras otro ruido que, puede ser considerado como un único evento que insertar un terremoto donde no lo hay. Equivalentemente, ignorar una explosión no tienen la misma importancia que ignorar un ruido.

- *Sesgo para clases predominantes* sobre otras. Usualmente, el número de eventos de ruido es al menos un 50 % de los eventos de una base de datos. La razón es simple: los eventos sísmicos están precedidos o antecedidos por ruido. Sí una base de datos consta de 10 clases y el sistema siempre clasifica correctamente los eventos de la clase ruido, que supongan el 50 % del total de todos los eventos, aunque falle en reconocer el resto de 9 clases el porcentaje de aciertos ya alcanza, al menos, el 50 %.

2.2.3. Requerimientos de los sistemas VSR

Los sistemas de reconocimiento aplicados a los eventos sismo-volcánicos se utilizan básicamente en estos dos escenarios:

1. *En centros de monitorización de volcanes activos*, cuyo objetivo es controlar la evolución de la actividad sísmica asociada al volcán para evaluar una posible situación de riesgo. Suelen recibir datos de distintas estaciones localizadas en el entorno del cráter demandando un sistema que segmente los eventos sísmicos dentro del flujo continuo de datos y los clasifique en *tiempo cuasi-real* y de forma no supervisada, prestando atención a aquellos eventos precursores de erupciones o que conlleven un peligro directo para la población (como caída de ceniza, lahares, colapsos, flujos piroclásticos y ríos de lava).
2. En los *centros de investigación*, que realizan estudios geofísicos sobre una región o un periodo de actividad sísmica para construir modelos geológicos de una zona y que requieren un conteo del número, tipo y localización de eventos sísmicos para ello. La clasificación se realiza en un entorno supervisado sin la necesidad de ejecutarse en tiempo real pero sí demanda una alta tasa de eficiencia en el reconocimiento que aporte la fiabilidad necesaria para fundamentar el estudio.

En función de su objetivo los sistemas VSR necesitan distintas funcionalidades, si bien, es conveniente que que cumplan estas propiedades:

- **Alta eficiencia** de clasificación, incluso para sistemas con múltiples clases de alta variabilidad.
- **Robustez**. Un concepto bajo el que se engloban diversas cualidades:
  - *Insensibilidad* al ruido registrado en los sensores que se mezcla con los eventos a clasificar, tal que no se vea seriamente afectada la tasa de reconocimiento.
  - *Insensibilidad* respecto al tamaño y variabilidad de la base de datos de entrenamiento y a muestras atípicas (*outliers*) respecto el patrón representativo de una clase, como, por ejemplo, eventos mal etiquetados que se usan al construir los modelos. Asociado a la capacidad de regularización que tenga el sistema para controlar su complejidad (Figura 2.1.3.1).

  - *Estabilidad* respecto al corpus de evaluación, tal que la tasa de reconocimiento se mantenga alta independientemente del conjunto de datos a clasificar. Relacionado con el problema del sobre-entrenamiento de modelos (Sección A.4) y el compromiso ajuste-variabilidad (Sección 2.1.3.1). Fundamental en entornos no supervisados.

- **Fiabilidad del etiquetado**. En situaciones que involucren evaluación del riesgo, es conveniente que el sistema entregue junto con la clase a la que pertenece cada evento una medida de confianza de que dicho evento realmente pertenezca a esa clase.
- **Escalabilidad** del sistema. Es deseable que el tiempo empleado en la construcción de modelos y en la clasificación de eventos sea linealmente proporcional al tamaño de las bases de datos y al número de clases a distinguir.
- Reconocimiento de eventos sobre **flujos continuos** de datos que permita la detección y segmentación automática de los eventos, sin necesidad de herramientas externas al sistema.

La mayoría de estas características son convenientes para cualquier sistema de clasificación en general, si bien en el área de VSR la escalabilidad, el reconocimiento en continuo y la robustez del sistema son imprescindibles conforme a un reconocimiento no supervisado. Propiedades como la fiabilidad del etiquetado y la alta tasa de clasificación cobran más importancia en el ámbito del análisis con fines investigadores.

2.3. Técnicas actuales de clasificación de sismos

Los primeros estudios sobre la clasificación automática de señales sísmicas datan de principios de los años '70s como una evolución de los algoritmos de detección automática o *picking* ((Allen, 1978; Anderson, 1978)) de las ondas de llegada de los terremotos. Bouvier (1972) trabajó en su tesis con distintos clasificadores para discriminar entre explosiones nucleares y terremotos de origen natural. Muchos de los trabajos posteriores continúan esta temática (Ives, 1975; Chen, 1978a), empezando a estudiar el efecto que el conjunto de características puede tener sobre en la eficiencia de clasificadores sencillos basados en probabilidades y similitud (Chen, 1977, 1978b). En los años '80s el modelado estructural de patrones aparece en los algoritmos de detección y segmentación (Faure et al., 1984; Gaby and Anderson, 1984; Joswig, 1990). Liu and Fu (1983) modela sintácticamente los sismogramas para distinguir entre terremotos y explosiones artificiales.

A principios de la década de los '90s las ANNs ganan popularidad y empiezan a usarse como clasificadores de terremotos tectónicos (Dowla et al., 1990; Fortuna et al., 1991; Bowman and Dowla, 1992; Falsaperla et al., 1992; Murphy and Cercone, 1993) En el inicio del siglo XXI, las redes bayesianas se emplean como clasificadores estructurados que permiten el reconocimiento de registros continuos en tiempo real

de eventos sismo-volcánicos (Ohrnberger, 2001; Riggelsen et al., 2007; Benítez et al., 2007; Cortés et al., 2014). Wassermann et al. (2007) demuestran sobre un registro de un año, que un reconocedor basado en HMMs alcanza el mismo nivel de eficiencia que un técnico experto al etiquetar eventos inducidos por la sismicidad volcánica. A mitad de la década las máquinas de vectores soporte (Langer et al., 2006; Curilem et al., 2014b) empiezan a desplazar a las ANNs debido a su sencillez de diseño e implementación. Recientemente se está apostando por esquemas combinados de reconocimiento que persiguen compensar las deficiencias de los clasificadores clásicos y adaptarse a entornos multicanal de monitorización (Curilem et al., 2009; Bicego et al., 2013).

A continuación examinaremos los algoritmos actuales de clasificación de eventos sismo-volcánicos, presentando los resultados de los trabajos de las técnicas más representativas. Una cobertura completa queda fuera del alcance de este capítulo dado el creciente interés de la comunidad científica en torno a este área. Los algoritmos a analizar los agruparemos en:

- Sistemas basados en instancias: kNNs.
- Clasificadores discriminativos: SVMs.
- Clasificadores basados en el perceptrón: MLP, ANNs, RNNs.
- Modelos probabilísticos: Bayes, HMMs, PGMs.
- Combinación de métodos o *ensembles*.
- Algoritmos de clasificación no supervisada: kMeans, SOMs.

Para cada tipo de técnicas haremos unas tablas comparativas indicando: los autores de cada trabajo; las clases y eventos del corpus de datos; la representación escogida, la duración y componentes de los eventos vectores de características; el clasificador y su complejidad y, finalmente, la tasa de eficiencia alcanzada.

2.3.1. Clasificadores basados en instancias

El aprendizaje basado en instancias (*Instance-Based Learning - IBL*) o *aprendizaje por memoria*, clasifica un vector comparándolo con un conjunto datos o *instancias* que ha guardado en la memoria en su etapa de aprendizaje. Se basa en la idea de que en situaciones parecidas se toman las mismas decisiones. Ejemplos de IBL son los métodos basados en núcleos, los kNNs y las redes de funciones radiales o RBFN. En el caso de describir mediante características o atributos simbólicos las instancias, la técnica más popular es el razonamiento basado en casos (*Case Base Reasoning - CBR*). El IBL se estructura en:

1. **Aprendizaje**: se seleccionan las *instancias*, generalmente son los pares $(\mathbf{x}_i, c(\mathbf{x}_i))$ de la DB_{tr}, agrupándolas por clases. Alternativamente pueden guardarse descripciones o listas de propiedades de clases.

2. **Clasificación**: se compara mediante alguna medida de similitud el vector incógnita con las instancias memorizadas, asignándole la misma clase de la instancia *más similar* a él.

IBL postula sus hipótesis o modelado en función de las instancias, y no de modelos generales propuestos a priori cuyos parámetros deben ajustarse en la fase de entrenamiento. Pertenece a los llamados *algoritmos perezosos* (o *lazy learning*), que retrasan la carga de procesamiento de datos a la fase de evaluación. Sus principales *ventajas* son:

- Carácter *incremental* del aprendizaje. Basta con almacenar nuevas instancias.
- Carácter *local*: que hace al algoritmo más flexible en el modelado, capaz de trazar fronteras no lineales para solucionar problemas complejos de clasificación.

La clave del IBL recae en la elección del criterio de *similitud* entre instancias que debe adecuarse a cada problema concreto. La mayoría de los casos se opta por definir una *métrica* o *distancia* como medida inversa de la similitud. Las distancias se agrupan en:

continuas definidas analíticamente sobre números reales como la euclídea, de Mahalanobis, de Manhattan o de Chebychev

discretas que miden la diferencia entre secuencias de símbolos como la de Hamming o Levenshtein para cuantificar cómo de diferentes son 2 palabras.

Debido a su diseño original, el IBL acarrea una serie de importantes *inconvenientes*:

- Alto coste de memoria, proporcional al tamaño de base de datos de entrenamiento.
- Evaluación lenta, al tener que comparar con todas las instancias guardadas.
- Es poco robusto a los datos de entrenamiento, siendo sensible a datos espúreos y ruidosos que puede incorporar como instancias. Tiende al sobreentrenamiento.

2.3.1.1. Clasificadores basados en vecindades del espacio de características

Los casos más relevantes son los kNNs y la estimación local o por vecindades de funciones de probabilidad o *núcleos de Parzen*. Los algoritmos más utilizados en el VSR son los basados en kNNs y sus variantes.

***k*-vecinos más cercanos (*k-Nearest Neighbors - kNNs*)** El algoritmo kNN (Silverman and Jones, 1989) define una vecindad $V_{\mathbf{x}}(k)$ entorno al vector $\mathbf{x}$ a clasificar que contiene los k-elementos $\{\mathbf{x}_k, c(\mathbf{x}_k)\}$ de entrenamiento más cercanos a $\mathbf{x}$ según la función distancia $d(\mathbf{x}, \mathbf{x}_k)$. Como vemos intuitivamente en la Figura 2.3.1, a $\mathbf{x}$ se

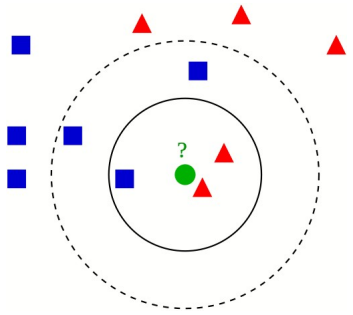


Figura 2.3.1.: Clasificación por vecindades (kNN). Dados los k-elementos más cercanos al vector a clasificar (circulo verde), se le asocia la clase *azul* o *roja* que tenga más elementos dentro de dicha vecindad. Si $k = 5$ el vector incógnita se asigna a la clase azul. Para $k = 3$ se etiquetaría como elemento de la clase roja.

le asigna la clase $w_x \in \{w_c\}_{c=1:C}$ más frecuente, que contiene más elementos dentro de la vecindad $V_{\mathbf{x}}(k)$:

$$d_{kNN:d}(\mathbf{x}) \equiv \arg\max_{w_c} \left\{ \sum_{\mathbf{x}_k \in V_{\mathbf{x}}(k)} \alpha_k \delta(w_c, c(\mathbf{x}_k)) \right\} \tag{2.3.1}$$

donde $\delta(a, b) = \delta_{ab}$ es la delta de Kronecker, de salida nula excepto cuando $a = b \Rightarrow \delta(a, b) = 1$. Los pesos $\{\alpha_k\}$ se usan para ponderar el efecto de los vecinos $\mathbf{x}_k \in V_{\mathbf{x}}(k)$ conforme a la distancia a $\mathbf{x}$ con el propósito de que los datos más cercanos influyan más en la Ecuación 2.3.1 que los más lejanos. kNN adolece de las mismas ventajas e inconvenientes que la familia de algoritmos de la que procede, los clasificadores basados en instancias. En concreto, existen varios puntos clave en su diseño:

- *La elección de k*. Si el número de datos en la vecindad es alto el sistema es menos sensible al ruido y espúreos. Si es muy alto, la evaluación será más lenta y las regiones más densas tienden a acaparar las menos densas.
- *Almacenamiento de demasiadas instancias*. Ralentiza la clasificación, aumenta la sensibilidad al ruido y al sobre-entrenamiento. Existen técnicas muy efectivas para seleccionar los mejores ejemplos como la indexar las instancias en función de la distancia entre ellas mediante árboles (*kd-trees*).
- *Alta sensibilidad al grupo de características*. Debido al uso de métricas en la función de decisión, es poco robusto a características irrelevantes que deben ser eliminadas mediante técnicas de reducción de dimensionalidad (Capítulo 4) o ponderadas convenientemente.

Debido a su sencillez de implementación, la clasificación por vecindades se ha usado desde los inicios del VSR (Chen, 1977, 1978a; Liu and Fu, 1983).

Modelado de lenguaje mediante *k*-vecinos (*kNN_dM*) y distancia de Levenshtein Liu and Fu (1983) presenta una interesante propuesta clasificando sintácti-

camente los sismogramas para distinguir entre terremotos y explosiones artificiales mediante un modelado lingüístico de las señales. Para ello halla de forma no supervisada 13 primitivas básicas en sus eventos y con solo 2 características logra unos resultados de reconocimiento de más del 90 %, comparable con otros autores que usan el mismo clasificador por vecindades pero 30 años después.

TEST ABIERTO + EVENTOS AISLADOS				
autores	*eventos/clases*	*vector {dur/avance}[s]*	*clasificador*	*eficiencia*
Liu and Fu (1983)	321/2	log.E+ZCC{6/6}	kNN.sintaćtico(13)	91 %
Galli et al. (2009)	291/3	8.stats(DWT){:/:}	kNN(10)/Bayes	90 %
Cárdenas-Peña et al. (2013)	948/4	15.mixtas {4/0.5}	kNN(3)	~76 %

Cuadro 2.3.1.: *VSR mediante modelos basados en similitud de patrones en el espacio de características.*

2.3.2. Clasificadores basados en análisis discriminativo

El objetivo es obtener las fronteras de decisión analíticamente mediante una *función discriminante* que tome como variables una combinación (lineal o no) de las características del espacio de descripción de los observables. Sus ejemplos más populares son las máquinas de vectores soporte (SVMs) y el vetusto análisis discriminante de Fisher (Subsubsección 4.4.2.4).

2.3.2.1. Máquinas de Vectores Soporte (*Support Vector Machines* - *SVM*s)

Entre los más populares clasificadores basados en análisis discriminativo se encuentran los *clasificadores de margen máximo*, cuyo máximo exponente son las SVM de Vapnik (2000), una generalización del perceptrón que colocan las fronteras entre clases de tal forma que se maximice los márgenes entre ellas conforme a una métrica concreta. Los discriminadores lineales SVM junto con el uso del *kernel trick* (núcleos o funciones de similitud que transforman el espacio de características original $\Omega_{\mathbf{X}}$ en uno Ω_ϕ de mayor dimensión pero donde es más fácil separar linealmente las clases, de Aizerman et al., 1964; Boser et al., 1992) forman un esquema capaz de solucionar problemas complejos no lineales y cuyos modelos tienen asegurada la convergencia iterativamente a un mínimo global (Ver la Figura 2.3.2). Por este motivo son muy populares en diversas áreas de reconocimiento de señales (Schuldt et al., 2004; Lotte et al., 2007).

Formalmente la función $d_{SVM:K}(\mathbf{x})$ de clasificación de las SVM es un simple discriminador lineal que separa 2 clases mediante el hiperplano $\mathbf{w}^T\phi(\mathbf{x})+b=0$ de máximo margen en el espacio Ω_Φ. Dado un conjunto de entrenamiento DB_{tr} consistente, es posible encontrar una transformación $\phi : \Omega_X \rightarrow \Omega_\Phi$ que convierta a Ω_Φ en un espacio

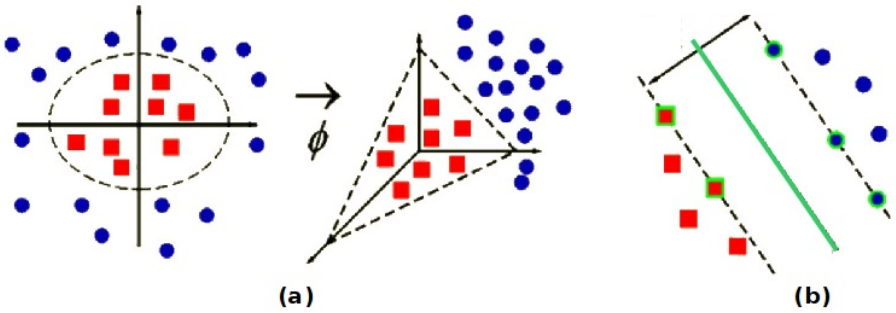


Figura 2.3.2.: SVM(kernel=ϕ) como clasificadores de máximo margen. (a) Transformación $\phi[\mathbf{\Omega_X}]$) desde el espacio $\Omega_\mathbf{X}$ de características original al espacio $\Omega_\mathbf{\Phi}$ linealmente separable de mayor dimensión. **(b)** Las fronteras en el nuevo espacio se construyen para maximizar el margen entre las clases. (Figura original en Masotti et al., 2008)

de Hilbert linealmente separable a expensas de aumentar lo necesario su dimensión. Definiendo una función de similitud o núcleo $K(\mathbf{x}_1, \mathbf{x}_2) \equiv \phi(\mathbf{x}_1) \cdot \phi(\mathbf{x}_2) = \boldsymbol{\phi}_1 \cdot \boldsymbol{\phi}_2$ posibilita no tener que hallar explícitamente $\phi(\mathbf{x})$ con lo que $d_{SVM:K}(\mathbf{x})$ puede desarrollarse como:

$$d_{SVM:K}(\mathbf{x}) \equiv sgn\{\mathbf{w}^T\phi(\mathbf{x}) + b\} = sgn\left\{\sum_s y_s\alpha_s K(\mathbf{x}_s, \mathbf{x}) + b\right\} = \tag{2.3.2}$$

$$= sgn\left\{\sum_s y_s\alpha_s \boldsymbol{\phi}_s \cdot \boldsymbol{\phi} + b\right\} \tag{2.3.3}$$

donde $sgn(\Delta)$ representa el signo de Δ, $\{y_s\} = \pm 1$ los valores numéricos asociados a las etiquetas de cada vector $\{\mathbf{x}_s\}$, $\{\alpha_s\}$ son escalares (multiplicadores de Lagrange) y b es el desplazamiento o *bias* de la frontera de decisión. Los núcleos más usados son el lineal $K_{lin}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1 \cdot \mathbf{x}_2$ si no hay transformación al espacio Ω_Φ y el de base radial gausiana (*Radial Basis Function - RBF*) $K_{RBF}(\mathbf{x}_1, \mathbf{x}_2) = \exp\left\{-0{,}5\,\|(\mathbf{x}_1 - \mathbf{x}_2)/\sigma\|^2\right\}$. El vector de pesos $\mathbf{w}$, o *weights*, es un vector normal asociado al hiperplano, cuya norma se busca minimizar en el proceso de entrenamiento para maximizar el margen y es una combinación lineal de los *vectores soporte* $\{\mathbf{x}_s\}$ (o $\{\boldsymbol{\phi}_s\}$ en Ω_Φ), esto es, los $s-$puntos de entrenamiento que caen en los márgenes $\mathbf{w}^T\phi(\mathbf{x}) + b = \pm 1$.

Cortes and Vapnik (1995) introducen técnicas de regularización que permiten una cierta tolerancia a errores de la base de datos de entrenamiento definiendo las SVM de *margen blando* y reduciendo así el sobreajuste. La robustez de las SVM de margen blando las convirtieron en las SVM estándar; son insensibles al sobre-entrenamiento, altamente escalables y tienen una alta capacidad de generalización. Su principal inconveniente es su lentitud en el entrenamiento y evaluación comparado con otros reconocedores, una elección de sus parámetros de diseño poco intuitiva y que requieren de técnicas adicionales para poder discriminar entre más de dos clases.

Curilem et al. (2014b) hace una interesante propuesta reconociendo 4 clases en segmentos de longitud variable, consiguiendo en torno al 80 % de exactitud con una parametrización mixta de 5 componentes detallada en Curilem et al. (2009). También estudia las 63 combinaciones posibles de 6 características (las mismas 5 de Curilem et al. (2009) y la fase instantánea de San-Martin et al. (2010)) extraídas directamente del sismograma y de su espectro.

TEST ABIERTO + EVENTOS AISLADOS				
autores	eventos/clases	vector {dur/avance}[s]	clasificador	*eficiencia*
Masotti et al. (2006)	425/4	62.STFT{:/:}	SVM(lineal)	94 %
Langer et al. (2009)	425/4	62.STFT{:/:}	SVM(rbf)	95 %
Curilem et al. (2014b)	1622/4	6.mixtas{variable/:}	SVM(rbf)	~80 %

Cuadro 2.3.2.: *Trabajos de reconocimiento automático mediante máquinas de vectores soporte (SVMs).*

2.3.3. Redes neuronales artificiales (*ANNs*)

Las redes neuronales pertenecen al grupo de los algoritmos bio-inspirados que pretenden emular la topología y forma de pensar del cerebro. Se basan en el concepto de neuronas de McCulloch and Pitts (1943) como unidades básicas de procesamiento y en la forma de interconexión entre ellas o topología para generar una respuesta ante un estímulo dado. Cada neurona representa una combinación lineal de entradas cuya salida binaria se discretiza en función de un valor umbral. Una respuesta originada por varias neuronas alimentadas por las mismas entradas componen la idea de *perceptrón* de Rosenblatt (1962), que posteriormente evolucionó al *perceptrón multicapa* (al que comúnmente nos referimos al utilizar el término red neuronal artificial) y luego dio paso a un variado tipo de estructuras y sistemas híbridos. Con la topología adecuada, las ANNs son *aproximadores universales*: pueden modelar cualquier función continua. Esto las hace unos clasificadores muy flexibles a la hora de abordar problemas complejos generales de los que a priori se tiene poca información.

Las ANNs han sido la técnica estándar durante los últimos 15 años en el reconocimiento de eventos sismo-volcánicos (Orozco-Alzate et al., 2012). De hecho, es probablemente el clasificador más usado en el área del reconocimiento de patrones. Arrastran el problema del diseño de la topología y un aprendizaje lento y poco robusto que suelen compensar con buenos resultados de clasificación si están bien construidas.

Perceptrón multicapa . El perceptrón multicapa (*Multi-Layered Perceptron - MLP*) es un tipo de ANN hacia adelante (*FeedForward Neuronal Nets - FF.NN*s) o separable en etapas de computación (*capas*) que se originó a partir del perceptrón

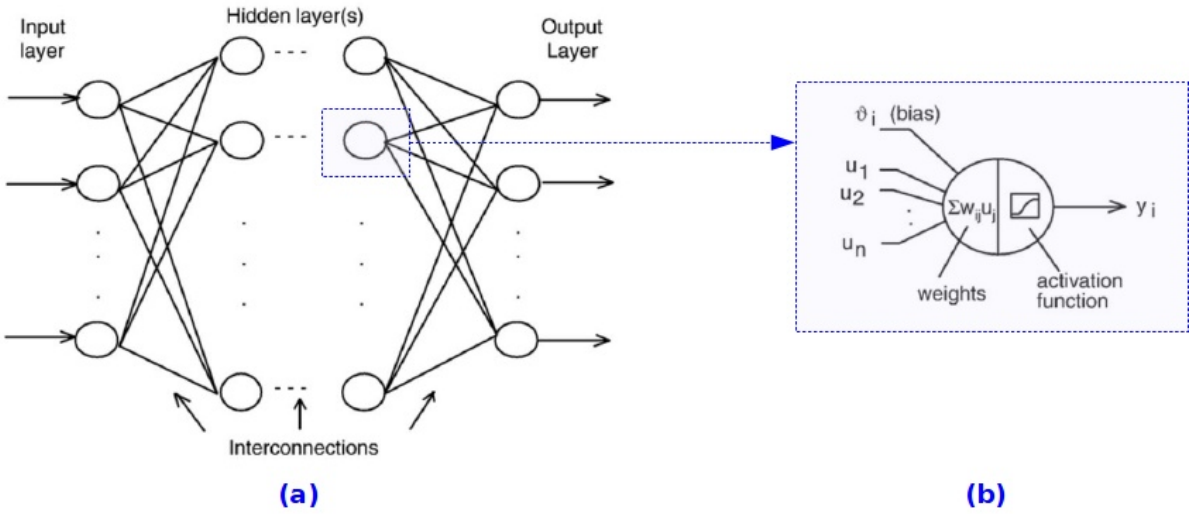


Figura 2.3.3.: Estructura del perceptrón multicapa. (a) Topología: 1 capa de entrada, varias ocultas y una de salida. **(b)** Neuronas (perceptrón simple) en cada capa. (Figura original en Langer et al., 2006)

incluyendo una o varias *capas ocultas* entre la entrada y la salida, lo que le otorga la capacidad de separar clases en espacios no lineales de características. La Figura 2.3.3 (a) muestra la topología típica en el caso de usar el MLP para clasificar: se asocia cada neurona de la capa de entrada a cada componente del vector de $K-$características y cada neurona de salida a una de las $C-$clases. El número H de capas o *layers* ocultas y el número $n(l)$ de nodos n en cada capa oculta l se escoge conforme a la complejidad y propiedades de cada tarea de reconocimiento. Cada capa tiene asociada un nivel umbral de ajuste o *bias* b_l concreto que limita la activación de sus neuronas. La Figura 2.3.3 (b) esquematiza una neurona n de una capa l, donde la señal $\mathbf{u}_{l,n}$ que es combinación lineal de las salidas de las neuronas de la capa $l-1$ anterior junto con la señal de ajuste alimentan su función activación $a(\mathbf{u}_{l,n})$. La función $d_{ANN:MLP}(\mathbf{x})$ de decisión asocia al vector $\mathbf{x}$ de entrada la clase correspondiente a la neurona c de la capa $H+2$ de salida que obtenga un mayor valor de activación:

$$d_{ANN:MLP}(\mathbf{x}) \equiv \arg\max_c\{a(\mathbf{u}_{H+2,c})\} = \arg\max_c\{a\left(\mathbf{w}_{H+1,c}\cdot\mathbf{a}_{H+1}+b_{H+2}\right)\} \tag{2.3.4}$$

donde $\mathbf{w}_{l,n}$es el vector de pesos que ponderan las conexiones de las neuronas de la capa l con la neurona n de la capa siguiente $l+1$ y $\mathbf{a}_l = \{a(\mathbf{u}_{l,1}), a(\mathbf{u}_{l,2}), ..., a(\mathbf{u}_{l,n(l)})\}$ es un vector que contiene las salida de las neuronas de la cada l.

Hay dos tipos básicos de funciones de activación $a(\mathbf{u}_{l,n})$, las más antiguas históricamente de tipo *escalón* se corresponden con las neuronas tipo *McCulloch-Pitts* que generan salidas binarias entre $[-1,1]$ o $[0,1]$. Actualmente se prefiere funciones tipo *escalón suavizado* que sean de derivadas continuas para facilitar el proceso de entrenamiento tales como sigmoides (tangente hiperbólica o la función logística) o

funciones *RBF* radiales (combinaciones lineales de núcleos, de salida proporcional a la distancia desde la entrada a un centro). La elección de la función de activación es fundamental: lo que realmente habilita al MLP para separar clases de forma no lineal es que el tipo de función de activación escogida no sea lineal. Nótese que no es necesario que todas las capas usen la misma función de activación.

El algoritmo clásico para entrenar las redes FF.NNs y concretamente el MLP es el de *retropropagación* (*backpropagation*) o propagación hacia atrás de errores. Básicamente es un problema de optimización de una función de coste o error (Subsubsección 2.1.3.1) que utiliza la técnica del descenso en gradiente para llegar iterativamente a un mínimo local.Actualmente se usan otros algoritmos más rápidos y robustos como Quasi-Newton, Levenberg-Marquardt o el gradiente conjugado escalado.

Otras arquitecturas de ANNs La gran versatilidad de las ANNs han originado multitud de variantes (Figura 2.3.4) y otros tantos métodos de entrenamiento. Respecto al VSR, las más interesantes son aquellas que permiten el reconocimiento de patrones en continuo, como las ANNs recurrentes y las ANNs recursivas. También son relevantes las redes de Kohonen o mapas auto-organizativos que permiten la clasificación de manera no-supervisada.

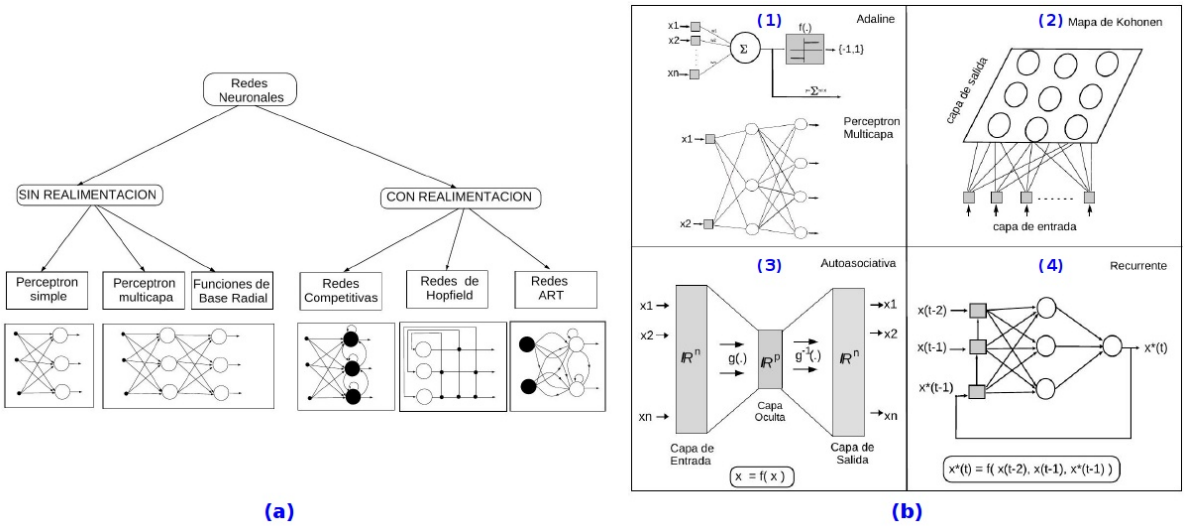


Figura 2.3.4.: Ejemplos de ANNs. (a) Taxonomía simplificada. **(b)** Topología según utilidad: **(1)** Perceptrón multicapa y Adaline para clasificación. **(2)** Mapas de Kohonen para agrupamiento. **(3)** Red autoasociativa para reducción de dimensionalidad. **(4)** ANN recurrente para modelado secuencial. (Figuras originales de Jiménez, 2003)

TEST CERRADO + EVENTOS AISLADOS				
autores	eventos/clases	vector {dur/avance}[s]	clasificador	*eficiencia*
Avossa et al. (2003)	353/2	15.PCA{24/24}	MLP(1c.3n)	100 %

TEST ABIERTO + EVENTOS AISLADOS				
autores	eventos/clases	vector {dur/avance}[s]	clasificador	*eficiencia*
Avossa et al. (2003)	353/2	15.PCA{24/24}	MLP(1c.3n)	98 %
Scarpetta et al. (2005)	881/2	6.LPC{2.56/?}+1.env.A{1/1}	MLP(1c)	98 %
Langer et al. (2009)	425/4	62.STFT{:/:}	MLP(6c)	82 %

Cuadro 2.3.3.: *VSR basado en redes neuronales.*

2.3.4. Clasificadores probabilísticos

Nos ocuparemos ahora de los sistemas que específicamente aprenden un modelo estadístico para cada tipo de eventos estimando su distribución de probabilidad, como los clasificadores bayesianos, los sistemas que combinan gausianas para estimar la probabilidad (GMMs) o los modelos gráficos (redes bayesianas y de Markov) que expresan la relación entre variables aleatorias mediante nodos conectados entre sí.

Clasificadores de Parzen (núcleos bayesianos) La familia de *núcleos bayesianos* o clasificadores de Parzen usan la regla de Bayes como función de decisión (Sección 2.1.2) y un modelado no paramétrico de las funciones de densidad $p(\mathbf{x}|w_c)$ de las clases $\{w_c\}$ basado en vecindades o núcleos (*Kernel Density Estimation - KDE*):

$$p_{KDE}(\mathbf{x}|w_c) \equiv \frac{1}{N_c B^D} \sum_{\mathbf{x}_c \to w_c} K\left(\frac{\mathbf{x}-\mathbf{x}_c}{B}\right) \equiv \frac{1}{N_c B^D} \sum_{\mathbf{x}_c \to w_c} K(\mathbf{z}) \tag{2.3.5}$$

donde D es la dimensión del espacio de descripción, N_c es el número de datos de entrenamiento $\{\mathbf{x}_c\}$ etiquetados con la clase w_c, $K(\mathbf{z})$ es una función núcleo o *kernel* no negativa, con media nula y cuya integral es la unidad. B un factor de suavizado en la estimación conocido como el ancho de banda, regulador del compromiso ajuste-variabilidad. La variable $\mathbf{z}$ es, por tanto, una versión estandarizada de $\mathbf{x}$. El criterio MAP de la estadística bayesiana junto la estimación no paramétrica definen la forma de la función de decisión de Parzen:

$$d_{B:MAP:KDE}(\mathbf{x}) \equiv \arg\max_{w_c} \{p_{KDE}(\mathbf{x}|w_c)P(w_c)\} \tag{2.3.6}$$

La estimación KDE es capaz de modelar distribuciones complejas del espacio de características y es bastante robusta a la implementación concreta de la función $K(\mathbf{z})$. Sin embargo requiere alto tiempo de procesamiento y la elección del parámetro B, que no siempre es fácil. En el caso de escoger un núcleo basado en la distribución

normal, existen aproximaciones concretas para B como la ley de Silverman o la de Scott. Una evolución de los KDE son los *KDE adaptativos*, que varia el valor de B dependiendo de localización de los datos de entrenamiento o del vector $\mathbf{x}$ a evaluar. Incluso se puede variar el ancho de banda de forma independiente para cada característica. En el caso de escoger B tal que $K(\mathbf{z})$ sea uniforme en una vecindad, se llega a una generalización de los kNN.

En el reconocimiento de eventos sismo-volcánicos la estimación no paramétrica de densidad por vecindades es usada indirectamente en clasificadores como los SVMs (*kernel trick*) y las ANNs (en las funciones de activación RBF). Hoogenboezem (2010) en su estudio de reducción de dimensionalidad con datos del volcán Nevado del Ruiz consigue con núcleos de Bayes un 81 % de eficacia.

Clasificadores de componentes gausianas (*GMM*s) Al igual que los clasificadores de Parzen, los GMMs (*Gaussian Mixture Models - GMMs*) se basan en el criterio MAP dentro de la estadística bayesiana pero estiman la probabilidad $p(\mathbf{x}|w_c)$ asociada a cada clase w_c mediante un GMM: una combinación lineal de G componentes gausianas multivariadas $N(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$. Para asignar el vector $\mathbf{x}$ a alguna de las clases $\{w_c\}$ se aplica:

$$d_{B:MAP:GMM}(\mathbf{x}) \equiv \arg\max_{w_c} \{P(w_c) p_{GMM}(x|w_c)\} = \arg\max_{w_c} \left\{ P(w_c) \sum_{g=1:G} \alpha_g N_g(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \right\} \tag{2.3.7}$$

Cada componente normal o gausiana $N_g(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ está ponderada por su correspondiente escalar α_g y tiene la siguiente forma:

$$N_g(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = \frac{1}{(2\pi)^{K/2}|\boldsymbol{\Sigma}_g|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_g)\boldsymbol{\Sigma}_g^{-1}(\mathbf{x} - \boldsymbol{\mu}_g) \right\} = \tag{2.3.8}$$

$$= \left((2\pi)^K |\boldsymbol{\Sigma}_g| \right)^{-1/2} \exp\left\{ -\frac{1}{2} d_M^2(\mathbf{x}, \boldsymbol{\mu}_g) \right\} \tag{2.3.9}$$

siendo G el número de componentes gausianas multivariadas de la combinación lineal pesadas por los coeficientes $\{\alpha_g\}$, K el número de características de los vectores, $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ la media y matriz de covarianza de la $g-$componente gausiana del modelo $GMM_G(\mathbf{x}_t)$ y $d_M(\mathbf{x}, \boldsymbol{\mu})$ es la distancia de Mahalanobis entre los vectores $\mathbf{x}$ e $\boldsymbol{\mu}$. En el Subsección 3.3.1 construiremos paso a paso especificando el proceso de entrenamiento y evaluación un sistema basado en GMMs capaz de reconocer secuencias de vectores.

Los GMM proporcionan un método estadístico sencillo capaz de modelar espacios de características no lineales que mediante la correcta elección del parámetro G controlan la complejidad y el nivel de ajuste del modelo a los datos. En la Subsección 3.3.1 detallaremos el uso de los GMMs como clasificadores. En el área de VSR, Avossa et al. (2003) los utiliza para modelar 2 clases de manera no-supervisada en un espacio

de vectores PCA, Álvarez et al. (2011) emplea los GMMs como clasificadores sencillos para seleccionar las mejores características mediante al algoritmo discriminativo de De la Torre et al. (1997) aplicado a los eventos volcánicos. Como aplicación práctica de la evolución del algoritmo DFS original de Álvarez et al. (2011) abordada en Capítulo 4 y del sistema en paralelo propuesto en el Capítulo 5 de esta tesis, Cortés et al. (2014) presentan un sistema en paralelo basado en GMMs que se adapta por independiente para cada tipo de evento en los llamados *canales*, donde se escoge la mejor configuración y la más efectiva descripción para cada clase posibilitando su análisis y clasificación por separado.

Modelos (probabilísticos) gráficos (*PGMs*). Los modelos gráficos (*Probabilistic Graphical Models - PGMs*) permiten esquematizar la interdependencia probabilística variables aleatorias mediante un grafo de nodos conectados por lazos. Cada variable es representada por un nodo, y cada lazo determina el tipo de relación entre nodos (Figura 2.3.5). Los lazos dirigidos jerarquizan el gráfico entre padres u origen del lazo e hijos o destino. Si el origen y el destino es el mismo nodo se habla de grafos cíclicos. En caso contrario de grafos acíclicos. Los PGM buscan *factorizar* la probabilidad conjunta de las variables en términos más sencillos. Se subdividen en dos vertientes principales que representan distintos tipos de relaciones: redes bayesianas y campos de Markov (Koller and Friedman, 2009).

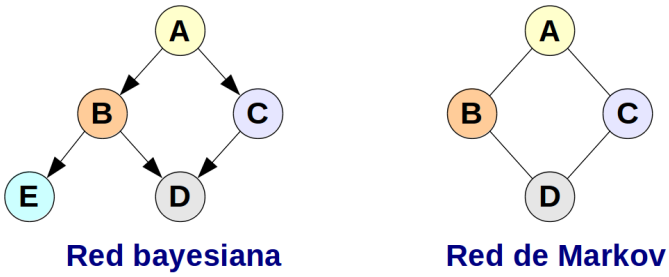


Figura 2.3.5.: PGMs como grafos de dependencia probabilística entre variables (A, B, C, D)**.** En la red bayesiana los nodos se conectan mediante lazos dirigidos y acíclica (no hay lazos formando un ciclo sobre el mismo nodo). En ella $P(A, B, C, D, E) = P(A)P(B|A)P(C|A)P(D|B, C)P(E|B)$. La red de Markov es no dirigida y puede ser acíclica. De su grafo se descompone en 4 cliqués que factorizan $P(A, B, C, D, E) = Z^{-1}\phi_1(A, B)\phi_2(B, D)\phi_3(D, C)\phi_4(C, A)$.

- **Las redes bayesianas** o **redes de creencia** (*Bayesian Networks - BNs*) son modelos gráficos acíclicos y dirigidos que descomponen la probabilidad conjunta como productos de probabilidades condicionales entre cada nodo X_n y sus respectivos nodos padres $\mathbf{X}_{pa(n)} = (X_{pa(n),1}, ..., X_{pa(n).P})$ tal que $P(X_1, ..., X_N) = \prod_{n=1:N} P(X_n|\mathbf{X}_{pa(n)})$ (Friedman et al., 1997). Las redes bayesianas dinámicas (*Dynamic Bayesian Networks - DBNs*) extienden la funcionalidad de las redes bayesianas para poder modelar secuencias de variables.

Pueden verse como una generalización de los HMMs y los de filtros de Kalman (Murphy, 2002).

- **Las redes de Markov** o **campos aleatorios de Markov** (*Markov Random Fields - MRFs*) son grafos no dirigidos, que pueden ser cíclicos o no, cuya factorización se hace en términos de funciones potenciales $\phi_c(\mathbf{X}_c)$ asociadas a cliqués o subgráficos completos cuyos nodos $\mathbf{X}_c = (X_{c,1}, ..., X_{c.C})$ están todos interconectados entre sí $P(X_1, ..., X_N) = Z^{-1} \prod_c \phi_c(\mathbf{X}_c)$, con Z conocida como la función de partición (Ver Figura 2.3.5).

A pesar de su potencialidad, los PGMs no han sido muy explotados en las señales sismo-volcánicas. Parpoula et al. (2013) construyen una red bayesiana estructurada mediante árboles (*Tree Augmented Naive Bayes - TAN*) para discriminar entre terremotos de gran magnitud o de magnitud pequeña. Para ello usa una gran base de datos con 10333 registros descritos inicialmente por 11 parámetros de los cuales se seleccionan los 9 mejores. Riggelsen et al. (2007) separan 286 registros en 2 clases, llegadas de ondas P y ruido ambiente, mediante una red bayesiana dinámica consiguiendo una tasa de efectividad del 95 %.

Modelos ocultos de Markov (*HMM*s) Los HMMs son clasificadores probabilísticos diseñados para reconocer patrones estructurados en secuencias. Sus inicios se remontan al final de los años '60 a partir de los trabajos de Baum (Baum and Petrie, 1966; Baum et al., 1967, 1970) como una evolución de las cadenas de Markov llegando a convertirse en la técnica más usada en el área del reconocimiento del habla (Rabiner and Juang, 1986, 1993; Young et al., 2006). El diseño de los HMMs permiten modelar estructuras secuenciales de una manera eficiente mediante una topología de S estados interconectados entre sí que van generando una secuencia $\mathbf{O} = \{\mathbf{O}_t\} = \mathbf{x} = \{\mathbf{x}_t\}_{t=1:T}$ de vectores observables. El resultado es un doble modelado estocástico, en el espacio de características y en el espacio que marca la secuencialidad del evento (en nuestro caso el tiempo). La Figura 2.3.6 muestra como un HMM de 3 estados emisores (Q_2, Q_3, Q_4) van generando secuencialmente un sismograma. Cada uno de estos estados Q_s tiene asociada una probabilidad $b_s(\mathbf{x}_t)$, normalmente un GMM, de emitir un vector de características $\mathbf{x}_t$ en un instante de tiempo t. El modelado temporal viene intrínsecamente ligado a los coeficientes (a_{ij}) de transición entre los estados $Q_i \rightarrow Q_j$. Al atravesar el el HMM desde el estado inicial Q_1 al final Q_5 siguiendo una secuencia concreta de estados $\{q_t\}$ se emite su correspondiente secuencia de vectores $\{\mathbf{x}_t\}$. La manera clásica de entrenar cada HMM es mediante el algoritmo Baum-Welch, un estimador de máxima verosimilitud (Sección 2.1.3.2) que solo asegura una convergencia a un mínimo local.

El camino $\{q_t\}$ de estados a seguir para atravesar el modelo está inicialmente *oculto* y se descubre o *decodifica* en la fase de evaluación. En ella se crea una red de decodificación construida a partir de los modelos previamente entrenados y de reglas gramaticales que gobiernan como se interconectan los modelos entre sí para formar estructuras. De esta forma se posibilita el reconocimiento sobre flujos *continuos* de

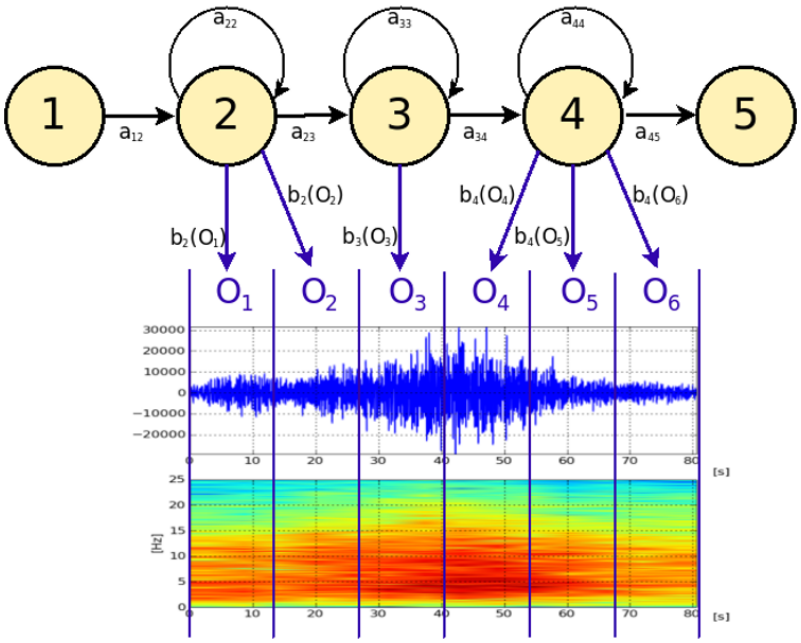


Figura 2.3.6.: Generación en un HMM de la secuencia de observables $\mathbf{O} = \{\mathbf{O}_t\} = \{\mathbf{O}_1, \mathbf{O}_2, \mathbf{O}_3, \mathbf{O}_4, \mathbf{O}_5, \mathbf{O}_6\}$ por la secuencia de estados $\mathbf{q} = \{q_t\} = \{q_1 = Q_2, q_2 = Q_2, q_3 = Q_3, q_4 = Q_4, q_5 = Q_4, q_6 = Q_4\}$. La probabilidad de que un estado Q_s genere el vector $\mathbf{O}_t$ se evalúa mediante la distribución de probabilidad $b_s(\mathbf{O}_t)$.

datos (Subsubsección 2.1.4.1) que el sistema trata como una estructura secuencial de eventos, cada un o de ellos generado por un modelo. Asumida la aproximación de Viterbi (Sección 3.3.2.1), la secuencia oculta de estados $\mathbf{q_x}$ que genera toda la secuencia $\mathbf{x}$ de vectores observados dado el macro-modelo M se decodifica mediante:

$$\mathbf{q_x} = \arg \max_{\forall \mathbf{q}=\{q_1,\ldots,q_T\} \in Q_\mathbf{x}} \{p(\mathbf{q}|\mathbf{x}; M)\} = \arg \max_{\forall \mathbf{q}=\{q_1,\ldots,q_T\} \in Q_\mathbf{x}} \{p(\mathbf{x}, \mathbf{q}; M)\} \quad (2.3.10)$$

con $Q_\mathbf{x}$ el conjunto de todas las posibles secuencias $\mathbf{q}$ que generen $\mathbf{x}$. En la Subsubsección 3.3.2.1 presentaremos de forma algo más detallada un sistema VSR basado en HMMs.

A pesar del éxito obtenido por los HMMs en ámbitos que requieren el procesamiento en tiempo real de patrones estructurados (Lee and Choi, 2003; Potamianos et al., 2004; Gales and Young, 2008), las principales limitaciones que ofrecen están relacionadas con el no cumplimiento de las suposiciones teóricas asumidas al aplicarse en determinadas situaciones prácticas (Rabiner, 1989):

- *Independencia estadística de los observables*: Se asume que los vectores de características se generan de forma independiente entre sí, lo que permite descomponer la probabilidad del suceso conjunto $(\mathbf{x}_1, .., \mathbf{x}_N)$ en el producto de probabilidades a priori de cada observable $\mathbf{x}_i$, tal que $P(\mathbf{x}_1, .., \mathbf{x}_N) = \prod_{i=\{1:N\}} P(\mathbf{x}_i)$. En sismología, por ejemplo, claramente existe una dependencia asociada a la

secuencialidad temporal de los frentes de onda; los observables que describan ondas tipo P se espera que lleguen antes de aquellos asociados con las ondas tipo S.

- *Procesos de Markov*: Un modelo de Markov (oculto o no) requiere que la probabilidad de que un observable sea generado en un estado q_u en el instante t solo dependa del estado q_v que estaba activo en el instante anterior $t-1$. O, equivalentemente, se asume que el estado futuro inmediato al cual va a evolucionar el modelo solo depende del estado actual en el que se encuentre. Si bien es útil para definir una estructura secuencial en las señales sísmicas, de nuevo, no tiene por qué cumplirse estrictamente: estados anteriores al estado actual pueden influir en el estado futuro al que evolucionará el evento.

TEST CERRADO + SEÑAL CONTINUA				
autores	eventos/clases	vector {dur/avance}[s]	clasificador	*eficiencia*
Benítez et al. (2007)	~2100/5	39.LFCC {4/0.5}	HMM(11e*24g)	91 %

TEST ABIERTO + EVENTOS AISLADOS				
autores	eventos/clases	vector {dur/avance}[s]	clasificador	*eficiencia*
Hoogenboezem (2010)	532/4	4.DCT(STFT) {2.56/1.92}	Parzen(1g)	81 %
Álvarez et al. (2011)	475/10	14.mixtas {4/0.5}	GMM(16g)	86 %
Cárdenas-Peña et al. (2013)	948/4	15.mixtas {4/0.5}	HMM(7e*4g)	88 %
Parpoula et al. (2013)	10333/2	9.stats {:/:}	TAN(11e)	89 %
Cortés et al. (2014)	328/5	5.mixtas {2/1}	GMM.paralelos(4g)	89 %

TEST ABIERTO + SEÑAL CONTINUA				
autores	eventos/clases	vector {dur/avance}[s]	clasificador	*eficiencia*
Riggelsen et al. (2007)	286/2	~256.log(Morlet.CWT) {15/15}	DBN(5c*6n)	95 %
Ibáñez et al. (2009)	1048/2	39.LFCC {4/0.5}	HMM(17e*16g)	84 %
Cortés et al. (2009a)	4687/10	39.LFCC {4/0.5}	HMM(~15e*15g)	81 %
Cortés et al. (2009b)	6788/10	39.LFCC {4/0.5}	HMM(~15e*12g)	78 %

Cuadro 2.3.4.: *VSR mediante clasificadores probabilísticos.*

En la última década, los HMM junto con los SVM han experimentando un auge importante en el área del VSR. Especial interés tienen los trabajos en los que se usan sobre registros continuos, como Benítez et al. (2007), Cortés et al. (2009a) y Beyreuther and Wassermann (2008). Cortés et al. (2009b) los aplica sobre un corpus de datos de más de 330 horas y 6788 eventos registrados por 4 estaciones de largo y corto periodo en los volcanes de Colima y Popocatéptl para evaluar la robustez de un sistema multi-estado. Ibáñez et al. (2009) estudia la portabilidad del sistema evaluando por separado y conjuntamente datos del Etna y el Estrómboli.

Aparte de como clasificadores, en sismología volcánica los HMM también se han utilizado para:

- Detectores (segmentación) de terremotos (Beyreuther et al., 2012).
- Modelado del estado de la actividad sísmica de un volcán y predicción de erupciones (Oliveros et al., 2008; Boué et al., 2015).

2.3.5. Combinación de técnicas de clasificación

La combinación de sistemas de reconocimiento de señales sismo-volcánicas pretende aumentar la eficiencia de clasificación a partir de distintas estrategias complementarias :

- *Usar distintos canales con el mismo tipo de clasificador*: Se analizan varios sismogramas de un mismo evento registrado en varias estaciones o en diferentes canales (sistemas *multi-flujo* o *multi-stream*). Debido a los efectos locales de sitio, la dispersión de energía y a la diversidad y localización de los sismómetros, la combinación de la información dada por cada clasificadores de cada sismograma puede llevar a una mejora general en la tasa de reconocimiento. Beyreuther and Wassermann (2008) solo consideran que se ha detectado un sismo si 2 de cada 3 estaciones lo clasifican como no-ruido con una probabilidad 3 veces superior a la del evento ruido.
- *Mezclar señales de distinta naturaleza correspondientes originadas por la misma fuente*. Una extensión lógica de la anterior técnica muy usada en la monitorización del estado del volcán y evaluación del riesgo (Carniel et al., 2006), que analizan, por ejemplo, señales sísmicas, químicas o de deformación del cono volcánico correspondientes a un mismo proceso de inserción de material en la cámara magmática. Se corresponde con los denominados sistemas *multi-modales* (Dupont and Luettin, 2000; Potamianos et al., 2004).
- *Reconocer el mismo canal con distintos tipos de clasificadores* (los propiamente conocidos como sistemas *combinación de clasificadores*): Se busca que los clasificadores se complementen entre sí, por ejemplo, para ser más eficientes ante el compromiso entre desviación y variabilidad (Sección 2.1.3.1). **?**Beyreuther and Wassermann (2008) promedian la probabilidad de cada clase con 3 modelos distintos de DHMMs, cada uno con distinto número de estados, para incrementar la robustez del sistema

En la Tabla 2.3.5 presentamos algunos trabajos que usan una combinación de técnicas aplicadas al mismo flujo de datos detalladas a continuación.

Modelado temporal en HMMs mediante transductores de estados finitos. Uno de los inconvenientes atribuidos a los HMM es un inadecuado modelado temporal de eventos proporcional a una exponencial negativa. Exportando técnicas de síntesis de voz, Beyreuther and Wassermann (2011) modelan explícitamente la duración de 4 clases sísmicas mediante distribuciones gausianas y el espacio de características mediante semi-modelos ocultos de Markov (*Hidden Semi-Markov Models* -

HSMMs, Oura et al. (2006)). Puesto que el algoritmo de Viterbi no es compatible con los HSMMs, la decodificación se realiza mediante transductores de estados finitos (*Weighted Finite-State Transducers - WFST*, de Mohri et al. (2002)). Cuando se evalúa en continuo en test ciego el sistema durante un mes completo, el esquema híbrido consigue, según sus autores, mejorar en torno al 40 % respecto la base clásica de los HMMs a costa de incrementar el coste computacional en un factor $x10$ en la decodificación. En un test cerrado de eventos aislados el esquema WFST{HSMM} alcanza clasificar correctamente el 88 % de los datos frente al 86 % de los HMM.

Espacios de puntuación generativa (métodos generativos incrustados). Los métodos generativos incrustados (*Generative Embedding Methods*) o espacios de puntuación generativa (*Generative Score Spaces*) pretenden incrementar la capacidad de clasificación de técnicas generativas usando modelos discriminativos. Se proyecta la información proporcionada por un modelado generativo para incrustarla en un nuevo espacio de características que es el punto de partida para usar métodos de clasificación discriminativos.

Bicego et al. (2013) presentan un interesante trabajo en el que combinando datos de diversas estaciones situadas en el volcán de Galeras, comparan distintas técnicas de proyección (FSE, LLE, SE y TE) de la información dada por HMMs en un espacio de características que es la entrada de un clasificador SVM(rbf). Con ello, transforman patrones secuenciales en patrones no estructurados, en el que independientemente de su duración, los eventos son descritos por un vector de características del mismo tamaño. Se comprueba como el esquema híbrido SVM(rbf){HMM+FSE} mejora ligeramente al sistema base basado en HMMs, incrementando además la capacidad de generalización al construir modelos con datos de una estación y reconocer eventos de otra.

Discriminantes lineales sobre algoritmos genéticos. Los algoritmos genéticos (*Genetic Algorithms - GA*) son métodos de búsqueda heurísticos que pretenden replicar los métodos de la selección natural en el área del aprendizaje automático (John, 1992; Fogel, 2006). Comprenden una parte de las técnicas inspiradas en la teoría de la evolución (*Evolutionary Algorithms - EA*), que seleccionan a un conjunto de individuos dentro de una población mediante estrategias de herencia, mutación, selección y cruce o recombinación. Son usados para solucionar problemas de optimización y búsquedas no lineales, en particular para selección de características (Bhanu and Lin, 2003; Zamalloa et al., 2008).

Orlic and Loncaric (2010) combinan las 3 componentes de eventos registrados en una red de 20 estaciones para construir una base de datos sobre la que un esquema incrustado de selección de 16 a 10 características (Subsección 4.1.5) basado en algoritmos genéticos. Posteriormente una sencillo discriminante lineal clasifica entre explosiones artificiales o terremotos tectónicos consiguiendo una eficacia del 85 %.

Redes neuronales auto-optimizadas. La idea es utilizar algoritmos de aprendizaje automático que permitan optimizar las redes neuronales. Dicha optimización abarca varios niveles de diseño de la ANN; la selección de características, la topología (entradas, número de capas ocultas y neuronas en cada una de ellas) y las técnicas para entrenar más eficientemente a la red.

TEST CERRADO + EVENTOS AISLADOS				
autores	*eventos/clases*	*vector {dur/avance}[s]*	*clasificador*	*eficiencia*
Beyreuther and Wassermann (2011)	450/4	12.mixtas{[1]~58/7}	HMM(6e*1g)	86 %
Beyreuther and Wassermann (2011)	450/4	12.mixtas{~58/7}	WFST{HSMM(6e*1g)}	88 %

TEST ABIERTO + EVENTOS AISLADOS				
autores	*eventos/clases*	*vector {dur/avance}[s]*	*clasificador*	*eficiencia*
Curilem et al. (2009)	1033/4	5.stats{30/30}	MLP(1c*14n){V.GA}	93 %
Orlic and Loncaric (2010)	200/2	10.stats+PSD{:/:}	Discriminante Lineal	85 %
Bicego et al. (2013)	400/4	~64.STFT{1/0.5}	SVM(rbf){HMM(2e*1g)}	85 %

TEST ABIERTO + SEÑAL CONTINUA				
autores	*eventos/clases*	*vector {dur/avance}[s]*	*clasificador*	*eficiencia*
Beyreuther and Wassermann (2008)	[2]~69/3	11.mixtas{3/0.05}	DHMM(~4e*1g)	62 %
Beyreuther and Wassermann (2011)	[3]~718/4	12.mixtas{~58/7}	HMM(6e*1g)	55 %
Beyreuther and Wassermann (2011)	~718/4	12.mixtas{~58/7}	WFST{HSMM(6e*1g)}	80 %

Cuadro 2.3.5.: *Reconocimiento automático de señales sismo-volcánicas usando esquemas combinados de clasificación.*

Curilem et al. (2009) usan algoritmos genéticos para configurar un perceptrón multicapa (MLP) y seleccionar las 5 mejores características de entre 8 posibles estadísticos extraídos los dominios temporal y espectral. Implementando la selección de Vasconcelos (V.GA), entrenan el MLP con una capa oculta de 14 neuronas mediante el algoritmo de Levenberg–Marquardt consiguiendo una eficiencia del 93 % al discriminar entre 4 clases.

2.3.6. Clasificación no supervisada (*clustering*)

Los algoritmos no supervisados son aquellos que no necesitan una partición de entrenamiento con eventos ya etiquetados para ejecutarse. Su utilidad básica es analizar las señales para descubrir estructuras o características que permitan asignar los datos con propiedades parecidos en grupos lo más homogéneos posibles. En nuestro caso, este agrupamiento o *clustering* se puede identificar con una clasificación no supervisada en el caso de que estos grupos coincidan con las clases de significado

sismo-volcánico. Podemos distinguir diversos propiedades en los métodos de clustering:

- *Algoritmos jerárquicos*. Van generando grupos menos generales a partir de conjuntos más homogéneos.
- *Métodos soft* o *hard*. En función de que a cada evento de le asigne o no una probabilidad de pertenecer a su grupo.
- *Algoritmos disyuntivos*. En el caso de que una muestra pueda pertenecer a más de un grupo.
- *Dependencia de los datos*. Se denomina métodos dependientes a aquellos que necesitan realizar un análisis previo sobre unos datos (de entrenamiento) para recabar información necesaria antes de poder ejecutarse.

Nótese que una vez creados los modelos, cualquier sistema supervisado puede actuar de manera no supervisada, por lo que en general, los algoritmos no supervisados se utilizan más como como una herramienta de análisis exploratorio y minería de datos más que de clasificación.

2.3.6.1. Clasificación no supervisada mediante el algoritmo k-medias (kMeans)

El método estrella de los métodos de clustering es el algoritmo *k-medias*. Directamente relacionado con los métodos basados en instancias (Subsección 2.3.1), predefine k grupos, $\{w_1, ..., w_k\}$, en el espacio de características donde se define una medida de similitud $m(\mathbf{x}, \boldsymbol{\mu}_k)$ entre un evento $\mathbf{x}$ y $\boldsymbol{\mu}_k$, la *media* o *centroide* del grupo w_k. En un proceso iterativo de optimización (Subsubsección 2.1.3.1) se actualizan conjuntamente los centroides de cada grupo y los eventos que pertenecen al grupo. Finalmente, un evento $\mathbf{x}$ se asigna al grupo w_k más cercano mediante la función:

$$d_{kM:m}(\mathbf{x}) \equiv \arg\min_{w_k} \{m(\mathbf{x}, \boldsymbol{\mu}_k)\} \tag{2.3.11}$$

Esta técnica comparte la mayoría de las ventajas y desventajas del algoritmo kNN. El gran inconveniente es la sensibilidad a la inicialización de los centroides y elementos que pertenecen a cada grupo en la primera iteración. Se utiliza en múltiples áreas dentro del aprendizaje automático, también para inicializar clasificadores supervisados (Young et al., 2006), aunque también como función de decisión en sistemas basados en similitud (Langer et al., 2009).

2.3.6.2. Clasificación no supervisada mediante mapas auto-organizativos (*SOM*)

Los mapas auto-organizativos (*Self Organized Maps - SOM*) o redes neuronales de Kohonen (Kohonen, 1988, 2001) agrupan patrones parecidos en nodos vecinos, de

modo que la capa de salida forma una estructura en 2D o 3D con el propósito de visualizar e interpretar el espacio de características de forma ordenada, lo que se consigue asignando un color concreto a los nodos que representan patrones similares.

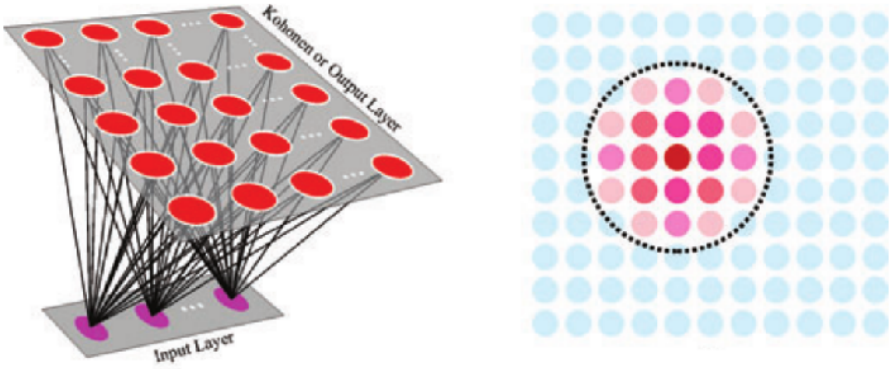


Figura 2.3.7.: Mapas auto-organizativos (Self-Organizing Maps - SOM). Son redes neuronales que agrupan patrones similares para reducir la dimensionalidad del espacio de entrada y poder visualizarlo en la capa de salida de la red. Figura de Langer et al. (2009)

En el VSR han sido muy utilizados para agrupar señales similares entre sí (Esposito et al., 2006; Langer et al., 2009; Messina and Langer, 2011), descubrir y seleccionar las mejores características para describir eventos (Köhler et al., 2009) y, ocasionalmente, para clasificar de manera no supervisada. Esposito et al. (2008a) usan mapas auto-organizativos para relacionar señales de muy baja frecuencia (*Very Long Period - VLPs*) en el Stromboli con las aberturas donde se producen explosiones. (Köhler et al., 2009) los emplean para distinguir entre las llegadas de ondas P, S e intervalos de ruido.

CLASIFICACIÓN NO SUPERVISADA + EVENTOS AISLADOS				
autores	*eventos/clases*	*vector {dur/avance}[s]*	*clasificador*	*eficiencia*
Langer et al. (2009)	425/4	62.STFT{:/:}	kMeans(Adapt. Dispersion)	74 %
Köhler et al. (2009)	152/3	20.mixtas{:/:}	SOM	66 %

Cuadro 2.3.6.: *Trabajos de clasificación automática de eventos sismo-volcánicos usando técnicas no supervisadas.*

2.4. Discusión sobre los sistemas VSR

Una vez hecha la introducción al reconocimiento de patrones y el aprendizaje automático en la Sección 2.1 y vista su aplicación práctica a nuestras señales de interés (Sección 2.2), para finalizar el tema actual haremos una pequeña reflexión del estado del arte de los sistemas VSR centrándonos en los requisitos exigidos sobre estos desde un punto de vista geofísico, relacionado con su integración y uso en los centros de monitorización de volcanes activos y de los retos abiertos que aún hay que satisfacer. Finalmente extraeremos algunas conclusiones respecto el estado actual en que está la tecnología desarrollada y las directrices a seguir según las necesidades expuestas.

2.4.1. Comparación entre técnicas de clasificación

Una competición directa a través de los resultados presentados en la Sección 2.3 es totalmente injustificada pues si bien sirven para hacerse una idea general de la evolución temporal de los clasificadores y la complejidad que aborda cada trabajo, incluso tareas que son aparentemente igual de complejas pueden no ser comparables debido al nivel de similitud que exista entre las clases a discriminar. Una evaluación entre sistemas solo tiene sentido si usan la misma base de datos. En este sentido, hay que resaltar que los eventos registrados en volcanes, aparte de incluir a los generados por la tectónica de placas, están sometidos a más efectos de sitio y a una topografía más variada, además de requerir más clases en los que agrupar las señales, por lo que el proceso de reconocimiento es inherentemente más complejo que en el caso de limitarnos a los sismos de origen exclusivamente tectónico.

En la Tabla 2.4.1 realizamos una comparación cualitativa de las principales técnicas de clasificación de señales sismo-volcánicas. Los criterios de evaluación especificados las columnas centrales están basados en los requerimientos de los sistemas VSR detallados en la Subsección 2.2.3: la *robustez* frente al corpus de datos con los que se construyen los modelos, la capacidad de usar fronteras *no lineales* entre clases en el espacio de características, la *escalabilidad* o capacidad del sistema para mantener su eficiencia conforme se incrementa el tamaño de las bases de datos y por último la capacidad de clasificar eventos sobre un flujo continuo de registro. La información mostrada es solo orientativa; dentro de cada categoría de clasificadores existen muchas implementaciones que pueden variar bastante (lo que se representa con el símbolo ☯) respecto las características generales expuestas, sobre todo en las ANNs, DBNs y la combinación o *ensembles* de clasificadores.

Como vimos en la Sección 2.3 todas las técnicas consiguen una suficiente tasa de clasificación a excepción de los algoritmos de agrupamiento o *clustering* si bien el objetivo principal de estos está más relacionado con descubrir el tipo y clase de relaciones existente entre eventos que con alcanzar una alto porcentaje de reconocimiento. En general los sistemas lineales tienen el inconveniente de que no pueden

clasificador	*modelado*	*robusto*	*no lineal*	*escalable*	*continuo*	*ventajas*	*inconvenientes*
kNNs	por vecindades	☹	☺	☹	☹	sencillez	escalabilidad
LDA	maximiza separabilidad	☺	☹	☹	☹	rapidez	sencillez
SVMs	maximiza márgenes convergencia global	☺	☺	☺	☹	robustez escalabilidad	lentitud binarios
ANNs	optimización iterativa topología flexible	☯	☺	☯	☯	generalidad	entrenamiento diseño
Bayes	estima PDFs / clase	☯	☯	☯	☹	sencillez	entrenamiento
Bayes-KDE	núcleos-Parzen	☯	☺	☯	☹	cuasi-óptimo	lentitud
GMMs	mezcla gausianas	☯	☺	☺	☯	sencillez	entrenamiento
HMMs	MLE:Baum-Welch	☯	☺	☺	☺	diseño	entrenamiento
DBNs	estructurado	☹	☺	☯	☺	topología	entrenamiento
mixtos	mezcla técnicas	☯	☯	☺	☯	flexibilidad	complejidad
cluster	kMedias, SOM	☯	☺	☯	☹	no-supervisado	baja eficiencia

Cuadro 2.4.1.: *Comparación cualitativa (☺=bueno, ☯=mixto, ☹=malo) de técnicas de clasificación de eventos sismo-volcánicos.*

modelar bien la complejidad que suele aparecer cuando tenemos datos de alta variabilidad. Para compensar este comportamiento se impone la técnica del *kernel trick*, o mapeo a espacios de características de mayor dimensión, usada sobre todo en SVMs y que permite usar discriminadores lineales. Las redes bayesianas en general y los HMMs en particular tienen la enorme ventaja de su naturaleza estructurada, lo que los hace especialmente útil en el reconocimiento en continuo. La equivalencia de los estados de los HMM con las diferentes llegadas de ondas de los eventos los hace especialmente atractivos y flexibles en su diseño; flexibilidad que comparten con las ANNs, pero en estas la elección de la topología es más complicada al no existir una relación directa con las características de las señales sísmicas. Ambos también comparten una delicada etapa de aprendizaje y una baja robustez a los datos de entrenamiento. La robustez de las SVMs frente a las ANNs junto a su capacidad de regularización y escalabilidad hacen de ellas una opción muy más interesante en el VSR aún requiriendo técnicas adicionales para poder discriminar entre más de 2 clases. Su lentitud respecto otros reconocedores no es un problema en VSR, pero sí lo es que no sean directamente aplicables sobre datos en continuo. Otra alternativa que está incrementando su popularidad son la combinación de clasificadores o *ensembles*, que permite abordar varios problemas a la vez: el compromiso ajuste-variabilidad, la clasificación en múltiples canales simultáneamente y la escalabilidad .

En cuanto al comportamiento de los reconocedores frente a aspectos concretos debemos destacar:

- En general, el *conjunto de características* que describen los eventos tiene un papel muy importante en la elección del clasificador. Los SVMs y ANNs alcanzan buenas tasas de reconocimiento con un vector de tamaño grande, mientras

que los clasificadores bayesianos tienden a funcionar mejor con vectores pequeños (Kotsiantis, 2007). Los kNNs y ANNs son muy sensibles a características superfluas.

- La *capacidad de generalización* es crucial en el caso de evaluación en abierto y no supervisada. Directamente relacionada con el compromiso entre el ajuste y variabilidad (Sección 2.1.3.1), los clasificadores sencillos (Naive Bayes, kNNs...) tienden a sufrir una mayor desviación respecto la predicción ideal, lo que equivale a una menor tasa de reconocimiento en abierto, pero son más robustos frente a fluctuaciones en los datos o espúreos. Los clasificadores más complejos (ANNs, SVMs, BNs, DTs, ...) son capaces de modelar estructuras más variables en los datos, pero corren el riesgo de estar sobre-ajustados a su corpus de entrenamiento y tener poco éxito sobre otros datos. Los métodos como la regularización para controlar la adecuada relación entre desviación y variabilidad, o de otra manera, el grado de complejidad en el modelado son muy importantes para aplicar los sistemas de manera no supervisada.
- Las *respuesta frente al ruido* en los registros puede ser capital dependiendo del emplazamiento del sensor. Los kNNs son muy sensibles al ruido que distorsionan las medidas de similitud usadas al clasificar. El ruido también tiene un especial efecto negativo en las ANNs, incrementando su riesgo al sobre-entrenamiento. Los clasificadores heurísticos y lógicos son más robustos en entornos ruidosos.

2.4.2. Conclusiones en torno a los sistemas VSR

Como *conclusión* a este capítulo cabe destacar que escogiendo una estrategia adecuada de configuración y dada una adecuada base de datos de entrenamiento (en términos de fiabilidad de etiquetado y tamaño) los resultados obtenidos con unos clasificadores o con otros llegan a ser bastante parecidos. Tal y como se aprecia en la comparación realizada en la Sección 2.3, la eficiencia de clasificación fácilmente ronda el 90 % incluso para trabajos de hace más de 30 años. La mayoría de los clasificadores pueden reconocer en tiempo real con una estación debido básicamente a las bajas frecuencias de muestreo (sobre $F_s = 100\ [Hz]$ o $F_s = 200\ [Hz]$ a lo sumo) que se usan con las señales sísmicas. La elección de una técnica concreta en lugar de otra debe hacerse más bien atendiendo a motivos como la correcta evaluación de la complejidad de la tarea de reconocimiento, el coste computacional que conlleva y el tipo, estructura y tamaño de los datos disponibles. Analizando las necesidades y limitaciones expuestos en la Sección 2.2 se concluye que los esfuerzos de investigación deben focalizarse en:

1. *Mejorar el etiquetado manual de bases de datos.* Una correcta clasificación del corpus de entrenamiento es crucial en el área del reconocimiento supervisado, concretamente para algunas técnicas que son muy sensibles a outliers. Un etiquetado tentativo semi-supervisado que ayude al técnico experto (García

et al., 2010) parece ser la mejor solución hasta ahora, al menos hasta que los sistemas no supervisados evolucionen lo suficiente para alcanzar una eficiencia de clasificación aceptable. Otra posible alternativa es construir bases de datos de alta fiabilidad o *maestras* para entrenar los modelos como se hace en Cortés et al. (2014). Es de vital importancia observar los sismogramas del mismo evento en más de una estación para evitar efectos de propagación y de sitio que lleven a una errónea clasificación por parte del experto (Sección 1.1.2).

2. Proporcionar *bases compartidas de datos*, de libre disposición entre la comunidad científica, como se hace en otras áreas (Garofolo et al., 1993; Patterson et al., 2002; Schuldt et al., 2004) con el objetivo de poder comparar distintos enfoques y avanzar adoptando técnicas comunes. Esta es una petición demandada ya por varios autores dedicados al reconocimiento de eventos sismo-volcánicos Beyreuther et al. (2012); Orozco-Alzate et al. (2012).

3. *El reconocimiento en tiempo real en continuo en varios canales y estaciones de monitorización a la vez*, que permita un rápido análisis de la sismicidad y la evaluación precoz del riesgo para la población (Bowman and Dowla, 1992; Benítez et al., 2007; Riggelsen et al., 2007; Beyreuther and Wassermann, 2008). En este sentido, llama la atención el poco uso de modelos que específicamente modelen la estructura de los eventos como los basados en redes probabilísticas. De hecho, en un estudio comparando distintos clasificadores modelando una gran base de datos de 10333 sismos, Parpoula et al. (2013) aconsejan el uso de algoritmos estructurados basados en árboles de decisión frente a SVMs, ANNs y BNs. Otro punto a tener en cuenta es el tiempo de respuesta del reconocedor, limitado inferiormente por la duración del avance de los segmentos en los que se particiona los datos al parametrizarlos. Algunos algoritmos incluso no particionan los datos, reconociendo fichero a fichero, lo que puede retardar en minutos los resultados e, incluso, comprometer las decisiones en caso de situaciones de alerta.

4. *Clasificación no supervisada de eventos* que permita clasificar señales en un volcán de forma satisfactoria con un sistema cuyos modelos hayan sido construidos con señales de otro volcán, o del mismo volcán pero en diferente periodos de actividad (Cortés et al., 2009b; Álvarez et al., 2010).

5. *Correcta gestión del riesgo volcánico* en varios niveles; desde la integración con los sistemas de monitorización de los observatorios hasta la capacidad de discriminación y especialización del sistema reconocedor de detectar aquellos eventos que sean especialmente relevantes para la evaluación del riesgo sobre poblaciones (Giacinto et al., 1997; Carniel et al., 2006; Cortés et al., 2009a, 2014).

La combinación de estos puntos nos sirven como base y motivación de las propuestas realizadas en este trabajo de tesis como desarrollaremos en el Parte II.

Parte II.

SISTEMA VSR DE RECONOCIMIENTO CONTINUO PROPUESTO

3. Sistema de clasificación de referencia

En este capítulo implementaremos un sistema de reconocimiento en continuo basado la tecnología actual de modelos ocultos de Markov (HMMs) para reconocer señales sísmicas registradas en los volcanes. Dicho sistema clásico o en serie (*Serial System Architecture - SSA*) será utilizado como punto de referencia para compararlo con las evoluciones teóricas que presentaremos en el Capítulo 5 e implementadas en el Capítulo 6 sobre una arquitectura en paralelo (*Parallel System Architecture - PSA*).

El paralelismo descrito por Ohrnberger (2001) entre las señales sísmicas registradas en los volcanes y la voz humana insta a la comunidad científica a adaptar las técnicas de reconocimiento automático de voz o *speech* (Rabiner and Juang, 1993; De la Torre et al., 1997; Young et al., 2006). No solo se comparte propiedades de las señales, si no que además requerimientos de procesado en tiempo real. Esto hace que se adopten tanto los esquemas de parametrización (basados en coeficientes cepstrales) como los clasificadores más populares en voz (HMMs y sus derivados) como un novedoso enfoque en los sistemas VSR (Alasonati et al., 2006; Benítez et al., 2007; Beyreuther and Wassermann, 2008; Ibáñez et al., 2009), que junto a los SVMs están desplazando progresivamente a otros clasificadores ya existentes basados en ANNs.

Con el objetivo de evaluar la robustez de nuestra propuesta sobre distintos escenarios, utilizaremos dos corpus de datos relativos a volcanes con diferentes estructuras y clases: el volcán de la isla Decepción en la Antártida y el volcán de Colima en México, que describiremos en la Sección 3.1. A partir de esquemas inspirados en el reconocimiento de voz representaremos de forma más compacta las señales volcánicas (Sección 3.2). En la Sección 3.3 estudiaremos los clasificadores probabilísticos que usaremos para reconocer señales sobre registros continuos (HMMs) y para el análisis de la reducción de dimensionalidad (GMMs) que abordaremos en el Capítulo 4. Posteriormente analizaremos el problema de la evaluación de los resultados y del sistema desde un punto de vista geofísico y seleccionaremos la métrica más adecuada a nuestro propósito (Sección 3.4). Finalmente, en la Sección 3.6 construiremos las bases de datos maestras para implementar experimentalmente en la Sección 3.7 el sistema de referencia y obtener los resultados base.

3.1. Origen y adquisición de datos: volcanes de Decepción y Colima

En esta sección describiremos el tipo de señales que utilizaremos para construir nuestro sistema de referencia. Una descripción más general de los eventos se presenta en el Capítulo 1. Estos registros conformarán dos corpus de datos iniciales, una para el volcán de Decepción y otra asociada al volcán de Colima con las que en la Sección 3.6 crearemos las respectivas base de datos *maestra*s, que serán las elegidas para generar los modelos de reconocimiento. Este es un concepto clave en el proceso de diseño de nuestros sistemas: la elección de unas bases de datos pequeñas, pero de alta fiabilidad que permitan construir modelos simples pero robustos, a costa quizás de perder algo de capacidad para describir estructuras complejas de datos.

Sistemas de adquisición: corto periodo frente a banda ancha. En base a nuestra experiencia, el uso de un sistema u otro de adquisición no condiciona demasiado los resultados de reconocimiento, ni siquiera los distintos filtros de ganancia que estos aplican sobre los datos. De hecho, todos nuestros sismogramas se formatean con un sencillo esquema de señales digitalizadas con números enteros de 2 bytes muestreadas a 50 Hz. ($i2.50Fs$). Donde si aparecen problemas graves es cuando la señal se satura porque el margen dinámico del sensor es insuficiente o si la SNR es muy baja y aparece mucho ruido solapado con los eventos. En sensores de 3 componentes disponer de los dos canales superficiales extra puede suponer una ventaja, sin embargo, la mayoría de los sistemas VSR solo usan el sismograma vertical, al igual que haremos nosotros.

Los efectos negativos susceptibles de ser inducidos por los sistemas de adquisición pueden ser minimizados mediante un buen acondicionamiento y descripción de la señal al ser convertida en una secuencia de vectores de características, tal y como veremos en la Sección 3.2.

3.1.1. Volcán de la isla de Decepción

Decepción es una isla emergida sobre la caldera de un volcán activo que forma parte del archipiélago de las Shetland del Sur situadas entre Sudamérica y el noroeste de la Península Antártica (ver la Figura 3.1.1). Siendo uno de los tres volcanes antárticos es objeto de numerosos estudios científicos e, incluso, destino turístico. La isla se encuentra deshabitada, excepto en el verano austral que acoge a grupos de investigadores en sus dos bases científicas, la española Gabriel de Castilla y la argentina Base Decepción. Decepción es el volcán más activo del estrecho de Bransfield, una de las regiones antárticas de mayor actividad sismo-volcánica que ha erupcionado al menos en seis ocasiones en los últimos 200 años. Actualmente pueden observarse numerosos eventos sísmicos e hidrotermales.

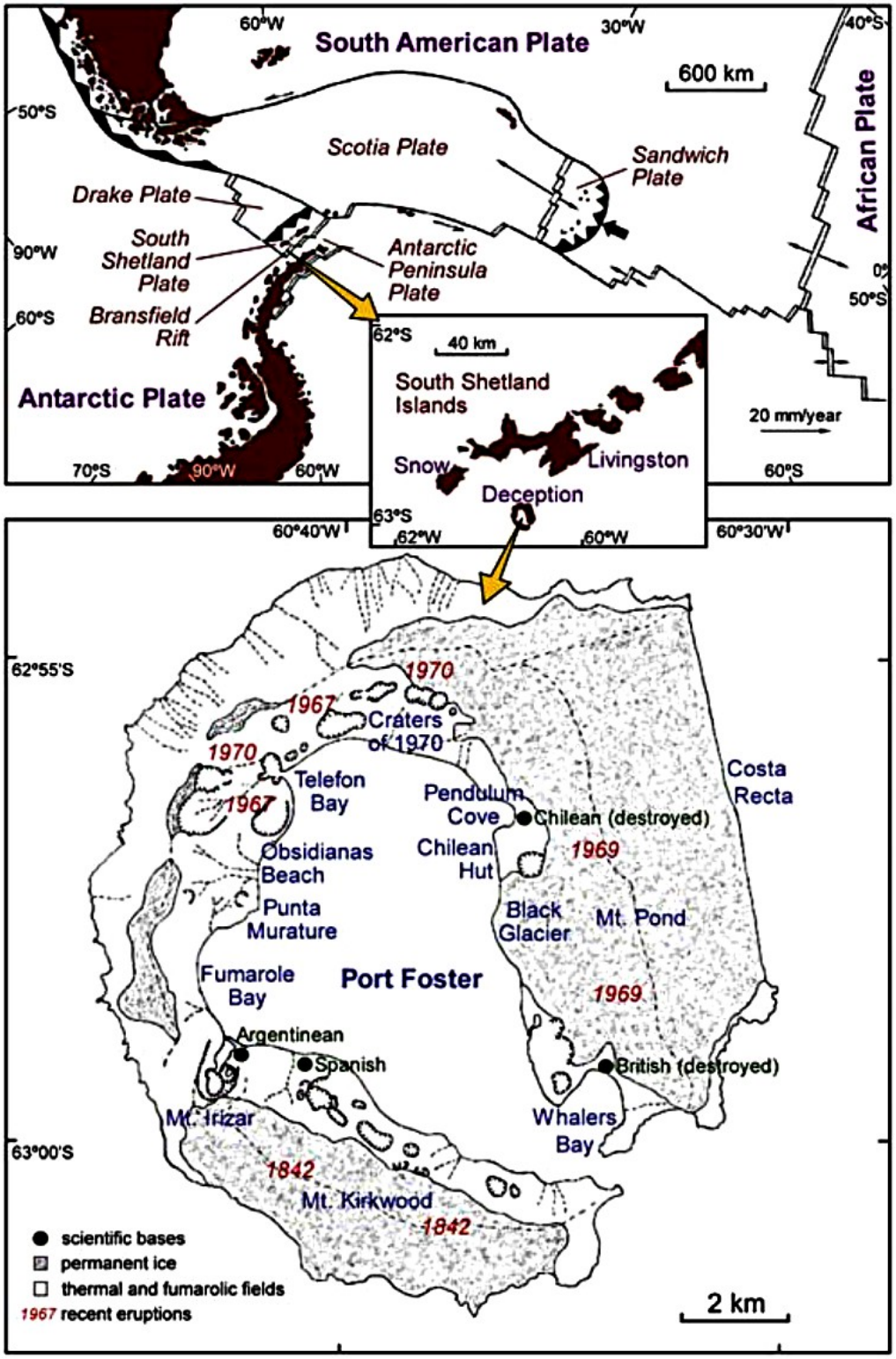


Figura 3.1.1.: Volcán Decepción: localización y sismicidad. Situación geográfica de las Islas Shetland del Sur, donde se localiza la isla Decepción. En la figura superior se indican las placas tectónicas que confluyen en la fosa oceánica de Bransfield. Abajo se detallan en *rojo* las erupciones históricas en Decepción, en *azul* los lugares de interés y en *negro* las bases científicas instaladas (Carmona et al., 2012).

3.1.1.1. Tectónica y sismicidad histórica

La isla Decepción es actualmente la caldera de un estratovolcán con un diámetro basal de 30 km. que se eleva 1400 m. sobre el fondo marino sobresaliendo unos 540 m. respecto el nivel del mar en su pico más alto, el monte Pond. La parte emergida de la isla tiene forma de herradura de unos 15 km. de diámetro con la bahía central formada por la inundación de la caldera del volcán tras una enorme erupción(mapa inferior de la Figura 3.1.1). La bahía Puerto Foster tiene un tamaño de 6 x 10 km. en su superficie y una profundidad máxima de 190 m. siendo accesible desde el mar a través de una brecha en los muros de la caldera de unos 150 m. de ancho conocida como los Fuelles de Neptuno. Los glaciares cubren aproximadamente la mitad de la superficie de la isla, concentrándose principalmente en el este (monte Pond) y sur (monte Kirkwood).

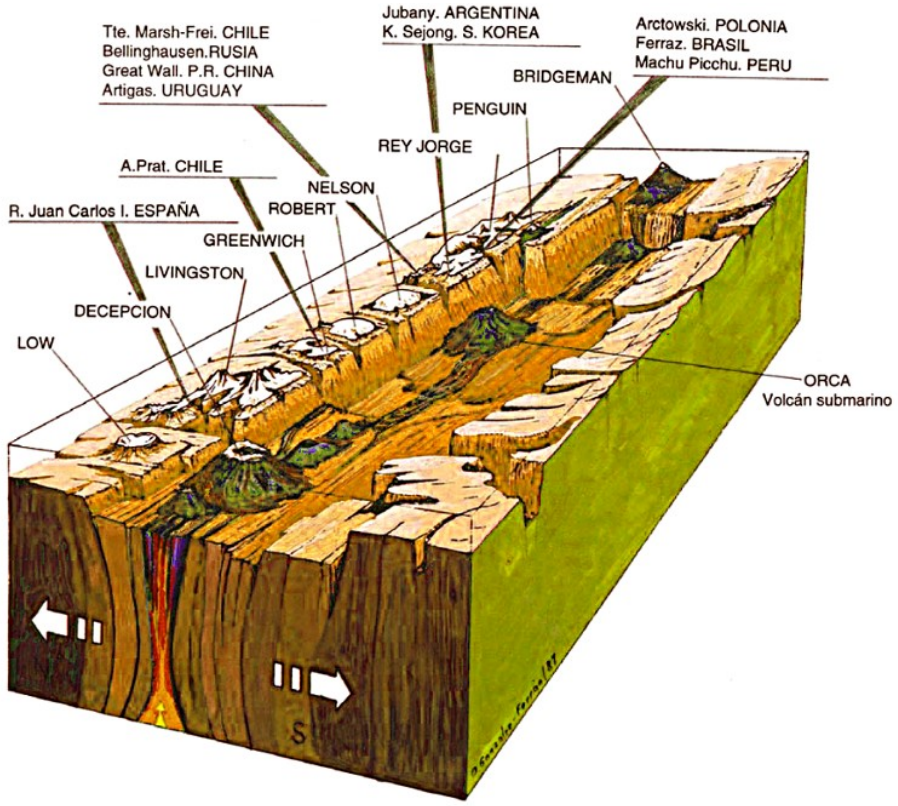


Figura 3.1.2.: Islas Shetland del Sur y fosa oceánica de Bransfield. Se muestran las bases de investigación y los países que las gestionan. Imagen modificada de González-Ferrán et al. (1995).

Marco tectónico. El esquema superior de la Figura 3.1.1 representa la zona tectónica en la que se localiza Decepción en la que confluyen las placas Sudamericana y Antártica y tres microplacas: la de Scotia, la del Drake y la de las Shetland del Sur

(Maurice et al., 2003). La subducción de la fosa de las Shetland originó su separación de la Península Antártica hace 2 Ma. creando el Rift de Bransfield, foco de sismos superficiales que junto con la subducción de la placa del Drake, origen de terremotos más profundos son los responsables de la actividad sísmica regional (Figura 3.1.2). Decepción se encuentra justo en el eje por donde se expande la fosa de Bransfield, lo que añadido a la tectónica regional explica su gran actividad sísmica.

Actividad volcánica. Decepción ha erupcionado al menos 6 veces desde la primera vez que fue visitada por Nathaniel Palmer en 1820 (González-Ferrán et al., 1995). La Figura 3.1.1 muestra en rojo las erupciones históricas, siendo en todos los casos de pequeño volumen y cercanas a la bahía interior. Las últimas 3 erupciones entre 1967 y 1970 fueron avistadas directamente (Baker et al., 1975). En 1967 ocurrieron simultáneamente 2 erupciones separadas entre sí por 2 km. que arrojaron ceniza, bombas y vapor, una de ellas fue submarina y originó un islote en la bahía Telefon. En 1969 tuvo lugar una segundo episodio eruptivo en la bahía Telefon ocasionó grietas en el hielo del Monte Pond junto con emisiones de material piroclástico destruyendo la base científica chilena. La última erupción tuvo lugar en agosto de 1970 en la orilla norte de la bahía Telefon y formó una serie de pequeños cráteres que modificó la costa.

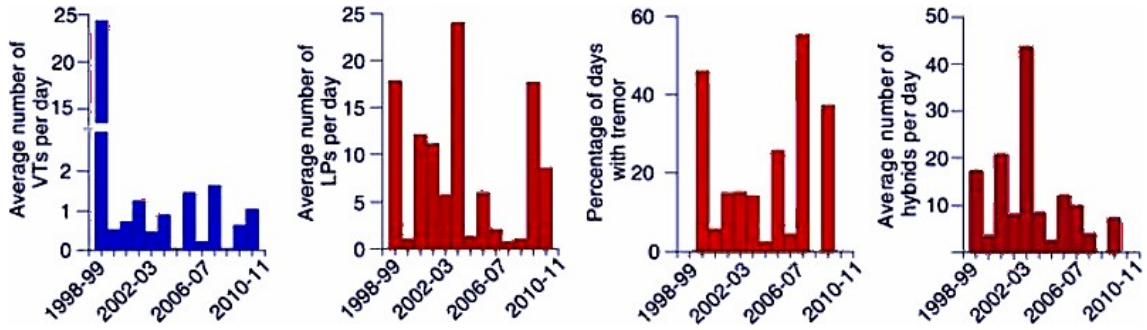


Figura 3.1.3.: Volcán Decepción: sismicidad desde 1998-2011. Los histogramas muestran el número de eventos / día en cada una de las campañas anuales realizadas por el IAG. De izq. a der: frecuencia de terremotos volcano-tectónicos (VTs), sismos de largo periodo (LPs), tremores(TRs) y eventos híbridos (HYs). Figura original de Carmona et al. (2012).

La actividad volcánica actual comprende diversos sismos y áreas hidrotermales, como las fumarolas y aguas termales que rodean al puerto Foster (Rey et al., 1995). Carmona et al. (2012) hacen un extenso resumen de la sismicidad reciente focalizándose en el periodo 1998-2011, destacando la actividad de las campañas del 1998-1999 y 2002-2003 (Figura 3.1.3). Los sistemas hidrotermales ejercen una gran influencia en la sismicidad de la isla. Ibáñez et al. (2003) estudian el episodio de 1998-1999 donde se detectaron más de 2000 VTs y multitud de eventos de bajo periodo localizados en distintas zonas. La sismicidad de los LPs se asoció al sistema hidrotermal

en la bahía de las Fumarolas. Los VT tienen una magnitud media de 0.5 y se originan a menos de 10 km. de profundidad bajo la bahía Puerto Foster. Los autores los asocian a la inyección profunda de magma. En el trabajo realizado por Stich et al. (2011) se demuestra como el ruido oceánico es capaz de generar eventos LP de forma sincronizada a los microseismos oceánicos.

3.1.1.2. Eventos del volcán de Decepción

La Figura 3.1.4 muestra los eventos más comunes registrados en la isla Decepción. Una descripción más general de los sismos podemos encontrarla en la Subsección 1.1.2.

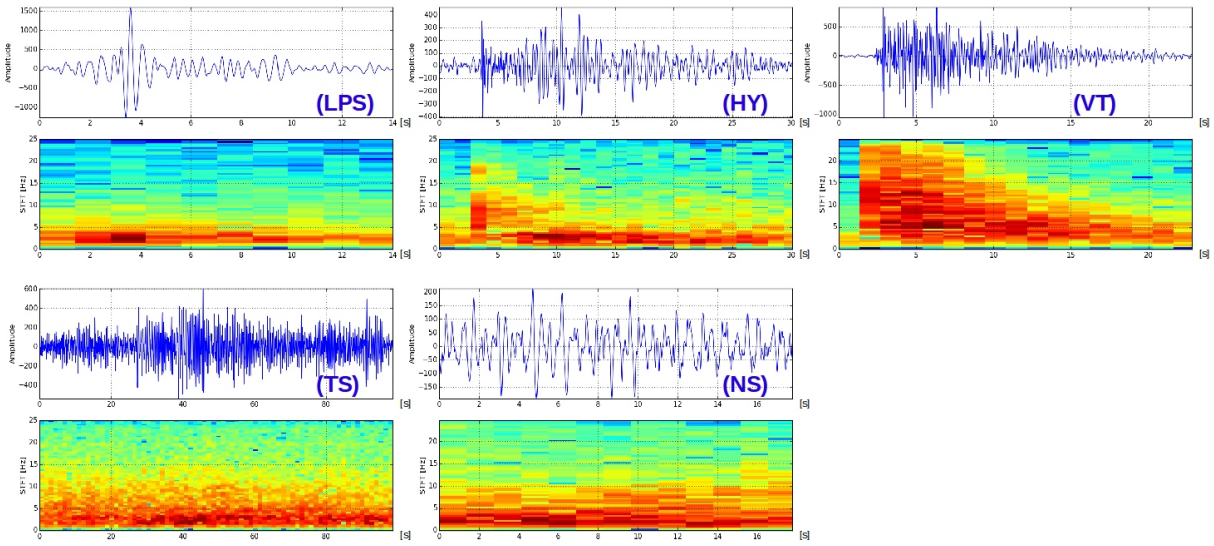


Figura 3.1.4.: Eventos en la isla Decepción. Se representa el sismograma y espectrograma de cada evento. De izq. a der.; 1ª fila: terremotos de largo periodo (LPS), eventos híbridos (HY) y de origen volcano-tectónico (VT). 2ª fila: tremor espasmódico (TS) y ruido (NS).

El abanico de eventos registrados en los sensores comprende tormentas, hielo desquebrajándose, viento y mareas, además de las señales de origen sismo-volcánico. Las peculiaridades de las clases más comunes de Decepción son:

- **Eventos volcano-tectónicos (VT).** Terremotos locales cuyos frentes de ondas P y S difieren menos de 4 segundos y con un amplio contenido espectral de hasta 25 Hz. Estos pueden ser interpretados como la fragmentación del cuerpo tectónico del volcán a esfuerzos locales. El origen de tales fuerzas se relaciona con procesos volcánicos en la isla y son los resultados de, por ejemplo, la interacción del agua con materiales calientes o de inyecciones de magma hacia

la superficie que generan fracturas en rocas frágiles o deformaciones por la inyección de fluidos (Ibáñez et al., 2003). En Decepción nos encontramos tanto VTs profundos (VT-A) como superficiales (VT-B) de pequeña magnitud que los integraremos en una sola clase.

- **Eventos de largo periodo (*LP*).** Señales con una envolvente en forma de huso y con duración menor de 60 s. cuyo contenido espectral tiene bandas casi monocromáticas centradas en frecuencias por debajo de 5 Hz. (Carmona et al., 2012) relaciona la actividad de los LP con sistemas hidrotermales. En Decepción algunos LPs con una alta frecuencia dominante son muy similares a VTs superficiales y es necesario un análisis más detallado (dirección y localización del sismo) que separe las componentes debido a la fuente y las debido a efectos del camino para poder diferenciarlos (Almendros, 1999).
- **Eventos híbridos (*HY*).** Eventos considerados *mixtos*: señales de baja frecuencia disparadas por un sismo local. Se caracterizan por una fase inicial de alta frecuencia, la cual se corresponde a un terremoto VT en el que las ondas P y S son difícilmente distinguibles, seguido por un tren de pulsos cuya energía espectral se concentra en torno a una frecuencia dominante, identificable con los LP. Ibáñez et al. (2003) constatan la dificultad existente en Decepción en clasificar cierto tipo de señales como LPs o HYs.
- **Tremor volcánico (*TR*).** Aunque se han observado tremores resonantes en torno a una $f_0 \approx 2$ Hz., lo tipos más comunes de tremores ne Decepción pertenecen a la subclase de tremores espasmódicos, con una duración que varía desde minutos a varias horas e incluso días (Ibáñez et al., 2003). Tremor y eventos LP son diferentes manifestaciones del mismo proceso: un evento LP es la respuesta a un cambio de presión repentino en una grieta llena de fluido, mientras que un tremor es la respuesta a fluctuaciones continuas de presión. Almendros et al. (1997) localizó los tremores y los LPs en la misma zona. Carmona et al. (2012) pone de manifiesto la importancia de los sistemas hidrotermales en la generación de eventos de bajo periodo (LF) e, incluso, la influencia de la presión atmosférica sobre ellos.
- **Ruidos (*NS*).** Aunque no es un evento sí mismo, si no más bien un conjunto de ellos, de una manera general es un conjunto de señales que engloba todas las señales que no puedan asociarse a las anteriores 4 clases. En la base de datos construidas, se suele restringir a los trozos de señal que anteceden y preceden a los eventos LP, HY, VT y TR, por lo que son señales de baja amplitud y con una distribución espectral acumulada en las frecuencias bajas de [0,10] Hz. que no se adapta a ningún patrón temporal concreto.

3.1.2. Volcán de Fuego de Colima

El volcán (de Fuego) de Colima forma, junto con los volcanes de el Cántaro y el Nevado de Colima el Complejo Volcánico de Colima (*CVC*), una cadena de estra-

tovolcanes andesíticos orientada de norte a sur que se encuentra en el límite de los estados de Jalisco y Colima, a unos 30 kilómetros de la ciudad de Colima en el oeste de México. Tiene una altura de unos 3860 m. habiendo erupcionado más de 40 veces desde el siglo XVI, por lo que es considerado el volcán más activo de Norteamérica (Bretón et al., 2002). El último episodio eruptivo importante se produjo entre los años 2004-2006. Actualmente se encuentra en erupción desde noviembre del 2014.

3.1.2.1. Marco sismo-tectónico

Allan and Carmichael (1984) datan el comienzo de actividad volcánica del Complejo Volcánico de Colima hace unos 1.7 Ma. con la formación de el volcán del Cántaro, actualmente inactivo. El CVC forma parte del Eje Volcánico Transversal o Cinturón Volcánico Trans-mexicano (CVT), formado como consecuencia de la subducción de las placas de Cocos (PC) y Ribera (PR) bajo la placa Norteamericana (Ver la Figura 3.1.5). El CVT es una zona de gran actividad sísmica formada por los principales volcanes activos mexicanos como los de las islas Revillagigedo en el Pacífico, el Climatorio, el Sanganguey, el Paricutín (nacido en 1943), el Nevado de Toluca, el Malinche, Popocatépetl e Iztaccíhuatl y el pico de Orizaba (la montaña más alta de México, con 5610 m.).

La Figura 3.1.5 enmarca al CVC en una zona de fragmentación del Bloque de Jalisco producida por el choque de tres grandes depresiones tectónicas o *rifts*: el graben de Tepic-Zacoalco, el graben de Colima y el graben de Chapala, si bien la evolución actual del CVC se asocia más ala interacción entre la falla de Tamazula y la depresión tectónica de Colima (Garduño Monroy et al., 1998).

La actividad eruptiva del CVC está sometida a un proceso cíclico de creación de grandes estratovolcanes y su posterior derrumbe formando grandes depósitos de avalanchas y escombros (Cortés et al., 2005). En 2008 el volcán de Colima tenía la probabilidad más alta de erupcionar de todo México con un $VEI \geq 4$ (*Volcano Explosivity Index* - *VEI*), comenzando un nuevo episodio en 2014 (Mendoza-Rosas and De la Cruz-Reyna, 2008; Arámbula-Mendoza et al., 2011). Las dos últimas grandes erupciones plineanas tuvieron lugar en 1818 y 1913, incluyendo esta última flujos piroclásticos que recorrieron hasta 15 km. A partir de 1991 se tienen datos registrados de la actividad del volcán, constatándose en las erupciones de 1991 y 1994 que sismos VT y enjambres de LPs antecedieron al derrumbe del domo (1991) generando flujos piroclásticos y a la explosión del 21 de Julio de 1994. En 1997 de nuevo comienza otro ciclo de crecimiento del domo con derrumbes de material que terminaría siendo destruido mediante las explosiones de 1999 (Saucedo et al., 2002), comenzando un nuevo periodo de actividad en 2001 que culmina con nuevas explosiones en verano de 2003 (Zobin et al., 2006).

En la Figura 3.1.6 apreciamos como el 25 de septiembre de 2004 se registran una sucesión de LPs, seguido un incremento de actividad en la noche del 30 provocada por pequeñas explosiones, fumarolas y colapsos de flujos piroclásticos. Estos eventos

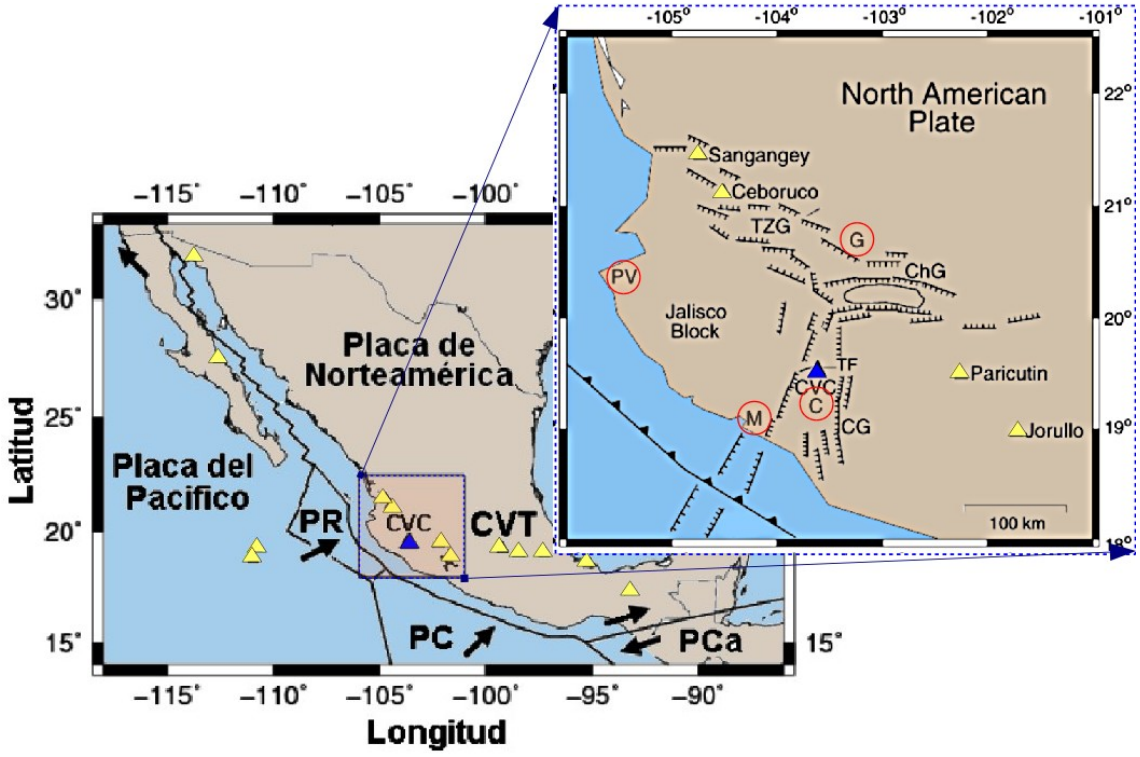


Figura 3.1.5.: Localización tectónica del volcán de Fuego de Colima. El triángulo azul indica donde se sitúa el Complejo Volcánico de Colima (*CVC*), perteneciente al Cinturón Volcánico Trans-mexicano (*CVT*) o Eje Volcánico Transversal. Los triángulos amarillos son volcanes de México cuyas erupciones están documentadas. Se muestran las placas colindantes (*PR* - Placa de Rivera, *PC* - Placa de Cocos y *PCa* - Placa del Caribe). La imagen ampliada detalla las 3 fosas tectónicas que confluyen en el CVC: Colima graben (*CG*), Chapala graben (*ChG*) y Tepic-Zacoalco graben (*TZG*) junto a la falla de Tamazula (*TF*) y las ciudades de Colima (*C*), Puerto Vallarta (*PV*), Guadalajara (*G*) y Manzanillo (*M*) señaladas entre círculos rojos. Figuras modificadas de Arámbula-Mendoza et al. (2011).

preceden el ciclo eruptivo más importante hasta la fecha de Colima en el siglo XXI que incluye derrumbes, explosiones vulcanianas escuchadas a 50 km. con columnas de material que se elevan a una altura de 10 km. con caída de cenizas en pueblos a un radio de 12 km. y flujos piroclásticos de hasta 6 km (Arámbula-Mendoza et al., 2011; Varley et al., 2010a). A finales de 2013 el volcán entra en una nueva fase eruptiva que comienza con actividad efusiva Zobin et al. (2015) que continúa un año después con explosiones de columnas de hasta 5 km. de altura y caída de cenizas en ciudades a 25 km. de distancia. en Julio del 2015 se producen erupciones que obligan a la evacuación de población.

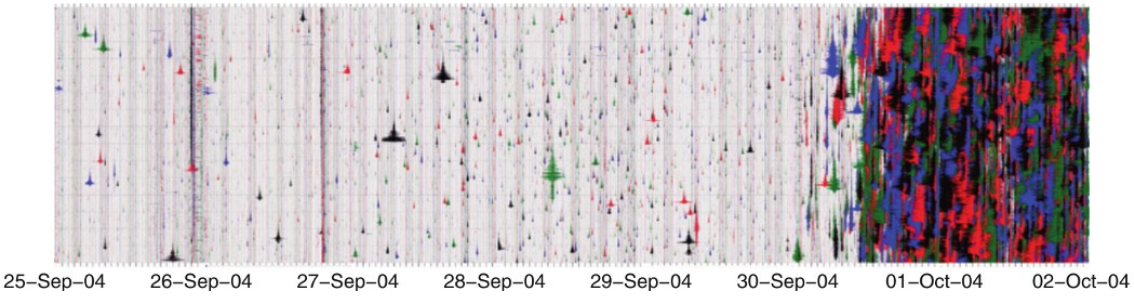


Figura 3.1.6.: Comienzo del episodio eruptivo 2004-2006 en Colima. Los sismogramas muestran enjambres de LPs que preceden explosiones, fumarolas y flujos piroclásticos en la noche del 30 de septiembre de 2004 (Arámbula-Mendoza et al., 2011).

3.1.2.2. Eventos del volcán de Colima

La Figura 3.1.7 muestra los eventos más representativos registrados en Colima. A continuación describimos aquellas clases que conforman nuestra base de datos maestras construidas en la Sección 3.6. Un estudio más detallado de las señales registradas en el volcán de Colima lo encontramos en Arámbula (2011) y González (2011):

- **Volcano-Tectónicos (*VT*)**. Son sismos con fases P y S bien definidas, con una energía espectral distribuida en un intervalo alto de frecuencias, hasta 25 [Hz]. La coda decae linealmente en el espectrograma. Están relacionados con fragmentaciones del cuerpo tectónico del volcán, debido a altas presiones sobre él ejercidas por fluido. Durante las erupciones de 1998 se produjeron enjambres de VTs y se relacionaron con nuevo magma entrante bajo el volcán.
- **Eventos de largo periodo (*LP*S)** en los que las fases P y S son indistinguibles. El intervalo de frecuencias más energéticas se localiza entre 1 y 5 Hz, acumulando el 95 % de su energía espectral en el intervalo de [1, 10] Hz. Estos eventos se relacionan con el crecimiento del domo y aparecen justo antes de explosiones en los registros de monitorización del 2005. También han aparecido próximos a la superficie en forma de enjambres. Es posible realizar una subclasificación de eventos LP basándose en criterios espectrales y en el frente de onda: subclases A, B o C se definen dependiendo del números de picos espectrales (frecuencia dominante y/o armónicos)Varley et al. (2010b).
- **Terremotos regionales (*REG*)**. Son sismos tectónicos generados en los bordes de las placas o en fallas cercanas, de origen no volcánico. Son fácilmente clasificados a causa que la onda S llega con bastantes segundos de retraso respecto a la onda P. Lo más común es que esta diferencia sea del orden de unas decenas, debido a que los sismos provienen de las zonas de subducción próximas a la ciudad de Colima. La energía espectral se concentra hasta los 10 Hz. En el espectrograma la coda del sismo se ajusta a una exponencial decreciente en sus frecuencias más energéticas.
- **Explosiones (*EXP*)**. Las explosiones son un evento muy frecuente en casi

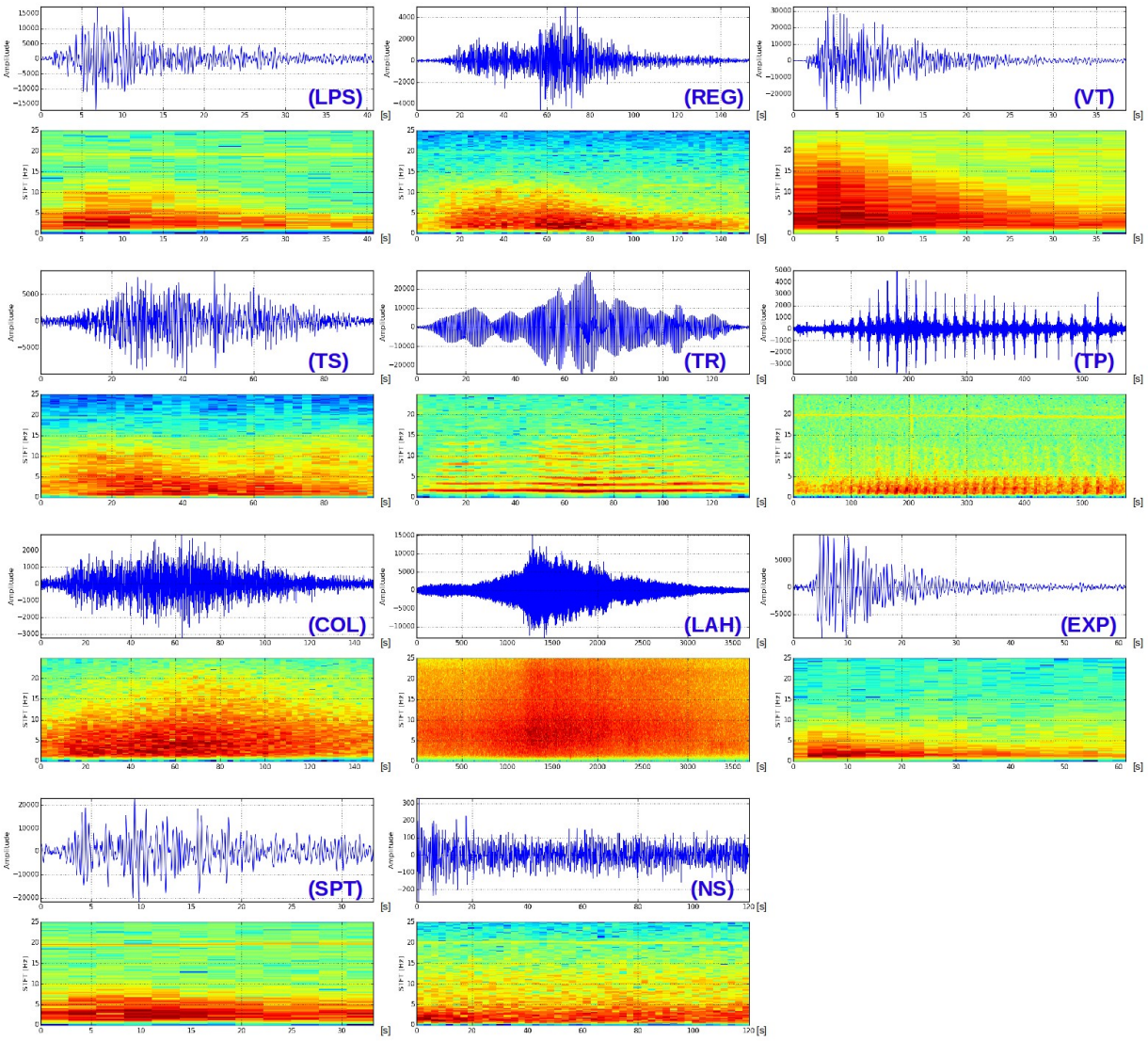


Figura 3.1.7.: Eventos del volcán de Fuego de Colima. Se representa el sismograma y espectrograma de cada evento. De izq. a der.; 1ª fila: terremotos de largo periodo (*LPS*), regionales (*REG*) y de origen volcano-tectónico (*VT*). 2ª fila: tremor espasmódico (*TS*), resonante (*TR*) y pulsante (*TP*). 3ª fila: colapsos de material (*COL*), lahar (*LAH*) y explosión (*EXP*). 4ª fila: pequeño tremor pulsante (*SPT*) y ruido (*NS*).

todos los periodos de actividad en el volcán Fuego de Colima, especialmente cuando el domo de lava se está enfriando en el cráter y comienza a ser destruido por la creciente presión bajo él (Arámbula, 2011). Las EXPs de Colima se estructuran en 2 fases: una inicial de baja frecuencia seguida de una llegada más energética de frecuencia mayor, relacionando la duración de la llegada inicial con su profundidad y fuerza de la explosión (Zobin et al., 2006) y la segunda fase con la fragmentación del domo (Varley et al., 2010a). Estos

eventos a menudo generan y se solapan con tremores espasmódicos y derrumbes de material responsables de un espectrograma de frecuencia alta.

- **Colapsos (*COL*)** o **derrumbes**. En Colima este tipo de eventos se producen normalmente por el crecimiento del domo de lava en el interior del cráter produciendo un desbordamiento de material por los lados del volcán (Arámbula, 2011). Incluyen flujos piroclásticos que pueden recorrer hasta 5 km de distancia como ha sido registrado en 1991, 1998 (Zobin et al., 2005) y 2004. La duración de la caída de rocas apenas supera los 5 minutos. Se relacionan con bandas de frecuencia alta (hasta 15 Hz.) en el espectrograma.

- **Lahares (*LAH*)**. Al igual que los colapsos, los lahares son también eventos superficiales no necesariamente asociados a la sismicidad de un volcán. Aparecen debido a lluvias torrenciales sobre el volcán que dan a lugar a la mezcla del agua con sedimentos volcánicos produciendo flujos de escombros y lodo que bajan por las laderas de las montañas sobre los cauces de ríos. La duración varia desde decenas de minutos a varias horas. Su banda de frecuencia cubre el total del intervalo representado por un sensor de bajo periodo (hasta 25 Hz.). A pesar de que descienden a una velocidad moderada (50 km/h) en Colima no ha habido víctimas mortales hasta la fecha por lahares.

- **Tremores (*T*)**. El tremor volcánico se relaciona con los cambios de presión en el fluido que se encuentra en la cámara magmática o en algunas de sus grietas (Chouet, 1981; Konstantinou and Schlindwein, 2003). Conserva la envolvente de amplitud del sismograma más o menos constante por un largo intervalo de tiempo (desde minutos a horas e incluso meses) en la banda espectral de [1,10] Hz, por lo que a veces se le conoce también como *ruido volcánico*. En la literatura se distingue usualmente entre tres tipos de tremores y, de cara al facilitar el proceso de reconocimiento automático hemos definido un tipo más:

  - **Tremor armónico** o **resonante *(TR)***. La forma de onda del sismograma está modulada con una o varias frecuencias dominantes (frecuencias *resonantes*), percibiéndose sus respectivos armónicos de forma clara. Las frecuencias de resonancia pueden variar su posición en el espectro incluso dentro del mismo evento debido a variaciones de la presión dentro de las cavidades resonantes (Sturton and Neuberg, 2003) que se van sellando progresivamente lo que aumentaba el valor de la frecuencia resonante. Si la presión es muy alta la desgasificación puede causar pequeñas explosiones o comportamiento efusivo del volcán. Esta es la causa que en el episodio eruptivo de 2002 se evacuaran poblaciones cercanas al cráter (Arámbula, 2011).

  - **Tremor espasmódico (*TS*)**, que tiene amplitudes variables sin una frecuencia dominante y se considera que está provocado por el solapamiento y encadenamiento de múltiples sismos de largo periodo (Almendros et al., 1997), por lo que suele abarcar un rango de frecuencias ([0,15] Hz.) más

amplio que otros tremores. Muchas de las explosiones en Colima a menudo finalizan con este tipo de tremor. Se denomina también *tremor de emisión* ya que se relaciona con procesos de emisión de cenizas o gas.

- **Tremor pulsante (*TP*)**. Se considera un evento mixto compuesto por un tremor espasmódico sobre el que se solapa unos pulsos que se repiten secuencialmente en un intervalo de tiempo que varía de 10 a 30 s. Su espectro abarca un intervalo de [1,5] Hz.
- **Pequeños pulsos de tremores (*SPT*)**. Sucesión de frentes de onda con un comportamiento de eventos de bajo periodo. Pueden considerarse como una secuencia solapada de eventos LP con una duración que suele ser menor de 50 segundos. Originalmente estaban incluidos dentro de la clase de tremores espasmódicos, pero con el objetivo de mejorar la efectividad en el reconocimiento y dada su menor duración se decidió crear una clase propia que englobase estos pequeños tremores.

- **Ruido (*NS*)** o **(*WNS*).** Al igual que ocurre en Decepción, denominamos como ruido cualquier evento que no pueda asociarse a ninguna de las clases anteriormente definidas. Se caracteriza por su baja amplitud y por que su energía espectral está contenida principalmente en la banda de frecuencia de [0,10] Hz, excepto cuando hay ráfagas de viento en el que el espectro mantiene una magnitud aproximadamente constante para en el intervalo de $[1, 20]$ Hz, pareciéndose al ruido blanco (*White NS - WNS*).

3.2. Descripción de datos

Proceso que comprende el procesamiento digital realizado sobre las señales que registran las estaciones hasta su transformación en un conjunto de muestras multidimensionales o vectores $\{\mathbf{x}\}$ del llamado *espacio de descripción* de datos, de representación de patrones o espacio de características $\Omega_{\mathbf{X}}$. Dado que nuestros eventos sísmicos se presentan secuencialmente en el dominio temporal apostamos por una representación secuencial de vectores para cada fichero de datos. La técnica de modelado escogida será la encargada de encontrar los patrones secuenciales para cada clase en el espacio de características y de describirlos adecuadamente en la etapa de construcción de modelos (Sección 3.3). La descripción de datos se divide en dos partes:

1. *Preprocesado* de datos o *acomodación* de estos al sistema: prefiltrado y formateado del sismograma antes de la parametrización
2. *Extracción de características* o *parametrización*: transformación de las muestras del sismograma ya acondicionado a secuencias de vectores de características

3.2.1. Preprocesamiento de la señal

A veces integrado en la etapa de adquisición de datos, el preprocesado nos permite preparar los sismogramas para la fase de extracción de características. En nuestros sistemas VSR solo usaremos la componente vertical del sismograma para reconocer eventos, transformándola según los siguientes pasos:

1. *Eliminar la componente continua del fichero.* Con el objetivo de facilitar las operaciones posteriores de filtrado y formateado. Puesto que estamos trabajando con señales estacionarias, este pre-filtrado no afecta a nuestro sistema.
2. *Datos en formato* $i2.50Fs$: enteros de 2 bytes (16 bits) muestreados a una frecuencia ,$F_S = 50[Hz]$. Nótese que el pasar de un formato con mayor margen dinámico al formato $i2$ puede ser un problema en el caso de eventos muy energéticos, y para evitar la saturación es conveniente realizar algún tipo de normalización. En nuestro caso si detectamos que el fichero de datos va saturar simplemente lo normalizamos al máximo de amplitud permitido para $i2$. Otra opción posible es la compresión logarítmica del rango dinámico. Si además es necesario reducir la frecuencia de muestreo lo más conveniente para evitar el aliasing es filtrar paso-bajo antes de decimar la señal.
3. *Filtrado espectral en la banda* $[f_L = 1, f_H = 25]\,Hz$, que es donde se concentra la mayor parte de la energía de los eventos sísmicos (Chouet, 1996a). Aplicamos un filtrado de respuesta infinita (*Infinite Impulsive Response - IIR*) de Butterworth de 2º orden para obtener una respuesta plana en la banda de paso. Con el objetivo de evitar desfases en la señal, la señal filtrada se invierte temporalmente y se vuelve a filtrar, consiguiendo un filtrado de desfase nulo.

La simplicidad del formato $i2.50Fs$ escogido se debe a razones prácticas: nos ahorramos tiempo de cómputo. Aunque pueden existir ciertas ventajas en usar otros formatos con mayor margen dinámico como $i3.100Fs$ consideramos que son muy pequeñas. En el caso de incluir eventos de muy largo periodo (*Very Long Period - VLP*) en el conjunto de clases a reconocer sería necesario reducir f_L, lo que conlleva inevitablemente utilizar sensores de banda ancha. Un valor de $f_L = 1$ es indicado para eliminar muchos tipos de ruido y parte del ruido oceánico de Decepción.

3.2.2. Parametrización: desde sismogramas a secuencias de vectores

Dada la similitud entre las señales de voz y los eventos sismo-volcánicos, para describir nuestros eventos usaremos un esquema de parametrización basado en la evolución temporal de la energía espectral. Los coeficientes cepstrales en escala MEL o MFCC (*MEL-Frequency Cepstral Coefficients*) son un estándar de codificación de voz inspirados en el modelo psico-acústico de la audición humana Rabiner and Juang (1993); Young et al. (2006), si bien han sido adaptados con éxito a señales

sismo-volcánicas (Benítez et al., 2007; Ibáñez et al., 2009; Cortés et al., 2009a) en la forma denominada LFCC (*Log-Frequency Cepstral Coefficients*).

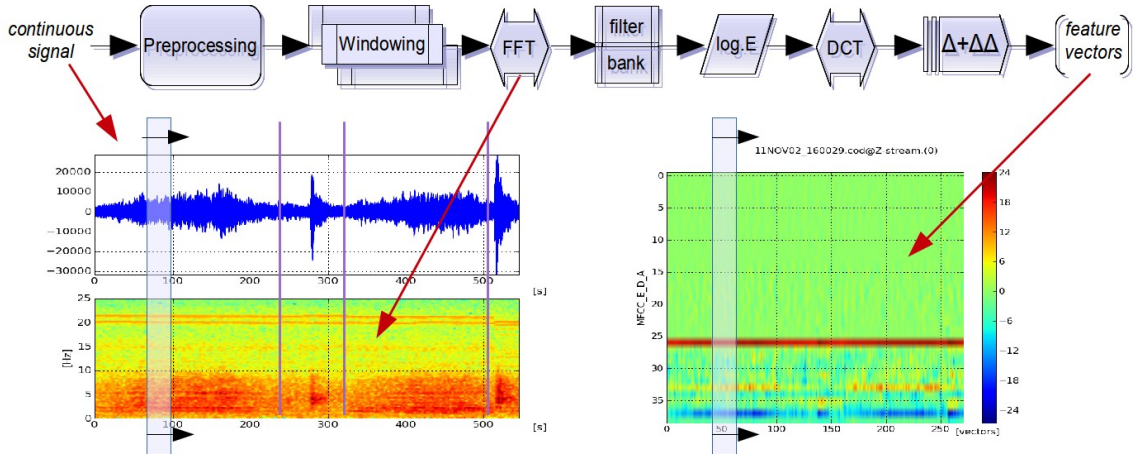


Figura 3.2.1.: Extracción de coeficientes LFCC.D.A a partir de los sismogramas. La señal en continuo se segmenta en trozos o *frames* solapados temporalmente (*Windowing*). De cada segmento se extrae la energía logarítmica en distintas bandas espectrales o *canales*. Los valores de energía logarítmica (*log.E*) forman las componentes de un vector que son decorrelacionadas aplicando la *DCT*. Posteriormente se añade información de contexto al vector incorporando las derivadas temporales de 1er (Δ) y 2º orden ($\Delta\Delta$) de sus componentes.

El esquema *MFCC.E.D.A.39* (renombrado como *LFCC.D.A.*) de parametrización de los datos se describe en la Figura 3.2.1 y consta de varias etapas:

1. **Fragmentación o ventaneo**: El flujo continuo de señal se se divide en segmentos o *frames* de duración determinada por el ancho temporal de la *ventana deslizante* o ventana de parametrización. Cada uno de esos segmentos se describirá mediante un vector de características. Para mejorar la robustez los segmentos se solapan al 50 %. Se suavizan las muestras de cada segmento mediante una ventana de Hamming (el uso de diferentes tipos de ventanas en el suavizado no varia significativamente los resultados, según Hoogenboezem (2010)) para evitar discontinuidades en los bordes que puedan enfatizar el efecto de Gibbs.

2. **Análisis mediante banco de filtros:** Se halla la energía espectral de cada segmento mediante la transformada rápida de Fourier (*Fast Fourier Transform - FFT*). Se normaliza la energía espectral en cada fichero o registro. El espectro se somete a un análisis de 16 *canales*: filtros triangulares que se solapan al 50 % sobre el intervalo $[1, 25]$ Hz en el eje de la frecuencia logarítmicamente escalado. Este escalado permite el énfasis del análisis de las bajas frecuencias sobre las altas, que es donde se acumula la mayor parte de la energía de los eventos sismo-volcánicos (y del habla). Los canales centrados en frecuencias

bajas serán más anchos que los centrados en altas frecuencias. Se construye el vector de características base con 16 componentes correspondientes al valor del logaritmo de la energía de cada canal.

3. **Decorrelación de componentes**: Dado que las componentes del vector están muy correlacionadas entre sí debido al solapamiento de los canales, se pretende eliminar información redundante mediante la aplicación de la transformada del coseno (*Discrete Cosine Transform - DCT*) sobre cada vector reduciendo su tamaño a 12 componentes. La reducción de la complejidad del vector es conveniente para evitar la *maldición de la dimensionalidad* (Subsección 4.1.1) y necesaria: propondremos un modelado del espacio $\Omega_{\mathbf{X}}$ de características generado por los vectores realizado por mezclas de probabilidades gausianas (Subsección 3.3.1) con matriz de covarianza diagonal, por lo que implícitamente se asume la independencia estadística entre las componentes del vector.
4. **Incorporación de información dinámica y energética**: La energía en forma logarítmica de cada segmento hallado en la etapa de ventaneo se añade al vector de características (coeficiente E). También se incorpora información contextual a cada vector: se calculan las derivadas temporales de cada componente de primer (Δ, coeficientes D) y segundo orden ($\Delta\Delta$, coeficientes D), triplicando así el tamaño del vector base (12+1 componentes base más 12+1 coeficientes delta, Δ, o *velocidades* más 12+1 coeficientes $\Delta\Delta$ o *aceleraciones*).

3.3. Clasificadores

Como ya vimos en el Capítulo 2, la similitud entre señales sísmicas y de locución de voz (Ohrnberger, 2001) y los requerimientos (Subsección 2.2.3) de detección y clasificación sobre un registro continuo en tiempo cuasi-real impuestos por la monitorización de volcanes activos, hacen de los reconocedores probabilísticos una interesante alternativa (Subsección 2.3.4). Los sistemas capaces de modelar estructuras secuenciales parecen a priori la mejor opción para modelar los eventos de origen tectónico de los que se esperan un orden temporal concreto en las llegadas de los frentes de onda (Subsección 1.1.2). Para construir el sistema base optaremos dos tipos de clasificadores, ambos basados en estadística bayesiana:

- Modelado del espacio de características mediante mezcla de gausianas (*Gaussian Mixture Models - GMMs*)
- Doble modelado, del espacio de características y de su evolución temporal mediante modelos ocultos de Markov (*Hidden Markov Models - HMMs*)

Los GMMs solo se utilizarán explícitamente en el capítulo de Capítulo 4. Gracias a la sencillez de su estructura se pretende minimizar la influencia que el clasificador pueda tener en determinados esquemas de selección de características (Álvarez et al., 2011; Cortés et al., 2014). Los HMMs serán los encargados de modelar los eventos de

cada clase. Ya han sido utilizados con éxito en múltiples sistemas de VSR (Benítez et al., 2007; Wassermann et al., 2007; Beyreuther et al., 2008; Ibáñez et al., 2009; Cortés et al., 2009b).

Junto a la definición formal de los clasificadores, hablaremos de cómo pueden reconocer eventos que aparecen en el flujo de datos en continuo registrado por un sensor sísmico. Para ello, es necesario que el sistema sea capaz de detectar y aislar temporalmente en el sismograma los *segmentos* o vectores de características consecutivos extraídos de la señal que el clasificador asigna a un mismo evento.

3.3.1. Modelado de características: GMMs

Aún siendo un clasificador sencillo y ampliamente estudiado, presentaremos los GMMs pues los usaremos en distintas partes de este trabajo:

- Como modelo del espacio de descripción de los datos y clasificador en el capítulo de reducción de características (Capítulo 4).
- En los HMMs: normalmente, la función de salida que estima la probabilidad de que un vector de características sea emitido por un estado de un HMM puede identificarse directamente con un GMM.
- En el sistema PSA: El decodificador construido para los GMM en el sistema base servirá como punto de partida para el módulo que en el sistema en PSA paralelo combinará los resultados de cada canal (Subsubsección 5.1.2.1).

Como vemos en la Figura 3.3.1, el esquema del sistema basado en GMM es bastante simple, y sirve como plantilla para diseñar cualquier sistema de clasificación automática supervisada. En los próximos apartados definiremos analíticamente el modelo al que aplicaremos el criterio MAP para construir un clasificador de eventos aislados.

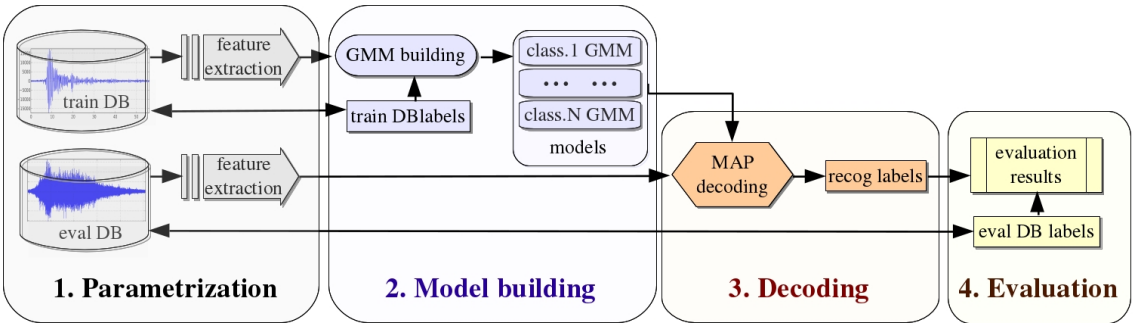


Figura 3.3.1.: Sistema de referencia ($VSR - SSA$) de clasificación de eventos aislados basado en GMMs.

3.3.1.1. Modelo mezcla de componentes gausianas

Los GMMs son modelos probabilísticos que estiman una densidad de probabilidad $p(\mathbf{x}|w)$ de un conjunto de vectores $\{\mathbf{x}_w\}$ previamente asignados a la clase w como una combinación lineal de G componentes gausianas $N_g(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$:

$$p_{GMM}(\mathbf{x}|w) \equiv \sum_{g=1:G} \alpha_g N_g(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = \sum_{g=1:G} \alpha_g \frac{\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_g)\boldsymbol{\Sigma}_g^{-1}(\mathbf{x}-\boldsymbol{\mu}_g)\right\}}{(2\pi)^{K/2}|\boldsymbol{\Sigma}_g|^{1/2}} \tag{3.3.1}$$

siendo:

$(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ media y matriz de covarianza de la componente g, computadas del subconjunto $\{\mathbf{x}_{w,g}\}$ de vectores de $\{\mathbf{x}_w\}$ que se han sido asociados a $N_g(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ en el proceso de inicialización.

α_g peso de la componente g en la combinación lineal. Usualmente representa la proporción de vectores $\{\mathbf{x}_{w,g}\}$ de $\{\mathbf{x}_w\}$ que han sido inicializados a $N_g(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$.

K es el número de características o *dimensionalidad* de los vectores $\{\mathbf{x}\}$ que describen a los segmentos de señal.

G número de componentes G del modelo. Debe seleccionarse adecuadamente acorde con la variabilidad del espacio de características a modelar y número de vectores asociados a la clase w para evitar el sub-entrenamiento o el sobre-entrenamiento(Figura 2.1.6).

El ajuste de los GMMs puede realizarse mediante análisis de histogramas de estimación de probabilidad o por estimadores de máxima verosimilitud (Sección 2.1.3.2) como el algoritmo EM de Dempster et al. (1977) que iteran hasta conseguir un máximo local. La manera más común para inicializar los GMMs es mediante algún método no supervisado basado en el algoritmo k-medias (Subsubsección 2.3.6.1). Los valores de inicialización pueden influir bastante en la posición del máximo local y en el número de iteraciones necesarias para llegar a él.

3.3.1.2. Clasificación mediante GMM: diseño del decodificador

El diseño del clasificador implica la evolución desde un sencillo clasificador de vectores basado el criterio de decisión MAP hasta un clasificador de eventos aislados, asumiendo para algunas simplificaciones sobre el modelo y la aplicación de filtros temporales sobre las clases. Esto divide de forma natural el proceso de clasificación en 2 etapas:

1. *(Pre-)clasificación*: en la que se aplica la función de decisión, ordenándose decrecientemente las probabilidades $p(\mathbf{x}, w_c)$ de asignar las etiquetas $\{w_c\}$ a un evento $\mathbf{x}$.

2. *Post-clasificación* o filtrado: En la que se aplican filtros sobre las probabilidades de clasificación, tras lo que se actualiza el orden de las etiquetas y se asigna la clase definitiva $x \rightarrow \hat{w}_{\mathbf{x}}$.

Clasificación de eventos aislados definidos como secuencias de vectores. En la Sección 2.3.4 desarrollamos el método para clasificar un vector de $\mathbf{x}$ de características mediante los GMMs. Vamos a definir ahora $\mathbf{x}$ como un evento descrito por una secuencia de T vectores $\mathbf{x} = \{\mathbf{x}_t\}_{t=1:T} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$. Asumiremos como ciertas las siguientes hipótesis:

1. Todas las clases $\{w_c\}$ de eventos tienen a priori la misma probabilidad de ser observadas en el volcán, tal que $P(w_c) = Cte = 1/C$, con C el numero total de clases, con lo que $p(\mathbf{x})$ puede escribirse como:

$$p(\mathbf{x}) = \sum_{c=1:C} p(\mathbf{x}, w_c) = \sum_{c=1:C} p(\mathbf{x}|w_c)P(w_c) = \frac{1}{C}\sum_{c=1:C} p(\mathbf{x}|w_c) \tag{3.3.2}$$

2. Todos los vectores de características son estadísticamente independientes entre sí ($\mathbf{x}_i \perp \mathbf{x}_j$):

$$p(\mathbf{x}) = p(\{\mathbf{x}_t\}) = p(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T) = \prod_t p(\mathbf{x}_t) \tag{3.3.3}$$

3. Nuestra base de datos está formada por eventos aislados, con lo que a la secuencia $\mathbf{x} = \{\mathbf{x}_t\}_{t=1:T}$ de segmentos le corresponde otra secuencia de etiquetas formadas tan solo por una única clase: $\mathbf{x} = \{\mathbf{x}_t\}_{t=1:T} \rightarrow \{\mathbf{w}\} = w_{\mathbf{x}}$.

Bajo estas suposiciones aplicamos el criterio MAP sobre el conjunto de modelos GMM $\{p_G(\mathbf{x}|w_c)\}_c$ que representan a cada una de las clases $\{w_c\}$ para discriminar el modelo con la mayor probabilidad de generar la secuencia $\mathbf{x} = \{\mathbf{x}_t\}_{t=1:T}$ mediante:

$$d_{B:GMM}(\mathbf{x}) \equiv \arg\max_c \left\{ \frac{p_G(\mathbf{x}|w_c)}{p(\mathbf{x})} P(w_c) \right\} = \arg\max_{w_c} \left\{ \frac{p_G(\mathbf{x}|w_c)}{\sum_c p_G(\mathbf{x}|w_c)} \right\} = \tag{3.3.4}$$

$$= \arg\max_c \left\{ \frac{\prod_t p_G(\mathbf{x}_t|w_c)}{\sum_c [\prod_t p_G(\mathbf{x}_t|w_c)]} \right\} \arg\max_c \left\{ \prod_t p_G(\mathbf{x}_t|w_c) \right\} \tag{3.3.5}$$

donde la probabilidad $GMM_{w_c}(\mathbf{x}_t) = p_G(\mathbf{x}_t|w_c) = p_{GMM}(\mathbf{x}_t|w_c)$ de que un vector $\mathbf{x}_t$ pertenezca a una clase w_c está dada por la Ecuación 3.3.1. Nótese que dada una secuencia concreta $\mathbf{x}$ de observaciones, $p(\mathbf{x})$, el denominador de las anteriores ecuaciones es constante e independiente de las clases, por lo que puede ignorarse al aplicar la función de decisión, quedando la Ecuación 3.3.5 como una fórmula similar a la que tenemos en la Ecuación 2.1.12, correspondiente con el modelo simplificado (*naive*) de Bayes bajo el criterio de máxima verosimilitud. Este denominador puede interpretarse como un factor de normalización paralelamente a la función de partición Z dada en los campos aleatorios de Markov (Sección 2.3.4).

Filtros de duración Con el objetivo de mejorar la tasa de clasificación, podemos aplicar filtros de duración mínima y máxima a los eventos en una fase de post-clasificación. Esta solución ya fue implementada implícitamente por Ohrnberger (2001) con los HMMs discretos y por Beyreuther and Wassermann (2011) en su modelado temporal de HMMs mediante transductores de estados finitos (WFST{HSMMs}, Sección 2.3.5), al reconocer en continuo con una ventana deslizante cuya duración depende de cada clase. Cortés et al. (2014) también usan explícitamente filtros de duración mínima en sus sistemas en serie y paralelo basado en GMMs.

En el caso de reconocimiento de eventos aislados, la implementación de estos filtros es inmediata; nótese que tras aplicar la función de decisión $d_{B:GMM}(\mathbf{x})$ de la Ecuación 3.3.4 en la etapa de pre-clasificación sobre las clases $\{w_c\}$ estas pueden ordenarse decrecientemente conforme a su probabilidad conjunta estimada por los GMMs, $p(\mathbf{x} = \{\mathbf{x}_t\}, w) = \prod_t GMM_w(\mathbf{x}_t)$. Sea $\{\hat{w}_{GMM,c}(\mathbf{x})\} = \{\hat{w}_{GMM,1}, ..., \hat{w}_{GMM,C}\}$ al conjunto de clases ya ordenadas decrecientemente que los modelos GMM asignan a la secuencia $\mathbf{x} = \{\mathbf{x}_t\}_{t=1:T}$ tal que $p(\mathbf{x}, \hat{w}_{GMM,i}) \geqslant p(\mathbf{x}, \hat{w}_{GMM,i+1})$. Debe cumplirse:

$$dur_m(w_i) \leqslant dur(\mathbf{x}) \leqslant dur_M(w_i) \tag{3.3.6}$$

donde $dur(\mathbf{x})$ representa la duración de la secuencia $\mathbf{x}$ y $\{dur_m(w_i), dur_M(w_i)\}$ son las duraciones mínima y máxima respectivamente que, en caso de existir, pueden permitirse en eventos de la clase w_i. Estos parámetros pueden asignarse en función de un análisis previo de los eventos de entrenamiento o basándose en el conocimiento geofísico que se tenga de cada clase. Si la duración de $\mathbf{x}$ no pasa los filtros temporales de la clase $\hat{w}_{GMM,1}$, $dur(\mathbf{x}) \notin [dur_m(\hat{w}_{GMM,1}), dur_M(\hat{w}_{GMM,1})]$, se asignaría $\mathbf{x} \rightarrow \hat{w}_{GMM,2}$. Si tampoco se pasan los filtros de $\hat{w}_{GMM,2}$ se probaría con $\hat{w}_{GMM,3}$ y así sucesivamente. Si ningún filtro temporal se pasa, se utiliza la asignación original $d_{B:GMM}(\mathbf{x}) = \hat{w}_{GMM,1}$.

3.3.2. Modelado de características y de la evolución temporal: HMMs

3.3.2.1. Clasificación mediante HMMs: los tres problemas.

Vamos a partir del reconocimiento aislado de eventos para estudiar los principios básicos de la teoría de propuesta por Baum (Baum and Petrie, 1966; Baum et al., 1970) sobre modelos ocultos de Markov (*Hidden Markov Models - HMMs*). El objetivo del sistema es asociar la secuencia de vectores de características con la correspondiente secuencia de eventos sismo-volcánicos. La dificulta de esta tarea reside en la variabilidad (tanto en el espacio secuencial como en el de descripción) de los eventos, los efectos de propagación y de efectos de sitio que afectan a la propagación de las ondas sísmicas registradas y las fuentes de ruido que afectan a la señal.

HMMs y cadenas de Markov. Un modelo de Markov es un modelo estadístico para un proceso aleatorio de Markov, esto es, un proceso estocástico que cumple la *propiedad de Markov*: "*el estado futuro al que evoluciona el modelo solo depende de su estado actual y no de los anteriores*". Cada *estado* representa unas propiedades estadísticas concretas del sistema a modelar. En nuestro caso, nos limitaremos a espacios discretos de estados (que tienen un número finito o, al menos, contable de estados) al asumir que los sismogramas han sido generados por un número finito de procesos físicos que asociaremos con los estados del modelo. También suponemos que el dominio secuencial (dado por la evolución temporal de los eventos sismo-volcánicos) es discreto, tal que el paso de un estado a otro se hace de forma síncrona y cada estado representa un intervalo temporal en el que las propiedades del sistema son semi-estacionarias, permitiendo que los observables sean correctamente descritos por los vectores del espacio de características. Un modelo de Markov con un espacio discreto de estados y en el que se evoluciona de una manera discreta entre ellos se denomina una *cadena de Markov* (*Markov Chain - MC*). En un modelo *oculto* de Markov, no se sabe el estado en el que se encuentra el sistema cuando se está observando su salida, permaneciendo la secuencia de estados desconocida u oculta hasta el momento de la decodificación. En la cadena de Markov siempre se tiene acceso a la secuencia de estados.

Un HMM también puede ser descrito como un caso simplificado de una red bayesiana dinámica (DBN) que estima la probabilidad de que un evento $\mathbf{x}$ descrito por una secuencia $\{\mathbf{x}_t\}_{t=1:T}$ de observables haya sido generado por una secuencia $\mathbf{q} = \{q_t\}_{t=1:T}$ no directamente observable u *oculta* de Q estados. El HMM va evolucionando su estado activo de manera *secuencial* sincronizado con la variable t. En cada instante t se pasa desde un estado anterior Q_i al estado actual Q_j y se genera el vector $\mathbf{x_t}$ con una probabilidad de observación $b_j(\mathbf{x}_t)$(Figura 3.3.2). El HMM genera de esta manera un doble modelado; respecto al espacio de características representado por los observables $\{\mathbf{x}_t\}$ y respecto al dominio secuencial representado por la evolución temporal dada por $\{q_t\}$. Formalmente, un modelo M_c de Markov asociado a la clase w_c queda definido como un conjunto de parámetros $\{\boldsymbol{\theta}_{M_c}\} = \{\pi, A, B\}$ siendo:

- $\pi = \{\pi_s\}$ es un vector contiene las *probabilidades de ocupación iniciales*; la probabilidad de que M_c se encuentre el estado Q_s en el instante $t = 1$, es decir, dada la secuencia de estados $\mathbf{q} = \{q_t\}_{t=1:T}$, $\pi_s = P(q_{t=1} = Q_s) = P(q_1 = Q_s)$
- $A = \{a_{ij}\}$ es la *matriz de probabilidad de transición* entre los estados $Q_i \rightarrow Q_j$. Los valores de los coeficientes a_{ij} definen la *topología* del modelo; tanto el número de estados, Q, como las transiciones (lazos) permitidas entre sus estados.
- $B = \{b_{st}\}$ es la *matriz de emisión o salida*, cuyos elementos indican la probabilidad de que el estado Q_s genere el observable $\mathbf{x}_t$. Según sean probabilidades discretas $\{b_{st}\}$ o continuas $\{b_s(\mathbf{x}_t)\}$ se habla de HMM discretos (*Discrete*

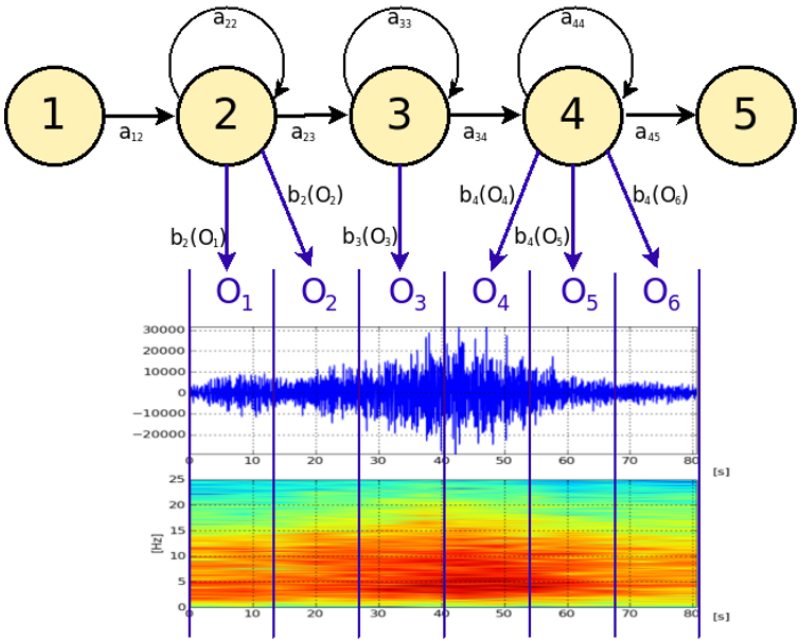


Figura 3.3.2.: Generación en un HMM de la secuencia de observables $\mathbf{O} = \{\mathbf{O}_t\} = \{\mathbf{O}_1, \mathbf{O}_2, \mathbf{O}_3, \mathbf{O}_4, \mathbf{O}_5, \mathbf{O}_6\}$ por la secuencia de estados $\mathbf{q} = \{q_t\} = \{q_1 = Q_2, q_2 = Q_2, q_3 = Q_3, q_4 = Q_4, q_5 = Q_4, q_6 = Q_4\}$. La probabilidad de que un estado Q_s genere el vector $\mathbf{O}_t$ se evalúa mediante la distribución de probabilidad $b_s(\mathbf{O}_t)$.

HMMs - DHMMs) o continuos (*Continuous HMMs - CHMM*). En este trabajo supondremos emisiones continuas de salida con G componentes gausianas multivariadas, correspondientes a modelos GMM de la Subsubsección 3.3.1.1: $b_s(\mathbf{x}_t) = GMM_G(\mathbf{x}_t; \boldsymbol{\mu}; \sum)$.

La construcción de un sistema de reconocimiento basado en HMMs implica la resolución de *los tres problemas*:

1. *Evaluación.* Dado un evento descrito por la secuencia de observaciones $\mathbf{x} = \{\mathbf{x}_t\}$ y un modelo M_c, se pretende evaluar la probabilidad $p(\mathbf{x}|M_c)$ de que $\mathbf{x}$ haya sido generada por M_c.
2. *Decodificación.* Dado un modelo M_c con Q estados y un evento $\mathbf{x} = \{\mathbf{x}_t\}_{t=1:T}$, se quiere encontrar la secuencia óptima de estados $\mathbf{q}_\mathbf{x} = \{q_t\}_{t=1:T}$ que genere el evento $\mathbf{x}$ mediante M_c.
3. *Estimación* o *entrenamiento.* Dado un modelo $M_c = \{\pi, A, B\}$ de una clase w_c y un conjunto de eventos de entrenamiento $\{\mathbf{x}\}$ de dicha clase, el objetivo es estimar los valores de los parámetros $\{\pi, A, B\}$ que mejor modelen a los datos $\{\mathbf{x}\}$, o, equivalentemente, que maximicen la probabilidad $p(\{\mathbf{x}\}|M_c)$.

Solamente el problema de evaluación tiene una solución analítica exacta. Los otros dos pueden ser resolubles bajo distintos enfoques que pueden llevar a distintas soluciones, partiendo en cada caso de hipótesis y aproximaciones (Rabiner, 1989).

1. Evaluación mediante M_c de la secuencia de observables $\{\mathbf{x}_t\}$: algoritmo Avance-Retroceso (*Forward-Backward, FB*). Una vez construido el modelo M_c de la clase w_c y observada una secuencia de vectores de características $\mathbf{x} = \{\mathbf{x}_t\}$, podemos estimar la probabilidad de que los observables $\{\mathbf{x}_t\}$ sean clasificados como un evento w_c como la probabilidad condicional de generar el evento $\mathbf{x}$ dado el modelo M_c; $\hat{p}(\mathbf{x}|w_c) \equiv p(\mathbf{x}|M_c)$. Como apreciamos en la Figura 3.3.2, una misma secuencia $\mathbf{x}$ de T observables puede ser generada por dichos estados mediante distintas secuencias de estados ocultas $\mathbf{q} = \{q_t\}$. Sea $Q_\mathbf{x}$ el conjunto de todas las secuencias $\mathbf{q}$ capaces de generar $\mathbf{x}$. Dado M_c como un conjunto de parámetros $\{\boldsymbol{\theta}_{M_c}\} = \{\pi, A, B\}$ la función verosimilitud del modelo se asocia a $p(\mathbf{x}|M_c)$ tal que $L(\boldsymbol{\theta}_{M_c}|\mathbf{x}) = L_{M_c}(\mathbf{x}) \equiv p(\mathbf{x}|M_c)$. Tratando $p(\mathbf{x}|M_c)$ como una probabilidad marginal sobre $\mathbf{x}$, $p(\mathbf{x}|M_c) \equiv p(\mathbf{x}; M_c)$ y las posibles secuencias $\mathbf{q} \in Q_\mathbf{x}$ como sucesos mutuamente excluyentes, podemos formular (Rabiner and Juang, 1986):

$$\hat{p}(\mathbf{x}|w_c) \equiv p(\mathbf{x}; M_c) = \sum_{\forall \mathbf{q} \in Q_\mathbf{x}} p(\mathbf{x}, \mathbf{q}; M_c) = \sum_{\forall \mathbf{q} \in Q_\mathbf{x}} p(\mathbf{x}|\mathbf{q}; M_c) p(\mathbf{q}; M_c) \tag{3.3.7}$$

Suponiendo independencia estadística entre los observables, para una secuencia de estados concreta $\mathbf{q} = \mathbf{q}_\mathbf{x} = \{q_1, ..., q_T\}$ que genere a los observables $\mathbf{x}$, tal que en cada instante t tenemos $q_t \rightarrow \mathbf{x}_t$, obtenemos:

$$p(\mathbf{x}|\mathbf{q}_\mathbf{x}; M_c) = \prod_{t=1:T} p(\mathbf{x}_t|q_t; M_c) = \prod_{t=1:T} b_{q_t}(\mathbf{x}_t) \tag{3.3.8}$$

Usando la regla de la cadena y definiendo $\pi_{q_1} \equiv a_{q_0 q_1}$, sabemos que la probabilidad de observar la secuencia $\mathbf{q}_\mathbf{x}$ es:

$$p(\mathbf{q}_\mathbf{x}; M_c) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T} = \prod_{t=1:T} a_{q_{t-1} q_t} \tag{3.3.9}$$

con lo que la Ecuación 3.3.7 queda como:

$$\hat{p}(\mathbf{x}|w_c) \equiv p(\mathbf{x}|M_c) \equiv p(\mathbf{x}; M_c) = \sum_{\forall \mathbf{q}_\mathbf{x} = \{q_1, ..., q_T\} \in Q_\mathbf{x}} \left\{ \prod_{t=1:T} a_{q_{t-1} q_t} b_{q_t}(\mathbf{x}_t) \right\} \tag{3.3.10}$$

La evaluación de $\mathbf{x} = \{\mathbf{x}_t\}_{t=1:T}$ directamente mediante la Ecuación 3.3.10 implica un coste computacional inviable, del orden de $2T \cdot Q^T$ operaciones, siendo Q el número de estados de M_c y T el número de vectores del evento $\mathbf{x}$. El algoritmo recursivo

Avance-Retroceso (*Forward-Backward - FB*) de Baum et al. (1967); Baum and Sell (1968) reduce esta complejidad a $T \cdot Q^2$, transformando la Ecuación 3.3.10 en:

$$\hat{p}(\mathbf{x}|w_c) \equiv p(\mathbf{x}; M_c) = \sum_{s=1:Q} \alpha_T(s) = \sum_{s=1:S} p(\mathbf{x}, q_T = q_s; M_c) \tag{3.3.11}$$

donde se definen las probabilidades de observaciones parciales de $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_t, ..., \mathbf{x}_T)$ respecto el instante de observación t:

$\alpha_t(s) = p(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_t, q_t = Q_s; M_c)$ es la probabilidad hacia *adelante,* de estar en el estado Q_s habiendo generado $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_t)$ observables de la secuencia $\mathbf{x}$ en el instante t.

$\beta_t(s) = p(\mathbf{x}_{t+1}, \mathbf{x}_{t+2}, ..., \mathbf{x}_T, q_t = Q_s; M_c)$ es la probabilidad hacia *atrás,* de estar en el estado Q_s si se observara la secuencia parcial $(\mathbf{x}_{t+1}, \mathbf{x}_{t+2}, ..., \mathbf{x}_T)$ de $\mathbf{x}$ a partir del instante t.

Ambas probabilidades se hallan de forma iterativa. Estrictamente, solo las probabilidades hacia adelante son necesarias para evaluar $p(\mathbf{x}|M_c)$, aunque se necesita de ambas para solucionar los problemas de decodificación y entrenamiento.

2. Decodificación: aproximación de Viterbi para encontrar la secuencia óptima de estados $\mathbf{q_x} \in Q_\mathbf{x}$. Se pretende seleccionar la secuencia *óptima* de estados $\mathbf{q_x} = \{q_t\}_{t=1:T}$ de entre todas las que sean capaces de generar los observables del evento $\mathbf{x} = \{\mathbf{x}_t\}_{t=1:T}$. Diferentes criterios de optimización pueden ser adoptados, tales como (Rabiner, 1989):

1. Escoger la secuencia formada por los estados $\{q_t\}_{t=1:T}$ que *individualmente* tengan la probabilidad más alta de estar activos en cada instante t. Matemáticamente:

$$q_t = \arg\max_{s=1:Q}\{\gamma_t(s)\} = \arg\max_{s=1:Q}\left\{\frac{\alpha_t(s)\beta_t(s)}{\sum_{s=1:Q}\alpha_t(s)\beta_t(s)}\right\} \tag{3.3.12}$$

   donde se hace uso de las probabilidades de observación parcial de $\mathbf{x}$ del algoritmo FB, definidas en el apartado anterior.

2. Escoger la secuencia de estados $\mathbf{q_x}$ que maximice la probabilidad condicional para una secuencia $\mathbf{x}$ conocida, lo que equivale (usando la regla de Bayes) a maximizar la probabilidad de obtener el suceso *conjunto* $(\mathbf{x}, \mathbf{q_x})$ dado el modelo M_c:

$$\mathbf{q_x} = \arg\max_{\forall \mathbf{q}=\{q_1,...,q_T\}\in Q_\mathbf{x}}\{p(\mathbf{q}|\mathbf{x}; M_c)\} = \arg\max_{\forall \mathbf{q}=\{q_1,...,q_T\}\in Q_\mathbf{x}}\{p(\mathbf{x}, \mathbf{q}; M_c)\} \tag{3.3.13}$$

Si bien existen otras posibilidades, la forma más empleada es maximizar la probabilidad conjunta $p(\mathbf{x}, \mathbf{q}; M_c)$. El *algoritmo de Viterbi* proporciona un método iterativo basado en la programación dinámica para resolver la Ecuación 3.3.13 de forma eficiente (Viterbi, 1967). Usando criterios como el de la Ecuación 3.3.12 puede llevar en la práctica a indeterminaciones topológicas si los coeficientes $\{a_{ij}\}$ de la matriz A de transición entre estados no permiten alcanzar el estado individualmente óptimo en el instante t.

Nótese que la solución dada por la Ecuación 3.3.13 proporciona una estimación en la evaluación de $\mathbf{x}$ por M_c, aproximando el valor de la sumatoria sobre todas las secuencias posibles por el valor de la secuencia que maximiza $p(\mathbf{x}, \mathbf{q}; M_c)$:

$$\hat{p}(\mathbf{x}|w_c) \equiv p(\mathbf{x}; M_c) = \sum_{\forall \mathbf{q} \in Q_{\mathbf{x}}} p(\mathbf{x}, \mathbf{q}; M_c) \approx \max_{\forall \mathbf{q} \in Q_{\mathbf{x}}} \{p(\mathbf{x}, \mathbf{q}; M_c)\} \tag{3.3.14}$$

Esta simplificación hace posible reducir el coste computacional del algoritmo a un orden de $\mathcal{O}\{T \cdot Q^2\}$.

3. Construcción de M_c: estimación de los parámetros (π, A, B) que mejor modelen los datos de entrenamiento $\{\mathbf{x}\}$. En la fase de entrenamiento un modelo de Markov $M_c = \{\pi, A, B\}$ asociado con la clase w_c se construye hallando el valor de sus parámetros a partir de un conjunto de datos etiquetados como $\{\mathbf{x}\} \rightarrow w_c$. Esta es la mayor dificultad que existe al diseñar un sistema basado en HMMs y la más estudiada. En la literatura se han propuesto distintos enfoques para ajustar estos parámetros:

1. Maximizar la *la probabilidad de generar los datos dado el modelo,* $p(\{\mathbf{x}\}|M_c)$. No hay manera analítica de resolver este problema. Ni tan siquiera se puede encontrar un máximo absoluto de forma recursiva (Rabiner, 1989). Se han usado métodos iterativos como el descenso en gradiente (GD) (Levinson et al., 1983), o estimadores de máxima verosimilitud (MLE) como el algoritmo esperanza-maximización (EM) (Dempster et al., 1977).
2. Maximizar la *información mutua* (*Maximum Mutual Information - MMI*) entre el modelo y las observaciones, $I(\{\mathbf{x}\}, \mathbf{q}_{\{\mathbf{x}\}}; M_c)$ (Merialdo, 1988; Normandin, 1996)
3. *Métodos mixtos*. Usualmente formulan problemas de optimización de parámetros definiendo funciones objetivo que sean convexas, llevando a un mínimo global. Autores como Oliver and Garg (2002) maximizan conjuntamente la verosimilitud y la información mutua entre observaciones y estados. Otros como Sha and Saul (2006); Jiang et al. (2006) usan técnicas de clasificadores de *amplio margen* (SVM) que definen las reglas de decisión maximizando una función distancia media entre clases para estimar los parámetros de los HMMs.

El método clásico más usado es la *re-estimación de Baum-Welch (BW)* (Baum et al., 1970). Es una particularización del algoritmo recursivo EM que permite usar estimadores MLE originalmente diseñados para tratar con pérdida de datos la cual se

identifica con las secuencias ocultas de estados. Utilizando las variables $\alpha_t(s)$, $\beta_t(s)$ y $\gamma_t(s)$ del algoritmo FB para representar la probabilidad de transición en el instante t desde los estados $Q_i \rightarrow Q_j$, $\xi_t(i,j) = P(q_t = Q_i, q_{t+1} = Q_j|\{\mathbf{x}\}; M_c)$, se pueden estimar los parámetros $\{\boldsymbol{\theta}_{M_c}\} = \{\pi, A, B\} \backsimeq \{\hat{\boldsymbol{\theta}}_{\hat{M}_c}\} = \{\hat{\pi}_s, \hat{a}_{ij}, \hat{b}_s(\mathbf{x}_t)\}$:

$$\hat{\pi}_s = \frac{n(q_1 = Q_s)}{n(q_1)} = \gamma_1(s) \tag{3.3.15}$$

$$\hat{a}_{ij} = \frac{n(Q_i \rightarrow Q_j)}{n(Q_i)} = \frac{\sum_{t=1:T-1} \xi_t(i,j)}{\sum_{t=1:T-1} \gamma_t(s)} \tag{3.3.16}$$

$$\hat{b}_s(\mathbf{x}_t) = \frac{n(\mathbf{x}_t, Q_s)}{n(Q_i)} = \frac{\sum_{q_t=Q_s, t=1:T-1} \gamma_t(s)}{\sum_{t=1:T-1} \gamma_t(s)} \tag{3.3.17}$$

donde $n(\Delta)$ se define como el número de veces (frecuencia) que ocurre el suceso Δ al analizar los datos $\{\mathbf{x}\}$. Baum et al. (1967) demostraron que cada vez que se hace esta estimación la probabilidad $p(\{\mathbf{x}\}|\hat{M}_c)$ de generar los eventos $\{\mathbf{x}\}$ a partir del modelo no empeora, lo que es condición suficiente para incrementar la verosimilitud del modelo $\hat{M}_c$ alcanzando la convergencia al máximo local más cercano. Esta *convergencia* se satisface cuando:

$$\Delta J\{M_c\} = p(\{\mathbf{x}\}|\hat{M}_c^{(r)}) - p(\{\mathbf{x}\}|\hat{M}_c^{(r-1)}) < T \tag{3.3.18}$$

siendo $\Delta J\{M_c\}$ el incremento en la función objetivo de la estimación MLE del modelo M_c y T un valor umbral previamente definido. Este algoritmo tiene la ventaja de converger rápidamente (tras un número r de iteraciones del orden de [10-100]). Al igual que los algoritmos FB y Viterbi tiene una complejidad de $\mathcal{O}\{T \cdot Q^2\}$ en cada iteración r. Sin embargo, presenta dos serios inconvenientes:

1. *El carácter local de la maximización* implica que el algoritmo es *muy sensible al punto de partida* de la recursión. Dado que experimentalmente la función objetivo $J\{M_c\}$ suele tener una forma compleja y multimodal, con varios máximos locales, una correcta inicialización de parámetros es crucial (Rabiner, 1989).
2. La *aproximación frecuentista*, basado en el cómputo de frecuencias de que los sucesos sean originados, del enfoque MLE requiere una *gran cantidad de datos* para una estimación fiable. Si el modelo es muy complejo al (tiene demasiados parámetros a estimar) se corre el riesgo de sobre-entrenarlo.

3.3.2.2. Extensión al reconocimiento en continuo mediante HMMs

Una de las ventajas de los HMM frente a otros clasificadores es que la fase de detección de eventos sobre un flujo continuo de datos y la de clasificación pueden realizarse

conjuntamente mediante la decodificación por Viterbi si se modela el ruido como un evento más. El paso de clasificación de eventos aislados a reconocimiento de señales continuas se soluciona simplemente construyendo un HMM general (*macro-HMM*, M_{RB}) o *red de búsqueda* que englobe a los HMMs de las C clases (M_{w_1}, M_{w_2}, ..., M_{w_C}) cuya estructura viene determinada por la interrelación que los modelos tienen entre sí, definida por unas reglas gramaticales de *modelado de lenguaje* como vimos en la Subsubsección 2.1.4.3 (Figura 3.3.3). En el caso del reconocimiento de señales

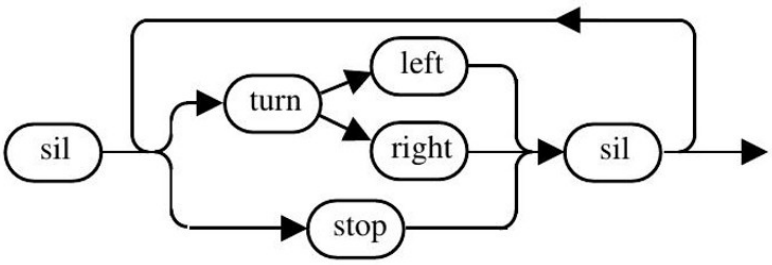


Figura 3.3.3.: Red de búsqueda o macro – HMM, $\mathbf{M_{RB}}$, construido a partir de los modelos $\{M_{sil}, M_{stop}, M_{turn}, M_{left}, M_{right}\}$ y reglas gramaticales que permiten frases con la estructura $(sil < (turn(left|right)|stop)sil >)$. Figura original de Young (1994).

sismo-volcánicas, estas reglas las usamos solo para discernir entre reconocimiento continuo o clasificación de eventos aislados. Construida la red, se halla el camino (o secuencia óptima de estados $\{\mathbf{q}_{\{\mathbf{x}\}}\}$) más probable de generar $\{\mathbf{x}\}$ a partir de la red M_{RB}. La secuencia óptima nos delimita temporalmente a cada evento, simplemente asociando cada subsecuencia de estados a los modelos de clase a los que pertenecen $\{\mathbf{q}_{\{\mathbf{x}\}}\} = \{q_1, ..., q_T\} = \{\mathbf{q}_{w_i}, \mathbf{q}_{w_j}, ..., \mathbf{q}_{w_k}\}$ y analizando el tiempo que se ha tardado en atravesar cada modelo M_{w_c} generado por la secuencia $\mathbf{q}_{w_c}$. Hallar $\{\mathbf{q}_{\{\mathbf{x}\}}\}$ equivale a resolver el problema de evaluación (Sección 3.3.2.1) del macro modelo M_{RB}. Se suele usar versiones de algoritmos adaptadas a modelos complejos y grandes espacios de búsqueda que también incluyan explícitamente modelos estocásticos de lenguaje en la decodificación (Gales and Young, 2008).

3.4. Criterios de evaluación

En el área del aprendizaje automático existen distintas funciones para evaluar los resultados de un sistema de clasificación. A continuación haremos un resumen de los métodos de evaluación más comunes, centrándonos en los que utilizaremos para medir la bondad de los sistemas VSR.

3.4.1. Criterios de evaluación generales

Como ya vimos en la Figura 2.2.2, la forma más común para evaluar un sistema de reconocimiento consiste en contabilizar el número errores debido a eventos ignorados

o borrados (D), insertados (I) o confundidos o sustituidos por otros (S), siendo N el número total de eventos en la partición de evaluación y H los eventos que se han reconocido correctamente. A partir de ellos se define:

$$\%Acc = 100\frac{H-I}{N} = 100\frac{N-D-S-I}{N} \tag{3.4.1}$$

$$\%Corr = 100\frac{H}{N} = 100\frac{N-D-S}{N} \tag{3.4.2}$$

El $\%Acc$ representa el porcentaje de *eficiencia* (o *accuracy*) de reconocimiento del sistema y $\%Corr$ el porcentaje *aciertos* o de eventos correctamente asignados a sus clases. Estas funciones pueden computarse como para eventos solo de una clase concreta como para todos los eventos existentes en el corpus de evaluación incluso puede utilizarse a nivel de ficheros.

```
------------------------- Overall Results ----------------------
SENT: %Correct=42.11 [H=16, S=22, N=38]
WORD: %Corr=93.89, Acc=71.76 [H=123, D=3, S=5, I=29, N=131]
------------------------ Confusion Matrix -----------------------
      HY  LPS SIL  TS  VT Del [ %c / %e ]   %Corr   %Acc
  HY  10   0   0   3   0   0  [76.9/ 2.3]    76.92   61.54
 LPS   0  26   0   0   0   1  [100.0/ 0.0]   96.30   77.78
 SIL   0   0  64   2   0   2  [97.0/ 1.5]    94.12   79.41
  TS   0   0   0  14   0   0  [100.0/ 0.0]  100.00   42.86
  VT   0   0   0   0   9   0  [100.0/ 0.0]  100.00   55.56
 Ins   2   5  10   8   4    ---------------------------------------
                              class_mean:    93.47   63.43
-----------------------------------------------------------------
SENT: %Ign=0.00, [ 0 /  38] data files ignored in recognition
WORD: %Ign  [ign / test / recog]  %Corr   %Acc   %Rel
  HY   0.00  [ 0 /  13 /  13 ]  ->  76.92   61.54  -14.98
 LPS   0.00  [ 0 /  27 /  27 ]  ->  96.30   77.78  -17.29
 SIL   0.00  [ 0 /  68 /  68 ]  ->  94.12   79.41  -17.71
  TS   0.00  [ 0 /  14 /  14 ]  ->  100.00  42.86  -18.18
  VT   0.00  [ 0 /   9 /   9 ]  ->  100.00  55.56  -24.08
-----------------------------------------------------------------
total  0.00  [ 0 / 131 / 131 ]  ->  93.47   63.43  -18.45
=================================================
```

Figura 3.4.1.: Resultados de evaluación de reconocimiento del sistema VSR. Se muestra la matriz de transición junto con la columna (*Del*) de eventos borrados y la fila (*Ins*) de insertados junto información adicional.

La Figura 3.4.1 muestra la salida típica de evaluación nuestro sistema VSR, que es una extensión de la información que da el HTK (Young et al., 2006). En ella se muestra la popular *matriz de transición* entre las clases consideradas, donde las filas de dicha matriz representa la distribución de la clasificación que el sistema hace para los eventos de una clase. La tasa de reconocimiento será mayor cuantos más eventos se acumulan en la diagonal de la matriz. La fila *Ins* representa las inserciones del

sistema en cada clase y la columna *Del* los borrados. La columna *%Rel* calcula la probabilidad media $p(\mathbf{x}|M_c)$ que el modelo M_c de la clase c genere la secuencia $\mathbf{x}$ de los vectores de características que describen a los eventos asignados a la clase c por el sistema.

3.4.2. Evaluación promediada por clase.

El propósito de este criterio es compensar la evaluación sesgada (Figura 2.2.2) que por defecto entrega el HTK para que todas las clases tengan el mismo peso en el cómputo final, independientemente de su número de eventos, para evitar situaciones en las que la mayor parte de *%Corr* y *%Acc* quedan asociadas a la clase con más eventos, usualmente el ruido, que normalmente engloba más del 50 % de los eventos de cualquier base de datos, enmascarando los resultados del resto de clases y dando una medida equivocada de la bondad del reconocedor.

La solución adoptada en este trabajo consiste en normalizar las funciones *%Corr* y *%Acc* computadas para cada clase por su respectivo número de eventos antes de promediarlas, proporcionando un operador general para todo el sistema: las funciones *%cCorr* y *%cAcc* detalladas en la Figura 3.4.1 como los valores que aparecen junto a *class_mean*. Pueden computarse como:

$$\%cAcc = mean_w\{\%Acc(w)\} \tag{3.4.3}$$

$$\%cCorr = mean_w\{\%Corr(w)\} \tag{3.4.4}$$

donde $\{w\}$ representan las clases a considerar del corpus de datos.

3.4.3. Re-evaluación geofísica de resultados de reconocimiento.

Como presentamos en la Subsección 2.2.2, una de las peculiaridades de aplicar los sistemas de reconocimiento a señales sismo-volcánicas es que la aplicación directa de los criterios de evaluación es bastante imprecisa (Figura 2.2.2). Concretamente, desde un punto de vista geofísico no todas las clases son igual de importantes (por lo que su evaluación debería pesarse según la relevancia de los eventos) y no todos los borrados e inserciones pueden ser considerados como errores al mismo nivel que las sustituciones de eventos.

Con el objetivo de incorporar estas hechos en la evaluación, hemos implementado un filtro de resultados que ignora los errores no relevantes geofísicamente hablando. Se aplica a la salida de evaluación del sistema VSR (Figura 3.4.1) actuando sobre eventos de las clases consideradas *no secuenciales* (Tabla 6.2.1), en referencia a que no presentan un patrón definido de evolución de sus características en el dominio temporal, tales como los tremores y ruido. No se contabilizan:

- Errores de borrado de eventos no secuenciales siempre que estos aparezcan juntos en el sismograma
- Errores de inserción de eventos no secuenciales siempre que dichos eventos aparezcan juntos

3.4.4. Otras medidas de evaluación.

Con el objetivo de hacer una evaluación más completa, otras funciones de puntuación han sido usadas en la literatura. La mayoría de ellas solo consideran 2 clases, una P positiva y otra N negativa, sobre los que se estudian los siguientes indicadores:

TP *True Positives* o positivos correctos, eventos correctamente reconocidos de la clase P.

TN *True Negatives* o negativos correctos, eventos correctamente reconocidos de la clase N.

FP *False Positives* o positivos incorrectos, eventos erróneamente reconocidos de la clase P.

FN *False Negatives* o negativos incorrectos, eventos erróneamente reconocidos de la clase N.

Donde los eventos incorrectos representan las inserciones en la correspondiente evaluación del HTK. A partir de ellos se definen otras funciones de evaluación como:

Eficiencia $= Acc = \frac{TP+TN}{TP+TN+FP+FN}$ asociada al éxito de reconocimiento sobre la clase conjunta $P + N$, donde el numerador representa el nº de eventos correctamente reconocidos y el denominador representa el nº total de eventos reconocidos, incluyendo las inserciones.

Sensibilidad $= Sen = \frac{TP}{TP+FN}$ asociada al éxito de reconocimiento sobre la clase P. $TP + FN$ es el nº de eventos de la clase P.

Especifidad $= Esp = \frac{TN}{TN+FP}$ asociada al éxito de reconocimiento sobre la clase N: la razón de los eventos bien reconocidos entre todos los que hay de la clase N.

La curva $ROC = Sen = f(1 - Esp)$, *Receiver Operating Characteristic*, engloba los conceptos de sensibilidad y especificidad, su área (*Area Under ROC* - *AUROC*) se considera un indicador de la eficacia generalizada de un reconocedor binario.

Otra medida muy interesante es el coeficiente K o Kappa de Landis and Koch (1977) que evalúa el grado de coincidencia que existe entre distintos etiquetadores (por ejemplo, entre el sistema VSR automático y un técnico experto), utilizado por Curilem et al. (2014b).

3.5. Metodología experimental

A lo largo de este capítulo y de los siguientes, a no ser que se especifique otra configuración, todas las pruebas experimentales usarán HMMs como modelos de predicción. Generalmente utilizaremos las versiones de eventos continuos de las bases de datos maestras *dec.95M.9c* y *col.04M.15c* definidas en la Sección 3.6. Cada corpus de la base de datos se divide en 3 partes complementarias y se promedia los resultados de 3 test abiertos, cada test usa 2 de esas partes para construir los modelos, referidas conjuntamente como la *partición de entrenamiento* y la otra, la *partición de evaluación* para evaluarlos.

Configuración estándar de los test

Dado que las bases de datos del volcán de Colima tienen eventos en general de más variabilidad y duración que las de la isla Decepción, escogeremos 5 estados (3 emisores) en los HMMs que modelen datos de Decepción y 7 (5 emisores) en el caso de Colima. Por la misma razón, en Decepción los segmentos (vectores de características u observaciones) tendrán una duración de 2 segundos frente a los 4 de Colima. El solapamiento entre segmentos queda fijado en un valor típico del 50 %. Excepcionalmente, en algunos tests con Decepción se usará un mayor solapamiento para evitar modelos sobre-entrenados (ver la Sección A.4 para una explicación más detallada). Por el mismo motivo, en Decepción el tamaño máximo del segmento será de 5 segundos en vez de los 10 de Colima.

Medida de evaluación de resultados

Como normal general, el criterio escogido para valorar los resultados de un esquema o configuración frente a otra lo basaremos en el porcentaje de eficiencia de reconocimiento promediado por clase ($\%cAcc$) definido previamente en en la Sección 3.4. Sin embargo, no siempre elegiremos el esquema cuyos resultados logren el mayor $\%cAcc$, pues como veremos en el Capítulo 4, atendiendo a los principios de eficacia (Ecuación 4.1.1), a veces un ligero aumento en la tasa de reconocimiento no justifica una elección de un vector con muchos componentes en lugar de otro de tamaño más pequeño pero con un $\%cAcc$ algo inferior.

Junto a la eficiencia promediada $\%cAcc$, usaremos el esquema de evaluación *base* o clásico frente a la interpretación geofísica o compensación de inserciones. Los motivos de esta decisión son varios:

- El esquema clásico es más utilizado en la literatura, lo que nos permite comparar de manera más equitativa nuestros resultados con otros trabajos.
- Consideramos que la interpretación geofísica es más adecuada para evaluar un sistema una vez estén configurados y construidos los modelos. La mayor parte del trabajo en esta tesis consiste precisamente en configurar y construir modelos, cuyo objetivo es mejorar los errores de reconocimiento cometiendo las mínimas inserciones y borrados posibles. El esquema geofísico ignora algunos

de esos errores, por lo que nos da una visión parcial en la evaluación que no ayuda a mejorar los modelos.

- La compensación de borrados e inserciones es una evaluación que necesita determinar a posteriori el mejor valor de penalización (HMM_PENAL) en la red de búsqueda. Aunque incrementa mejora de forma notable los resultados tiene el inconveniente de que varía no solo del tipo de eventos a considerar, si no también del nº de eventos en cada fichero de datos de cada corpus, parámetro que depende de cada realización de cada corpus de datos y, por tanto, poco exportable y generalizable de un volcán a otro e, incluso, de una base de datos a otra.

3.6. Bases de datos maestras

En respuesta a los problemas acarreados por la falta de fiabilidad de las bases de datos (Subsección 2.2.2) en este trabajo optamos por usar bases de datos *maestras* que contienen eventos típicos de cada clase para minimizar el error en el etiquetado manual y contribuir a la construcción de modelos robustos. La creación de estos corpus se realiza a partir de una gran base de datos inicial DB_0 de forma semi-supervisada para minimizar los errores subjetivos (Subsección 2.2.2) al etiquetar. Tanto como para Decepción como para Colima construiremos 2 versiones de la misma base maestra: una con registros en continuo (usadas por defecto) y otra con eventos aislados que se utilizará en el Capítulo 4 de reducción de dimensionalidad. El proceso iterativo de construcción viene detallado en el Algoritmo 3.1.

La DB_0 puede estar no etiquetada si se tienen unos modelos adecuados para realizar un primer reconocimiento no-supervisado o puede ser un corpus ya etiquetado para construir los modelos a a partir de dichas etiquetas en el proceso de entrenamiento del ciclo inicial. Hay que tener especial cuidado en no construir un corpus demasiado pequeño con el que es más posible sobre-entrenar los modelos. Estas bases de datos maestras se han usado en numerosos trabajos sobre reconocimiento VSR y reducción de dimensionalidad Álvarez et al. (2011); Cortés et al. (2014, 2015).

3.6.1. Base de datos maestra de Decepción: *dec.95M*

3.6.1.1. Adquisición de datos

Nuestra base de datos maestra proviene del mismo corpus usado en Benítez et al. (2007), donde se puede encontrar una información detallada del proceso de adquisición de datos. Fueron registrados en la campaña de 1994-1995 organizada por el Instituto de Geofísica de Andalucía (IAG) durante el verano antártico.

Algoritmo 3.1 Construcción de bases de datos maestras.

Sean:

$DB_0(W_0)$ una base de datos (etiquetada o no) con un conjunto $W_0 = \{w_{0,0}, ..., w_{0,C(0)}\}$ con $C(0)$ clases, donde la clase $w_{0,c,}$ tiene $n_E(0, c)$ eventos

$\{HMM_{C(0)}\}$ un conjunto de modelos, cada uno de ellos describiendo una clase de W_0

Se realiza iterativamente el siguiente ciclo hasta considerar que el nivel de eficiencia $\%cAcc$ de reconocimiento de la base de datos ha alcanzado un mínimo exigido, o, en su defecto, que alguna clase tenga un número de eventos demasiado pequeño como para seguir eliminando eventos de ella:

1. Se particiona equitativamente la base DB_i en 2 conjuntos: $DB_{i,1}$ y $DB_{i,2}$
2. Se construye 2 respectivamente 2 conjuntos de modelos $HMM_{i,1}$ y $HMM_{i,2}$ cada uno respectivamente con los eventos de $DB_{i,1}$ y $DB_{i,2}$
3. En un test abierto, se reconocen los eventos de $DB_{i,2}$ con los modelos de $HMM_{i,1}$ y $DB_{i,1}$ con $HMM_{i,2}$, obteniendo un promedio $\%cAcc_i$ de reconocimiento
4. Las etiquetas obtenidas del reconocimiento en abierto son revisadas por el técnico experto, modificando la base DB_i según:
   a) Se eliminan eventos que no sean fácilmente asignables a alguna de las clases del conjunto W_i considerado o que tengan una tasa de efectividad $\%cAcc(c)$ más baja que un nivel mínimo requerido
   b) Se eliminan clases con un número de eventos demasiado pequeño o se integran en otras clases con propiedades parecidas
   c) Ocasionalmente se pueden crear nuevas clases (o *subclases*) si existe un número suficiente de eventos que compartan un mismo patrón de características no claramente asignable a las clases de W_i.

   DB_{i+1} tendrá un nuevo conjunto de clases W_{i+1}, cuyo elemento $w_{i+1,c,}$ tiene $n_E(i+1, c)$ eventos.

La Figura 3.6.1 detalla la localización de 3 arrays de corto periodo con 8 canales cada uno usado en la adquisición de datos, entre las bases argentina y española. Un array se compone de:

- 1 sismómetros Mark - L4C de 3 componentes y una frecuencia de banda inferior $f_L = 1$ Hz.
- 5 sensores Mark - L25 de una componente vertical, con una frecuencia natural de 4.5 Hz. extendida electrónicamente a 1 Hz.

Tras un análisis de los 24 canales, se escoge aquel que tiene la mayor SNR (canal $5C$,

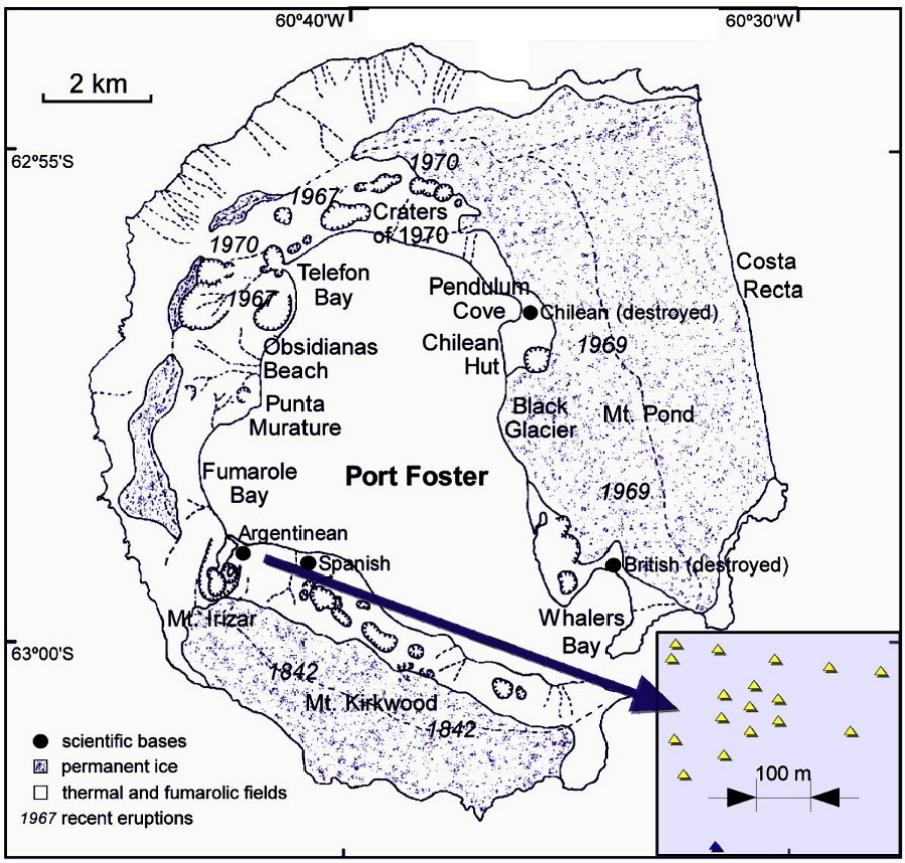


Figura 3.6.1.: Array de adquisición de datos en la isla Decepción. La localización de los sismómetros del array se señalan mediante triángulos. El triángulo azul corresponde a la estación escogida (la de mayor relación señal/ruido) para extraer nuestras señales. Se muestran además fechas relativas a erupciones históricas y su localización en la isla.

marcado como un triángulo azul en la Figura 3.6.1) para extraer las señales menos ruidosas. Todos los sensores usan un formato de enteros de 16 bits para digitalizar una muestra a una frecuencia de muestreo de 200 Hz. Con el objetivo de eliminar la influencia del ruido oceánico sobre las señales (Ibáñez et al., 2000) los datos se filtran en la banda $[1, 25]$ Hz. mediante un filtro de Butterworth de fase nula y orden 2+2. Posteriormente se submuestrea a 50 Hz. para evitar complejidad computacional. La grabación de la señal se disparaba por un sistema de detección de eventos basado en STA/LTA que solo permitía registros de una duración máxima de 150 segundos.

3.6.1.2. Construcción de la base de datos maestra

La base de datos maestra de Decepción se construye bajo las directrices dadas por la Sección 3.6, a partir de un corpus inicial $dec.95I$ de unos 2200 eventos registrados

en continuo que es una versión del original que Benítez et al. (2007) utilizan en su sistema de reconocimiento. Las clases y eventos de la base inicial se listan en la Tabla 3.6.1.

clase	eventos	min[s]	media[s]	max[s]	total[s]	std[s]	%std/media
HY	54	7.8	29.4	136.8	1587.1	18.90	64.32
LP	765	2.4	9.8	30.7	7469.8	3.81	38.98
NS	1222	0.3	15.4	128.2	18835.2	11.80	76.54
TS	77	10.4	93.3	150.0	7184.2	43.63	46.77
VT	75	5.4	19.1	80.9	1434.5	12.88	67.34

2193 eventos/520 ficheros	4.22 eventos/fichero
10.14 [horas]	70.21 [s]/fichero, 16.65 [s]/evento

Tabla 3.6.1.: *Estadísticas de la base de datos inicial de Decepción dec.95I.*

Dados los datos de *dec.95I*, aplicamos el proceso descrito en el Algoritmo 3.1 creando nuevos tipos de clase en el etiquetado semi-supervisado tales como subclases de eventos de largo periodo, terremotos regionales, varios tipos de tremores y otros eventos que difícilmente podían ser asignados a un tipo de clase concreta, designándose como eventos basura (*GAR*). Tras la larga y ardua tarea de re-etiquetado, finalmente construimos la base de datos maestra *dec.95M* con solo 5 tipos de eventos: híbridos (*HY*), de bajo periodo (*LP*), ruidosos (*NS*), tremores (*TR*) y sismos volcano-tectónicos (*VT*) (Tabla 3.6.2).

clase	eventos	min[s]	media[s]	max[s]	total[s]	std[s]	%std/media
HY	41	10.0	21.9	37.0	897.0	6.95	31.77
LP	58	4.0	12.2	21.0	707.0	3.59	29.44
NS	187	3.0	15.4	72.0	2880.0	9.94	64.55
TS	47	5.0	53.4	143.0	2508.0	38.68	72.49
VT	41	6.0	14.8	28.0	608.0	4.85	32.70

374 eventos/115 ficheros	3.25 eventos/fichero
2.11 [horas]	66.09 [s]/fichero, 20.32 [s]/evento

Tabla 3.6.2.: *Estadísticas de duración de la base de datos maestra dec.95Mc.*

La Figura 3.6.2 muestra los histogramas de duración de cada clase. La complejidad del modelado temporal está relacionada con la variabilidad en la duración de los eventos, y, como puede observarse, ninguno de ellos se adapta a una distribución normal.

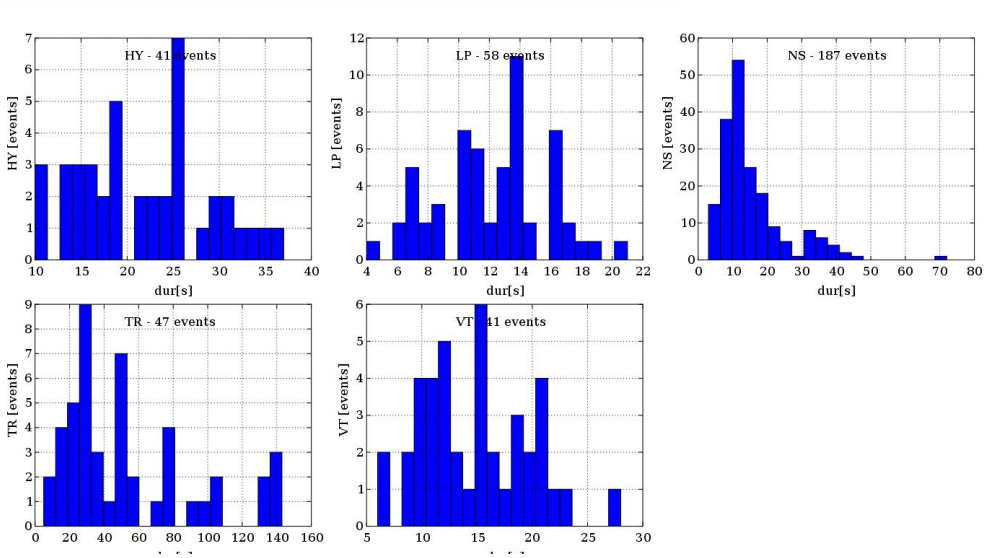


Figura 3.6.2.: Distribución temporal del corpus maestro *dec.95M* para los eventos híbridos (*HY*), de bajo periodo (*LP*), ruidosos (*NS*), tremores (*TR*) y sismos volcano-tectónicos (*VT*).

3.6.2. Base de datos maestra de Colima: *col.04M*

3.6.2.1. Adquisición de datos

La Figura 3.6.3 indica la localización de las estaciones de monitorización supervisadas por la Red Sismológica del Estado de Colima (*RESCO*). En el registro de datos se utilizan 3 sensores repartidos en 2 estaciones (Arámbula-Mendoza et al., 2011):

- Estación Soma: a 1.7 km. del cráter, dotada con un sensor identificado como $EZV4$ de corto periodo (SS-1 Ranger, $T_s = 1$ s.).
- Estación Fresnal: a unos 5 kms., con un sensor, $EZV5$, de corto periodo (SS-1 Ranger, $T_s = 1$ s.) y $EZ5V$ de banda ancha (Guralp CMG-40 TD, desde $T_s = 30$ s. hasta $f_H = 50$ Hz.).

Las señales fueron analizadas mediante la GUI de Matlab desarrollada por Lesage (2009). Los datos se recolectaron durante los años 2004-2006 en un episodio de gran actividad que incluyó explosiones, derrumbes y lahares. El sistema de adquisición usado es el Earthworm (Johnson et al., 1995) desarrollado por el USGS (United States Geological Survey) que está continuamente monitorizando el volcán y cada 2 a 5 minutos digitaliza los datos en formato SEISAN a una frecuencia de 100 muestras/s. Los sensores de corto periodo usan un conversor analógico/digital de 16 bits, mientras que el de banda ancha digitaliza cada muestra con 24 bits. El etiquetado supervisado se llevó a cabo por 3 técnicos expertos. Al final del proceso todos los ficheros se limitan a una banda espectral de [1,25] Hz. Cada muestra en el sismograma se discretiza con un formato de enteros de 16 bits y se submuestrea a

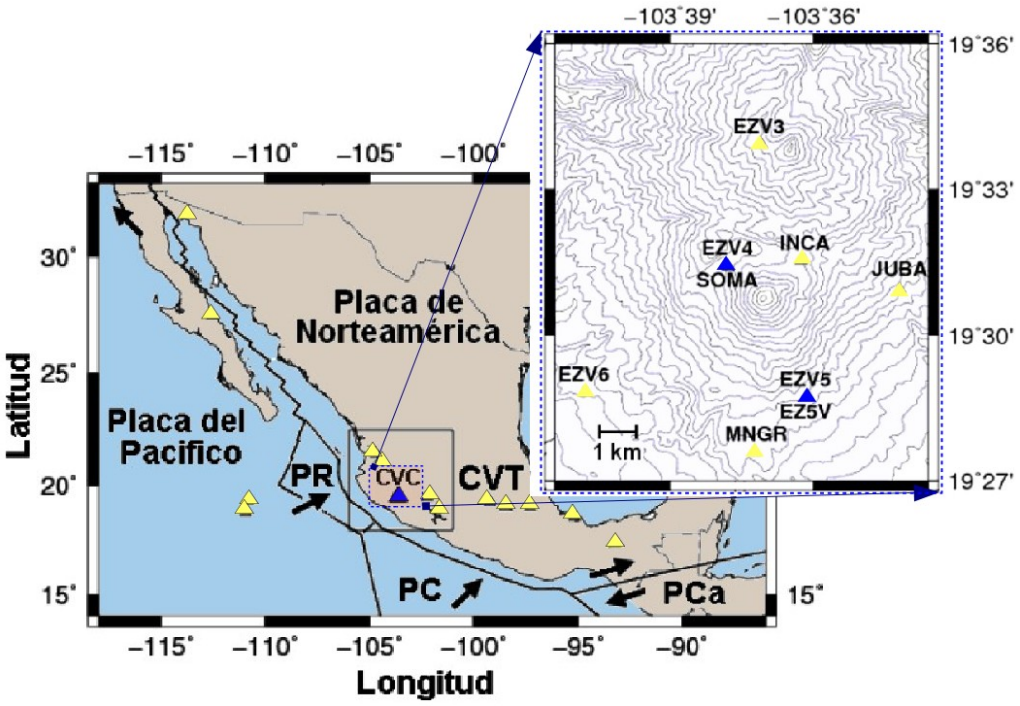


Figura 3.6.3.: Localización de estaciones de monitoreo en el volcán de Fuego de Colima. La localización de las estaciones se señalan mediante triángulos. Los triángulos azules corresponden a las estaciones utilizadas en la creación de la base de datos: Soma (sismómetro de corto periodo $EZV4$) y Fresnal ($EZV5$ de corto periodo y $EZ5V$ de banda ancha). Figuras modificadas de Arámbula (2011).

50 Hz.

3.6.2.2. Construcción del corpus a partir de la base inicial *col.04I*

La creación del corpus maestro comienza con el análisis, re-etiquetado supervisado de una colección de 200 horas de señales registradas durante el periodo eruptivo de 2004-2006 usada en Cortés et al. (2009b). La base inicial $col.04I$ se compone de 10 clases listadas en la Tabla 3.6.3:

A pesar de ser estar registrada en continuo, el número de eventos en cada fichero es muy pequeño, solo de 1.25. esto se explica porque en las transcripciones originales de esta base la mayoría de eventos estaban precedidos por eventos basura (GAR) que había que eliminar en el proceso de entrenamiento de los modelos, y, por tanto, hubo que trocear los ficheros en continuo cada vez que aparecía un evento GAR. Aún así, la enorme variabilidad temporal y en el espacio de características de las señales de Colima, las hace difíciles de modelar. A esto hay que añadir dos propiedades que la hace muy particular:

1. La amplitud de las señales está normalizada en cada fichero al máximo de su

clase	eventos	min[s]	media[s]	max[s]	total[s]	std[s]	%std/media
COL	669	11.5	107.1	298.2	71675	54.53	50.90
EXP	581	26.5	98.5	876.6	57237	51.88	52.67
LAH	80	1331.0	2715.1	5076.3	217210	792.69	29.20
LPS	554	26.9	48.0	502.1	26582	22.96	47.85
REG	357	62.8	149.3	363.1	53282	44.62	29.90
TP	691	10.5	54.3	4273.4	37544	221.28	407.28
TR	222	18.9	194.0	2457.8	43066	282.94	145.85
TS	28	144.4	548.1	1067.6	15345	242.82	44.31
VT	384	18.9	47.5	134.9	18219	14.15	29.82
WNS	1120	0.5	144.5	1788.9	161885	200.72	138.87

4686 eventos/3751 ficheros	1.25 eventos/fichero
195.01 [horas]	187.16 [s]/fichero, 149.82 [s]/evento

Tabla 3.6.3.: *Estadísticas de duración de la base de datos col.04I.*

margen dinámico (16 bits, en muestras de enteros)

2. Existen muchos eventos que saturan el sismogramas, por lo que tienen que ser eliminados

El corpus *col.04I* se reduce desde 195 horas a unas 50 y de 4686 eventos a 669. En el proceso de re-etiquetado supervisado se definen clases intermedias como distintos tipos de eventos de largo periodo: resonantes con múltiples frecuencias (*LPR*) o con una sola dominante (*LP*0) y espasmódicos (*LPS*).

Finalmente, la base de datos maestra *col.04M* contiene 669 eventos repartidos en 11 clases: colapsos (*COL*), explosiones (*EXP*), lahares (*LAH*), eventos de largo periodo (*LPS*); 2ª fila: sismos regionales (*REG*), tremores pulsantes pequeños (*SPT*), pulsantes (*TP*) y resonantes (*TR*); 3ª fila: tremores espasmódicos (*TS*), sismos volcano-tectónicos y ruidos (*WNS*). Sus estadísticas de duración se muestran en la Tabla 3.6.4. La Figura 3.6.4 detalla la distribución temporal de las clases de la base *col.04Mc*. Se observa una gran variabilidad en la mayoría de ellas.

3.7. Construcción de los sistemas base

En este apartado implementaremos de forma práctica nuestros modelos base y los evaluaremos para ver su eficiencia. Seguiremos las directrices teóricas y prácticas presentadas en los trabajos previos de Benítez et al. (2009); Ibáñez et al. (2009); Cortés et al. (2009a) de reconocimiento mediante HMMs haciendo hincapié en los aspectos técnicos relacionados con HTK.

El esquema del sistema nuestro sistema de reconocimiento en continuo basado en HMMs se presenta en la Figura 3.7.1 y se estructura en las mismas etapas que ri-

clase	eventos	min[s]	media[s]	max[s]	total[s]	std[s]	%std/media
COL	77	35.0	123.6	294.0	9519	48.25	39.05
EXP	45	35.0	82.8	168.9	3726	30.44	36.75
LAH	30	1331.0	2750.9	3952.3	82528	684.74	24.89
LPS	62	30.0	44.7	70.0	2772	7.51	16.80
REG	45	82.7	139.1	249.0	6285	33.81	24.31
SPT	86	16.0	30.5	48.0	2624	8.47	27.76
TP	37	65.0	382.3	1065.0	14146	256.41	67.07
TR	42	22.0	167.6	1228.0	7038	197.96	118.12
TS	32	91.3	337.5	843.5	10798	192.74	57.12
VT	57	24.0	41.6	66.0	2369	10.42	25.06
WNS	156	10.0	232.7	1782.0	36304	208.5	120.53

669 eventos/552 ficheros	1.21 eventos/fichero
49.47 [horas]	322.62 [s]/fichero, 266.20 [s]/evento

Tabla 3.6.4.: *Estadísticas de duración de la base de datos col.04M.*

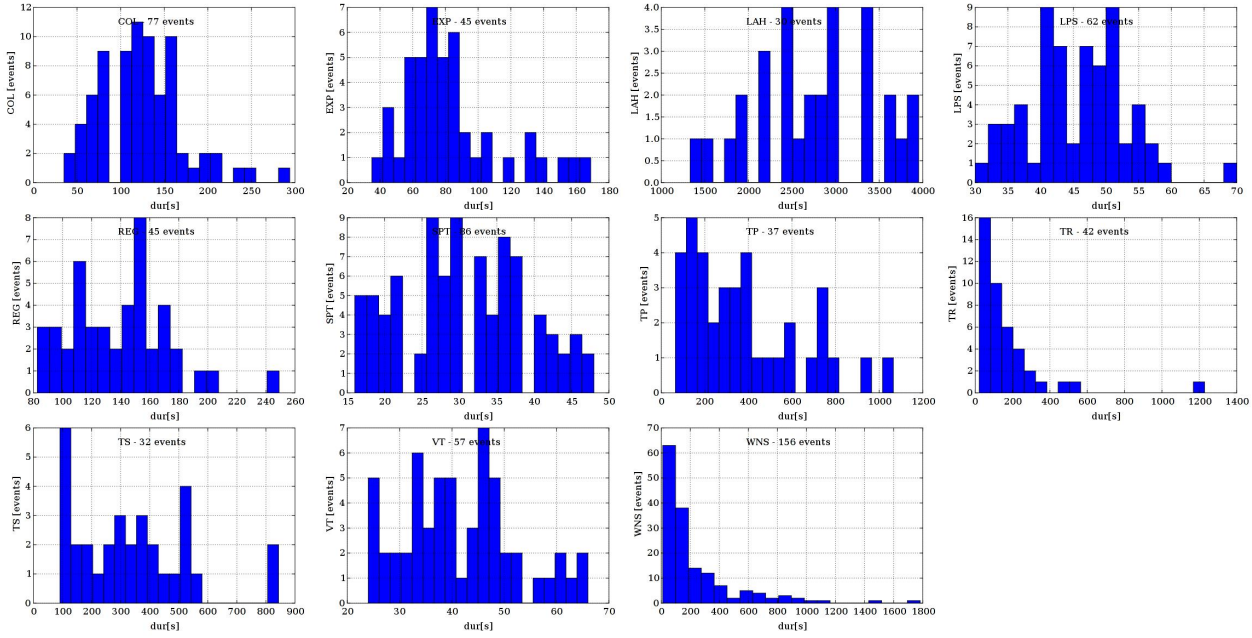


Figura 3.6.4.: Distribución temporal del corpus *col.04Mc* para las clases (de izq. a der.): 1ª fila: colapsos (*COL*), explosiones (*EXP*), lahares (*LAH*), eventos de largo periodo (*LPS*); 2ª fila: sismos regionales (*REG*), tremores pulsantes pequeños (*SPT*), pulsantes (*TP*) y resonantes (*TR*); 3ª fila: tremores espasmódicos (*TS*), sismos volcanotectónicos y ruidos (*WNS*).

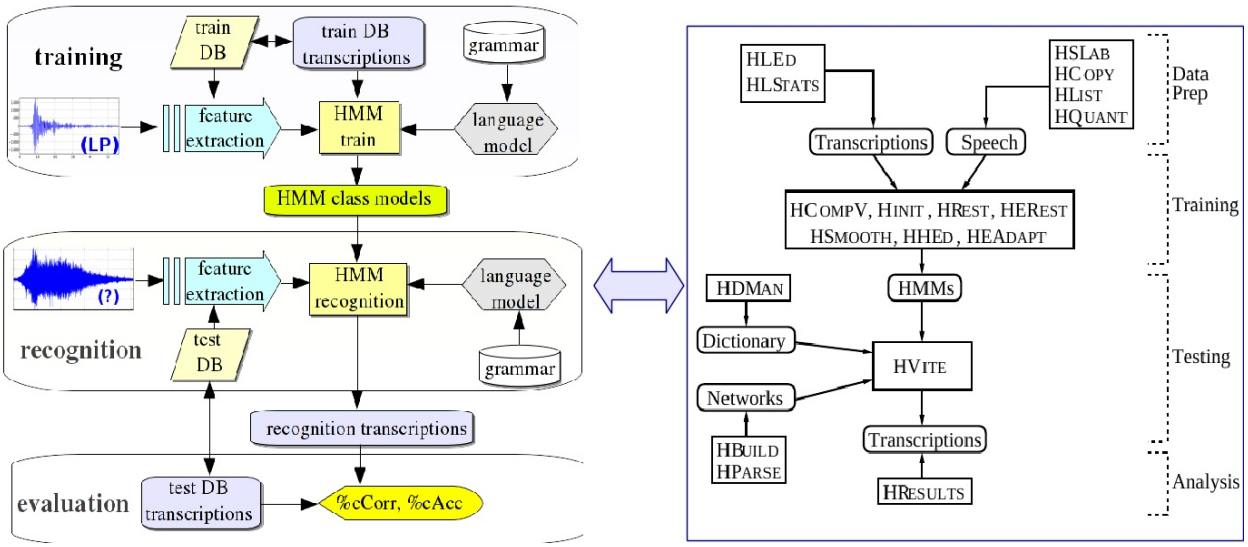


Figura 3.7.1.: Sistema de reconocimiento de patrones secuenciales VSR-Python basado en HMMs y herramientas HTK. Se crea un interfaz en Python para las utilidades de HTK que permiten usar los HMMs como modelos se secuencias. Se modela tanto el espacio de características de los eventos como la estructura existente entre ellos (modelado del *lenguaje*) conforme a unas reglas gramaticales.

gen cualquier sistema de reconocimiento supervisado tal y como describimos en la Subsección 2.1.1: parametrización de la señal en vectores, construcción de modelos (aprendizaje), clasificación o reconocimiento y evaluación. La única diferencia respecto a un sistema de básico de clasificación supervisada (Figura 2.1.4) radica en los bloques de modelado del lenguaje que se encargan de describir en base a las reglas gramaticales la estructura existente a nivel de eventos. En nuestro caso la gramática se limita a definir si se modelan señales en continuo o solo eventos aislados.

El sistema de reconocimiento $VSR - Python$ ha sido desarrollado como un interfaz implementado en Python de las librerías HTK (Young et al., 2006) al que se le han añadido otras funcionalidades . Esta herramienta se presentó en el *Taller sobre clasificación automática de señales sismo-volcánicas* de la *Conferencia internacional Cities On Volcanoes, 7th Ed.*, (`http://www.citiesonvolcanoes7.com/`) organizado por la Universidad de Savoie y el grupo IAG+TSTC de la Universidad de Granada. Se ha utilizado en la prevención de erupciones volcánicas (Boué et al., 2015) y actualmente está integrado en el sistema de monitorización en tiempo cuasi-real del volcán de Colima (Gonzalez-Amezcua et al., 2012).

3.7.1. Descripción de los datos

Una vez construidas las bases de datos maestras Decepción $dec.95Mc$ y Colima $col.04Mc$ en la Sección 3.6 anterior, describimos de una manera eficiente la for-

ma de onda proporcionada por los sismogramas según el esquema de parametrización $LFCC.D.A$,39 presentado en la Sección 3.2. Seguimos las directrices de la Subsección 3.2.1 realizando:

1. *Formateado de datos*: Los ficheros que contienen los datos de Decepción tenían muestras de 4 enteros a una frecuencia de muestro de 200 Hz. ($i4.200Fs$), por lo que se transformaron al formato $i2.50Fs$ requerido eliminando su componente continua, filtrando paso-baja con una frecuencia de corte a 25 Hz. y finalmente submuestreando a 50 Hz. El corpus de Colima fue suministrado ya en formato $i2.50Fs$, por lo que simplemente se fijo la media nula de amplitud de señal en sus ficheros.

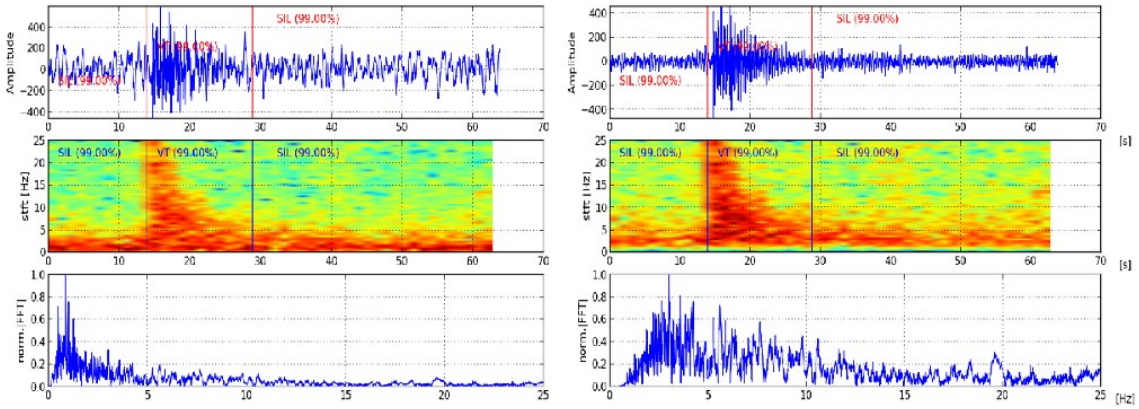


Figura 3.7.2.: Filtrado espectral de eventos en la banda [1,25] Hz. Se muestra el efecto del filtrado de un fichero etiquetado como $\{SIL|VT|SIL\}$. Observamos como en la señal filtrada (a la derecha) es más fácil distinguir el sismo volcano-tectónico en cada uno de las 3 representaciones: sismograma, espectrograma y el espectro de Fourier.

2. *Filtrado espectral de desfase nulo en la banda* $[f_L = 1, f_H = 25]\ Hz.$, con el objetivo facilitar la discriminación de los eventos de interés sobre el ruido de fondo (Figura 3.7.2).

3. *Parametrización* al esquema $LFCC.D.A.39$. El proceso de codificación en el sistema base se lleva a cabo directamente por las utilidades del HTK (para otros esquemas con vectores geofísicos o mixtos desarrollamos nuestro propio sistema de parametrización en Python). Puesto que HTK fue diseñado en sus orígenes para trabajar con señales de voz, espera trabajar con señales en muestreadas en un rango de frecuencia concreto (entorno a $F_S \approx 16kHz.$). Nuestras señales están de 2 a 3 órdenes de magnitud por debajo de ese valor, por lo que es necesario introducir un factor de escala de 1000 "engañando" al HTK que cree trabajar con señales muestreadas a $F_S = 50 * 1000 = 50\ kHz$. Sin este pequeño truco el HTK falla al parametrizar. Esta es la razón del porqué del cambio de notación en el esquema de $MFCC.E.D.A.39$ a $LFCC.D.A.39$: La

expansión logarítmica que el HTK hace en el eje de las frecuencias a la hora de diseñar los canales se realiza conforme a una aproximación de la escala MEL ideada para enfatizar las bajas frecuencias y suavizar las más altas a partir de $1\,kHz.$, donde es más sensible el oído humano (Stevens et al., 1937). Sin embargo, en el rango en el que los canales actúan sobre nuestros eventos ($[1000, 25000]\ Hz.$) comparado con la percepción humana ($[20, 20000]\ Hz.$) la escala MEL no actúa para lo que fue diseñada, pudiendo aproximarse casi por una transformación lineal (o logarítmica muy suavizada), por lo que renombramos el esquema como *LFCC* (*Log-Frequency Cepstral Coefficients*) en lugar de *MFCC* (*MEL-Frequency Cepstral Coefficients*).

3.7.2. Construcción de los modelos

A la hora de diseñar y entrenar nuestros modelos hay varias cuestiones que debemos considerar:

- Topología de los HMM
- Componentes de la mezcla de gausianas de los GMM para describir el espacio de características en cada estado.
- Parámetros de inicialización y entrenamiento de los HMM
- Mecanismos de poda del HTK

La topología es la definida en la Sección 3.5, 3 estados emisores para los modelos de Decepción y 5 para Colima, con una estructura simple de izquierda a derecha en el HMM y estados enlazados consigo mismo y con el predecesor, tal y como se muestra en la Figura 3.3.2. El script VSR-Python de construcción del sistema va incrementando el número de componentes gausianas en $[1, 2, 4, 8, 16]$ en cada estado emisor HMM siguiendo este proceso iterativo:

1. *Redefinición* de modelos HMM: se aumenta su complejidad incrementando las componentes de la probabilidad de emisión en cada estado
2. *Entrenamiento* del modelo, reestimando sus parámetros el algoritmo de Baum-Welch (Sección 3.3.2.1)
3. *Clasificación* mediante el algoritmo de decodificación de Viterbi (Sección 3.3.2.1)
4. *Evaluación* de los resultados calculando la eficiencia $\%cAcc$ promediada por clase (Subsección 3.4.2)

Antes de entrar en este bucle es necesario inicializar los HMMs. A continuación detallamos todo el proceso de construcción del sistema paso a paso (se pueden consultar más detalles del proceso en el manual de HTK por Young et al. (2006)):

0. **Inicialización de modelos**. Se define la topología de los HMM creando prototipos. La orden HCompV computa estadísticos de la base de datos (medias

y varianzas) que luego serán usadas por otras utilidades. Puesto que contamos con las transcripciones con las segmentaciones temporales de cada evento en los ficheros podemos inicializar los modelos independientemente para cada clase mediante las HInit+HRest. HInit aplica el algoritmo de agrupamiento k-medias para asociar los vectores de características (u observaciones) a cada estado haciendo una estimación inicial de los parámetros que ajusta iterativamente con el algoritmo de Viterbi. Posteriormente HRest actualiza los modelos empleando el algoritmo iterativo de Baum-Welch.

Nótese que una buena inicialización es fundamental pues el método de Baum-Welch es muy sensible al punto inicial de la estimación MLE de máxima verosimilitud cuando se maximizan funciones objetivo multimodales, típicas cuando se quiere optimizar los parámetros de definición de los HMM. (Sección 3.3.2.1).

1. **Incremento de las componentes de salida**. En todos los HMMs y en todos sus estados emisores se incrementa progresivamente el número de componentes gausianas en las probabilidades GMM (Subsección 3.3.1) para modelar las observaciones. HHEd examina las componentes con mayor varianza y las desdobla hasta alcanzar el número de gausianas requerido. Las componentes con una varianza menor que un umbral calculado por HCompV se eliminan.

2. **Reentrenamiento de modelos**. HERest realiza un entrenamiento embebido al construir un macro-HMM por cada fichero es capaz de actualizar todos modelos aplicando Baum-Welch a todos los datos de entrenamiento a la vez. No necesita etiquetas de tiempo para ejecutarse.En teoría, Baum-Welch converge rápidamente en menos de 100 iteraciones consiguiendo que el modelo estimado $H\hat{M}M_c$ se adapte mejor (o, al menos, no peor) a los eventos asignados a la clase c tras terminar de ejecutarse (Elemento 3.3.2.1). Esta mejora se puede medir gracias al incremento medio de la probabilidad de salida por frame que cada modelo HMM_c alcanza al evaluar sus datos $\{\mathbf{x}_c\}$ de entrenamiento. Promediando entre todos los modelos definimos $\Delta\bar{p} \equiv mean_{c,\mathbf{x}}\left\{\Delta p(\{\mathbf{x}_c\}|H\hat{M}M_c)\right\}$ como un indicador de la mejora en el entrenamiento. Nuestro sistema fuerza la ejecución iterativa de HERest un mínimo de 2 veces consecutivas y un máximo que viene controlado por el umbral $\Delta\bar{p} = 0{,}15$.

3. **Clasificación de datos**. HVite se aplica para encontrar la secuencia óptima de estados en el macro-HMM o red de búsqueda M_{RB} (Subsubsección 3.3.2.2), lo que nos da las transcripciones de salida con la conveniente clasificación de los datos de evaluación. Tanto HVite como HERest utilizan los *mecanismos de poda* cuando están analizando sobre la red M_{RB}. Básicamente consiste en ignorar en el análisis de los caminos (secuencia de estados, modelos o *paths*) que sean muy poco probables: que en un instante t tengan una probabilidad acumulada menor que un umbral $PRUNNING$ respecto del máximo de probabilidad en ese mismo instante. Estos mecanismos aceleran el proceso de decodificación y entrenamiento al mismo tiempo que actúan contribuyen a no construir modelos demasiado complejos regularizando los HMM. Nuestro

sistema fija el $PRUNNING = 350$, valor que por defecto usa HTK.

4. **Evaluación del sistema**. Una vez obtenidas las transcripciones de reconocimiento, HResults se encarga de compararlas con las transcripciones originalmente supervisadas por los técnicos y de generar la matriz de transición. Dicha información es complementada y presentada en la forma que vemos en la Figura 3.4.1. El sistema VSR-Python se encarga además de generar las tasas de robustez de clase para cada evento como se describe en la **??**.

3.7.3. Evaluación del sistema: resultados base

Para comprobar los efectos que sobre el análisis del sistema pueden tener distintos criterios de evaluación detallados en la Sección 3.4, presentaremos los resultados experimentales desde 3 puntos de vista:

- Evaluación *base (base)*: la que utilizaremos por defecto a lo largo de toda la tesis.
- Evaluación *geofísica (geo)*: que tiene como objetivo no tener en cuenta aquellas inserciones y borrados que no tienen importancia desde el punto de vista geofísico (Subsección 3.4.3) en clases *no secuenciales* con una evolución temporal no muy marcada tales como ruidos, tremores y derrumbes (concretamente NS y TR en Decepción y COL, LAH, TP, TR, TS, y WNS en Colima).
- Evaluación *compensada (penal)*: igualando las inserciones y borrados mediante la penalización de la probabilidad asociada a cada camino de la red de búsqueda cada vez que se pasa de un modelo a otro (Subsubsección 3.3.2.2).

SSA-BASE.39 : evaluación del sistema base con el vector MFCC.D.A.39

	dec.95Mc					*col.04Mc*				
Eval	**%cCorr**	**%cAcc**	(-)	(?)	(+)	**%cCorr**	**%cAcc**	(-)	(?)	(+)
base	90.31	30.60	2	7	60	95.65	56.71	1	8	62
geo	90.37	42.22	3	9	43	95.65	77.34	1	7	32
penal	83.98	63.13	8	11	17	93.28	86.05	5	9	10

Tabla 3.7.1.: *$SSA - BASE.39$: evaluación del sistema BASE con arquitectura serie (SSA)* con el vector $MFCC.D.A.39$ y modelos HMM de 2 gausianas para los esquemas *base*, geofísico (*geo*) y compensado (*penal*). Se promedia entre clases los porcentajes de eventos *%cCorr* correctamente reconocidos y el *%cAcc* de eficiencia, indicando los errores de borrados (−), sustituciones (?) e inserciones (+).

La Tabla 3.7.1 muestra los resultados de reconocimiento del sistema base en Decepción y Colima para los criterios de evaluación base (*base*), geofísico (*geo*) y compensado (*penal*). La calidad del reconocimiento se mide mediante los promedios por clase de *%cCorr* de aciertos y *%cAcc* eficiencia, detallando los errores de borrados

(−), sustituciones (?) e inserciones (+). Nótese que los HMMs se han construido con solo 2 componentes gausianas para evitar el efecto de sobre-entrenamiento en el modelo VT de Decepción (Sección A.4). Observamos:

- *Los resultados de un mismo reconocimiento se pueden interpretar de manera muy diferente según que criterio de evaluación consideremos.* La variación en la eficiencia *%cAcc* alcanza hasta los 30 puntos en ambas bases de datos.
- *La importancia de las inserciones es definitiva en la eficiencia %cAcc.* Mientras los borrados y sustituciones aproximadamente se mantienen estables en los esquemas *base* y *geo*, al compensar las inserciones con los borrados en el esquema *penal* la eficiencia mejora de forma asombrosa, sobre todo en Colima, cuyos modelos de clases no secuenciales tienden a generar demasiadas inserciones. El parámetro *%cCorr* no se ve afectado por las inserciones, puesto que no las contabiliza.
- *Los errores debido a sustituciones se mantienen estables sea cual sea el criterio de evaluación.* Técnicamente, los errores por sustituciones son siempre errores del reconocedor, no atribuibles o re-interpretables según que esquema de evaluación, por lo que son los que más deben preocuparnos. Nótese que al compensar inserciones en la evaluación *penal*, tanto los borrados como las sustituciones se incrementan.
- *No siempre un mal resultado en eficiencia %cAcc no siempre tiene que estar asociado a unos resultados malos de reconocimiento*, tal y como vimos de forma teórica en la Figura 2.2.2 y tal y como constatamos de forma práctica en la Tabla 3.7.1. De igual manera, no todos los errores de inserciones y borrados son igual de importantes desde un punto de vista geofísico (no es lo mismo ignorar un ruido que una explosión) y no todas las inserciones son realmente errores; algunas de ellas tras un re-etiquetado y observación detenidos de la base de datos pueden ser eventos no detectados por el experto geofísico Benítez et al. (2007); Ibáñez et al. (2009) o inserciones de ruidos.

Bajo estas consideraciones, los resultados del reconocimiento en continuo no son un mal punto de partida. Aunque hay un amplio margen para mejorar, tarea de la que nos ocuparemos en los próximos capítulos.

4. Reducción de dimensionalidad

En capítulos anteriores hemos comprobado la conveniencia de describir las señales sísmicas de forma compacta mediante el proceso de parametrización obtenemos una representación en vectores de características más eficaz para la tarea de reconocimiento automático.

En este capítulo presentaremos varios esquemas de parametrización y usaremos diversos tipos de técnicas clásicas, actuales y otras que hemos desarrollado específicamente para reducir la *dimensionalidad* (el número de componentes) del vector de características. Mediante un estudio experimental seleccionaremos la parametrización más conveniente y sobre ella aplicaremos distintas técnicas de reducción para escoger la más eficaz dentro de cada categoría. El binomio (parametrización, técnica de reducción) escogido sera usado en los próximos capítulos para describir los datos de nuestro sistema VSR de reconocimiento.

El capítulo se estructura de la siguiente manera: Empezaremos (Sección 4.1) con una pequeña introducción a la reducción de dimensionalidad, exponiendo sus conceptos fundamentales y las técnicas principales que existen para reducir de tamaño un vector de características. En la Sección 4.2 propondremos distintos tipos de parametrizaciones, desde esquemas clásicos basados en características estadísticas y transformaciones sobre los sismogramas a esquemas que explotan la naturaleza geofísica de las señales. Tras configurar y comparar los esquemas propuestos nos quedaremos con un vector de componentes mixtos sobre el que aplicaremos distintos métodos para seleccionar sus mejores características en la Sección 4.3. En la Sección 4.4 transformaremos el espacio de características usando distintas técnicas para obtener otros espacios donde la información esté más ordenada y se requiera una menor dimensión para describir los datos, permitiendo mantener o incluso mejorar las tasas de reconocimiento. Finalmente, en la Sección 4.5 compararemos los métodos de reducción propuestos esquematizando sus ventajas e inconvenientes y concluiremos el capítulo con algunas observaciones obtenidas tras este estudio.

4.1. Introducción a la Reducción de Dimensionalidad

En esta sección plantearemos el problema de la maldición de la dimensionalidad como los efectos que sobre el sistema VSR tienen describir los datos a modelar mediante un vector con demasiadas componentes. Posteriormente plantearemos de manera formal el proceso de reducción de dimensionalidad y clasificaremos los algoritmos propuestos para resolver este problema.

4.1.1. Motivación: *la maldición de la dimensionalidad*

Una vez que los datos han sido descritos por un conjunto de características mediante el proceso de parametrización, a veces es necesario reducir la dimensión del espacio generado por dichas características. Esta necesidad viene originada por la denominada *maldición de la dimensionalidad*: a medida que el espacio aumenta de dimensión es necesario una gran cantidad de datos para que no queden distribuidos de una forma dispersa en él. Esta dispersión de los datos propicia la pérdida de interrelación estadística entre ellos dificultando la estructuración en grupos de datos parecidos entre sí y, por tanto, su modelización. La reducción de dimensionalidad pretende extraer la información subyacente más relevante contenida en los datos sin pérdida significativa de esta. Su necesidad viene motivada por varios aspectos:

• *Ahorro del coste computacional en el procesamiento de los datos.* El coste computacional de la tarea de aprendizaje se incrementa con la dimensión de los datos a los que se aplica, así, la mayoría de las técnicas de aprendizaje no son efectivas con datos de alta dimensión. Muchas operaciones presentes en los métodos de reconocimiento de patrones tienen un coste computacional exponencial con la dimensión del espacio de características, por ejemplo la estimación de la matriz de covarianza, de modo que si la dimensión es suficientemente alta el problema puede llegar a ser intratable o demasiado costoso en lo relativo a tiempo de ejecución y recursos utilizados.

• *Necesidad de disminuir el tamaño de las bases de datos.* La cantidad de muestras necesarias para que los resultados tenga significado estadístico a menudo también crece exponencialmente con la dimensionalidad: en general, si tenemos N muestras que quedan correctamente descritas por 1 variable, necesitaremos N^D muestras para describirlos con la misma precisión en un espacio con D dimensiones. La capacidad de generalización de lo aprendido, es decir, la eficacia que obtendremos al aplicar nuestros modelos de una forma no supervisada a un corpus de datos desconocido, mejora cuanto menor es el cociente entre el número de muestras y la dimensión del espacio de características. Un número de muestras insuficiente conduce al fenómeno de sobreajuste (Bishop, 1995): una adaptación excesiva del modelo de aprendizaje a los datos de entrenamiento, lo que conlleva una degradación de la eficiencia en el reconocimiento cuando se realiza sobre nuevos datos.

• *Descripción más compacta de los datos.* El hecho de contar con una representación de los datos descritos con pocas características facilita la interpretación de

dichas características así como su visualización, permitiendo simplificar los modelos. Además, la eliminación de la redundancia así como del posible ruido presente en los datos suele tener efectos positivos en los resultados de reconocimiento.

- *Eficiencia de almacenamiento y recuperación de los datos*, que aumenta al estar descritos de una manera más comprimida.

La reducción de dimensionalidad se lleva a cabo mediante dos estrategias básicas: los métodos de *selección* y los métodos de *extracción* de características. Mientras los primeros intentan encontrar el mejor subconjunto a partir del conjunto inicial formado por las componentes del vector de características, los segundos transforman este conjunto inicial en otro vector con menor dimensión o número de componentes.

4.1.2. Planteamiento del problema

Dado un vector de parametrización con K componentes (o características), la *Reducción de Dimensionalidad* (RD) consiste en obtener un nuevo conjunto de características de menor tamaño pero que mantenga o mejore su eficacia desde el punto de vista del modelado de datos. Definido matemáticamente:

Algoritmo 4.1 Objetivo de la reducción de dimensionalidad.

Sean:

$C = \{C_1, ..., C_K\}$ un conjunto con K características

$T = \{T_1, ..., T_L\}$ un conjunto de funciones definidas en $E_C \rightarrow E'_{C'}$ que transforman E_C, el espacio de características generado por C, en un nuevo espacio $E'_{C'}$, generado por C', más ordenado, según diversos criterios, siendo J la dimensión de $E'_{C'}$ tal que: $dim\{E'_{C'}\} = J \leq dimensión\{E_C\} = K$.

S un subconjunto de C' con $I \leq J$ elementos.

El objetivo es escoger la tupla $\{T_i, S\}$ que maximice una función de medida $M : C' \rightarrow \mathbb{R}$ que represente el incremento de la eficacia del conjunto S respecto al conjunto original C.

A menudo, a la transformación del espacio de características se le conoce con el nombre de *Reducción* (*extracción* o *transformación*) *de Características* (RC) que, en sí misma, puede conllevar una proyección del espacio C' sobre un subespacio deseado implicando una reducción en la dimensionalidad. Ejemplos clásicos de estas transformaciones que son comunes en el proceso de parametrización los encontramos en la transformada del coseno (DCT) o el análisis de componentes principales (PCA) (Sección 4.4).

El proceso de seleccionar el subconjunto S con las características más eficaces de C' (o de C si definimos T_i como la función identidad) se denomina *Selección de Características* (SC) e implica una reducción de dimensionalidad al proyectar el espacio

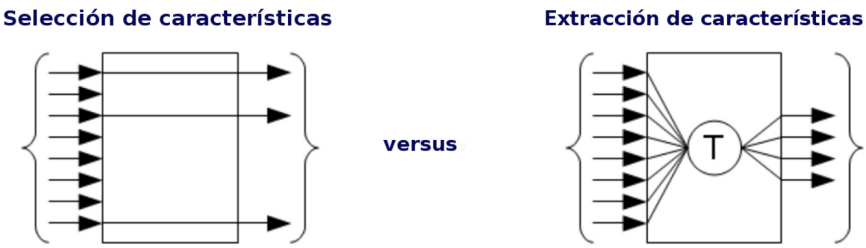


Figura 4.1.1.: Estrategias básicas de reducción de dimensionalidad al describir los datos: selección de las mejores características o reducción de características a partir de una transformación T. Figura original de Hoogenboezem (2010).

$E'_{C'}$ sobre los ejes definidos por S, pero no un cambio de espacio (Figura 4.1.1). En general la *eficacia* de un conjunto C con K características se puede medir como:

$$eficacia(C) = \frac{bondad(C)}{K} \tag{4.1.1}$$

Esta fórmula adopta diferentes formas según la expresión analítica que usemos para la *bondad* y suele ser utilizada de una forma no lineal, requiriendo un umbral mínimo de la bondad para finalmente escoger un conjunto C concreto. Nótese que existen $2^K - 1$ posibles subconjuntos S que pueden formarse a partir de un vector con K componentes. A menudo se define la bondad como alguna medida basada en la tasa de reconocimiento de un sistema que use C como vector de características, aunque existen varios tipos de medidas (de teoría de la información, de distancia, de dependencia estadística, de consistencia...) igualmente populares (Yu et al., 2007). Generalmente, las distintas definiciones de bondad pretenden cuantificar estos dos conceptos:

1. La *redundancia* existente entre los elementos de C.
2. La utilidad o *relevancia* que los elementos de C tienen en la descripción de los datos.

Una buena selección de C implica características relevantes conforme a criterios basados en máxima relevancia, $max\{rel(C)\}$, o no redundantes entre sí acorde a criterios de mínima redundancia, $min\{red(C)\}$, como los definidos en la Ecuación 4.1.3 y la Ecuación 4.1.2. También existen algoritmos mixtos, que integran medidas de redundancia y relevancia conjuntamente (Peng et al., 2005). Si bien existen varias definiciones en la literatura, lo más común es usar funciones basadas en correlación o información mutua para la redundancia. En lo referido a la relevancia, la capacidad para describir los datos se suele definir en base a la capacidad de separar los datos en clases, tomando medidas inspiradas en información mutua entre características y clases, probabilidad condicional de modelos estadísticos o, simplemente, la eficacia de reconocimiento dados unos modelos de clasificación.

Definiciones de redundancia y relevancia de características. Ejemplos de definiciones de redundancia y relevancia para un conjunto C de características en el campo de la teoría de la información las encontramos en Peng et al. (2005):

$$red(C) = \frac{1}{K^2} \sum_{c_i, c_j \in C} MI(c_i, c_j) \tag{4.1.2}$$

$$rel(C) = \frac{1}{KH} \sum_{c_i \in C, \lambda_j \in \lambda} MI(c_i, \lambda_j) \tag{4.1.3}$$

$$mRMR(C) = rel(C) - red(C) \tag{4.1.4}$$

Siendo $MI(x, y)$ la información mutua entre las variables x e y, o, al menos, una estimación de de ella, H el número de clases $\{\lambda_1, .., \lambda_H\}$ en las que están etiquetados los datos y $mRMR$ el criterio mixto de mínima redundancia y máxima relevancia.

Alternativamente, podríamos utilizar definiciones basadas en medidas de probabilidad que Zhao et al. (2010) usa para la característica C_i y el subconjunto S_i formado por el conjunto total C excluyendo a S_i, tal que $S_i = C \backslash C_i$:

- C_i es *estadísticamente relevante* para C si*:*

$$\exists S_i' \subseteq S_i \ / \ P(\lambda | C_i, S_i') \neq P(\lambda | S_i') \tag{4.1.5}$$

  C_i es *estadísticamente redundante* en C si*:*

$$\exists S_i' \subseteq S_i \ / \ P(\lambda | C_i, S_i') \neq P(\lambda | S_i') \ ; \ P(\lambda | C_i, S_i) = P(\lambda | S_i) \tag{4.1.6}$$

La Ecuación 4.1.5 indica que según la estadística, la característica C_i es *débilmente relevante* (John et al., 1994) para el conjunto C si al excluirla de C decrece la capacidad de predicción de eventos de la clase λ (dada por la probabilidad condicional $P(\lambda, C)$) cuando se describen mediante C. Esto puede ser debido a que C_i esté correlacionada con λ, o debido a que C_i forme parte de un subconjunto S_i' que esté correlacionado con C o a ambas cosas a la vez. Asimismo, C_i es redundante si su inclusión en C no aumenta la capacidad de predicción. Como veremos en Subsección 4.1.4, la eliminación directa de una característica estadísticamente redundante no siempre conlleva un incremento en la eficiencia de predicción (Guyon and Elisseeff, 2003), por lo que algunos autores prefieren reducir la dimensionalidad usando medidas de correlación entre características o agrupando características con patrones similares (Hall, 1999; Haindl et al., 2006).

4.1.3. Clasificación de algoritmos de reducción de dimensionalidad

Existen diferentes maneras de clasificar los algoritmos de reducción de dimensionalidad según el criterio adoptado.

- Algoritmos *dependientes* o *no dependientes de los datos*. Las técnicas dependientes usan transformaciones que necesitan extraer algún tipo de información estadística de los datos para ser aplicadas. Un ejemplo lo encontramos en las matrices de proyección de la PCA o en cualquier algoritmo que necesite construir modelos para seleccionar características. Los algoritmos independientes como la DCT no necesitan ningún análisis previo de las señales para definir sus transformaciones.
- Metodologías *supervisadas* o *no supervisadas*. Según se necesiten o no conocer las etiquetas que separan en clases a los datos para implementar el algoritmo. Esta clasificación previa se realiza manualmente, requiriendo para ello tener una información intrínseca de la naturaleza de los eventos proporcionada, en nuestro caso, por características geofísicas como la evolución temporal de la señal, su energía, la localización espacial o la impulsividad de la forma de onda. (Ohrnberger, 2001; Álvarez et al., 2009).

Como veremos en las próximas secciones, aparte de estas clasificaciones generales, es común usar clasificaciones específicas para los algoritmos de selección. Las más aceptadas son (Zhao et al., 2010):

1. Clasificación según el *conjunto de análisis*: de forma independiente para cada característica o conjuntamente en grupos.
2. Clasificación según los *métodos de evaluación de la bondad* de un conjunto de características.

4.1.4. Selección de características según el conjunto de análisis

Hay dos enfoques principales para realizar la selección de características según las analicemos conjuntamente o por separado (Yu et al., 2007):

- Estudiando las características de forma *I*ndependiente entre sí (*SC_I* o *Variable Ranking*)
- Evaluando *S*ubgrupos de características (*SC_S* o *Subset Selection*)

Los métodos *SC_I* tienen como principal ventaja un bajo coste computacional con orden de complejidad $\mathcal{O}(K)$ siendo K el número de elementos del conjunto de características. Sin embargo, no pueden usar medidas que engloben a más de una variable al mismo tiempo (como correlaciones), por lo que son ineficaces para eliminar la redundancia entre características.

Separabilidad en el espacio de descripción de datos: redundancia, relevancia y correlación de características. Es conveniente tener en mente algunas consideraciones importantes en torno a la reducción por selección en general y que, en particular, esbozan las desventajas de los métodos *SC_I* frente a los *SC_S*. Las figuras originales se encuentran en Guyon and Elisseeff (2003) y muestran un espacio generado por un conjunto $C = \{C_1, C_2\}$ de características que modelan a las clases $\{\lambda_1, \lambda_2\}$ donde las probabilidades condicionales de las clases descritas por cada característica ($P(\lambda_{i,}|C_j)$, con $i, j = 1, 2$) se proyectan, en forma de histogramas, en cada eje del espacio. Usando los criterios estadísticos de redundancia y relevancia de la Sección 4.1.2 y midiendo la correlación entre variables mediante el coeficiente de *Pearson* (Subsubsección 4.3.1.2) se puede observar como:

- *Características no relevantes por sí mismas pueden serlo cuando se usan conjuntamente.*
- *Características (aparentemente) redundantes pueden ayudar a una mejor separación entre clases.*
- *Una alta correlación entre 2 componentes no implica que no tengan un grado de complementariedad en el espacio de características.*

Una consecuencia directa de estas ideas es que se puedan obtener resultados bastante diferentes según la metodología empleada. Los algoritmos *SC_S* intentan solucionar los inconvenientes de los métodos *SC_I*, pero conllevan, al menos, un coste computacional $\mathcal{O}(K^2)$. A cambio, suelen obtener mejores resultados, pues son capaces de computar las relaciones entre características, en vez de evaluarlas por independiente. Algunos autores optan por estrategias mixtas; Köhler et al. (2009) realizan una selección no supervisada de características mixtas evaluando primero las caracterís-

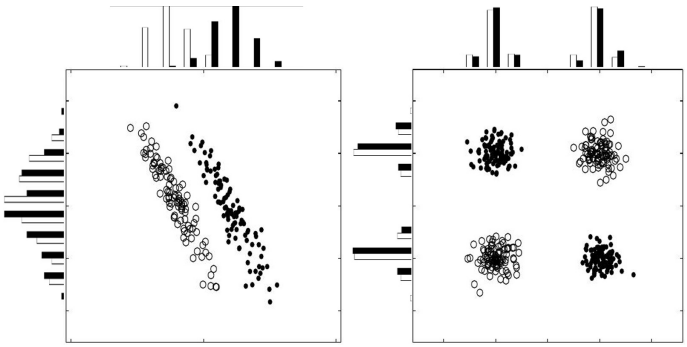


Figura 4.1.2.: Espacio de descripción: utilidad de características dependientes. *Fig. de la izquierda*: A pesar de tener $P(\lambda_{1,}|C_1) \backsimeq P(\lambda_{2,}|C_1)$, la característica C_1 aumenta la capacidad de predicción cuando se usa conjuntamente con C_2. *Fig. de la derecha*: Problemas *XOR* (OR eXclusive - O Exclusivo) o de *tablero de ajedrez*; C_1 y C_2 apenas sirven para clasificar por sí mismas pues $P(\lambda_{i=1,2,}|C_1) \backsimeq P(\lambda_{i=1,2,}|C_2)$, pero conjuntamente adquieren la capacidad de separabilidad (siempre que no se use un clasificador lineal). Figuras de Guyon and Elisseeff (2003).

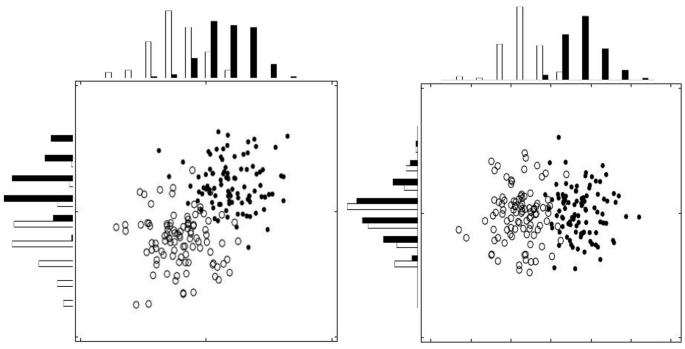


Figura 4.1.3.: Espacio de descripción: características aparentemente redundantes. Dos características que parecen redundantes (*izquierda*) dejan de serlo tras la rotación de 45 grados del espacio de características (*derecha*). Detalles en Guyon and Elisseeff (2003).

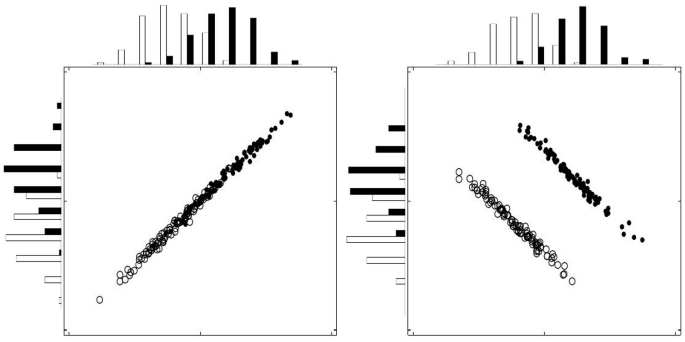


Figura 4.1.4.: Espacio de descripción: utilidad de características correlacionadas. En ambas figuras la (anti)correlación entre las características es muy alta. Las características son complementarias y redundantes en la fig. de la izquierda, y no redundantes en la de la derecha (Guyon and Elisseeff, 2003).

ticas de forma independiente, agrupándolas en subconjuntos después y hallando la correlación de los subconjuntos entre sí.

4.1.5. Selección de Características según el método de evaluación

Atendiendo a la forma en la que se evalúan las características los podemos clasificar los algoritmos en (Saeys et al., 2007):

Filtros (*SC_F*) que evalúan a los subconjuntos sin necesidad de modelos predictivos.

Guiados (*SC_M*) o *Wrappers*, que construyen modelos para cada subconjunto de características a evaluar.

Incrustados (SC_I) que seleccionan las características y estiman los modelos conjuntamente en en la etapa de construcción del sistema clasificador.

La selección por filtros es más rápida y es independiente del clasificador, lo que los hace muy populares cuando hay involucrada una gran cantidad de datos de dimensiones elevadas (Biesiada and Duch, 2007; Van Hulse et al., 2009; Arauzo-Azofra et al., 2011). Nótese que para se pueden formar $K!/J!(K-J)!$ subconjuntos de J componentes a partir de un conjunto inicial con K características. El uso de modelos predictivos persigue unos mejores resultados de selección en detrimento del coste computacional. Los algoritmos guiados suelen usar unos modelos sencillos durante la selección para evitar su influencia en los resultados, pero, el entrenamiento y evaluación de cada modelo para cada subconjunto analizado suele requerir demasiado tiempo (Hall, 1999). Diversas técnicas de búsquedas selectivas han sido desarrollados recientemente para mejorar su eficacia (Yu et al., 2007). Los algoritmos incrustados pretenden solucionar las desventajas de los guiados a cambio de unos resultados totalmente dependientes del clasificador (Cárdenas-Peña et al., 2013). Esta integración y, a menudo, una función de evaluación de subgrupos ligada a la tasa de reconocimiento permiten alcanzar unos altos resultados (Orlic and Loncaric, 2010). En los últimos años está tomando fuerza el uso de estrategias mixtas de evaluación (Bermejo et al., 2011; Curilem et al., 2014b).

4.1.6. Metodología experimental

De manera general seguiremos adoptando la metodología especificada en la Sección 3.5. Con el objetivo de comprobar los efectos que la complejidad del clasificador pueda tener en algunos esquemas de selección de características, en las pruebas experimentales usaremos en el modelado junto a los HMMs los más sencillos GMMs. Para focalizar el estudio exclusivamente en las cuestiones de reducción de dimensionalidad, utilizaremos las versiones de eventos aislados *dec.95Ms* y *col.04Ms* correspondientes a las bases de datos maestras de *dec.95M* y *col.04M* definidas en la Sección 3.6.

Configuración estándar de los test

El solapamiento entre segmentos es por defecto el 50 % excepto en los test de *dec.95Ms* con HMMs en los que se estudia la duración del segmento, en los que usaremos un solapamiento del 90 % para evitar modelos sobre-entrenados (ver la Sección A.4 para una explicación más detallada). Por el mismo motivo, en *dec.95Ms* el tamaño máximo del segmento es de 5 segundos en vez de los 10 de *col.04Ms*.

4.2. Características propuestas

En este apartado propondremos y evaluaremos distintos esquemas de parametrización para construir un vector inicial de características y unos resultados base sobre

los que estudiar las técnicas de reducción de dimensionalidad. Cuatro tipos de características serán analizados:

1. Características de naturaleza geofísica
2. Transformaciones del espacio original de los datos
3. Características de naturaleza estadística
4. Vectores de características mixtas

Las parametrizaciones que usaremos están basadas en esquemas previamente usados en trabajos anteriores. Empezaremos por evaluar las tres primeras categorías por independiente, obteniendo las configuraciones óptimas para cada una de ellas. Posteriormente, construiremos un vector híbrido uniendo las mejores características obtenidas de cada categoría y, finalmente, eliminaremos de ese vector aquellos componentes que por independiente tienen menor capacidad discriminatoria para quedarnos con un vector base de 30 elementos.

Con el objetivo de escoger esquemas de parametrización que sirvan para obtener modelos *exportables*, esto es, que puedan reutilizarse en volcanes distintos de donde se construyeron, cabe distinguir entre:

1. Características *propias* o *locales* de una base de datos: aquellas cuyos valores estadísticos (media, varianza, ...) cambian demasiado como para ser usados en otros volcanes o, incluso, como para ser usados en la misma zona con diferentes equipos de adquisición o en diferentes etapas de la actividad sísmica.
2. Características *exportables*: que a pesar de conllevar cierta variabilidad, pueden ser útiles para construir modelos generales de clase.

La obtención de parametrizaciones y modelos exportables es un tema de investigación aún en desarrollo (Ibáñez et al., 2009; Cortés et al., 2009b; Duin et al., 2010) y un reto difícil de resolver debido a la complejidad de las señales sísmicas (Subsección 2.2.1). En este trabajo, nos centraremos en las características exportables.

Información contextual o *dinámica* de las características. Para incluir información contextual en cada segmento añadiremos a una parametrización básica de K componentes estáticos, *<par>_K*, K elementos más correspondientes a las derivadas temporales de orden 1 (deltas, $\_D$) de los coeficientes estáticos, otros K de orden 2 (aceleraciones, $\_A$) y otros tantos de orden 3 ($\_T$) generando los esquemas *<par>_D_<K*2>*, *<par>_D_A_<K*3>* y *<par>_D_A_T_<K*4>* respectivamente.

4.2.1. Características de naturaleza geofísica

Una descripción de los datos que proporcione información sobre la naturaleza de los eventos (energía, localización, evolución temporal,...), acerca de sus fuentes de

generación o del medio en el que se propagan es muy útil para la clasificación manual supervisada de datos hecha por técnicos expertos, paso previo de la mayoría de los estudios científicos. Por ello, son numerosos los trabajos que usan una descripción geofísica de los datos para construir sistemas de clasificación automática no supervisada. Ejemplos de descripciones geofísicas de los datos son:

- Autocorrelación espacial (Aki, 1957).
- Análisis de frecuencia y número de onda (Kvaerna and Ringdal, 1992).
- Sonograma discretizado (Ohrnberger, 2001).
- Medidas espectrales del elipsoide de polarización (Pinnegar, 2006).
- Medidas de impulsividad y periodicidad en el domino temporal y espectral (Álvarez et al., 2009).
- Medidas de localización: latitud, azimut, epicentro, profundidad, intensidad, etc. (Parpoula et al., 2013).

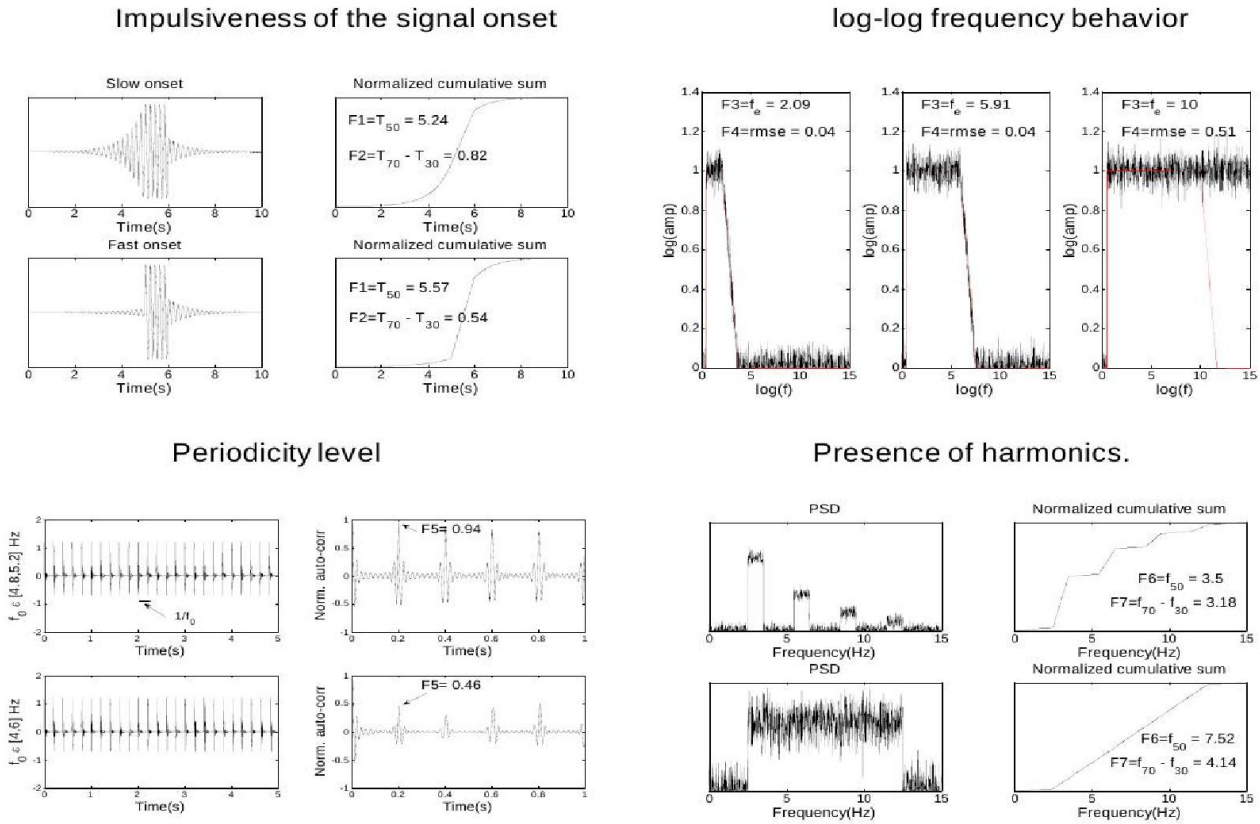


Figura 4.2.1.: Extracción de características geofísicas para describir propiedades de (de izq. a der. y de arriba a abajo): impulsividad, comportamiento en frecuencia, periodicidad y existencia de armónicos (Álvarez et al., 2009) .

Aunque es indudable la utilidad de propiedades geofísicas en el análisis manual de datos, en general, la tasa de reconocimiento que se alcanza con ellas no es espe-

cialmente mayor que la que se puede alcanzar usando otros esquemas, por lo que es bastante común usarlas junto a otros tipos de características en vectores mixtos, tal y como presenta Orozco-Alzate et al. (2012) en su compendio sobre trabajos dedicados a la identificación automática de eventos sismo-volcánicos.

Algunas desventajas de las parametrizaciones geofísicas son:

- Suelen tener características no exportables, como las basadas en localizaciones (2D y 3D) de eventos y las basadas en la polarización.
- A veces requieren para su extracción las componentes horizontales de los sismogramas y ser adquiridas a la vez en más de una estación.

Las características que estudiaremos han sido escogidas con el propósito de ser lo más exportables posible y están basadas en el esquema propuesto por Álvarez et al. (2009). El esquema básico *geo_13* de componentes geofísicos que parametriza cada segmento del sismograma se define en la Tabla 4.2.1. El vector tiene 6 características extraídas en el dominio temporal y 7 obtenidas tras pasar al dominio de la frecuencia. Las características *t.NN* y *t.Max* pretenden modelar la envolvente del sismograma y, paralelamente, las *f.NN* y *LPFerr* la curva básica del espectro. *t.Slo* está diseñada para distinguir entre eventos de impulsividad baja (LPs, REGs, NSs, TRs, COLs, LAHs,...) y alta (EXPs, VTs, HYs,...). *A.nac* mide la periodicidad de algunos eventos como los pulsos de los TP, los lóbulos de la secuencias de LPs, e, incluso, la autocorrelación de señales de periodo largo (LPs y TRs). La frecuencia de mayor energía espectral (*f.Max*) permite discernir entre los eventos de largo periodo (LPs, TRs) y los de contenido espectral más energético en frecuencias altas (EXPs, VTs, HYs, LAHs, COLs,...).

identificador	*característica*	*propiedad a medir*
t.20	posición del 20 % de amplitud	envolvente: llegada
t.50	posición del 50 % de amplitud	envolvente de amplitud
t.80	posición del 80 % de amplitud	envolvente: coda
t.Slo	pendiente de la amplitud: t.70-t.30	impulsividad del sismograma
t.Max	posición temporal del máximo de amplitud	envolvente: máxima energía
A.nac	2º pico de la autocorrelación normalizada	periodicidad
f.20	frecuencia del 20 % de energía normalizada	modelado espectral: banda inferior
f.50	frecuencia del 50 % de energía normalizada	modelado espectral: banda media
f.80	frecuencia del 80 % de energía normalizada	modelado espectral: banda alta
f.Slo	pendiente del espectro: f.70-f.30	modelado espectral: ancho de banda
f.Max	frecuencia de la máxima energía espectral	modelado espectral: energía máxima
f.nac	2º pico de la autocorrelación normalizada	existencia de frecuencias resonantes
LPFerr	error respecto a un espectro paso-baja	forma paso-baja del espectro

Tabla 4.2.1.: *Características de origen geofísico del esquema de parametrización geo_13.*

El proceso de extracción de estas características se resume en el Algoritmo 4.2.

Algoritmo 4.2 Extracción de las componentes geofísicas del vector *geo_13*.

Sean:

$s(n)$ un segmento (o *frame*) de señal con N muestras

$SC_{s(n)}(m)$ la suma acumulativa de $s(n)$

$nac_{s(n)}(m)$ autocorrelación normalizada de $s(n)$

$PN_{s(n)}(P)$ posición i normalizada donde se alcanza el porcentaje P de la suma acumulativa normalizada del módulo de $s(n)$:

$$PN_{s(n)}(P) = \frac{1}{N} \arg_i \left\{ \frac{SC_{|s(n)|}(m=i)}{SC_{|s(n)|}(m=N)} = \frac{P}{100} \right\} \tag{4.2.1}$$

Definimos las características:

t.P posición normalizada en el dominio temporal donde el módulo de la amplitud acumulada alcanza el $P\,\%$

f.P frecuencia normalizada a la cuál se alcanza el $P\,\%$ de la energía espectral acumulada. Extraída usando el operador *PN(P)* para un segmento obtenido tras aplicar la STFT

A.nac valor del 2º pico de autocorrelación normalizada, siendo $s(n)$ un segmento del sismograma, en el dominio temporal. La autocorrelación se normaliza usando su primer pico o máximo, en $nac(m=1)$

f.nac valor del 2º pico de autocorrelación normalizada, siendo $s(n)$ un segmento en el dominio de la frecuencia tras aplicar la STFT al sismograma

LPFerr Módulo de la diferencia en el dominio espectral de un segmento STFT respecto a un filtro paso-baja con frecuencia de corte f_c

Los resultados de las pruebas de configuración del esquema *geo_13* se muestran en la Figura 4.2.2. El modelado temporal de los HMMs se muestra especialmente útil en *col.04Ms*, con una mejora cualitativa de unos 10 puntos de $\%cCorr$ frente a los GMMs. Las características geofísicas se modelan bastante mejor con los GMMs en *dec.95Ms* que en *col.04Ms*, existiendo entre ambas bases de datos una alta similaridad respecto a los HMMs . La información dinámica del vector es más relevante en *dec.95Ms*, pero la diferencia de resultados quizás no justifique usar órdenes mayores de 2 (vectores con parámetros $\_A$ y $\_T$) para no comprometer la complejidad del sistema. Unos modelos de 4 componentes gausianas en *dec.95Ms* y de 8 en *col.04Ms* parece ser lo más efectivo en este sentido.

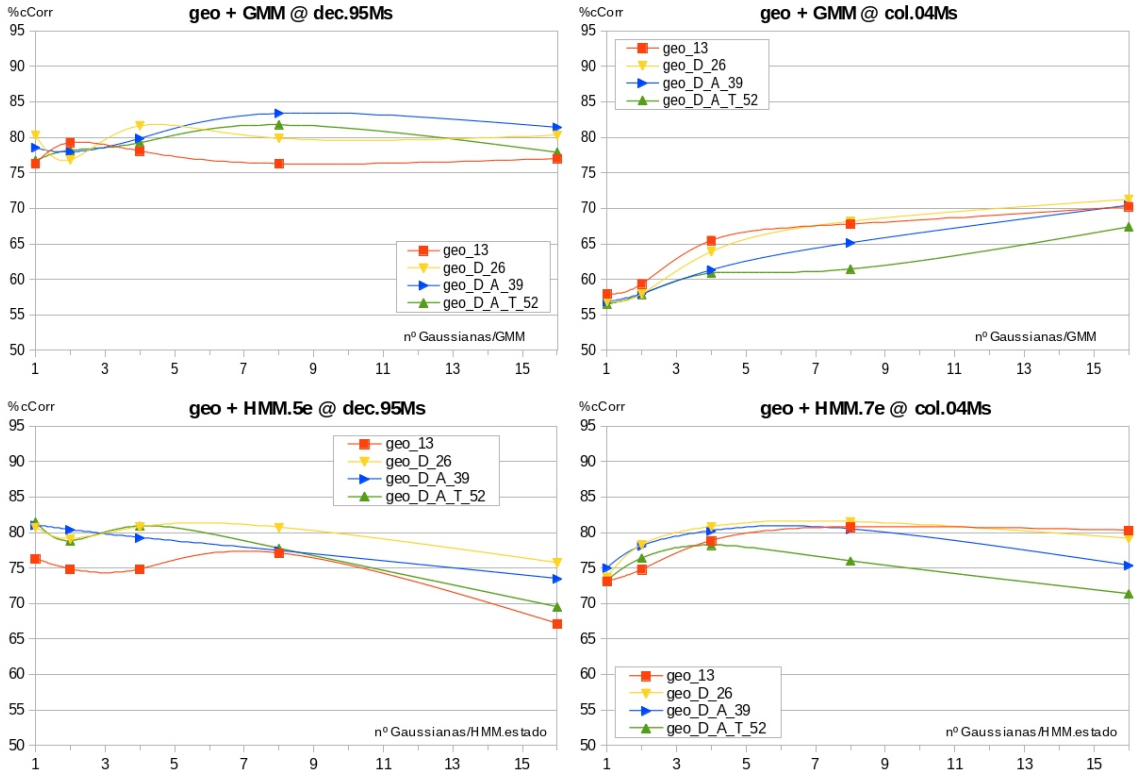


Figura 4.2.2.: Curvas de configuración del vector con características geofísicas *geo_13*. Tasa de reconocimiento (*%cCorr*) frente a distintos órdenes de información contextual y distintas componentes gausianas de los GMMs y HMMs.

4.2.2. Características basadas en transformaciones del sismograma

Muchos autores se decantan por un esquema clásico de parametrización basado en una transformación espectral sobre la forma de onda seguida de algún tipo de normalización e incluso, algún algoritmo sencillo para reducir dimensionalidad. Las combinaciones más empleadas son:

- coeficientes de predicción lineal (*LPC*, detallados en la Subsubsección 4.4.1.2) más normalizaciones logarítmicas (Del Pezzo et al., 2003; Masiello et al., 2006).
- coeficientes basados en la evolución temporal de la energía espectral, como STFT y DCT para reducir dimensionalidad en Avesani et al. (2012) o MFCCs (Figura 3.2.1) también referenciados como *LFCCs*, *Log-Frequency Cepstral Coefficients* (Benítez et al., 2007; Ibáñez et al., 2009; Cortés et al., 2009a; Álvarez et al., 2011; Gutiérrez Espinoza, 2013).
- transformadas Wavelets (Galli et al., 2009; Alasonati et al., 2006; Hloupis, 2009).

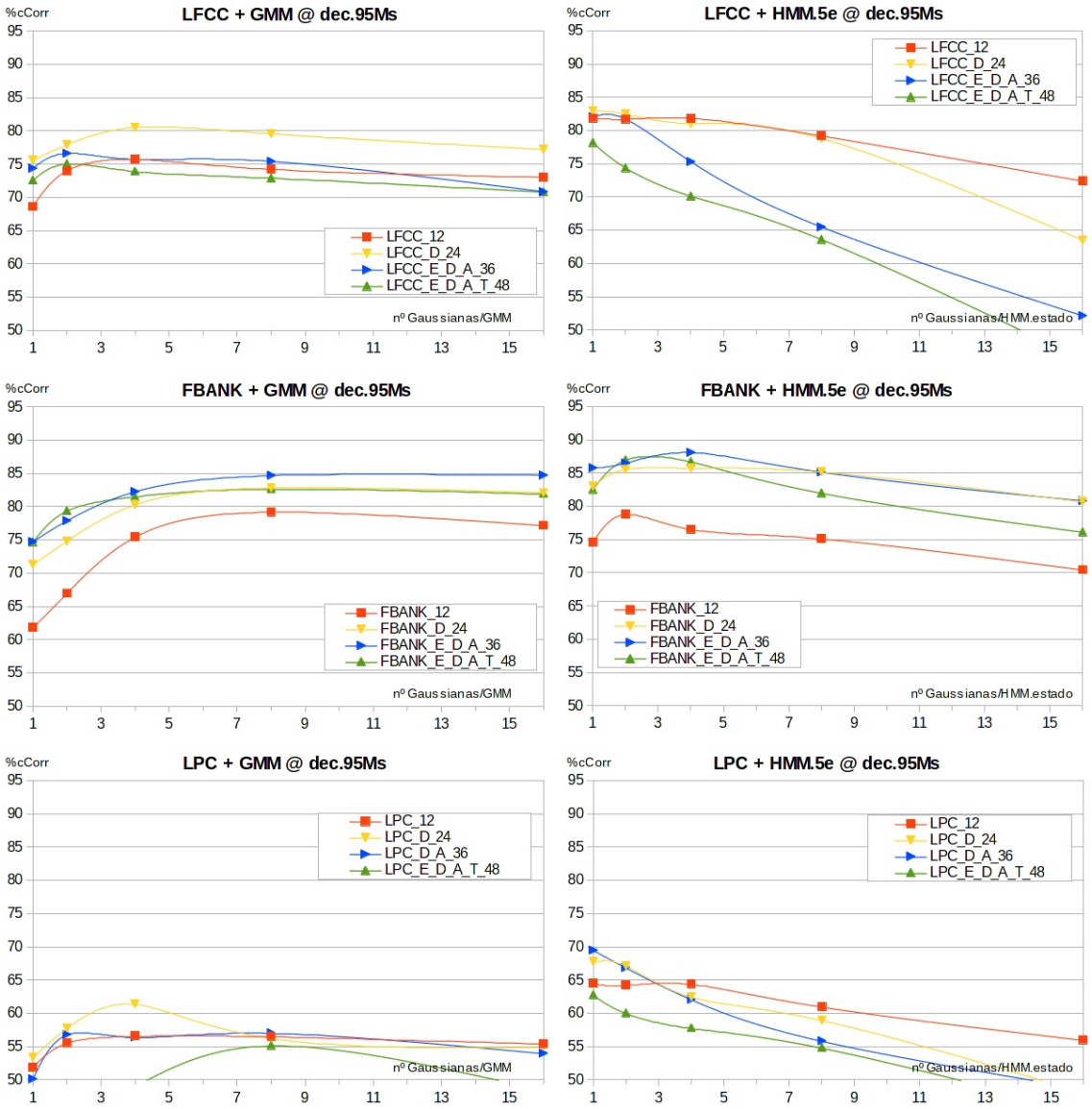


Figura 4.2.3.: Evaluación de parametrizaciones mediante transformación del sismograma en *dec.95Ms*. Eficiencia de reconocimiento (*%cCorr*) del modelado GMM y HMM.

Basándonos en las similitudes entre las señales sismo-volcánicas y las de voz establecidas en Ohrnberger (2001) y Benítez et al. (2007), configuraremos y evaluaremos esquemas de parametrización heredados del área de reconocimiento de habla: coeficientes LPC, LFCC y de análisis mediante banco de filtros (*FBANK*). Como describimos en la Figura 3.2.1 al construir el sistema de referencia en el Capítulo 3, LFCC describe la evolución temporal de la energía espectral en determinados intervalos de frecuencia denominados canales que luego decorrelaciona mediante la DCT. FBANK también se basa en análisis de banco de filtros, pero dispone linealmente los

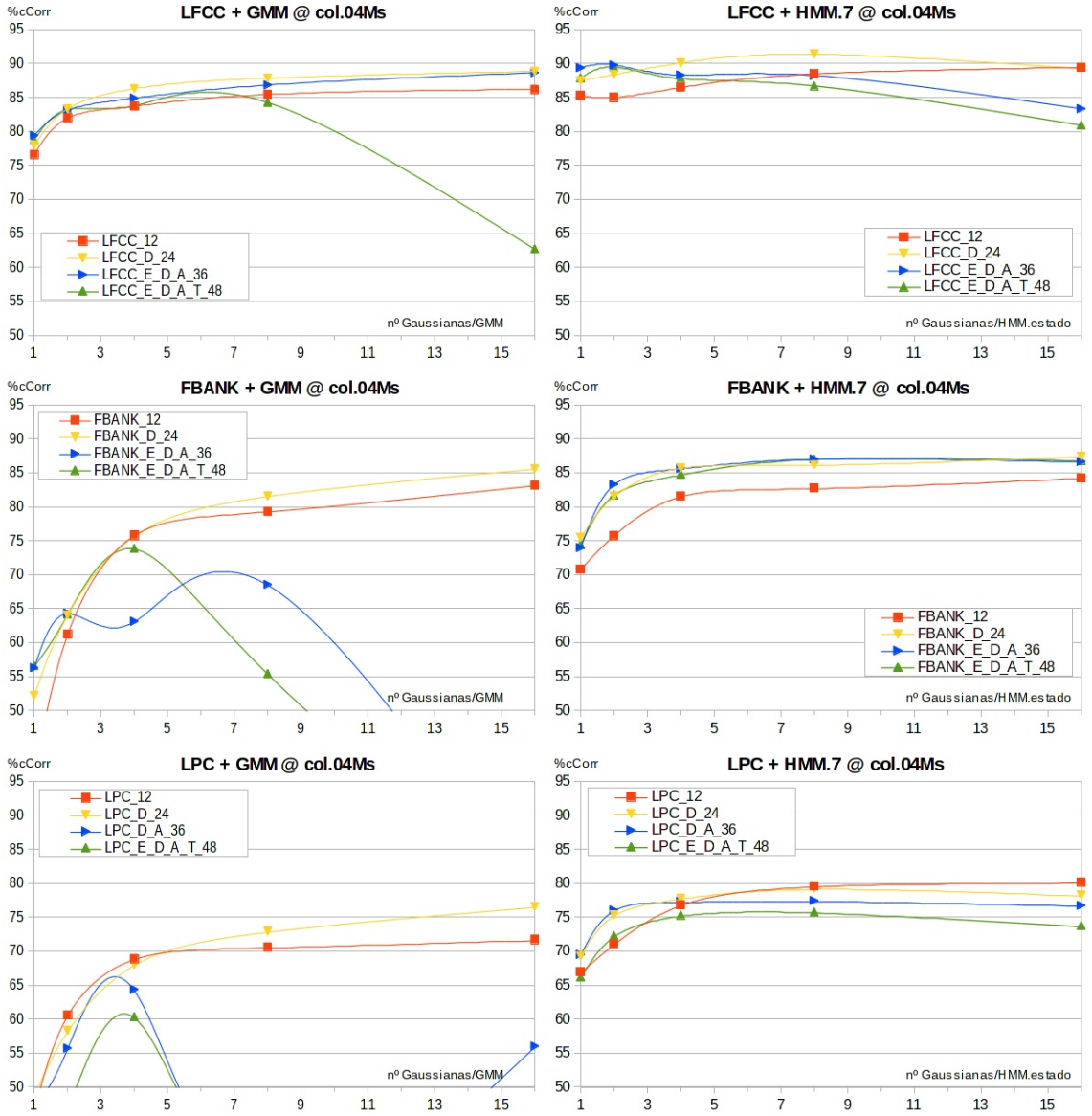


Figura 4.2.4.: Evaluación de parametrizaciones mediante transformación del sismograma en *col.04Ms*. Eficiencia de reconocimiento (*%cCorr*) del modelado GMM y HMM.

canales en el eje de la frecuencia en lugar de escalarlo logarítmicamente ni aplica la DCT para decorrelar las componentes de energía tal y como se hace en el esquema LFCC (Young et al., 2006).

4.2.2.1. Evaluación de los esquemas LPC, FBANK y LFCC

Las pruebas de evaluación se hacen variando el número de componentes gausianas de los modelos y el orden de la información dinámica del vector con el objetivo de

escoger la mejor parametrización. Los resultados se presentan en la Figura 4.2.3 y en la Figura 4.2.4. La Figura 4.2.3 nos revela el comportamiento de *dec.95Ms* respecto al número de gausianas para distintas configuraciones. Estudiando las gráficas encontramos que un valor de 4 gausianas es un buen compromiso entre un tasa de reconocimiento *%cCorr* alta y unos modelos con una complejidad moderada. El valor promedio de eficiencia no varía demasiado en los GMMs respecto los HMMs. En cuanto a las parametrizaciones, LPC obtiene significativamente peores resultados que LFCC y FBANK. FBANK parece ser algo mejor que LFCC.

La Figura 4.2.4 muestra la eficiencia de reconocimiento de *col.04Ms*. De nuevo, podemos apreciar una tendencia similar entre GMMs y HMMs, con la salvedad de que los GMMs parecen no poder modelar correctamente el espacio de características para vectores de tamaño elevado. En los HMMs la caída de las curvas debido al incremento en gausianas no es tan evidente como en el caso de *dec.95Ms* (Figura 4.2.3), probablemente porque en *col.04Ms* el número de clases es mayor (11 frente a 5) y también la duración media de los datos (detalles en la Sección 3.6), lo que manifiesta la necesidad de usar unos modelos más complejos. Un valor de 8 gausianas parece lo más adecuado para los HMMs y de 4 en los GMMs. La parametrización LPC sigue siendo más ineficiente que LFCC y FBANK. Esta vez LFCC obtiene mejores resultados que FBANK.

dec.95Ms	GMM				HMM.5e			
	LFCC	*FBANK*	LPC	*media*	LFCC	*FBANK*	LPC	*media*
<par>_12	73.13	72.14	55.23	66.83	79.40	75.12	62.05	72.19
<par>_D_24	78.18	78.29	56.79	*71.09*	77.77	84.06	60.94	*74.26*
<par>_D_A_36	74.62	80.89	54.91	70.14	71.36	85.30	60.83	72.50
<par>_D_A_T_48	73.07	80.06	49.65	67.59	66.39	82.89	56.19	68.49
media	74.75	*77.85*	54.14	68.91	73.73	*81.84*	59.95	71.86

col.04Ms	GMM				HMM.7e			
	LFCC	FBANK	LPC	*media*	*LFCC*	FBANK	LPC	*media*
<par>_12	82.81	67.89	63.66	71.45	86.95	79.04	74.93	80.31
<par>_D_24	84.87	71.85	64.90	*73.87*	89.34	83.30	75.99	*82.88*
<par>_D_A_36	84.62	54.78	51.18	63.53	87.83	83.34	75.43	82.20
<par>_D_A_T_48	78.58	54.46	37.96	57.00	86.58	82.99	72.66	80.74
media	*82.72*	62.25	54.43	66.46	*87.68*	82.16	74.75	81.53

Tabla 4.2.2.: *%cCorr (promediado sobre el número de gausianas) para transformaciones del dominio temporal de la señal.*

Los resultados promediados para el número de gausianas aparecen la Tabla 4.2.2. Haciendo un análisis general, se comprueba que la tasa de reconocimiento más alta se alcanza cuando la información de contexto se describe añadiendo las derivadas

de orden 1 (esquemas *<par>_D_24*). Una vez fijada la configuración óptima para la información contextual, nos centramos en seleccionar la parametrización más adecuada. FBANK se proclama como la mejor opción en *dec.95Ms*, mientras que LFCC lo hace en *col.04Ms*. Sin embargo, globalmente parece que LFCC es la elección correcta, pues promediando sobre el conjunto de resultados es más estable que FBANK.

4.2.2.2. Configuración del esquema *LFCC_D*

Una vez escogida la parametrización *LFCC_D* a partir de la Tabla 4.2.2, nos proponemos encontrar su tamaño de vector y duración del segmento más adecuados. La Figura 4.2.5 nos muestra las curvas de configuración.

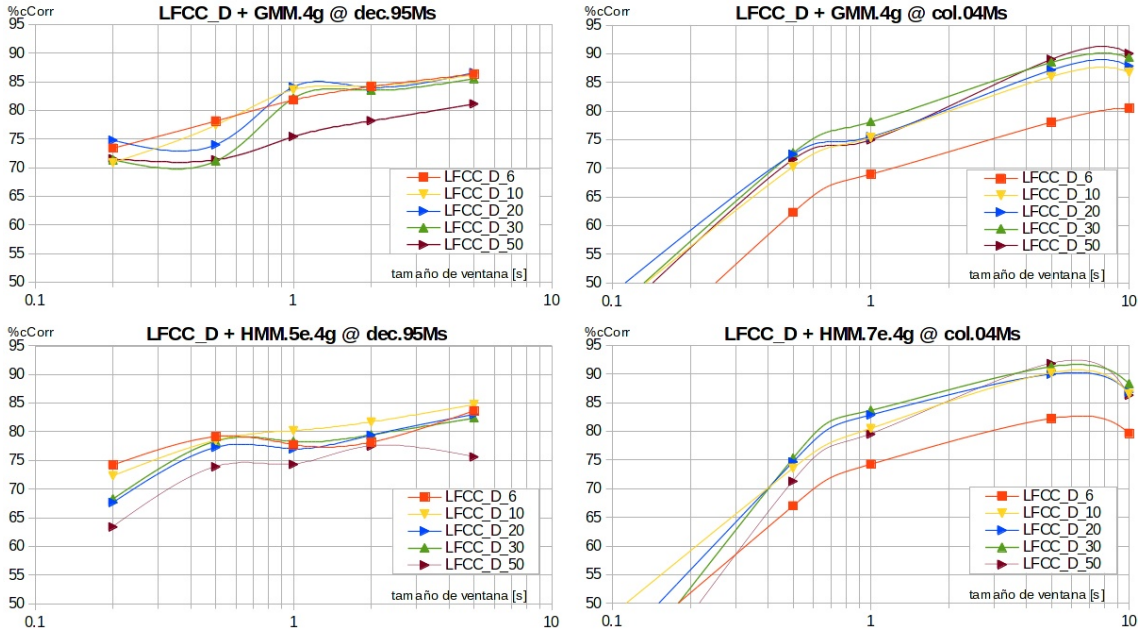


Figura 4.2.5.: Curvas de configuración (*%cCorr* frente a duración de ventana) para el esquema de parametrización *LFCC_D*.

Los resultados son muy similares entre GMMs y HMMs. Respecto al tamaño de ventana, se sigue una clara tendencia: las ventanas grandes (de 5 segundos) son más eficientes que las pequeñas. En ventanas muy grandes (de 10 segundos) el *%cCorr* empieza a decaer en *col.04Ms*.

Un comportamiento similar es obtenido en Avesani et al. (2012) con STFT+DCT y un vector de 4 componentes. Con el mismo esquema de parametrización, Hoogenboezem (2010) estudiando rangos de duración del segmento de 0.64 a 20.48 [*s*] obtiene un máximo de 81 % de eficiencia de reconocimiento para ventanas de 2.56 [*s*] y 75 % de solapamiento, pero con muy poca diferencia respecto al peor valor (78 % para 0.64 [*s*] y 25 % de solapamiento). De hecho, Hoogenboezem (2010) cuando combina las

parametrizaciones STFT y LPC con los algoritmos de reducción de dimensionalidad (DCT, PCA y el análisis discriminante lineal descritos en la Sección 4.4) demuestra que el tamaño de ventana y el solapamiento tienen muy poca influencia, si bien se obtienen los mejores resultados con ventanas grandes (de 10 a 20 [*s*]). Para Avesani et al. (2012) el tamaño de segmento sí que influye en la tasa de reconocimiento (hallamos diferencias de 15 puntos para segmentos que van desde 0.64 hasta 10.24 [*s*]), pero el solapamiento influye más levemente (apenas 5 puntos para valores desde el 0 % hasta el 75 % de solapamiento).

	dec.95Ms		*col.04Ms*		
	GMM.4g	HMM.5e.4g	GMM.4g	HMM.7e.4g	*media*
LFCC_D_6	80.77	78.59	64.63	68.70	73.17
LFCC_D_10	80.47	79.51	72.73	75.80	*77.13*
LFCC_D_20	80.70	76.88	74.24	75.18	76.75
LFCC_D_30	78.75	77.38	74.75	74.79	76.42
LFCC_D_50	75.54	72.94	73.80	71.68	73.49
media	79.25	77.06	72.03	73.23	75.39

Tabla 4.2.3.: *Configuración del esquema LFCC_D. %cCorr promediando la duración del segmento.*

En cuanto al tamaño del vector, las curvas de Figura 4.2.5 muestran que vectores muy grandes (de 50 componentes) son menos efectivos en *dec.95Ms* y que tamaños pequeños funcionan peor en *col.04Ms*, no existiendo una gran diferencia entre el resto de esquemas. Con el objetivo de ayudar a tomar una decisión, sintetizamos las gráficas de Figura 4.2.5 en la Tabla 4.2.3, que constata la equidad entre GMMs y HMMs. Finalmente, aunque por poco margen se alcancen mejores valores de reconocimiento para otros esquemas en los GMMs, la mejor opción es dividir el espectro en 5 bandas e incluir sus derivadas temporales de orden 1 (vector *LFCC_D_10*).

4.2.3. Características basadas en estadística de los datos

Varios parámetros estadísticos pueden ser considerados de forma general para describir series temporales. Interpretando el sismograma como la representación en el dominio temporal de las señales sísmicas, nos encontramos diversos estadísticos para describir eventos sismo-volcánicos:

- *Estadísticos sobre el sismograma*: desviación estándar, media, mediana, máximo, curtosis y asimetría. Estadísticos espectrales: media y energía (Curilem et al., 2009).
- *Estadísticas circulares*: varianza y oblicuidad (San-Martin et al., 2010).

El vector *stat_8* que proponemos en la Tabla 4.2.4 contiene medidas estadísticas básicas extraídas para cada segmento en el dominio temporal y en el espectro de

identificador	*característica*	*propiedad a medir*
htkE	energía HTK	energía logarítmica normalizada
A.std	desviación de amplitud	variabilidad del sismograma
A.kur	curtosis de amplitud	distribución de variabilidad gausiana
A.skew	asimetría de amplitud	asimetría respecto la media de amplitud
f.meanE	energía media espectral	energía media espectral
f.std	desviación espectral	variabilidad espectral
f.kur	curtosis espectral	distribución de la variabilidad espectral
f.skew	asimetría espectral	asimetría respecto la media espectral

Tabla 4.2.4.: *Características estadísticas de la parametrización stat_8.*

Fourier. Su curvas de eficiencia se dibujan en la Figura 4.2.6. A grandes rasgos se aprecia la misma tendencia en las gráficas obtenidas con las características solo geofísicas, pero unos 5 puntos por debajo. En este caso, el punto de inflexión de la tasa de reconocimiento parece estar ligeramente desplazado hacia la derecha, entorno a las 8 gausianas, probablemente debido a que el vector base tiene menos componentes

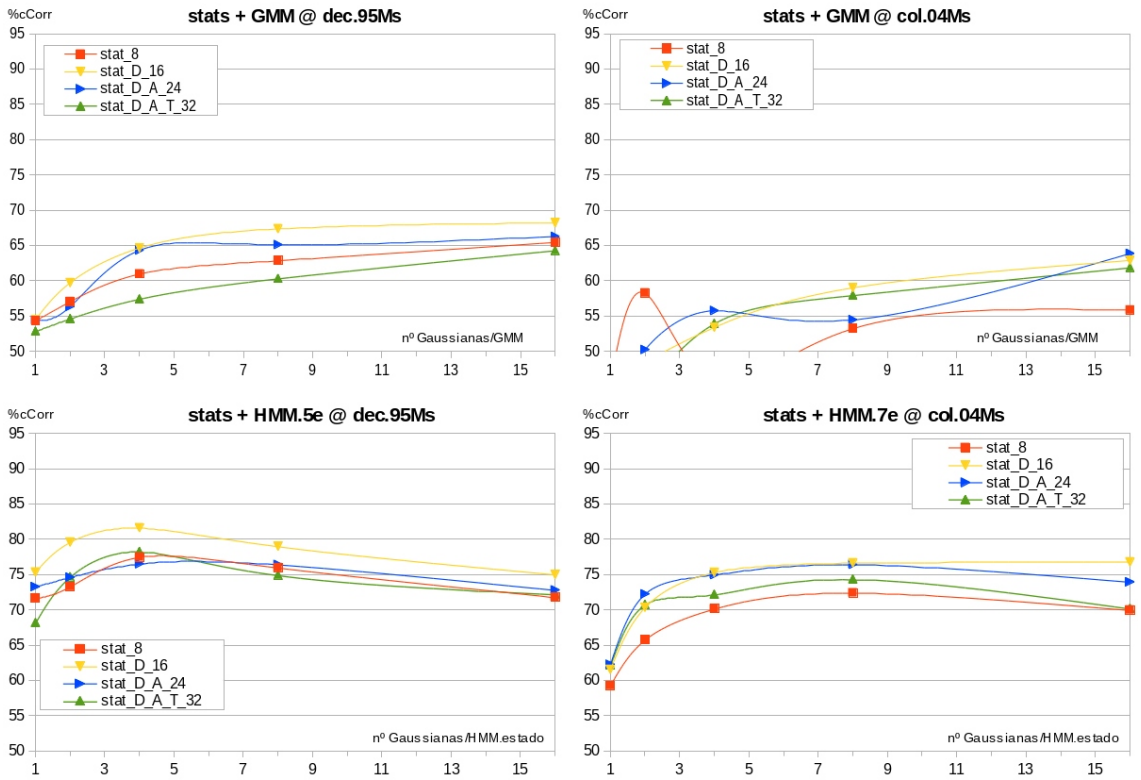


Figura 4.2.6.: Configuración del vector base estadístico *stats_8*. Reconocimiento (*%cCorr*) para distinta información contextual y distintas complejidad de los modelos GMMs y HMMs.

(8 frente a las 13 de la parametrización geofísica *geo_13* definida en el apartado anterior), necesitándose modelos algo más complejos para modelar los mismos datos. Este hecho nos revela que virtualmente, la capacidad de discriminación de estos dos vectores es bastante similar, a igualdad del tamaño del vector.

En cuanto a la información dinámica, observamos que no hay una gran diferencia (unos 5 puntos en la tasa $\%cCorr$) entre utilizar un orden 0, dado por el vector base, y un orden 3, lo que sugiere describir los datos de la forma más sencilla posible, quizás mediante un esquema que incluya solo las derivadas temporales de orden 1. Una comparativa más detallada entre los esquemas basados en propiedades geofísicas, estadísticas y esquemas mixtos la realizaremos en la Subsección 4.2.4.

¿Se deben normalizar en energía los registros sísmicos?. La propiedad *htkE* normaliza la amplitud de todo el registro sísmico y luego lo escala logarítmicamente. La normalización en amplitud del sismograma es una estrategia que persigue encontrar características exportables, pero en contrapartida destruye la información de la energía absoluta de un evento que suele ser bastante importante para los geofísicos. Una discusión del tema la encontramos en la Sección A.5.

4.2.4. Características basadas en esquemas mixtos

La descripción heterogénea de las señales mediante propiedades de diferente naturaleza persigue aumentar la robustez de la parametrización. Con el objetivo de caracterizar geofísicamente los datos y de integrar una información más generalizada proporcionada por transformadas clásicas, podemos encontrar que son muchos los trabajos que optan por esquemas mixtos:

- Medidas espectrales, de autocorrelación, de fase y frecuencia instantáneas, de polaridad y de desfase en Kubichek and Quincy (1985).
- Estimaciones espectrales mediante LPC e información de amplitud (Scarpetta et al., 2005).
- Características geofísicas y estadísticas exportables junto coeficientes espectrales (Álvarez et al., 2011; Cortés et al., 2014; Curilem et al., 2014b,a).
- Componentes de polarización, de la STFT, de autocorrelación y de la traza compleja (Leprettre et al., 1998; Beyreuther and Wassermann, 2008; Köhler et al., 2009).
- Estadísticos tomados sobre el sismograma y energías de canales espectrales (Orlic and Loncaric, 2010).

La construcción de nuestro propio vector mixto la dividiremos en 2 etapas:

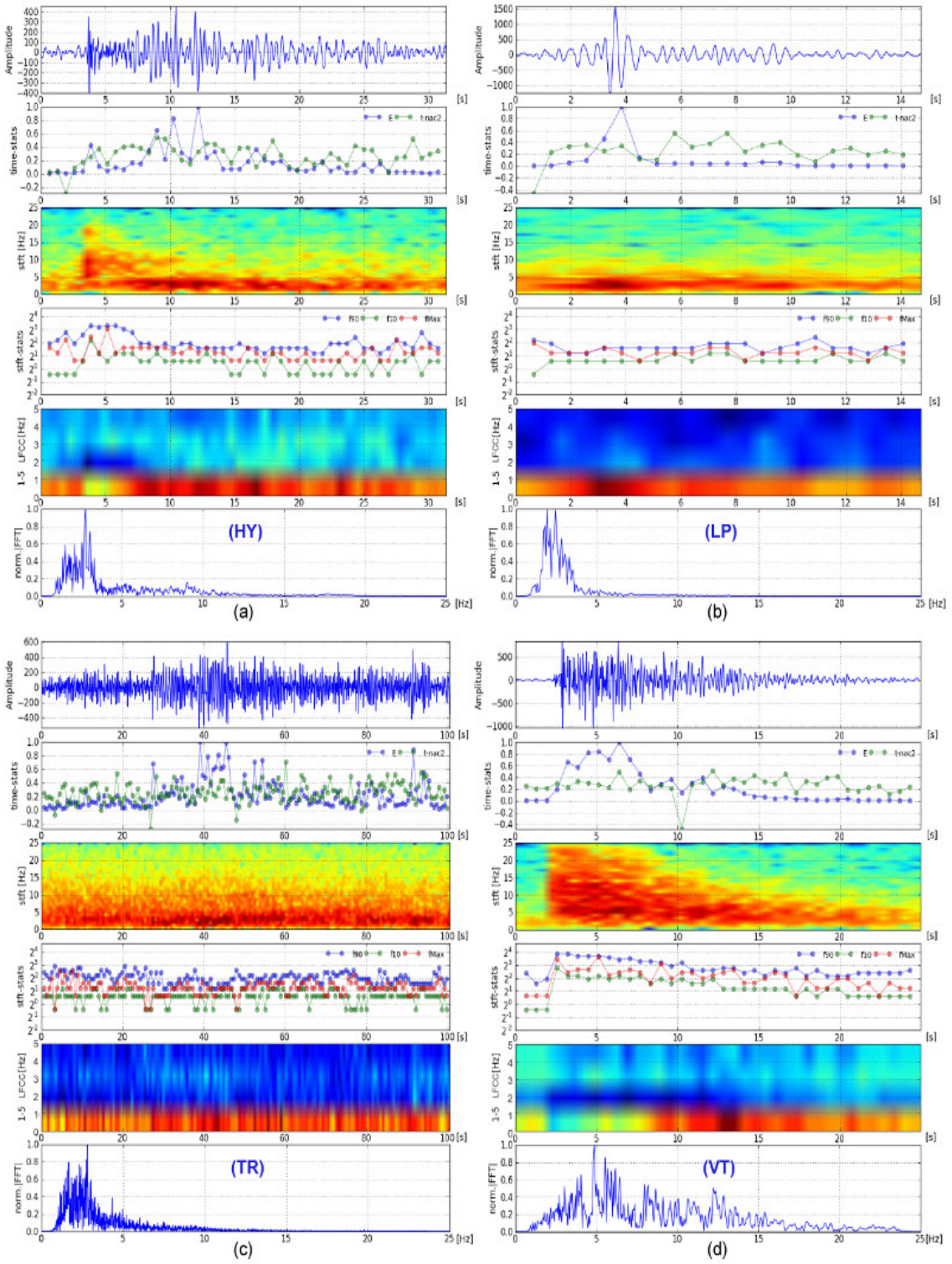


Figura 4.2.7.: Distintas representaciones para los eventos de Decepción: sismograma, estadísticas temporales para cada segmento, espectrograma, estadísticas espectrales de segmentos, coeficientes LFCCs y espectro del sismograma.

1. Propondremos un nuevo esquema, *geoStat_21* uniendo la información geofísica y estadística de los esquemas geofísicos (*geo_13*) y estadísticos (*stat_8*) ya presentados en los apartados anteriores. Analizaremos qué nivel de información de contexto es más eficiente añadir.
2. Estudiaremos por independiente cada componente geo-estadística del vector propuesto anteriormente para seleccionar aquellas que potencialmente pueden ser más discriminativas y construir con ellas un nuevo vector añadiendo información espectral del esquema *LFCC_D_10* ya estudiado en la Subsección 4.2.2.

El resultado será una parametrización mixta con características geofísicas, estadísticas y espectrales sobre el que aplicaremos técnicas de selección (Sección 4.3) y reducción (Sección 4.4) en los próximos apartados.

4.2.4.1. Construcción del vector geo-estadístico

En este test evaluamos un esquema mixto creado al juntar características geofísicas y estadísticas a partir vectores ya propuestos y configurados en la Subsección 4.2.1 y la Subsección 4.2.3. Los resultados nos guiarán para seleccionar cuanta información de contexto es necesaria para modelar nuestros datos.

dec.95Ms	GMM.4g			HMM.5e.4g			
	geo.13	stats.8	geoStats.21	geo.13	stats.8	*geoStats.21*	*media*
<par.coefs>	77.38	60.14	73.59	74.00	74.06	77.99	72.86
<par.coefs>_D	79.79	62.90	78.82	79.38	78.13	85.14	*77.36*
<par.coefs>_D_A	80.23	61.31	58.80	78.28	74.75	84.61	73.00
<par.coefs>_D_A_T	78.77	57.90	42.36	77.66	73.66	83.45	68.97
media	*79.04*	60.56	63.39	77.33	75.15	*82.80*	73.05

col.04Ms	GMM.4g			HMM.7e.4g			
	geo.13	stats.8	*geoStats.21*	geo.13	stats.8	*geoStats.21*	*media*
<par.coefs>	64.14	51.67	70.67	77.53	67.52	81.09	68.77
<par.coefs>_D	63.61	53.56	70.52	78.66	72.15	84.34	*70.47*
<par.coefs>_D_A	62.33	53.36	64.66	77.81	71.99	81.45	68.60
<par.coefs>_D_A_T	60.81	52.04	59.71	75.00	69.93	80.75	66.37
media	62.72	52.66	*66.39*	77.25	70.40	*81.91*	68.55

Tabla 4.2.5.: *Eficiencia para esquemas geo-estadísticos en función de la información dinámica.* Eficiencia de reconocimiento de clase, *%cCorr* (promediado sobre el número de gausianas) cuando se añade información contextual a esquemas de parametrización *<par.coefs>* de distinta naturaleza y número de componentes.

La tasa de reconocimiento de los esquemas geo-estadísticos viene representada en las gráficas de la Figura 4.2.8. En ellas observamos que, al contrario de como ocurría

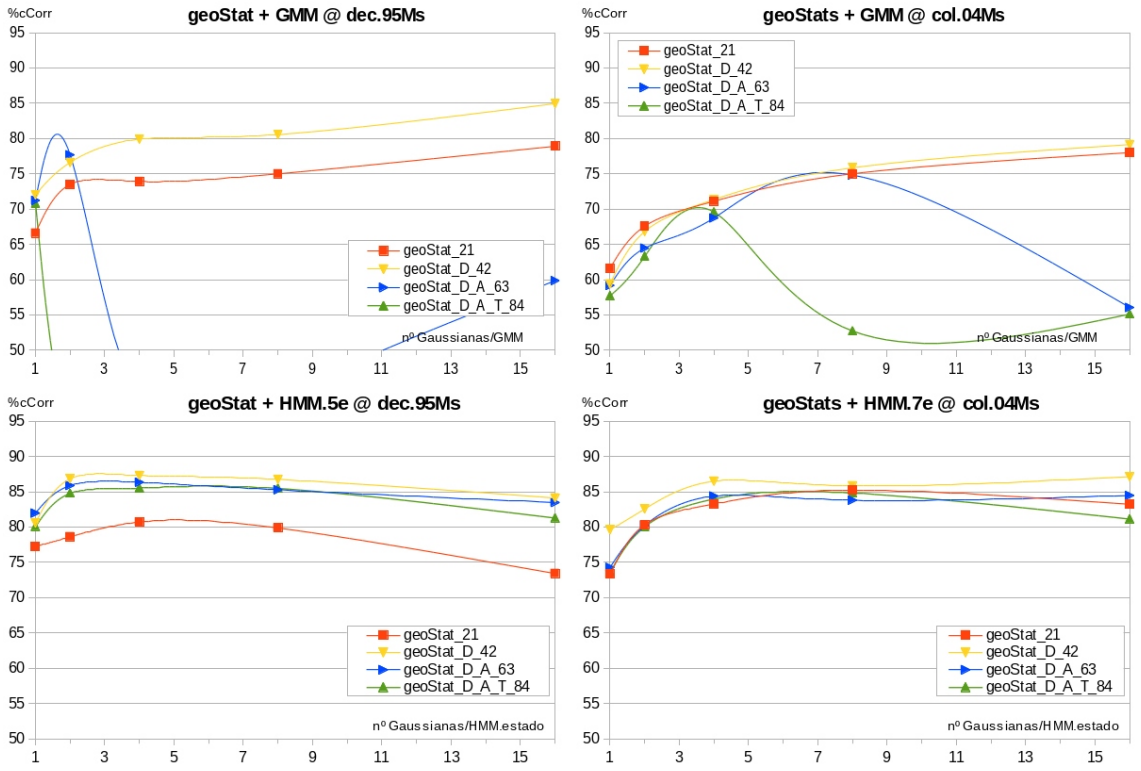


Figura 4.2.8.: Comportamiento del vector geo-estadístico base *geoStats_21* frente a distintas complejidades de los modelos y diferentes niveles de información contextual.

en las parametrizaciones basadas en el espectro (Subsección 4.2.2), a priori, existe una diferencia considerable entre usar GMMs o HMMs.

La Tabla 4.2.5 promedia para el número de gausianas los resultados de la Figura 4.2.8 y los compara con los vectores geofísicos y estadísticos que forman la parametrización híbrida. Apreciamos que no siempre los esquemas mixtos parecen ser la mejor opción (véanse los resultados para *dec.95Ms* con GMM.4g), sin embargo, conjuntamente ofrecen el mayor porcentaje de reconocimiento. Cuestión aparte es si en algunos casos concretos vale la pena agrandar el vector con características híbridas cuando esquemas más simples ofrecen buenos resultados.

Atendiendo a la información de contexto, los valores de la Tabla 4.2.5 claramente apuntan a escoger una solución geo-estadística que incluya los parámetros delta (*geoStats.21_D*), hecho que también se constata cualitativamente al analizar las curvas de la Figura 4.2.8. En el caso particular de *dec.95Ms* parece que la información estadística solo aporta mejoras al modelar explícitamente la evolución temporal de los eventos vía HMMs, siendo la mejor opción un esquema solo geofísico con información contextual de orden 1 o 2 al usar los GMMs.

4.2.4.2. Evaluación de las características geo-estadísticas por independiente

En este apartado vamos a analizar el poder discriminatorio de cada una de las 42 componentes del vector *geoStats.21_D* (renombrado como *geoSTATS.D.42*). La Figura 4.2.9 presenta el promedio de la tasa de reconocimiento que cada característica alcanza por independiente cuando describe los eventos segmentados en tamaños de 0.2, 0.5, 1, 2 y 5 segundos en *dec.95Ms* y de 0.2, 0.5, 1, 5 y 10 segundos en *col.04Ms* (un estudio detallado para cada componente lo encontramos en la Sección A.1). Se observa como el *%cCorr* de reconocimiento de los HMMs es levemente mejor en

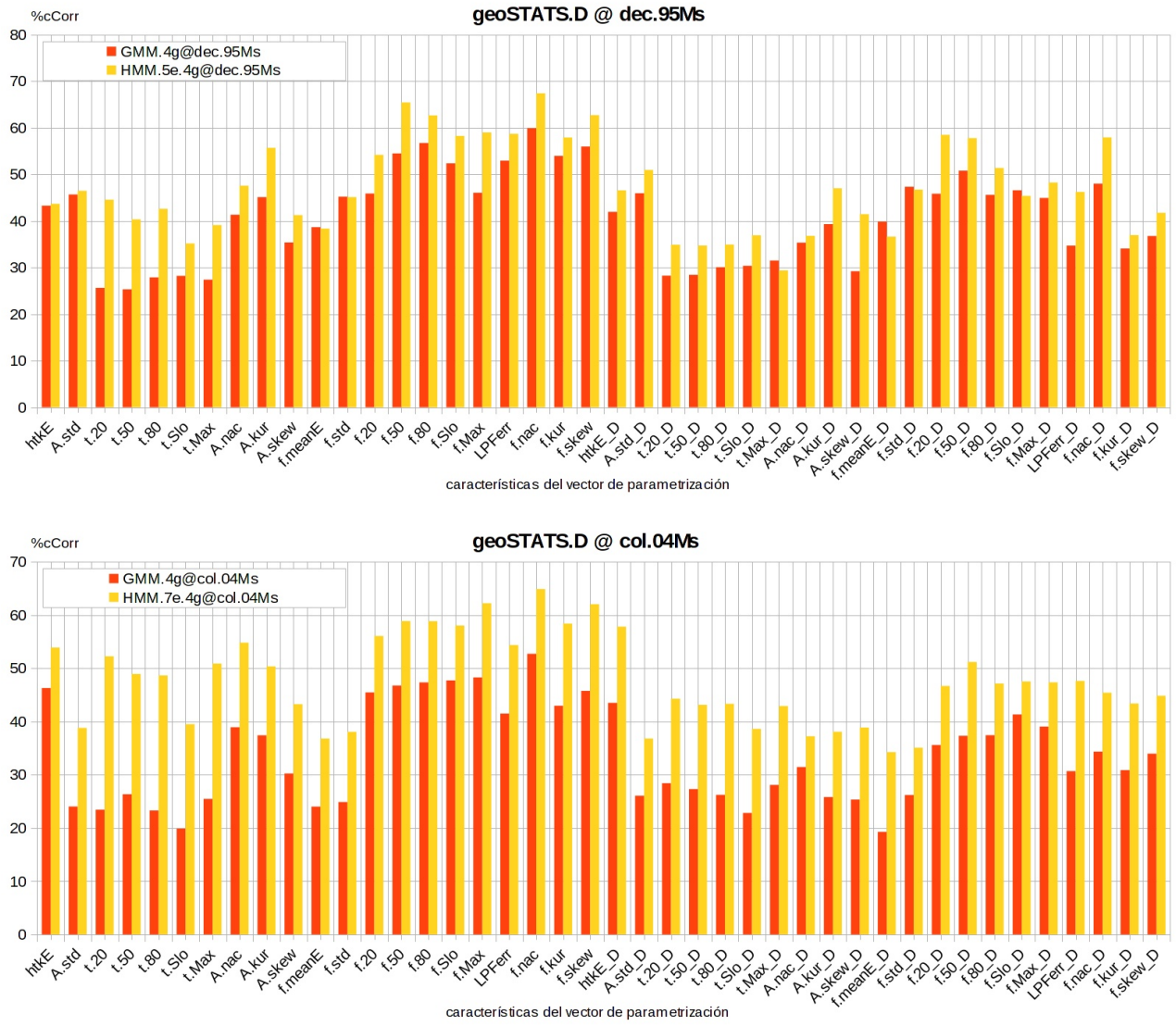


Figura 4.2.9.: Eficiencia de reconocimiento promediado respecto a la duración del segmento para cada característica del esquema geo-estadístico *geoSTATS.D.42*.

dec.95Ms que la obtenida con GMMs. En *col.04Ms* esa diferencia es aún más acusada.

Se vislumbra un patrón en la distribución de los resultados que es seguido con pequeñas variaciones tanto por GMMs como por HMMs en los datos de *dec.95Ms* y *col.04Ms*:

- Las componentes extraídas a partir del espectro obtienen resultados sensiblemente mejores que aquellas que se definen directamente sobre la forma de onda.
- Este comportamiento también se aprecia en la información de contexto. En general, las características delta (*_D*) aportan el mismo grado de discriminación que sus homólogas estáticas en el dominio temporal y algo menor en la frecuencia.

La Tabla 4.2.6 ordena las características del vector *geoSTATS.D.42* de forma decreciente respecto a su tasa de reconocimiento promediada. Analizando sus resultados podemos hacer las siguientes observaciones:

- Las componentes más relevantes son independientes del corpus de datos y de los modelos usados: 7 de las 10 mejores características son comunes para todas las combinaciones, estando las 8 mejores de cada combinación dentro del primer cuartil (delimitado en las 10 mejores).
- Se comprueba la tendencia presentada en Figura 4.2.9: las características basadas en el dominio temporal son bastante menos discriminativas que aquellas extraídas a partir del espectro; la información de contexto es menos útil que la información base en el dominio de la frecuencia, pero igual de efectiva en el dominio del sismograma.

Mediante la representación de la Tabla 4.2.6 pretendemos reducir el tamaño del vector seleccionando las mejores componentes. Para ello escogeremos las 10 características estáticas que obtienen los mejores resultados (las 9 primeras de la Tabla 4.2.6 junto con *htkE*) y sus correspondientes 10 características de contexto, formando el vector *geoSTATS.D.20* con información geo-estadística de 20 componentes.

4.2.5. Resultados experimentales: elección del vector mixto *geoLFCC.D.30*

Tendencia de los resultados experimentales. En general todas las pruebas de configuración que impliquen un aumento de complejidad en los modelos tienden a seguir un *patrón general en los resultados*: la tasa de reconocimiento empieza creciendo hasta un valor máximo para disminuir suavemente a partir de él. Este punto de inflexión viene marcado por una combinación de varios factores que indica el grado óptimo de complejidad en el modelado:

		dec.95Ms		*col.04Ms*		
relevancia	característica	GMM.4g	HMM.5e.4g	GMM.4g	HMM.7e.4g	*media*
1	*f.nac*	*59.93*	*67.41*	*52.73*	*64.89*	61.24
2	*f.skew*	*55.98*	*62.72*	*45.76*	*62.04*	56.62
3	*f.80*	*56.74*	*62.65*	*47.35*	*58.84*	56.39
4	*f.50*	*54.49*	*65.44*	*46.76*	*58.87*	56.39
5	*f.Slo*	*52.37*	*58.25*	*47.71*	*58.03*	54.09
6	*f.Max*	46.05	*58.99*	*48.28*	*62.22*	53.88
7	*f.kur*	*53.96*	*57.92*	*42.96*	*58.40*	53.31
8	*LPFerr*	*52.95*	*58.71*	41.48	*54.37*	51.88
9	*f.20*	45.87	54.19	*45.46*	*56.07*	50.40
10	*f.50_D*	*50.79*	57.78	37.32	51.18	49.27
11	*htkE_D*	41.95	46.57	*43.50*	*57.82*	47.46
12	*A.kur*	45.12	55.69	37.41	50.35	47.14
13	*htkE*	43.27	43.68	*46.28*	53.89	46.78
14	*f.20_D*	45.84	*58.48*	35.59	46.68	46.65
15	*f.nac_D*	*48.00*	*57.93*	34.34	45.40	46.42
16	*A.nac*	41.33	47.57	38.93	54.81	45.66
17	*f.80_D*	45.60	51.37	37.42	47.13	45.38
18	*f.Slo_D*	46.59	45.38	41.33	47.52	45.20
19	*f.Max_D*	44.94	48.24	39.04	47.35	44.89
20	*A.std_D*	45.94	50.94	26.06	36.80	39.93
21	*LPFerr_D*	34.71	46.23	30.68	47.62	39.81
22	*f.skew_D*	36.77	41.75	33.96	44.83	39.33
23	*f.std_D*	*47.35*	46.72	26.18	35.08	38.83
24	*A.std*	45.67	46.48	24.03	38.80	38.75
25	*f.std*	45.22	45.12	24.87	38.06	38.32
26	*A.kur_D*	39.31	47.01	25.81	38.08	37.55
27	*A.skew*	35.38	41.26	30.23	43.25	37.53
28	*t.20*	25.62	44.56	23.43	52.25	36.47
29	*f.kur_D*	34.09	36.97	30.87	43.38	36.33
30	*t.Max*	27.39	39.12	25.45	50.88	35.71
31	*t.80*	27.87	42.60	23.28	48.67	35.60
32	*t.50*	25.31	40.35	26.33	48.93	35.23
33	*A.nac_D*	35.34	36.80	31.45	37.23	35.20
34	*f.meanE*	38.67	38.34	24.01	36.80	34.45
35	*t.20_D*	28.25	34.89	28.41	44.29	33.96
36	*A.skew_D*	29.19	41.46	25.33	38.86	33.71
37	*t.80_D*	30.05	34.93	26.21	43.33	33.63
38	*t.50_D*	28.42	34.73	27.30	43.13	33.40
39	*t.Max_D*	31.49	29.37	28.08	42.90	32.96
40	*f.meanE_D*	39.84	36.63	19.25	34.26	32.50
41	*t.Slo_D*	30.36	36.92	22.81	38.61	32.18
42	*t.Slo*	28.20	35.16	19.90	39.50	30.69

Tabla 4.2.6.: *%cCorr para cada característica del vector geo-estadístico geoSTATS.D.42*. Los 10 mejores valores de *%cCorr* de cada columna están en *cursiva*.

- El número de componentes (gausianas en nuestro caso) de la distribución de probabilidad que modela a las características.
- El tamaño del vector de parametrización, o, equivalentemente, la dimensionalidad del espacio de características.
- La efectividad del vector de parametrización para describir los datos.

Las variaciones de este comportamiento típico vienen motivadas por las diferencias entre las bases de datos así como por el tipo de modelado escogido. Concretamente:

1. **Diferencias de GMMs respecto HMMs**. La caída en los HMMs comienza antes, lo que puede explicarse teniendo en cuenta que un HMM con E estados tiene $E-2$ estados emisores frente solo uno de un GMM, multiplicando por $E-2$ sus componentes respecto a los GMMs. Esto se traduce en una degradación más temprana de los HMMs en gráficas que evalúen la tasa de reconocimiento frente al número de gausianas por estado.
2. **Diferencias respecto al modelado de *col.04Ms* frente a *dec.95Ms*.** Los datos de *col.04Ms* están agrupados en más clases y sus eventos tienen más variabilidad que los de *dec.95Ms* (Sección 3.6), lo que implica que su descripción resultará más compleja en dos sentidos:
   a) Los modelos estadísticos requerirán más componentes para describir la mayor variabilidad en las características de *col.04Ms*.
   b) Usualmente, los HMMs serán más efectivos que los GMMs al modelar explícitamente la variabilidad temporal. Dicha mejora es más notable en *col.04Ms* que en *dec.95Ms*.

Esquema geoLFCC.D.30 propuesto. Finalmente, tras el análisis y configuración de distintos esquemas, proponemos como parametrización de referencia para ser usada en las secciones siguientes en el estudio de la reducción de dimensionalidad el vector *geoLFCC.D.30* con 30 componentes: las 20 geo-estadísticas del vector *geo.STATS.D.20* seleccionadas en la Subsección 4.2.4 más las 10 del vector *LFCC_D_10* con información espectral escogido en la Subsección 4.2.2.

4.3. Reducción de dimensionalidad mediante Selección de Características

En esta sección aplicaremos distintos métodos para ordenar las componentes de un vector de características de acuerdo a su poder de discriminación. Compararemos técnicas clásicas existentes con otras nuevas que propondremos. Estudiaremos primero la selección de características mediante filtros (SC_F), más simples de implementar y con menor carga computacional para luego pasar a analizar la selección guiada por modelos de predicción (SC_M), más compleja pero que suele obtener mejores resultados.

Configuración básica de los test de evaluación. En los test de reducción de dimensionalidad usaremos los GMMs para modelar el espacio de características descrito por el vector mixto *geoLFCC.D.30*. Aunque en el estudio hecho en la Sección 4.2 se observan mejoras para esquemas concretos dependiendo deunas configuraciones u otras, seguiremos utilizando los valores estándar propuestos en la Subsección 4.1.6, dejando para próximos capítulos una configuración adaptada a cada tipo de evento.

Elección de los GMM como modelos para reducir dimensionalidad. El hecho que la tendencia en los resultados sea aproximadamente la misma en GMMs y HMMs unido a que las características más discriminativas lo son con independencia de los modelos y de las bases de datos (Subsección 4.2.4), justifica el uso de solo los GMMs (o HMMs) en el estudio de la reducción de dimensionalidad. Esta elección viene determinada por la simplicidad de los GMMs frente a los HMMs para reducir el gasto computacional que conlleva evaluar distintas técnicas con ambos modelos. Aunque los resultados no saldrán idénticos, podemos esperar conclusiones parecidas entre GMMs y HMMs a partir de las conclusiones obtenidas en la Subsubsección 4.2.4.2.

4.3.1. Selección de Características por Filtros

El uso de filtros garantiza una independencia de los resultados de selección respecto al modelado de datos al definir la medida de bondad. Pueden evaluar subgrupos de características conjuntamente obteniendo una medida para cada subconjunto que luego se comparan entre ellas.

Nuestra implementación de los filtros incluye diferentes tipos de medidas estadísticas, de correlación y de probabilidad definiendo un operador de *medida* de la distancia M dentro del espacio de características (M no siempre será una distancia en el estricto sentido algebraico). Como se esquematiza en el Algoritmo 4.3, cada filtro ordenará una característica según su distancia media al resto de ellas, tal que los elementos que globalmente más se parezcan entre sí (estén más cerca) serán los últimos en seleccionarse, siendo los más alejados los primeros en incluirse en el vector.

Nótese, que esta implementación tiene una gran ventaja y un gran inconveniente:

- La complejidad computacional del algoritmo es relativamente baja: $\mathcal{O}(K(K-1)t_F)$, siendo t_F el tiempo medio que un filtro requiere para evaluar un subgrupo de 2 elementos.
- Solo se estudian la relación de subgrupos con 1 o 2 características, ignorando las posibles relaciones de subgrupos con más elementos y de subgrupos entre sí.

Aunque es posible el estudio de subgrupos con varias componentes, nos decidimos a limitar su tamaño debido al aumento del coste computacional que conlleva, incluso mayor que el de la selección guiada por modelos de predicción: $\mathcal{O}(\frac{1}{2}K(K-1)(t_M+t_E)$,

Algoritmo 4.3 Selección de características mediante filtros (*SC_F*).

Sean:

$C = \{C_1, ..., C_K\}$ un vector de parametrización con K características iniciales

$M(C_i, C_j) = d_{ij}$ la medida de la distancia entre las características C_i y C_j, siendo M el operador medida del filtro F

El objetivo es reordenar las características de C, de mayor a menor poder discriminatorio según la bondad de cada característica C_i definida como:

$$B(C_i) = \frac{1}{K} \sum_{j=1}^{j=K} M(C_i, C_{j \neq i}) \tag{4.3.1}$$

siendo t_M el tiempo promedio requerido en construir los modelos en cada iteración y t_E el tiempo en evaluarlos.

A continuación definiremos primero los filtros propuestos para realizar una comparativa al final en la Subsubsección 4.3.1.4. La reordenación de las características del vector C se hace a partir de la partición de entrenamiento y los resultados se obtienen en el corpus de evaluación.

4.3.1.1. Filtros basados en estadísticos simples

En este caso usaremos estadísticos sencillos para evaluar cada característica de forma independiente, convirtiéndose la Ecuación 4.3.1 en $B(C_i) = M(C_i)$. Probaremos con distintas medidas de la característica C_i, $M(C_i)$ definidas en la Tabla 4.3.1.

identificador	*operador distancia (M)*	*propiedad a medir*
std	$std(C_i)$	variabilidad de la característica C_i
std.norm	$\frac{std(C_i)}{1+RD^{(*)}(C_i)}$	variabilidad normalizada de C_i
stand	$mean\left(\frac{\lvert C_i - \bar{C_i}\rvert}{1+std(C_i)}\right)$	variabilidad de C_i estandarizada
sampleE	$mean(E(C_i))$	media energética de C_i

$^{(*)}RD \equiv max(C_i) - min(C_i)$ = Intervalo de amplitud (Rango Dinámico) de C_i

Tabla 4.3.1.: *Filtros estadísticos usados en la selección de características.*

La variabilidad de una característica puede ser un buen indicativo de su capacidad discriminatoria (ver la discusión en la Sección A.6). La desviación estándar (*std*) es el estadístico que mejor muestra esa variabilidad. Normalizando la *std* entre el rango dinámico, *RD(*C_i*)*, se persigue poder comparar en unos intervalos equiparables

los valores obtenidos para cada característica, independizándolos de su naturaleza. Valores altos de la media de C_i estandarizada (*stand*, o *unidad tipificada z* de C_i) se relacionan con una mayor dificultad para modelar C_i mediante una distribución gausiana de una componente, por tanto, nos dan otra medida alternativa de la variabilidad. La energía por muestra (*sampleE*) discrimina a las características que tienen más energía frente a las menos energéticas, y puede ser relevante entre propiedades de la misma naturaleza.

4.3.1.2. Filtros basados en correlación

Estos filtros tienen como base distintas medidas de correlación entre 2 características, C_i y C_j. A pesar de que la correlación no parece ser la mejor medida para seleccionar las bases del espacio de características (Guyon, 2008), nos puede servir como primera aproximación. Matemáticamente:

- *C_i y C_j no están correlacionadas si su covarianza es nula*, o equivalentemente, si:

$$E\{C_iC_j\} - E\{C_i\}E\{C_j\} = 0 \tag{4.3.2}$$

La Tabla 4.3.2 presenta los filtros que evaluaremos. Para una interpretación más geométrica denotaremos las características $C_i = \vec{x}$ y $C_j = \vec{y}$, entendiendo que son vectores cuyos elementos son las componentes i, j de los vectores de parametrización que contiene la partición de entrenamiento. Todas las medidas M están en el intervalo $[0, 1]$ para poder aplicarse el Algoritmo 4.3, donde 0 significa máxima correlación y 1 mínima.

identificador	*operador distancia(M)*	*propiedad a medir*
ccc0	$1 - \left\lvert \frac{\vec{x}.\vec{y}}{\lvert\lvert\vec{x}\rvert\rvert_2 \lvert\lvert\vec{y}\rvert\rvert_2} \right\rvert$	dependencia lineal entre $\vec{x}$ e $\vec{y}$
pearson	$1 - \left\lvert \frac{cov(\vec{x},\vec{y})}{\sigma_{\vec{x}}\sigma_{\vec{y}}} \right\rvert$	dependencia lineal entre $\vec{x}$ e $\vec{y}$
srcc	$1 - \left\lvert \frac{cov(\vec{x_R},\vec{y_R})}{\sigma_{\vec{x_R}}\sigma_{\vec{y_R}}} \right\rvert$	dependencia lineal entre $\vec{x}_R$ e $\vec{y_R}$ (*)
krcc	$1 - \left\lvert \frac{PC_{xy} - PD_{xy}}{0{,}5N(N-1)} \right\rvert$	asociación estadística entre $\vec{x}$ e $\vec{y}$ (*2)

(*) $\vec{x_R}, \vec{y_R} \equiv rank(\vec{x}), rank(\vec{y})$ = ordenación creciente de los valores de $\vec{x}$ e $\vec{y}$

(*2) PC_{xy}, PD_{xy} = Pares Concordantes y Pares Discordantes entre las muestras del vector ($\vec{x}$,$\vec{y}$), con N observaciones

Tabla 4.3.2.: *Filtros de correlación propuestos para la selección de características.*

La correlación lineal cruzada de retraso nulo (*ccc0*) nos da el valor del coseno del ángulo que forman los vectores, $\theta(\vec{x}, \vec{y})$, tal que $\cos\theta(\vec{x}, \vec{y}) = 0 \Leftrightarrow \vec{x} \perp \vec{y}$, lo que implica que nuestras características no están linealmente correlacionadas. El coeficiente de correlación de Pearson (*pearson*) es equivalente a *ccc0* para características de media

nula. Numerosos trabajos sobre selección de características optan por estos filtros (Hall, 1999; Haindl et al., 2006; Biesiada and Duch, 2007; Hsu and Lu, 2008).

Uno de los inconvenientes de los filtros basados en correlación cruzada es que no son sensibles a la independencia no-lineal. Para compensar los efectos de no-linealidades y valores espúreos es común pre-procesar los datos mediante algún filtrado o escalado como reducir su margen dinámico. El coeficiente de correlación de Spearman (*srcc*) es una alternativa a este escalado: sustituye los valores de las características $\vec{x}, \vec{y}$ por su rango $\vec{x}_R, \vec{y}_R$ (la posición de cada valor una vez ordenados crecientemente) antes de aplicar el filtro de Pearson (Guyon, 2008).

El coeficiente de correlación de Kendall (*krcc*) o coeficiente tau (τ) es una medida de la asociación estadística entre $\vec{x}$ e $\vec{y}$ basada en la correlación del rango entre 2 variables, sobre las que no se presupone ninguna restricción sobre sus distribuciones. Kendall estudia la relación entre pares concordantes y discordantes de muestras del vector bivariado $(\vec{x}, \vec{y})$; un par $\{(x_i, y_i), (x_j, y_j)\}$ es *concordante* si $(x_i - x_j)$ tiene el mismo signo que $(y_i - y_j)$ (Noether, 1981). Lutu and Engelbrecht (2010) usan el coeficiente τ para hallar correlaciones entre características y entre características y clases como una parte de su algoritmo de selección heurística. Van Hulse et al. (2009) lo utilizan precisamente para comparar la correlación entre diversos filtros.

4.3.1.3. Filtros basados en funciones de probabilidad

Otra manera de estudiar la dependencia estadística entre 2 características C_i y C_j es mediante sus funciones de densidad de probabilidad (*Probability Density Function, PDFs*): marginales ($p(C_i)$, $p(C_j)$), condicionales ($p(C_i|C_j)$, $p(C_j|C_i)$) y conjunta ($p(C_i, C_j)$). Si asociamos las características C_i y C_j a las variables aleatorias X e Y es conocido que si hay independencia estadística entre ellas la probabilidad del suceso conjunto (X, Y) equivale al producto de las probabilidades de los sucesos por independiente (Ecuación 4.3.4).

Medidas basadas en distribución de la probabilidad y en la teoría de la información.

- *Teorema de Bayes*:

$$p(X, Y) = p(X|Y)p(Y) = p(Y|X)p(X) = p(Y, X) \tag{4.3.3}$$

- *El suceso aleatorio X* es *estadísticamente independiente del suceso Y* si y solo si se cumple*:*

$$p(X, Y) = p(X)p(Y) \quad p(X|Y) = p(X) \quad p(Y|X) = p(Y) \tag{4.3.4}$$

- *Entropía* de la variable X (incertidumbre media asociada a X, o valor medio ponderado de la información de X, $I(x_i)$):

$$H(X) = \sum_i p(x_i) I(x_i) = \sum_i p(x_i) \log \frac{1}{p(x_i)} \tag{4.3.5}$$

- *Información mutua media* entre X e Y (con $I(x_i; y_j) = \log \frac{p(x_i)p(y_j)}{p(x_i,y_j)}$ la información mutua del suceso conjunto (x_i, y_j)):

$$MI(X;Y) = \sum_i \sum_j p(x_i, y_j) I(x_i; y_j) = \sum_i \sum_j p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \tag{4.3.6}$$

El teorema de Bayes (Ecuación 4.3.3) junto con las definiciones de independencia (Ecuación 4.3.4), redundancia y relevancia de la página 133 (Sección 4.1.2), sustentan la base teórica de los filtros basados directamente en probabilidades y de otras tantas medidas definidas en el campo de la teoría de la información a partir de las probabilidades. Este tipo de filtros cuentan a su favor con que son capaces de cuantificar la dependencia estadística lineal y no lineal, pero tienen la gran desventaja de que la estimación de las funciones de probabilidad a partir de los datos de entrenamiento no es una tarea inmediata y suele requerir demasiado tiempo computacional, sobre todo en funciones multivariadas (Guyon, 2008).

Nótese que dos variables estadísticamente independientes implican que no están correlacionadas. Lo contrario no es necesariamente cierto.

Estimación de distribuciones de probabilidad en el espacio continuo de características. La extracción en una forma analítica de las distribuciones de probabilidad de las características a partir de los datos sigue siendo una rama abierta a la investigación actualmente. El problema principal recae en la integración en el espacio continuo de características cuando solo tenemos muestras limitadas de datos. La literatura ofrece estas soluciones básicas (Peng et al., 2005):

- *Discretización del espacio de características*. Lo que permite pasar de la integración a la suma, facilitando la estimación. Dentro de este apartado entrarían los estimadores de PDF basados en histogramas (Drugman et al., 2007).
- *Estimación directa en el espacio continuo*. Estimadores paramétricos (modelos, descritos analíticamente una vez definidos los parámetros), semi-paramétricos (combinación de paramétricos, como los GMM) y no paramétricos (Leiva, 2007). Los más populares en teoría de la información son los basados en el estimador no paramétrico de Parzen, también conocidos como los estimadores de densidad

basada en núcleos (estimadores *KDE - Kernel Density Estimators*). Los KDE tienen la ventaja de que solo dependen de los datos, pero los inconvenientes de que hay que seleccionar el núcleo (o ventana) y su anchura para definir el estimador y un alto coste computacional en bases de datos grandes.

En nuestro caso optamos por usar la estimación no paramétrica mediante núcleos gausianos implementada en el software libre *SciPy* (Jones et al., 2001).

El Tabla 4.3.3 esquematiza los filtros usados. La medida *int.pdf* pretende evaluar la similitud entre $p(X)$ y $p(Y)$. Para ello antes se estandarizan las variables X e Y obteniendo x_E e y_E y se estiman sus PDFs, $\hat{p}(x_E)$ y $\hat{p}(y_E)$ mediante la partición de entrenamiento. Posteriormente se integra la curva $g(x_E, y_E) \equiv \hat{p}(x_E)\hat{p}(y_E)$ y se normaliza para obtener un operador distancia acorde con el Algoritmo 4.3. Integramos en el intervalo $[-4\sigma_E, 4\sigma_E]$, con $\sigma_E = 1$ para señales estandarizadas, que comprende la mayor parte de la energía de las variables aleatorias.

La información mutua normalizada o *J.MI* (versión de la distancia de Jaccard según la teoría de la información) es una métrica que computa la cantidad de información común a las variables X e Y.

identificador	*operador distancia(M)*	*propiedad a medir*
int.pdf	$1 - \frac{2I_{XY}}{I_X + I_Y}$	similitud entre $p(X)$ y $p(Y)^{(*)}$
J.MI	$1 - \frac{I(X;Y)}{H(X,Y)}$	información compartida por X e Y

(*) $I_Z = Int[p(z_E)] \equiv \int_{-4\sigma_E}^{+4\sigma_E} \hat{p}(z_E)dz_E$ integral de la PDF de z_E, siendo z_E la versión estandarizada de la variable Z.

$I_{XY} = Int[\hat{p}(x_E)\hat{p}(y_E)]$

Tabla 4.3.3.: *Filtros basados en estimación de la densidad de probabilidad usados en la selección de características.*

4.3.1.4. Comparación de filtros

Para hacer un estudio de los distintos filtros propuestos nos serviremos de la Figura 4.3.1 y la Tabla 4.3.4.

Analizando las curvas de resultados de la Figura 4.3.1 observamos:

- Todos los filtros convergen hacia valores muy parecidos (dentro del intervalo $[80, 85]$ $\%cCorr$) a partir de un determinado tamaño del vector (unas 16 componentes). Esto indica que:

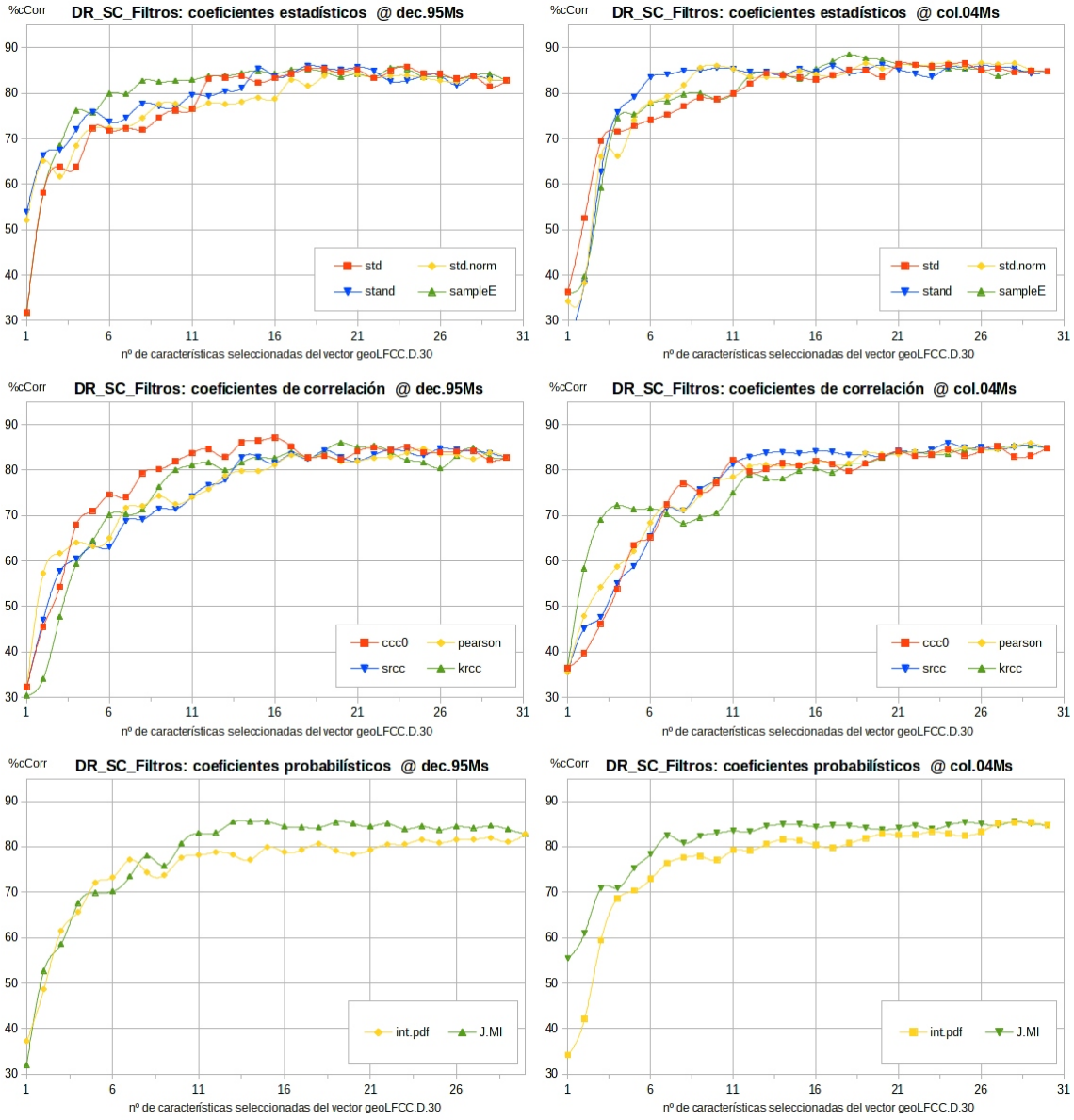


Figura 4.3.1.: Selección de características mediante filtros.

  - Las diferencias entre filtros se encuentran fundamentalmente en la selección de las primeras componentes más significativas.
  - Globalmente, la selección por filtros es robusta para eliminar las componentes menos significativas y depende poco del filtro elegido.

- Los filtros estadísticos parecen tener un mejor comportamiento seleccionando las componentes más discriminativas, o sea, para vectores con pocas características.
- La selección no implica una mejora sustancial en la tasa de reconocimiento

(84,40 %*cCorr* para el vector *geoLFCC.D.30*, con 30 componentes), sin embargo, se dobla la eficacia del sistema: se mantiene unos valores por encima del 80 %*cCorr* para vectores con la mitad del tamaño que el original.

SC_F: estadísticos	std	std.norm	*stand*	E	*media*
dec.95Ms	77.62	77.64	79.40	*80.06*	78.77
col.04Ms	79.27	79.89	*80.21*	79.48	79.97
media	78.44	78.76	*79.80*	79.77	*79.37*

SC_F: correlación	*ccc0*	pearson	srcc	krcc	*media*
dec.95Ms	*78.18*	75.57	75.04	75.47	76.68
col.04Ms	75.28	76.11	76.21	*76.64*	77.05
media	*76.73*	75.84	75.63	76.05	*76.87*

SC_F: información	int.pdf	*J.MI*	*media*
dec.95Ms	75.40	*78.25*	76.83
col.04Ms	76.74	*81.07*	78.91
media	76.07	*79.66*	*77.87*

Tabla 4.3.4.: *Selección de características mediante filtros.* %*cCorr* promedio de cada filtro.

Promediando las curvas anteriores entre el tamaño del vector construimos la Tabla 4.3.4. Su análisis nos muestra que:

- La elección de un tipo de filtros u otro no es determinante: las curvas de resultados son bastante parecidas y sus promedios apenas se diferencian 3 puntos en %*cCorr*. Los filtros de estandarización (*stand*), de energía por muestra (*sampleE*) y de información mutua normalizada (*J.MI*) parecen ser las mejores opciones.

- Los filtros basados en correlación de características obtienen tasas ligeramente peores. Una explicación posible la podemos encontrar en la Figura 4.1.2 y la Figura 4.1.4, donde gráficamente se demuestra que bajo ciertas condiciones, componentes correlacionadas pueden complementarse para particionar mejor el espacio de características (Guyon and Elisseeff, 2003).

La elección del mejor filtro no está nada clara, en parte por la similitud de sus resultados. Aún así, nos quedaremos con el filtro estadístico de la media estandarizada, *stand*; no solo por obtener la mejor tasa de reconocimiento, sino por ser de fácil implementación y de rápida ejecución.

4.3.2. Métodos guiados por Modelos de predicción

En esta parte analizaremos como ejemplo de selección guiada por modelos (*SC_M)* una generalización de la propuesta realizada por Álvarez et al. (2011). Esta generalización nos lleva a una familia de métodos SC_M sobre los que haremos un pequeño estudio para escoger el más eficaz.

4.3.2.1. Algoritmo Discriminante de Selección de Características original

El algoritmo Discriminante de Selección de Características, *DSC* (o *Discriminant Feature Selection, DFS*) desarrollado por Álvarez et al. (2011) es una versión aplicada a señales sismo-volcánicas del algoritmo discriminatorio de selección propuesto por primera vez por De la Torre et al. (1997) que usan como medida un criterio de mínimo error de clasificación.

Álvarez et al. (2011) logran unos resultados sorprendentes aplicando el DFS a eventos del volcán de Colima con un esquema de clasificación de frames usando unos modelos GMM de reconocimiento sencillos. El sistema es capaz de reducir un vector basado en componentes cepstrales desde 39 a 10 elementos con un decremento del 16 % en el error de clasificación y, de reducir de 84 a 14 elementos un vector de características mixtas (cepstrales y geofísicas) con una disminución del 14 % en el error.

El DFS va disminuyendo iterativamente el tamaño de un vector de características quitando la componente menos importante según una función de medida *M=-L* basada en unas funciones discriminantes *d*. Siendo un algoritmo SC_M, usa unos modelos para evaluar la bondad de cada subconjunto *A* de características analizado, construyendo las funciones *d* mediante la PDF de dichos modelos. Usualmente la evaluación de características se hace mediante test cerrado sobre bases de datos maestras con eventos aislados, claramente representativos de las clases y cuyos modelos proporcionen valores fiables de la función de medida. Matemáticamente podemos esquematizarlo en el Algoritmo 4.4.

La función de coste L puede definirse de distintas formas, si bien Álvarez et al. (2011) usan una función logística o sigmoide (Figura 4.3.2) para suavizar la función d_k:

$$L(C_k) = \frac{1}{E}\sum_{\forall f \in DB_{ev}} l_k(f) = \frac{1}{E}\sum_{\forall f \in DB_{ev}} sig(\alpha d_k(f)) = \frac{1}{E}\sum_{\forall f \in DB_{ev}} \frac{1}{1+exp(-\alpha d_k(f))} \tag{4.3.7}$$

siendo α un parámetro de suavizado (*slope* o *pendiente*), entre 0 (línea constante en 0.5) e ∞ (función escalón) y:

$$d_k(f) = -\left[log(G_{c(f)}(f)) - \max_{\lambda_j}\{log(G_{j\neq c(f)}(f))\}\right] \tag{4.3.8}$$

donde:

Algoritmo 4.4 Algoritmo discriminante de selección de características o *DFS* de Álvarez et al. (2011).

Sean:

C un conjunto de K características, tal que $C = \{C_1, ..., C_K\}$.

DB_{tr}, DB_{ev} son las particiones de entrenamiento y evaluación de la base de datos, con T y E frames respectivamente, previamente clasificados en $\{\lambda_1, .., \lambda_H\}$ clases.

$A^{b)}$ un subconjunto de C con $J = J^{b)} \leq K$ características a analizar en el bucle b, con b en $\{1..K-1\}$

$S^{b)}$ una secuencia de C con las $b-1$ características ordenadas de menor a mayor capacidad discriminatoria seleccionadas previamente en los $b-1$ bucles anteriores.

$L(C_k) = \frac{1}{E}\sum_{\forall f \in BD_{ev}} l_k(f)$ la función de coste de la característica C_k, promedio de evaluar cada frame f de BD_{ev} descrito mediante el subconjunto de características $A_k^{b)} = A^{b)} \setminus C_k$.

Al finalizar cada bucle de análisis b, se elimina de A la característica $C^{b)}$ menos discriminatoria para añadirla a S:

$\forall b$ en $1..K-1$:

$$C^{b)} = \arg\min_{C_k \in A^{b)}}\{L(C_k)\}$$

$$S^{b+1)} = S^{b)} \cup C^{b)};\ A^{b+1)} = A^{b)} \setminus C^{b)}$$

$G_j(f)$ es la evaluación en el frame f del modelo de la clase λ_j dado por la PDF de una gausiana multivariada diagonal.

$c(f)$ función de *clasificación*: asocia a cada frame su correspondiente clase; $G_{c(f)}$ es el modelo *correcto* de la clase del frame f, mientras que $G_{j \neq c(f)}$ son los modelos del resto de clases, que representan una clasificación *errónea* de f.

Dado que $G_{c(f)}(f) > G_{j \neq c(f)}(f)$, $d_k(f)$ es $< 0 \Rightarrow l_k(f) \to 0$ si el frame f ha sido correctamente clasificado y $d_k(f) > 0 \Rightarrow l_k(f) \to 1$ en caso contrario. Cuando f es *difícil de clasificar* (hay 2 o más modelos que dan probabilidades muy parecidas al evaluar f) entonces $d_k(f) \approx 0 \Rightarrow l_k(f) \approx 0{,}5$. El valor de α es escogido experimentalmente.

Objetivos del suavizado de las funciones discrimantes. El tipo de función que usamos para suavizar y su grado de suavizado controlado por el parámetro α parámetro debe ser escogido para cumplir un doble objetivo:

1. Minimizar la contribución a la función de coste de los frames cuya clasificación puede ser *poco fiable,* o, al menos, poco discriminatoria ($|d_k(f)| \to 0$).
2. Maximizar la de aquellos que son altamente discriminantes ($|d_k(f)| \gg 1$).

Una correcta selección de la función de suavizado y de α tiende a reducir el error sobre la partición de entrenamiento, incrementando la capacidad de generalización del algoritmo al usar las características seleccionadas para describir otra partición.

4.3.2.2. Generalización del DFS: algoritmos discriminativos guiados por modelos

La generalización del DFS original surge con los objetivos principales de:

- Usar BDs de eventos, no de frames, que permitan una más fácil interpretación de resultados.
- Estudiar distintas funciones M de medida.
- Probar diversos tipos de funciones discriminantes d_k y adecuar su suavizado a su rango de valores de los datos.
- Minimizar la influencia excesiva que las *clases dominantes* (aquellas como laha-res, tremores y colapsos con una duración total acumulada considerablemente mayor respecto a la de otras clases, y por tanto, también con más frames) pueden tener sobre la función de coste L, y consecuentemente, sobre los resultados.
- Posibilitar la selección de características en ambos sentidos: con *direccionalidad decreciente* (reduciendo el tamaño del vector de características) y con *direccionalidad creciente* (aumentando desde cero el conjunto de las mejores características).
- Poder utilizar cualquier clasificador probabilístico, GMMs y HMMs inclusive, para estudiar la influencia de los modelos sobre los resultados de selección.

La consecución de estos objetivos originan toda una familia de algoritmos discriminativos SC_M y requiere la modificación sustancial del Algoritmo 4.4 tal y como describimos a continuación.

Paso de frames a eventos:

DB_{tr}, DB_{ev} son las particiones de entrenamiento y evaluación de la base de datos, con $T = \sum_h T_h$ y $E = \sum_h E_h$ eventos respectivamente, siendo T_h y E_h el número de eventos de entrenamiento y de evaluación de la clase λ_h.

$G_h(e)$ es el valor de la probabilidad media por frame obtenida al evaluar el evento e mediante el modelo de la clase λ_h. Nótese que la forma analítica de $G_h(e)$ depende del número de componentes gausianas y del

tipo de modelo usado en cada caso, GMMs (Subsección 3.3.1), HMMs (Subsección 3.3.2) o cualquier otro modelo estadístico.

$c(e)$ función de *clasificación*, equivalente a $c(f)$ pero asociando cada evento a su clase en vez de cada frame.

Uso de distintas funciones de medida, paso de $L(C_k)$ a $M(A_k)$: Para cada subconjunto de características A_k usado por los modelos para parametrizar los datos definimos la función de medida $M(A_k)$ como una generalización de $L(C_{k)}$ de la siguiente forma:

$$M(A_k) = \frac{1}{H}\sum_h m_k(\lambda_h) = \frac{1}{H}\sum_h \frac{1}{E_h}\sum_{e=1:E_h} m_k(e) \tag{4.3.9}$$

que simplemente es el promedio entre clases del promedio de las funciones de medida m_k evaluadas para los eventos e de una misma clase. Dichas funciones son la generalización de las funciones $l_k(C_k)$ de la Ecuación 4.3.7 siendo $A^{b)} = A^{b-1)} \cup C^{b-1)}$ para una direccionalidad creciente o $A^{b)} = A^{b-1)} \setminus C^{b-1)}$ en caso de direccionalidad decreciente. Las funciones m_k que vamos a usar se pueden dividir en dos categorías, que originan los dos tipos de algoritmos DFS generalizados, *DFS-rsv* y *DFS-recog*:

1. $m_k = dfs_k\{smooth_\alpha(d_k(e))\}$ funciones *dfs* discriminantes generalizadas con distintas funciones *smooth* de suavizado controlado por el parámetro α. A su vez, $d_k(e) = rsv_k(e)$, donde $rsv(e)$ (*r*eliability *s*core *v*alue) cuantifica la fiabilidad en el reconocimiento del evento e a modo de generalización de la Ecuación 4.3.8. Analizaremos distintas formas para $d_k(e)$:

   a) *rsv(e)≡max(e).* $d_k(e)$ es la versión original de Álvarez et al. (2011):

   $$d_k(e) = g_{c(e)}(e) - \text{máx}_{\lambda_h}\{g_{h\neq c(e)}(e)\} \tag{4.3.10}$$

   rsv(e)≡mean(e). Un promedio que tiene en cuenta los valores de las etiquetas no correctas:

   $$d_k(e) = g_{c(e)}(e) - mean_{\lambda_h}\{g_{h\neq c(e)}(e)\} \tag{4.3.11}$$

   rsv(e)≡rel(e). Una versión normalizada por la suma de las probabilidades de los $\{\lambda_1, .., \lambda_H\}$ modelos evaluados:

   $$d_k(e) = \frac{H\left[g_{c(e)}(e) - mean_{\lambda_h}\{g_{h\neq c(e)}(e)\}\right]}{sum_{\lambda_h}\{g_h(e)\}} \tag{4.3.12}$$

   En todas ellas $g_h(e)$ representa la probabilidad logarítmica media del evento, $\log G_h(e)$, normalizada entre $[0, 1]$. Esta normalización es necesaria para que las funciones de suavizado actúen de igual manera independientemente del vector de características usado.

2. $m_k = recog_k(e)$ funciones *recog* que evalúan resultados de reconocimiento definidas en la Subsección 3.4.2, también normalizadas entre $[0, 1]$:

   a) $recog_k(e) \equiv cAcc(e_\lambda) = cCorr(e_\lambda)$. Tasa de eficiencia de reconocimiento promedio para los eventos e de la clase λ.

   b) $recog_k(e) \equiv cPout_\lambda(e_\lambda)$. Probabilidad de que el evento e pertenezca a la clase λ.

   c) $recog_k(e) \equiv cRel_\lambda(e)$. Fiabilidad de que el evento e haya sido correctamente evaluado por la clase λ.

Suavizado de las funciones discriminantes d_k: Aparte de la función sigmoide, pueden definirse otras curvas con la forma idónea para el suavizado hecho en el DFS (Sección 4.3.2.1) con el objetivo de minimizar el efecto de los eventos poco discriminantes frente a los más discriminantes. Esto se consigue enfatizando la contribución de los valores de d_k en el entorno de ± 1 frente a aquellos cercanos a 0. Con este propósito, definimos dos nuevas funciones $smooth_\alpha(d_k)$ de suavizado (Figura 4.3.2):

$$bisigmoide_\alpha(d_k) \equiv \begin{cases} sig_\alpha(d_k - \alpha_N) - sig_\alpha(-\alpha_N) & d_k \geqslant 0 \\ -sig_\alpha(d_k - \alpha_N) + sig_\alpha(-\alpha_N) & d_k < 0 \end{cases} \tag{4.3.13}$$

con $\alpha_N = 2/3$ y $sig_\alpha(x) = sigmoide(\alpha x) = (1 + e^{-\alpha x})^{-1}$

$$norm\_sinhh_\alpha(d_k) \equiv \frac{e^{\alpha d_k} - e^{-\alpha d_k}}{e^\alpha - e^{-\alpha}} = \frac{e^{\alpha d_k} - e^{-\alpha d_k}}{e^\alpha - e^{-\alpha}} \tag{4.3.14}$$

Con el propósito de estandarizar la salida del suavizado en el intervalo $[-1, 1]$ de la $sig_\alpha(x)$, usaremos a partir de ahora su versión escalada y desplazada:

$$sigmoide_\alpha(d_k) \equiv 2(sig_\alpha(d_k) - 0{,}5) = \frac{1 - e^{-\alpha d_k}}{1 + e^{-\alpha d_k}} \tag{4.3.15}$$

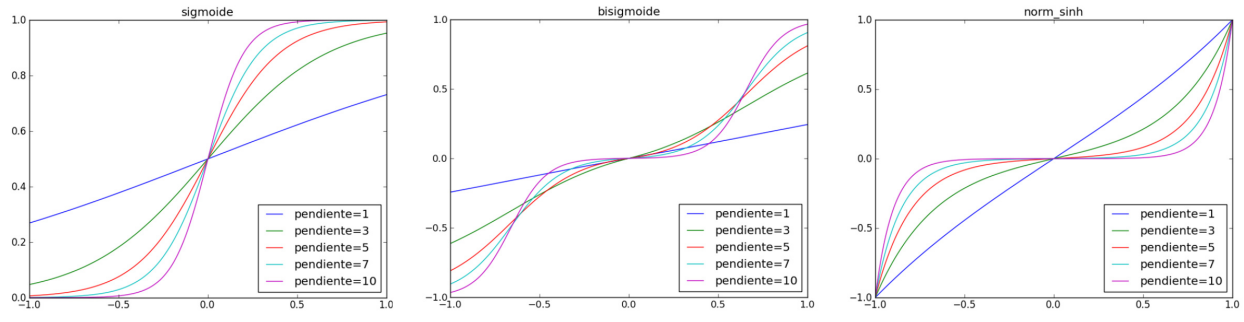


Figura 4.3.2.: Funciones de suavizado o *smooth* de las funciones discriminantes para distintos valores de la pendiente α.

Configuración del algoritmo *DFS-rsv*: Como paso previo al uso de la selección mediante el *DFS-rsv* necesitamos encontrar los mejores valores de la tupla $(rsv, smooth, \alpha)$ definida en apartados anteriores. El proceso completo de configuración requiere de varias pruebas detalladas en la Sección A.3. En ella utilizamos un vector mixto de 26 componentes con características geofísicas, estadísticas y coeficientes LFCC dado por Cortés et al. (2014) sobre las bases de datos *dec.95Ms* y *col.04Ms*. Finalmente, la mejor configuración se obtiene para los valores $(rsv = max, smooth = sigmoide, \alpha = 5)$.

4.3.2.3. Comparación de algoritmos DFS

Para finalizar el apartado dedicado a la selección de características guiada por modelos vamos a comparar las distintas familias del DFS generalizado para los esquemas *DFS-rsv* y *DFS-recog* con las funciones de medida $m_k = dfs_k\{sigmoide_{\alpha=5}(max)\}(e)$ y $m_k = recog_k\{cAcc, cPout, cRel\}(e)$ respectivamente. Distinguiremos además entre una direccionalidad creciente $(+)$ y decreciente $(-)$ a la hora de ordenar las componentes del vector respecto su capacidad de discriminación.

La Figura 4.3.3 presenta las gráficas de selección del DFS generalizado. Podemos extraer las siguientes conclusiones:

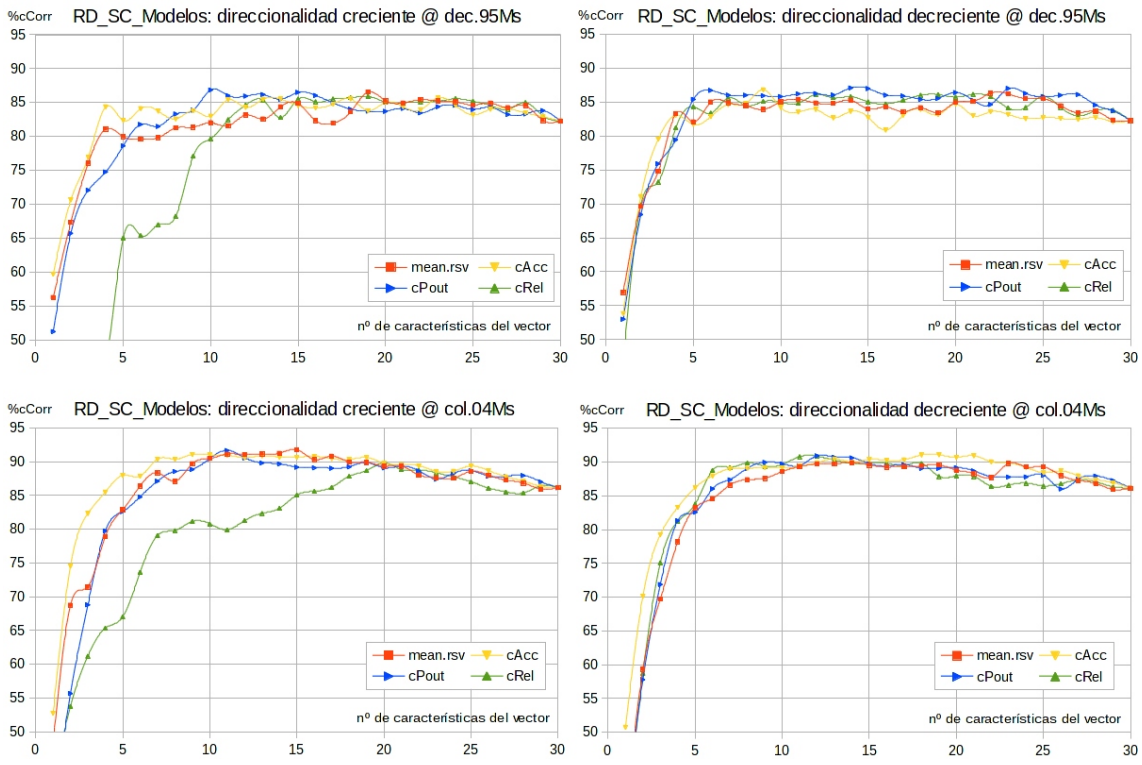


Figura 4.3.3.: Selección de características guiada por modelos: algoritmos DFS generalizados.

- La selección guiada por modelos parece ser una solución muy eficaz para la selección de características: para un tercio del tamaño del vector original (10 componentes) se mantienen valores estables hasta 5 puntos por encima de los que se consiguen con todo el vector.
- El comportamiento para una misma implementación no varía significativamente entre las distintas bases de datos.
- Excepto para *DFS-recog=cRel*, la tendencia de los esquemas es bastante similar en un mismo corpus e, incluso, entre ellos.
- La direccionalidad decreciente parece ser ligeramente más eficaz que la creciente para un tamaño pequeño del vector ($[5, 10]$ componentes).

Para escoger entre las diferentes opciones es necesario analizar los valores promedio de la Tabla 4.3.5. Comprobamos como efectivamente los resultados son relativamente parecidos, destacándose levemente la solución basada en el *cAcc* en ambas bases de datos. De igual manera, constatamos que la direccionalidad decreciente funciona algo mejor que la creciente, sin embargo, la curva más eficiente es para el esquema creciente que usa *cAcc* como función discriminante.

SC_M: +	dfs-rsv	*recog-cAcc*	recog-cPout	recog-cRel	*media*
dec.95Ms	81.49	*82.61*	81.66	76.04	80.45
col.04Ms	85.78	*87.34*	84.66	80.12	84.48
media	83.64	*84.98*	83.16	78.08	*82.46*

SC_M: -	dfs-rsv	*recog-cAcc*	recog-cPout	recog-cRel	*media*
dec.95Ms	82.66	81.73	*83.55*	82.58	82.63
col.04Ms	84.50	*86.75*	84.85	84.88	85.25
media	83.58	*84.24*	84.20	83.73	*83.94*

Tabla 4.3.5.: *Selección de características guiada por modelos de direccionalidad creciente (+) y decreciente (-).* Valores promedio de *%cCorr*.

A pesar de la similitud entre resultados, llama la atención el comportamiento del esquema *DFS-recog=cRel* frente al *DFS-recog=cPout*. Al parecer, las operaciones que se hacen para transformar la probabilidad de salida dada por los modelos guiados por probabilidades (tasa *cPout*) en una tasa de fiabilidad que cada modelo tiene al evaluar un evento (*cRel*) destruyen información útil de cara a la selección. También es posible que el indicador *cRel* no sea el más adecuado para medir esta fiabilidad.

A priori, igualmente destaca que pese a la idea del *DFS-rsv* de ponderación según el grado de discriminación de unos modelos frente a otros, las medidas basadas en el *cCorr* funcionen algo mejor. Quizás la explicación se sustente sobre un compendio de varias razones:

- La medida *cCorr* está dada por el clasificador (Subsección 3.3.1) el cual se basa en probabilidades, como *cPout,* pero que tiene en cuenta además otros factores como la duración mínima y máxima que pueden tener los eventos.
- La ponderación hecha por el esquema *DFS-rsv* parece que es demasiado débil como para diferenciarse del esquema *DFS-recog=cPout*, sobre el que se basa.

4.4. Reducción de dimensionalidad mediante transformación de características

En este apartado analizaremos técnicas clásicas de reducción de dimensionalidad transformando el espacio original de las características en otro donde la información esté estructurada de tal forma que se necesiten menos variables del en el espacio transformado para describirla con el mismo nivel de detalle.

Muchos de los métodos presentados son transformadas clásicas de extracción de características que ya hemos analizado la Subsección 4.2.2 para parametrizar los datos, mientras otras son técnicas específicamente diseñadas para reducir la dimensionalidad.

La metodología experimental es la misma que la explicada en la Sección 4.3. Distinguiremos entre algoritmos dependientes, que necesitan analizar unos datos de entrenamiento para definir la transformación y aquellos cuya transformación no necesita de ningún análisis previo. El software usado para evaluar estos métodos se engloba dentro de las librerías científicas de Python: *Modular toolkit for Data Processing* (*MDP*), descrito en Zito et al. (2008).

4.4.1. Transformaciones no dependientes de datos

4.4.1.1. Transformada Discreta del Coseno (*DCT)*

La Transformada Discreta del Coseno (*Discrete Cosine Transform*, DCT) es una transformación lineal ortogonal con una gran capacidad de compactación de la energía convirtiéndose en la base de muchos estándares en algoritmos de la compresión de datos (Pennebaker, 1992; Reznik et al., 2007). La DCT usa funciones coseno como bases ortonormales del espacio transformado para convertir una señal discreta $\{x_n\}$ de N muestras en otra $\{X_k\}$ de la siguiente forma:

$$X_k = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right] \qquad k = 0, \ldots, N-1. \tag{4.4.1}$$

La DCT tiene varias cualidades que la hacen especialmente atractiva:

- Tiende a la transformada óptima para la preservación de la energía, la PCA (Subsubsección 4.4.2.1), pero con menos coste computacional (Feig and Winograd, 1992).
- Sus bases no dependen de los datos.

Sin embargo, comparte con la PCA algunas debilidades: solo puede decorrelacionar linealmente, es sensible a la magnitud relativa entre variables y está orientada a describir las direcciones de máxima varianza en el nuevo espacio.

En extracción de características geofísicas es típico usarla como decorrelador de coeficientes espectrales (Benítez et al., 2007; Ibáñez et al., 2009; Hoogenboezem, 2010; Beyreuther and Wassermann, 2011; Avesani et al., 2012).

4.4.1.2. Predicción Lineal (*LPC*)

El modelo de Predicción Lineal (*Linear Predictive Coding*, LPC) es una técnica paramétrica que pretende estimar el valor de una muestra x_m en una secuencia discreta $\{x_n\}$ a partir de sus p muestras anteriores (Makhoul, 1975).

$$\widehat{x}(m) = \sum_{k=1}^{p} a_i x(m-k) \qquad ; \qquad \{z_n\} \equiv \{x_n - \hat{x}_n\} \tag{4.4.2}$$

En la Ecuación 4.4.2 p es el orden del modelo y a_i son los coeficientes de predicción, que pueden ser hallados de diferentes maneras. El método más común consiste en minimizar el valor esperado del cuadrado del error de predicción $\{z_n\}$ mediante la matriz de autocorrelación de la señal, conocido como el análisis autoregresivo de Yule-Walker o método de la autocorrelación. Los coeficientes de predicción permiten además estimar la envolvente espectral de la señal $\{x_n\}$ mediante la Ecuación 4.4.3:

$$P\hat{S}D[x_n](w) = \frac{\sigma_{z_n}^2}{|1 + \sum_{k=1}^{p} a_k e^{-iwk}|^2} \tag{4.4.3}$$

El análisis de predicción lineal es usado en múltiples aplicaciones, destacando en el área de codificación y síntesis de voz (Rabiner and Juang, 1993). En geofísica, Del Pezzo et al. (2003); Masiello et al. (2006); Hoogenboezem (2010) parametrizan directamente la forma de onda de los eventos sismo-volcánicos mediante coeficientes LPC.

4.4.2. Transformaciones dependientes de los datos

4.4.2.1. Descomposición en Componentes Principales (*PCA*)

La descomposición en componentes principales o PCA (*Principal Component Analysis*, Pearson, 1901) es el método más popular de reducción de dimensionalidad. Realiza una transformación lineal del espacio de características original con el objetivo

de que las las bases del espacio transformado, las llamadas *componentes principales*, se correspondan con las direcciones ortogonales de máxima varianza en el espacio original.

La rotación del espacio dada por la PCA persigue preservar la máxima energía posible de los datos sobre las mínimas posibles bases ortogonales (Jolliffe, 2002). Para reducir la dimensionalidad se escoge un subconjunto de las primeras componentes principales en función de la energía que se quiera conservar tras proyectar los datos sobre dicho subconjunto.

La descomposición en componentes principales de un corpus de datos representado por la matriz X, con N observaciones de K variables aleatorias (en nuestro caso N vectores de K características cada uno), es equivalente a realizar la transformada discreta de Karhunen-Loève (KLT) de $X_0 = X - \boldsymbol{\mu_X}$, siendo $\boldsymbol{\mu_X}$ la media de cada característica en X (Gerbrands, 1981). Este centrado en la media es necesario para que la primera componente principal marque la dirección de máxima varianza.

$$Y = KLT\{X_0 = X - \boldsymbol{\mu_X}\} = W_{PCA}^T X_0 \tag{4.4.4}$$

W_{PCA} es la matriz de proyección cuyas columnas son los vectores base del espacio de características transformado. La matriz W_{PCA} puede hallarse directamente a partir de la descomposición en valores singulares de X (la forma más eficiente computacionalmente) o factorizando una matriz D de dispersión o *scattering* en sus autovectores y autovalores de la forma:

$$DW = \Lambda W \tag{4.4.5}$$

siendo Λ una matriz diagonal con $\{\Lambda_k\}_{1:K}$ valores propios y W la matriz de proyección, cuyas columnas se corresponden a los vectores propios estructurados en orden decreciente según los valores de sus correspondientes autovalores. En el caso de la PCA, la matriz de dispersión se define como la covarianza del espacio original de las variables, Σ_X, o una estimación suya a partir de los datos experimentales, $\hat{\Sigma_X}$.

$$D_{PCA} = \hat{\Sigma_X} = \frac{1}{N-1}(X - \boldsymbol{\mu_X})^T(X - \boldsymbol{\mu_X}) \tag{4.4.6}$$

Sin embargo, no está garantizado que las direcciones de mayor varianza definan la mejor partición del espacio para la separación de las clases, en parte debido a las limitaciones que la PCA presenta:

- *Suposición de linealidad y ortogonalidad.* Solo puede decorrelar características que han sido generadas mediante combinación lineal de vectores ortogonales entre sí. A veces, la estructura de los datos puede ser representada más naturalmente usando ejes no necesariamente ortogonales (Bartlett, 2001).
- *Suposición de distribución gausiana de las variables aleatorias.* PCA solo garantiza que los vectores base representan las direcciones de máxima varianza suponiendo que las características originales se adaptan a una distribución gausiana (Draper et al., 2003).

- *Es sensible a la magnitud relativa de las variables cuando se usa el método de la covarianza*, por lo que es una práctica común estandarizar cada característica antes de la descomposición, o, equivalentemente, definir la matriz de dispersión en la Ecuación 4.4.6 como la matriz de correlación de los datos.

PCA ha sido usado en numerosos trabajos para describir eventos sísmicos y seleccionar características: Avossa et al. (2003) extraen características de 2 clases eruptivas en el Strómboli para clasificar. Masiello et al. (2006) comparan la representación en 2 dimensiones dada por PCA, análisis de componentes curvilíneos y mapas auto-organizativos del espacio original de características de coeficientes LPC y de amplitud. Hoogenboezem (2010) usa PCA para reducir la dimensionalidad de parametrizaciones espectrales.

4.4.2.2. Análisis de Componentes Independientes (*ICA*)

El análisis de componentes independientes (*Independent Component Analysis*, ICA) es una técnica usada en la separación ciega de señales que pretende obtener las K fuentes independientes que mezcladas aditivamente generan N señales dependientes (Comon, 1994a). El típico ejemplo de separación ciega es el de una reunión donde hay hablando K personas al mismo tiempo grabadas por N micrófonos y se quiere poder reconocer cada conversación independientemente. Matemáticamente puede ser formulado mediante:

$$Y = ICA\{X\} = W_{ICA}X \tag{4.4.7}$$

donde las columnas de Y, $(\boldsymbol{y}_1, .., \boldsymbol{y}_K)$, representan las componentes estadísticamente independientes, también llamadas variables latentes, factores o señales fuente. Las columnas de X son las variables observables. W_{ICA} es la matriz de transformación de dimensión (K, N), con $K \leqq N$ a estimar iterativamente maximizando la independencia estadística de las señales fuente entre sí, medida mediante alguna función objetivo $J(W)$. Existen varias opciones para definir $J(W)$, agrupándose fundamentalmente en tres tipos de criterios (Hyvarinen et al., 2001):

1. *Minimización de la información mutua entre las fuentes.*
2. *Medidas basadas en la no-gausianidad de las fuentes.* Las señales con distribuciones normales no son adecuadas para estimar la matriz W, pues, según el Teorema del Límite Central, el promedio de variables aleatorias independientes entre sí tiende a una distribución de Gauss. La no-gausianidad se suele medir mediante la curtosis o la entropía negativa.
3. *Modelando las fuentes independientes mediante la estimación de su función de probabilidad.* El método Infomax (Cardoso, 1997) hace un modelado equivalente, basado en la maximización de la información mutua entre las variables de entrada y salida de una red neuronal.

Teóricamente, ICA requiere cumplir unas suposiciones previas:

- *Es una mezcla lineal de señales fuente.* Tal y como se observa en la Ecuación 4.4.7. Sin embargo existen modificaciones del algoritmo básico que son no lineales y otras que incluyen fuentes de ruido gausianas.
- *Las componentes independientes (excepto una a lo sumo) han de ser no-gausianas.*
- *No puede haber menos variables observables que fuentes independientes.* Es una condición para poder separar las fuentes. En el caso de usar ICA para reducir la dimensionalidad siempre se cumple.

En los últimos años ICA se ha convertido en una técnica ampliamente usada en sismología para analizar señales complejas: Acernese et al. (2003) usan 6 componentes independientes para separar la señal fuente de los efectos de difusión en explosiones del volcán Strómboli; ICA, MSD-ICA (*Multiresolution Subband Decomposition* ICA, propuesto por Cichocki and Georgiev, 2003) y otros algoritmos basados en estadísticos de orden superior (mayor que 2) han sido aplicados por Cabras et al. (2008, 2010) para separar con éxito de una forma no supervisada eventos volcánicos del ruido oceánico en los volcanes Merapi y Etna, inseparables por métodos de 2º orden; De Lauro et al. (2009) descomponen en fuentes independientes sismogramas para caracterizar explosiones en el volcán Erebus, mejorando los resultados dados por análisis clásicos espectrales y de polarización.

La aplicación práctica de ICA conlleva algunas limitaciones (Hyvärinen and Oja, 2000; Tuncer et al., 2008):

- *Es común una etapa de pre-procesamiento en los datos.* Para maximizar la eficacia del algoritmo recursivo y asegurar que todas las variables son tratadas con la misma relevancia, las observaciones de los datos deben tener media nula y no estar correlacionadas entre sí, preferiblemente con una matriz de covarianza identidad.
- *No existe un único método de implementar ICA,* por lo que existe una gran variedad de algoritmos (lnfomax, FastICA, JADE,...) cuyos resultados pueden variar significativamente entre sí.
- *No se establece un orden de importancia de las fuentes independientes.* El peso relativo de cada fuente en el modelo generativo de las variables observadas no se determina. Tampoco su varianza.

Debido a la abundancia de algoritmos disponibles en el análisis de componentes independientes, vamos a usar dos de ellos para reducir dimensionalidad:

FastICA. Propuesto por Hyvarinen (1999), FastICA usa una estimación de la entropía negativa para medir la no-gausianidad inspirado en redes neuronales. Toma como entrada datos centrados de covarianza unitaria iterando de forma eficiente hasta que se proyecta en la misma dirección.

CubICA. Ideado por Blaschke and Wiskott (2004), mejora la técnica original de Comon (1994b) usando los momentos de tensores de orden 3 y 4 conjuntamente para estimar las componentes independientes. Permite obtener

distribuciones de probabilidad simétricas y no simétricas de las señales fuente.

4.4.2.3. Análisis Factorial (*FA*)

El análisis de factores (*Factorial Analysis,* FA) es una técnica del análisis exploratorio de datos que data de finales del siglo XIX propuesta por Charles Spearman para la medida de la inteligencia. FA trata de describir la relación subyacente que existe entre las variables observadas de una señal multidimensional X mediante una combinación lineal de variables ocultas o latentes, llamadas *factores*. Formalmente, los datos de $X = (\mathbf{x}_1, \ldots, \mathbf{x}_M)^T$, representados por su matriz de media nula X_0, son generados mediante (Harman, 1967):

$$X - \boldsymbol{\mu} = X_0 = LS + E \tag{4.4.8}$$

donde $\boldsymbol{\mu} = (\mu_m)_{m=1:M}$ es la media de la variable $m - ésima$ de los datos, L es la matriz base o de mezcla y S es el vector que contiene a las K variables latentes que modelan las correlaciones entre las variables observadas. E es una matriz de error que describe el ruido independiente asociado a X. Definida de esta forma, Shalizi (2009) esquematiza la descomposición en factores como un caso particular de los modelos gráficos generativos (Figura 4.4.1).

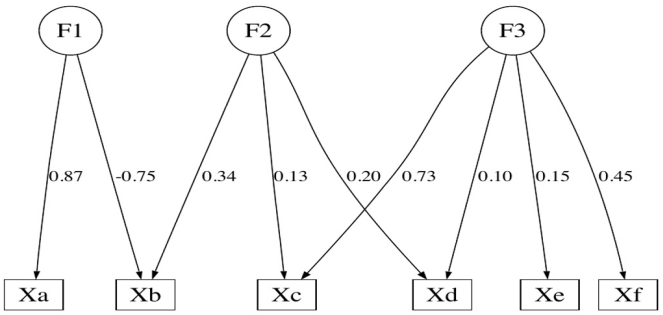


Figura 4.4.1.: Equivalencia entre descomposición factorial y modelos gráficos generativos. Las variables observadas X_m son descritas por unas variables ocultas F_k mediante una combinación lineal definida en la matriz de mezcla L. Figura de Shalizi (2009).

El modelo de la Ecuación 4.4.8 asume que:

- S y E son independientes
- El valor esperado de los factores es nulo: $E[S] = 0$
- Los factores no están correlacionados entre sí: $\Sigma_F = I$

Bajo estas hipótesis, la resolución del problema queda indeterminada, incluso suponiendo que el número de variables ocultas es menor o igual que el de observadas

($K \leq M$). En la literatura existen varios procedimientos para aproximar los factores basados en descomposición en autovalores o en estimaciones de máxima verosimilitud, siendo el más popular el de mínimos cuadrados. Ghahramani et al. (1996) resuelven mediante el algoritmo EM (*Expectation- Maximization*) suponiendo que los factores se adaptan a una distribución gausiana.

FA ha sido ampliamente aplicada como técnica de reducción de dimensionalidad (Filzmoser et al., 2009). En Wagner and Owens (1996) encontramos una comparativa entre PCA y FA en el ámbito de las señales sísmicas.

Las principales ventajas del FA son:

- Tiene en cuenta el error aleatorio inherente a la medición, por lo que es menos sensible al ruido en los datos (Chen and Wang, 2006).
- Es un modelo sencillo y poco restrictivo respecto a los datos (Cabras, 2011).
- Puede ser usado como modelo de predicción (Shalizi, 2009).

Sin embargo, presenta también algunas limitaciones:

- Su solución no es única y depende del algoritmo de optimización. Su cálculo es costoso computacionalmente y necesita de hipótesis adicionales para la estimación de los factores.
- Un número incorrecto de factores dificulta la interpretación y análisis de los datos. Aunque se han desarrollado algunos procedimientos para estimar el número óptimo de factores, para ello se requiere un conocimiento previo de las señales (Filzmoser et al., 2009).

4.4.2.4. Análisis Discriminante Lineal (*LDA / FDA*)

En contraposición con otras técnicas como PCA y DCT que pretenden preservar la máxima varianza de los datos con el mínimo número de dimensiones, el análisis discriminante lineal (*Fisher Discriminant Analysis - FDA*, o su derivado *Linear Discriminant Analysis - LDA*) es una transformación del espacio de características focalizada en maximizar la separabilidad entre datos de distintas clases (Figura 4.4.2). FDA proyecta linealmente las muestras de $X = (\mathbf{x}_1, .., \mathbf{x}_N)^T$ mediante la matriz W_{FDA} a un subespacio donde es más fácil discriminar entre clases:

$$Y = FDA\{X\} = W_{FDA}^T X \tag{4.4.9}$$

El criterio usado por Fisher (1936) para hallar W_{FDA} es maximizar la separación entre datos de diferentes clases, representada por la matriz S_B de dispersión entre clases, al mismo tiempo que minimiza la separación entre los datos de una misma clase, descrita por la matriz de covarianza propia de clases S_W. Definimos:

$$\begin{aligned} S_B &= \sum_{\lambda_i \in \lambda} N_i (\boldsymbol{\mu}_{\boldsymbol{\lambda_i}} - \boldsymbol{\mu})(\boldsymbol{\mu}_{\boldsymbol{\lambda_i}} - \boldsymbol{\mu})^T \\ S_W &= \sum_{\lambda_i \in \lambda} \left(\sum_{\mathbf{x_i} \in \lambda_i} (\mathbf{x_i} - \boldsymbol{\mu}_{\boldsymbol{\lambda_i}})(\mathbf{x_i} - \boldsymbol{\mu}_{\boldsymbol{\lambda_i}})^T \right) = \sum_{\lambda_i \in \lambda} (\Sigma_{\lambda_i}) \end{aligned} \tag{4.4.10}$$

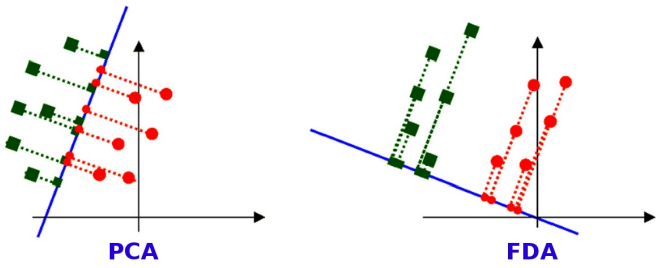


Figura 4.4.2.: Separabilidad de PCA vs. FDA. Proyección unidimensional para separar las clases R (**roja**) y V (**verde**) en un espacio de características bidimensional. La 1ª componente de PCA proyecta en la dirección de máxima varianza de los datos. FDA busca la proyección que mejor discrimine entre clases. Figura original en Veksler (2004).

donde $\lambda = \{\lambda_1, .., \lambda_K\}$ es el conjunto de clases, N_i es el número de muestras $\mathbf{x_i}$ de entrenamiento de la clase λ_i, $\boldsymbol{\mu_{\lambda_i}}$ el vector promedio de cada característica para λ_i, Σ_{λ_i} su matriz de covarianza y $\boldsymbol{\mu}$ el vector promedio de cada característica, definida en cada columna de la matriz X de datos. La idea básica de Fisher (Figura 4.4.2) puede implementarse maximizando una función objetivo $J(W)$ en $W = W_{FDA}$. Fukunaga (1990) propone varias funciones objetivo, siendo la más clásica la dada en la Ecuación 4.4.11:

$$J(W) \equiv traza\left\{\frac{s_B}{s_W}\right\} = traza\left\{\frac{W^T S_B W}{W^T S_W W}\right\} = \frac{det\left\{W^T S_B W\right\}}{det\left\{W^T S_W W\right\}} \tag{4.4.11}$$

donde s_B y s_W se definen en el espacio transformado de forma homóloga a S_B y S_W. Existen dos estrategias básicas para maximizar $J(W)$:

1. **LDA** supone que las características de $X = (\mathbf{x}_1, .., \mathbf{x}_M)$ están normalmente distribuidas (*condición de gausianidad*) y que las covarianzas de las clases son iguales entre sí (*condición homoscedástica*).
2. **FDA** usa el cálculo diferencial para maximizar $J(W)$, desembocando en un problema de autovalores generalizado de la matriz $S_W^{-1} S_B$:

   $$S_B W = \Lambda S_W W \tag{4.4.12}$$

   resoluble si S_W es invertible, proporcionando como máximo $K-1$ autovectores independientes en $W = W_{FDA}$.

El análisis discriminante como clasificador. Sobre la Figura 4.4.2 se puede dibujar un hiperplano (en este caso una línea) que separe a una clase V (verde) de otra R (roja) descritas en el espacio de 2 características. Nótese que es muy intuitivo usar LDA / FDA como clasificador definiendo un umbral en el espacio proyectado. De hecho, es común usar FDA / LDA para separar entre clases.

La aplicación experimental de FDA presenta algunos inconvenientes:

- *Es una proyección lineal, con una dimensionalidad limitada a $(K-1)$ hiperplanos.* El rango de S_B es a lo sumo $K-1$, lo que define el máximo de autovalores distintos de cero y con ello, el de la dimensión. Esto implica que el número de proyecciones distintas a la salida viene definido por la dirección de los autovectores que se corresponden con los $K-1$ autovalores más altos. El resto de autovectores estará contenido en esos hiperplanos. En el caso de tener datos complejos, este límite puede resultar insuficiente.
- *En algunos casos no se puede maximizar eficazmente la función de coste* (Veksler, 2004). Puede ocurrir si $J(W) \simeq 0$, o, equivalentemente cuando los vectores media de las clases son muy parecidos entre sí: $\mu_{\lambda_i} \simeq \mu_{\lambda_j}\ \forall i,j$. También se da en casos en los que las clases están muy solapadas entre sí en el espacio de características: $J(W) \gg 1$.
- *El resultado es sensible a la escasez de datos de entrenamiento* (Leiva, 2007), lo que puede llevar a que S_W no sea de rango máximo, por lo tanto no invertible, y quede sin solucionar directamente la Ecuación 4.4.12. Es corregible regularizando la matriz S_W (Li et al., 2006).
- *Es un algoritmo supervisado y dependiente de los datos*, que requiere una clasificación previa del corpus de entrenamiento.

4.4.2.5. Análisis de Características Lentas (*SFA* / *ISFA*)

Inspirada en el campo de la neurociencia computacional, el análisis de características lentas (*Slow Feature Analysis* - SFA, propuesto por Wiskott, 1999) es una técnica de aprendizaje no supervisada enfocada a replicar el modelado que el cerebro humano hace ante múltiples entradas sensoriales cuya información varía rápidamente. Sirva como ejemplo el modelado visual que se hace de una persona en una escena en movimiento; en general, el cerebro reconoce a objetos asociándoles propiedades estadísticas que no cambian o que cambian lentamente a partir de la descripción de la escena que obtiene mediante la vista y el oído (Figura 4.4.3).

Una formulación más formal del problema equivale a diseñar un algoritmo no lineal que sea capaz de aprender de manera no supervisada características invariantes en un intervalo temporal τ (Sprekeler and Wiskott, 2008):

$$Y = SFA\{X\} = G_{SFA}(X) \tag{4.4.13}$$

Siendo Y la matriz que define las variables latentes $(\boldsymbol{y}_1, .., \boldsymbol{y}_K)$ y X nuestra matriz de datos con M variables de características $(\boldsymbol{x}_1, .., \boldsymbol{x}_M)$. $G_{SFA}(X)$ es un vector de K funciones sobre X tal que: $\mathbf{y}_k = \boldsymbol{g}_k(X)$. La propiedad de *lentitud* viene impuesta al minimizar los operadores J_k definidos como:

$$J_k \equiv \langle y_k'^2 \rangle_\tau \tag{4.4.14}$$

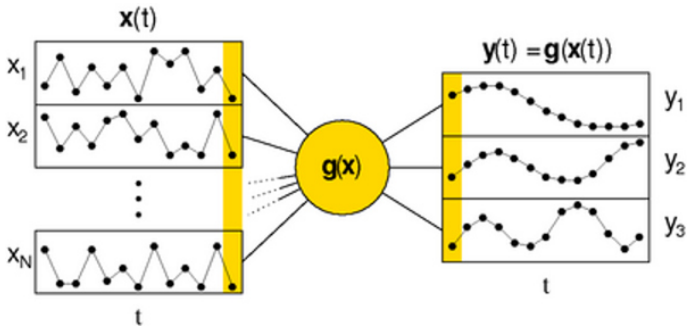


Figura 4.4.3.: Análisis de componentes lentas (SFA). La idea es obtener a partir de $\{x_i\}_{1:N}$ entradas que oscilan rápidamente las salidas $\{y_j\}_{1:3}$ latentes independientes entre sí que contienen una información más estable según unos criterios estadísticos. Figura de Wiskott et al. (2011).

donde y'_k es la derivada temporal de y_k y $\langle u \rangle_\tau$ es el promedio temporal de u en el intervalo τ. Tras filtrarse mediante $\langle \cdot \rangle_\tau$, las salidas deben cumplir las condiciones de media nula, varianza unidad y deben estar decorrelacionadas entre sí.

Una vez planteado, el proceso de optimización puede resolverse por cálculo variacional. Definiendo las funciones $g_k(X)$ como combinaciones lineales de funciones no lineales, Wiskott and Sejnowski (2002) presentan un método más sencillo gracias a un algoritmo iterativo que aproxima la solución mediante una descomposición generalizada de autovectores.

SFA disfruta de unas ventajas considerables sobre otros métodos:

- No se hacen suposiciones a priori sobre la distribución de probabilidad de las características.
- Las salidas están decorreladas y ordenadas por su grado de invarianza.
- Es un método no supervisado capaz de extraer relaciones no lineales entre variables.

La generalidad del método le ha otorgado una creciente popularidad en los últimos diez años, aplicándose en distintas áreas como reconocimiento de objetos (Franzius et al., 2008), segmentación (Kuhnl et al., 2011), reconocimiento de acciones Zhang and Tao (2012), reducción de dimensionalidad y reconocimiento de patrones (Escalante et al., 2012).

Sin embargo, SFA presenta ciertas limitaciones (Wiskott and Sejnowski, 2002):

- Suele consumir más recursos que otras técnicas, aconsejándose su uso en datos con una dimensionalidad moderada. Este último inconveniente puede compensarse implementándose jerárquicamente en redes neuronales.
- La extracción de las funciones no lineales puede complicarse en presencia de ruido.

- No siempre un análisis de componentes lentas puede ser adecuado, en concreto cuando se describen eventos que tengan una rápida variabilidad.

Con el objetivo de paliar estos inconvenientes y de explorar el concepto subyacente de lentitud, SFA ha evolucionado recientemente a otras técnicas como SFA incremental de Kompella et al. (2011) y el análisis independiente de características lentas (*Independent Slow Feature Analysis, ISFA*) de Tobias Blaschke (2007). Dado que la independencia estadística es insuficiente en el caso de separación ciega de fuentes no lineales, Tobias Blaschke mezcla el análisis de componentes independientes (ICA) con el análisis de variables lentas para separar distintas señales de música.

4.4.3. Comparación de métodos basados en transformaciones del espacio de características

La Figura 4.4.4 nos muestra las curvas de tasa de reconocimiento para cada técnica estudiada en la sección anterior para distintos tamaños del vector de parametrización tras la transformación del espacio de características. Debido al enorme tiempo que requerían los experimentos, se ha ido aumentando el tamaño del vector de 3 en 3 componentes. En ambas bases de datos, los resultados, a excepción de DCT y LPC, siguen un patrón similar: la eficacia de reconocimiento crece respecto el tamaño del vector, llega a un máximo y se mantiene más o menos estable en una zona donde todas las técnicas ofrecen valores muy parecidos. En *col.04Ms* el máximo se alcanza después que en *dec.95Ms*, debido probablemente a que contiene señales más complejas.

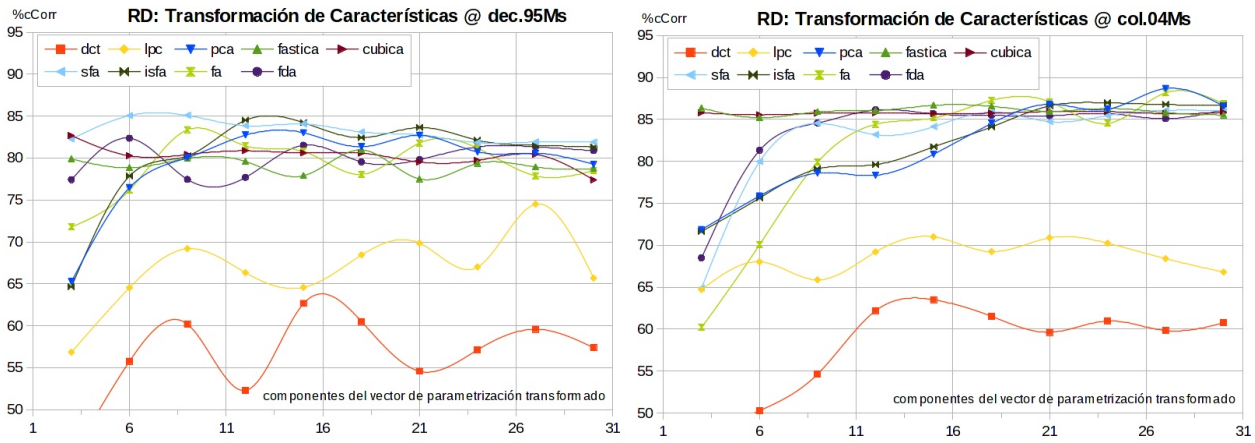


Figura 4.4.4.: Reducción de dimensionalidad mediante distintas transformaciones del espacio de características.

Los resultados promedio para cada técnica se presentan en el Tabla 4.4.1. Se constata la similitud entre los métodos, a excepción de DCT y LPC.

	DCT	LPC	PCA	fastICA	*cubICA*	SFA	ISFA	FA	FDA
dec.95Ms	56.44	66.70	79.22	79.18	80.26	*83.19*	80.26	79.10	79.94
col.04Ms	58.06	68.44	81.84	*86.05*	85.78	82.50	81.89	81.37	83.41
media	57.25	67.57	80.53	82.61	*83.02*	82.84	81.08	80.24	81.67

Tabla 4.4.1.: *Reducción de dimensionalidad mediante reducción de características.* Promedio de *%cCorr* para cada método.

Analizando conjuntamente los resultados de la Figura 4.4.4 y la Tabla 4.4.1 observamos:

- Las transformaciones DCT y LPC obtienen con diferencia los peores resultados. Quizás se deba a que no se hace ningún tipo de estandarización sobre las características, con lo que las variables de menor energía contribuyan menos que aquellas que describen más variabilidad en los datos. De hecho, estas transformaciones han demostrado su capacidad para reducir la dimensionalidad en áreas como el reconocimiento del habla y la codificación de imágenes donde las variables tienen órdenes de magnitud similares (Tekalp and Tekalp, 1995; Rabiner and Schafer, 2007). Otra razón podemos encontrarla en que son las únicas transformaciones cuyas matrices de proyección no dependen de los datos. LPC es una técnica de predicción y en la práctica se aplica a segmentos con un alto grado de correlación entre sus muestras, lo que explica los malos resultados al reducir la dimensionalidad de un vector que a priori, es modelado suponiendo que sus componentes están poco correlacionadas entre sí.
- Al igual que ocurre con DCT y LPC, los resultados de PCA demuestran que la conservación de la energía no es la mejor estrategia para discriminar entre clases. Aún siendo la transformación ortogonal óptima para describir la variabilidad en los datos, sus limitaciones la sitúan entre las peores opciones.
- El análisis de factores, FA, alcanza una puntuación ligeramente similar a PCA. La sensibilidad a una elección correcta del número de factores (Filzmoser et al., 2009) o la generalidad de su planteamiento le confinan a una efectividad baja en nuestro test de reducción, acorde con las conclusiones de Shalizi (2009).
- En oposición a las transformaciones para conservar la energía, el análisis discriminante de Fisher, FDA, obtiene unos resultados relativamente modestos para estar focalizado a la discriminación entre clases y ser la única técnica supervisada. Las causas pueden estar en una relativa escasez de datos de entrenamiento para *dec.95Ms*, lo que explicaría la diferencia en eficacia respecto a col.04Ms, o, más probablemente, en que la matriz de proyección de FDA solo tiene $C-1$ vectores independientes, siendo C el número de clases. Esta última restricción justifica la forma de las curvas en la Figura 4.4.4; con un buen comportamiento para vectores pequeños pero con un progresivo declive a partir de 6 y 11 componentes respectivamente para *dec.95Ms* y *col.04Ms*, con 5 y 11 clases.

- En conjunto, los algoritmos fastICA y cubICA basados en la maximización de la independencia estadística consiguen unos resultados ligeramente mejores que el resto, particularmente en *col.04Ms*. Curiosamente, estos métodos extraen características cuyas funciones de probabilidad son no-gausianas.
- Las técnicas basadas en análisis de patrones que varían lentamente, SFA e ISFA, funcionan mejor en *dec.95Ms* que en *col.04Ms*, seguramente por que sea más difícil extraer las componentes lentas en un sistema cuyas características varían a velocidades muy diferentes entre sí, como ocurre con las clases de muy diferente dinámica que conforman *col.04Ms*. Obtienen tras ICA los mejores resultados, probablemente por la semejanza existente entre ambos conjuntos de técnicas, como detallan Blaschke et al. (2006).

La Tabla 4.4.2 nos esquematiza las propiedades más relevantes de las técnicas de reducción de dimensionalidad que hemos analizado en este apartado con el objetivo de escoger la más adecuada a nuestras necesidades. Damos una valoración cualitativa de 1 a 3 estrellas (*) para facilitar la comparación. Por *eficacia* nos referimos al tamaño del vector de características necesario para que las curvas de la Figura 4.4.4 alcancen unos resultados cuasi-estables en la tasa de reconocimiento. La columna *rapidez* evalúa el tiempo de computación requerido por cada técnica. En el caso de que la matriz de proyección requiera un análisis de los datos la columna *dependiente* estará marcada afirmativamente.

	%cCorr	eficacia	rapidez	supervisada	dependiente	suposición sobre datos
DCT	*	**	***	no	no	ninguna
LPC	*	*	***	no	no	ninguna
PCA	**	*	***	no	sí	variables gausianas
fastICA	***	***	**	no	sí	variables no gausianas
cubICA	***	***	**	no	sí	variables no gausianas
SFA	***	**	*	no	sí	correlación lenta
ISFA	**	*	*	no	sí	correlación lenta
FA	**	**	**	no	sí	ninguna
FDA	**	**	**	sí	sí	variables gausianas

Tabla 4.4.2.: *Evaluación cualitativa de los métodos de reducción de dimensionalidad mediante transformación del espacio de características.*

Teniendo en cuenta la información dada en la comparativa de la Tabla 4.4.2 y en los resultados presentados anteriormente parece que la mejor opción pasa por las técnicas cubICA o SFA. Aún no obteniendo los mejores resultados de clasificación, el análisis de características lentas tiene algunos puntos a su favor como la no suposición previa sobre la distribución probabilística de los datos y su capacidad para extraer información no lineal de las variables originales. En contra, es algo más costoso computacionalmente que ICA.

4.5. Comparación de métodos y conclusiones sobre la reducción de dimensionalidad

4.5.1. Resultados experimentales

La Figura 4.5.1 presenta las gráficas de tasa de reconocimiento en función del tamaño del vector de características para las dos mejores técnicas de reducción de dimensionalidad de cada categoría: selección de características por mediante filtros (SC_F) de estandarización y de energía; selección por modelos (SC_M) guiados por eficacia de reconocimiento que progresivamente van incluyendo nuevas componentes (+cAcc) o quitando componentes poco discriminativas (-cAcc) y, finalmente, reducción de dimensionalidad por transformación o reducción del espacio de características (RC) mediante análisis de características lentas (SFA) y de componentes independientes (cubICA). En ella destaca el buen comportamiento de las técnicas de selección mediante modelos frente a selección por filtros, mientras la reducción por transformación se queda entre ambas, diferencias que son más fácilmente apreciables en los resultados de *col.04Ms*.

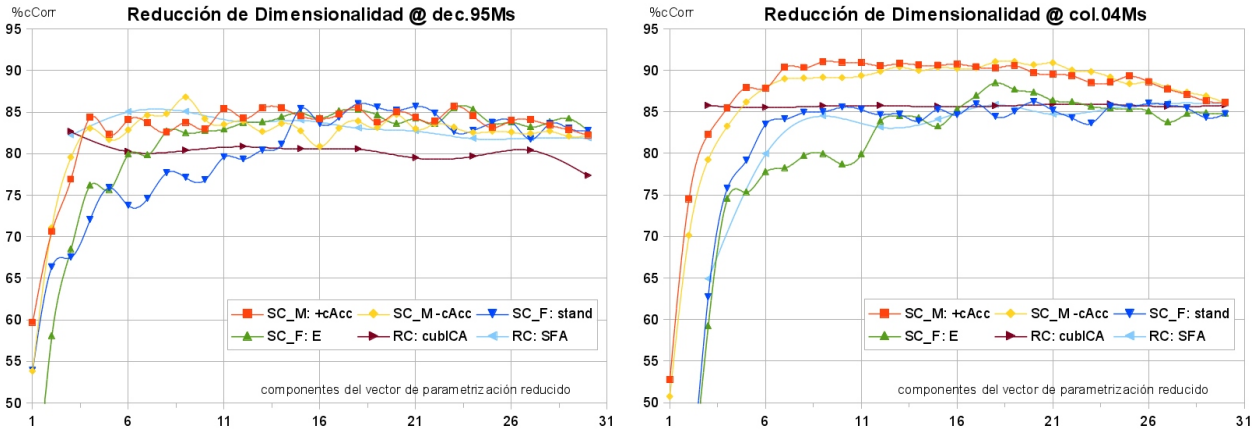


Figura 4.5.1.: Resultados finales de reducción de dimensionalidad. Curvas de los dos mejores métodos en las categorías de: selección por modelos (*SC_M*), selección por filtros (*SC_M*) y reducción de características (*RC*).

Aparte de una alta tasa de reconocimiento, sobrepasando en unos 5 puntos al resto de curvas en *col.04Ms*, los métodos guiados también destacan por su eficacia: obtienen buenos resultados rápidamente para un vector con pocas características, dentro del intervalo de [5,10]. Esta propiedad también es compartida por las técnicas de transformación, que mantienen sus tasas de reconocimiento casi constantes a partir de 3 componentes en *dec.95Ms*. Los filtros tardan más en alcanzar sus mejores puntuaciones, requiriendo sobrepasar el tamaño medio del vector para ello.

Los resultados promedio de la Figura 4.5.1 respecto al tamaño del vector de parametrización se muestran en la Tabla 4.5.1: los algoritmos de selección guiados por modelos mejoran en unos 2 puntos a los métodos de reducción que a su vez superan por unos 3 puntos a la selección por filtros.

	SC_M: +cAcc	SC_M: -cAcc	SC_F: stand	SC_F: E	RC: cubICA	RC: SFA
dec.95Ms	*82.61*	81.73	79.40	80.06	80.26	83.19
col.04Ms	*87.34*	86.75	80.21	79.48	85.78	82.50
media	*84.98*	84.24	79.80	79.77	83.02	82.84

Tabla 4.5.1.: *Resultados finales para los 2 mejores métodos de reducción de dimensionalidad en cada categoría.* Promedio de *%cCorr* para cada técnica.

A juzgar por los resultados obtenidos, parece que la mejor opción es usar un esquema de selección de características guiado por modelos, que progresivamente vaya incluyendo características al vector siguiendo un simple criterio definido por la tasa de reconocimiento. Los algoritmos SFA y cubICA se manifiestan especialmente eficaces cuando pretendemos describir los datos con muy pocas características.

Discusión de resultados. En el apartado anterior hemos escogido la técnica de reducción de dimensionalidad que mejor se adapta a nuestras necesidades según el estudio experimental realizado en este capítulo. Hay que destacar que no todos los sistemas se diseñan con las mismas necesidades y que no todas las experimentaciones comparando los mismos métodos extraen las mismas conclusiones. En este contexto, creemos necesario comentar algunas cuestiones:

- La tasa de reconocimiento promediado por clase con el esquema de parametrización mixto *geoLFCC.D.30* definido en la Subsección 4.2.5 es de 82.79 %cCorr para *dec.95Ms* y de 84.80 %cCorr para *col.04Ms.* Observando la Figura 4.5.1 comprobamos como la mayoría de las técnicas superan ese valor en ambas bases de datos para vectores con un tamaño de incluso un tercio respecto al vector original. Esto nos indica que, en general, todos los métodos aplicados obtienen en su conjunto unos buenos resultados, cumpliendo su objetivo original: reducir la dimensionalidad manteniendo la información relevante de los datos.
- La limitación de una evaluación con subgrupos de solo 2 componentes hecha en nuestra implementación de los filtros lleva a una selección de características que se torna poco eficaz comparada con la selección guiada o con la reducción mediante transformadas. Existen otros filtros más avanzados que utilizan estimaciones rápidas de distribuciones de probabilidad y de medidas de información que con un coste computacional relativamente bajo pueden ser una opción más que interesante frente a otras metodologías (Guyon, 2008).

- Por razones de extensión, caen fuera del objetivo de esta tesis algunas técnicas más actuales de reducción de dimensionalidad tales como:

  1. **Métodos de análisis local**. Segmentan el espacio de características y analizan los datos mediante núcleos o *kernels* que evalúan cada segmento:

     - El análisis de la Correlación Canónica (*Canonical Correlation Analysis, CCA*). Define una transformación lineal maximizando óptimamente la covarianza entre dos corpus de datos. También cuenta con su versión local, Kernel-CCA (Hardoon et al., 2004).
     - Técnicas basadas en el algoritmo *RELIEF* (Kira and Rendell, 1992). RELIEF realiza una selección mediante filtrado local estimando la dependencia condicional entre variables, es robusto al ruido y muy rápido. En contraposición no es capaz de eliminar efectivamente la redundancia entre componentes.
     - Otros métodos que han adquirido una creciente popularidad: *Kernel PCA* (Scholz et al., 2008), *Locally Linear Embedding* (Roweis and Saul, 2000), *Isomaps* (Tenenbaum et al., 2000).

  2. **Algoritmos no lineales de separación de señales**. Los mapas autoorganizativos (Subsubsección 2.3.6.2) o de Kohonen (*Self Organizing Maps - SOM*) que usan una red neuronal para reducir la dimensionalidad de los datos, proyectando el espacio de características a un plano Masiello et al. (2006); Esposito et al. (2008b); Köhler et al. (2009).

  3. **Métodos heurísticos de optimización**. Como los algoritmos evolutivos (EA) y, concretamente los algoritmos genéticos (Orlic and Loncaric, 2010), usados con éxito en distintas áreas (Bhanu and Lin, 2003; Zamalloa et al., 2008). Cortés et al. (2015) obtienen unos resultados de selección de características con los algoritmos genéticos ligeramente inferiores a los logrados con el DFS generalizado.

- En lo relativo los esquemas de parametrización nos queda por plantear algunas propuestas que pueden ser muy interesantes de cara a futuras investigaciones:

  - Quizás unas mismas características no sean igual de discriminatorias en datos con eventos aislados que en datos continuos. Características específicamente diseñadas para segmentar eventos, inspiradas en técnicas de detección (Withers et al., 1998; Ladd et al., 2000), serían muy útiles en un sistema que se aplique en datos continuos.
  - La incorporación de parámetros usados en otras disciplinas de monitorización de volcanes al vector de características podría aportar una información útil para la clasificación. Medidas sobre la composición química de gases emitidos en el cráter o sensores de actividad mediante cámaras infrarrojas quizás podrían ayudar a discriminar entre eventos externos (colapsos, lahares, flujos piroclásticos) y sismos internos (eventos de bajo

periodo, eventos híbridos, sismos tectónicos). Esta idea básica aplicada a describir el estado de la actividad eruptiva de un volcán ya ha sido aplicada con éxito por Carniel et al. (2006) dentro del proyecto europeo *MULTIMO* (*Multi-disciplinary monitoring, modelling and forecasting of volcanic hazard*).

- Una cuestión abierta es cómo de recomendable es el uso de características geofísicas con una dependencia local como azimut, polarización o magnitud. Si bien es posible que mejoren la eficiencia de reconocimiento del sistema posibilitando distinguir entre sismos de origen tectónico y origen volcánico, eventos superficiales o generados en la corteza, también es cierto que empeoran la generalidad del sistema. El objetivo de nuestro trabajo es extraer características generales, no ligadas a efectos de sitio o localizaciones concretas de volcanes. Nótese que, incluso buscando propiedades generales, la estadística de las variables pueden cambiar apreciablemente simplemente moviendo de lugar el sismómetro dentro del mismo volcán.
- Un tema de discusión interesante (Hoogenboezem, 2010) es si realmente es necesario usar el ventaneo de eventos, o, si se podría analizar segmentos de señal más grandes (de minutos) para aprovechar la ventaja que las wavelets ofrecen sobre otras parametrizaciones a costa de comprometer el reconocimiento en tiempo real. Nótese que un tamaño pequeño (de máximo 10 segundos) de los segmentos no es lo suficientemente grande como para aprovechar la resolución espacio-temporal ofrecida por las wavelets.

4.5.2. Conclusiones

En este capítulo hemos realizado un profundo análisis de las técnicas clásicas de reducción de dimensionalidad, presentando sus ventajas e inconvenientes, con un enfoque hacia aquellas que han sido más utilizadas en el área de la sismicidad volcánica. En este sentido, hemos contribuido en una doble vertiente:

1. Hemos revisado la mayoría de las características usadas en la literatura para describir las señales sismo-volcánicas, aprendiendo de sus ventajas e inconvenientes. Asimismo hemos propuesto y analizado varias parametrizaciones de distinta naturaleza, presentando un vector de características híbrido que consigue mejores resultados que otros esquemas homogéneos (Álvarez et al., 2009; Cortés et al., 2014).
2. Realizamos una generalización de una técnica de selección de características discriminativa guiada por modelos estadísticos, obteniendo el mejor promedio en la tasa de reconocimiento respecto las otras opciones analizadas.

A pesar de que por razones de extensión no hemos podido analizar las técnicas más recientes de reducción de dimensionalidad, la profundidad del estudio teórico acompañado siempre por resultados experimentales nos lleva a poner de relieve las siguientes conclusiones:

- El uso de métodos de reducción de dimensionalidad queda totalmente justificado en la mayoría de los algoritmos analizados; mantienen e incluso mejoran el nivel de aciertos en clasificación reduciendo el vector de descripción de datos hasta 2/3 de su tamaño original.
- La elección de un método sobre otro no debe basarse solo en la tasa de reconocimiento que puedan alcanzar. La mayor parte de estos métodos logran una tasa mayor que en el espacio original reduciendo a la mitad la dimensionalidad. Otros factores de implementación y aplicación han de tenerse en cuenta tales como:
  - La *rapidez* de ejecución y coste computacional de cada algoritmo.
  - Lo *eficaz* que sea al clasificar eventos con las menos características posibles.
  - Las *suposiciones* estadísticas que se requieran sobre los datos para aplicar las técnicas correctamente.
  - La *dependencia* de los datos para extraer las nuevas bases del espacio reducido de características, que obliga a contar con una partición de entrenamiento para reducir la dimensionalidad.
  - La ejecución *supervisada* de un algoritmo, que implica una dependencia de una partición de entrenamiento de datos con su respectivo etiquetado previo.

  Estas propiedades se sintetizan cualitativamente en la Tabla 4.5.2, que otorga tres estrellas (***) a las mejor calificadas y una (*) a las peores. En general, la elección de una técnica sobre otra dependerá de la importancia relativa que se le dé a cada una de estas cualidades y de las especificaciones del sistema a diseñar. Dado que el esquema final de parametrización y reducción de dimensionalidad debe fijarse en la etapa de inicialización antes de que el sistema se use como reconocedor, el coste de hacer un estudio que requiera una dependencia de datos e incluso una supervisión previa es bastante razonable frente

		%cCorr	eficacia	rapidez	supervisada	dependiente	suposición sobre datos
SC_F:	*stand*	*	*	***	no	no	ninguna
	E	*	*	***	no	no	ninguna
SC_M:	*+cAcc*	***	**	*	sí	sí	ninguna
	-cAcc	***	**	*	sí	sí	ninguna
RC:	*SFA*	**	**	*	no	sí	correlación lenta
	cubICA	**	***	**	no	sí	variables no gausianas

Tabla 4.5.2.: *Comparación cualitativa de los mejores métodos analizados de reducción de dimensionalidad en las categorías*: selección por modelos (SC_M), por filtros (SC_F) y por reducción de características (RC).

al beneficio obtenido si la parametrización es eficaz y con un vector pequeño podemos obtener altas tasas de reconocimiento. En este sentido, a nuestro juicio, parece más lógico apostar por técnicas guiadas, aún teniendo que preparar una base de datos etiquetada para hacer la selección previa. Estos métodos cuentan a su favor con dos grandes factores:

1. El paradigma de la selección frente a reducción: que permite una interpretación geofísica más directa de las características que describen los eventos.
2. No necesitamos asumir suposiciones estadísticas sobre los datos, a menudo requeridas por las técnicas de reducción o por filtros.

Estas ventajas junto a los resultados experimentales decantan nuestra decisión de escoger el DFS generalizado como técnica de reducción de dimensionalidad en la descripción de los datos de nuestro sistema.

- A raíz de los resultados experimentales, podemos afirmar que, con pocas excepciones, las características con mayor poder descriptivo lo son tanto como para HMMs como para GMMs en las dos bases de datos evaluadas. Sin embargo, en la Sección 4.2 queda demostrado la influencia que distintas configuraciones (tamaño de segmento, componentes de los modelos, estados de los HMMs...) tienen en la eficiencia de reconocimiento de cada característica. Este hecho justifica experimentalmente abordar por separado la configuración y selección de características independientemente para cada tipo de evento en las arquitecturas en paralelo que realizaremos en el Capítulo 5 y Capítulo 6.

5. Diseño del sistema de reconocimiento en paralelo VSR-PSA

Una vez analizadas las necesidades que los centros de monitorización de volcanes activos requieren de los sistemas de reconocimiento automático de sismos *(Volcano-Seismic Recognition - VSR)* en el Capítulo 2, comprobamos que las soluciones actuales aún quedan lejos de satisfacer las demandas. En los últimos 20 años ha avanzado notablemente en áreas como la detección y la clasificación de eventos una vez aislados, pero aún queda por mejorar en los sistemas de reconocimiento en tiempo real sobre registros continuos.

La sismicidad juega un papel fundamental en los sistemas de alerta temprana, concretamente, la frecuencia e intensidad de ciertos tipos de eventos o *precursores* como tremores y señales de (muy) baja frecuencia que se asocian a episodios eruptivos. En este capítulo proponemos un sistema con una arquitectura en paralelo (*Parallel System Architecture - PSA*) de *canales* enfocados a analizar y reconocer en tiempo real cada uno de estos tipos de señales y otras clases de eventos, como derrumbes y lahares, que suponen un riesgo inminente en poblaciones cercanas al lugar de la erupción.

En la Sección 5.1 presentaremos las bases teóricas y desarrollaremos el diseño de un sistema VSR-PSA como una evolución del sistema base en serie VSR-SSA que estudiamos en el Capítulo 3. Las adaptaciones que en cada canal se hacen en torno a un tipo concreto de evento, la llamada *clase propia* del canal es la clave en este nuevo enfoque. Esta especialización brinda la oportunidad de analizar con un alto nivel de eficacia y robustez las mejores configuraciones y el conjunto de características más útiles para describir cada tipo de evento, lo que contribuye al objetivo más ambicioso de construir un sistema de reconocimiento no supervisado de carácter universal. La flexibilidad que conlleva el sistema VSR-PSA le dota para realizar numerosos tipos de funcionalidades y análisis que serán detallados en la Sección 5.2.

5.1. Paralelización: canales de reconocimiento específicos para cada clase

En los próximos apartados describiremos la arquitectura en paralelo propuesta en el sistema VSR-PSA a partir de los sistemas de reconocimiento clásicos en serie. Empezaremos presentando los VSR-PSA como una evolución lógica de los esquemas en serie VSR-SSA. Definiremos cualitativamente nuestra propuesta, centrada en el concepto de *canal* de reconocimiento específico para cada clase y de un *clasificador conjunto* que combina los resultados de los canales. Una vez presentada la propuesta, diseñaremos cada bloque del sistema en paralelo.

Sistemas en paralelo como evolución de los esquemas en serie. En el Capítulo 2 estudiamos las propuestas que el reconocimiento automático de patrones da a las necesidades de los actuales centros de monitorización de volcanes activos. Las conclusiones listadas en la Subsección 2.4.2 nos hacen reflexionar sobre las propiedades deseables en un sistema VSR (Subsección 2.2.3) que nos inspiran para realizar nuestra propuesta de un sistema basado en HMMs con distintos canales en paralelo, cada uno de ellos especializado en un tipo de evento concreto, denominado clase propia del canal. Básicamente, las áreas sobre las que se que centra la mayor parte de la investigación actual en el reconocimiento automático de sismos se sintetizan en:

1. Reconocimiento en tiempo real sobre registros de datos continuos
2. Capacidad de discriminar y analizar eventos especialmente relevantes para los sistemas de alerta temprana y de protección de la población, tales como precursores de erupciones y eventos que implican un elevado riesgo de la población
3. Sistemas de reconocimiento multiclase robustos a efectos locales, de una alta fiabilidad y que funcionen de manera no supervisada.

El primer y último de estos requisitos nos llevan a elegir los HMMs como clasificadores base de nuestra propuesta. Los HMMs ya han demostrado su eficacia en estos aspectos en el área del reconocimiento del habla (Garofolo et al., 1993; Rabiner and Schafer, 2007; Gales and Young, 2008) y cuentan con varias herramientas fiables disponibles (Young et al., 2006). La segunda de las demandas inspira la arquitectura VSR-PSA propuesta; un esquema de canales en paralelo, cada uno conteniendo un sistema clásico en serie VSR-SSA, donde cada etapa se diseña específicamente para reconocer una clase de eventos determinada. La paralelización del sistema base nos permite ofrecer las siguientes funcionalidades clave que nos acercan a satisfacer los objetivos deseados:

- Análisis por independiente de clases especialmente relevantes
- Estudio de las mejores configuraciones y grupo de características para cada canal, que contribuyen a incrementar la eficacia de detección y reconocimiento

- Búsqueda de conocimiento científico entorno a las clases, que permitan universalizar un sistema VSR (SSA o PSA) de forma no supervisada para facilitar su integración en los centros de monitorización

La idea de paralelizar sistemas tiene sus orígenes en la combinación de clasificadores como ya vimos en la Subsección 2.3.5 del Capítulo 2. En la bibliografía encontramos ejemplos de sistemas en los que algunas de sus etapas son diseñadas específicamente para cada tipo de eventos. Ohrnberger (2001); Beyreuther and Wassermann (2008) en una fase de post-procesado utilizan filtros de mínima duración que dependen de cada clase. Beyreuther and Wassermann (2011) usan técnicas heurísticas en unos modelos HSMM multi-estado que segmenta solo aquellos eventos cuya probabilidad es múltiplo de la asignada al ruido. Cortés et al. (2014) implementan el diseño VSR-PSA sobre clasificadores GMMs que modelan con un conjunto distinto de características a cada clase.

Descripción del sistema PSA en paralelo: canales específicos y reconocimiento conjunto. La Figura 5.1.1 muestra el esquema en paralelo propuesto (*Parallel System Architecture - PSA*). El sistema se divide en 2 bloques principales; uno de *procesado* o análisis en paralelo del registro continuo y otro de *clasificación conjunta* tal y como se hace en un sistema serie VSR-SSA:

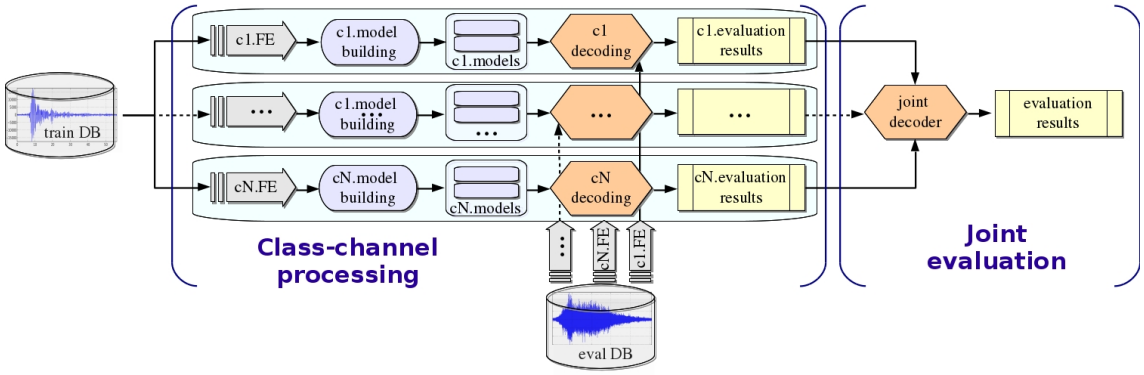


Figura 5.1.1.: Estructura en paralelo del sistema (PSA) de clasificación propuesto. El bloque de análisis se compone de canales de procesamiento; cada uno de ellos es un sistema de reconocimiento en serie (SSA) especializado en una clase concreta de eventos. Su salida alimenta el bloque de reconocimiento conjunto o multiclase.

- **Análisis mediante canales específicos**: Para cada clase w_{ch} del conjunto $\{w_c\}$ a reconocer, existe un subsistema o *canal* de reconocimiento construido especialmente para detectar y clasificar los eventos pertenecientes a w_{ch}, denominada la *clase propia* del canal *ch*. Un canal no es más que un sistema en serie VSR-SSA, estudiado en el Capítulo 3, diseñado específicamente para conseguir una alta eficacia y robustez al reconocer eventos de un tipo concreto.

Por lo tanto, hereda el diseño y etapas de los sistemas VSR-SSA: parametrización, selección de características, construcción de modelos, clasificación y, en el caso de contar con las transcripciones de la base de datos de evaluación, evaluación de resultados. La única salvedad radica en el diseño del decodificador que aparte de las transcripciones con las segmentaciones temporales y las etiquetas asociadas a cada segmento, debe proporcionar las secuencias de probabilidades $\{p(\mathbf{x}, w_c)\}_{ch}$ para cada frame $\mathbf{x}$ y para cada modelo de la clase $w_c$en cada canal ch. Estas probabilidades son necesarias como referencia para normalizar, dentro de cada canal, las secuencias $\{p(\mathbf{x}, w_{ch})\}_{ch}$ que constituyen la entrada del bloque de clasificación conjunta.

- **Clasificación conjunta usando todos los canales**: En el caso de que quiera usarse el sistema para dar una salida de reconocimiento clásica, como la proporcionada por el sistema base VSR-SSA (Capítulo 3), el decodificador conjunto requiere de cada canal ch la secuencia filtrada $\{p(\mathbf{x}, w_{ch})\}_{ch}$ de probabilidades de la clase propia w_{ch} del canal. Dado que cada canal tiene su propio esquema de parametrización, es necesario que estas C secuencias estén en el mismo intervalo de probabilidad para poder ser comparadas en el proceso de decodificación (Sección A.7). Una vez filtradas, las C secuencias son usadas para reconocer en continuo y, posteriormente, evaluar los resultados si se tiene acceso a las transcripciones de la base de datos de evaluación. Este módulo no es más que una evolución semi-heurística del clasificador base GMM de eventos aislados presentado en la Subsección 3.3.1 con el objetivo de reconocer sobre datos continuos. Puesto que el proceso se hace con probabilidades que provienen de canales especializados, a priori, el resultado obtenido con el sistema VSR-PSA debe ser más robusto y fiable que el dado por el sistema serie VSR-SSA.

La flexibilidad del sistema VSR-PSA comienza desde su misma etapa de configuración, posibilitando múltiples opciones de diseño en función del uso. Esta arquitectura en canales paralelos especializados incrementa las posibilidades del esquema clásico VSR-SSA bajo un único sistema que engloba (Sección 5.2):

- Detector de eventos sismo-volcánicos.
- Discriminador de eventos de una clase concreta.
- Análisis estadístico de cada tipo de eventos.
- Selección del mejor conjunto de características para cada clase.
- Reconocimiento automático. Tanto para eventos específicos de una clase concreta como para eventos de diverso tipo (reconocimiento multiclase).

Tipos de canales. Con objeto de aumentar la capacidad de análisis y la flexibilidad, la implementación realizada de nuestro sistema nos permite usar 2 tipos de canales en paralelo:

1. **Canales *biclase* o binarios *(PSA.bin)***. En cada canal se distingue tan solo entre 2 clases $\{w_{ch}, \bar{w}_{ch}\}$; la clase w_{ch} o propia del canal y el resto de clases agrupadas bajo una sola, $\bar{w}_{ch}$. La mayoría de los clasificadores de aprendizaje automático (ANNs, SVMs, LDAs,...) son en su origen sistemas binarios que necesitan ser extendidos para poder discriminar entre múltiples clases mediante esquemas (Bishop, 2007) que aplican el mismo sistema a distintas combinaciones de 2 clases ('uno contra el resto', 'uno contra uno',...). Estos canales son especialmente útiles en el análisis, para encontrar las propiedades típicas de una clase concreta que permitan discriminarla sobre otras, tal y como el mejor set de características de parametrización.
2. **Canales multiclase *(PSA.mul)***. Aunque cada canal *ch* se diseña para discriminar su clase propia w_{ch}, se construyen modelos para poder reconocer eventos de todas las clases $\{w_c\}$. Están orientados a su uso como clasificadores, pues, en general, es más efectivo construir un modelo para cada clase $w_c \neq w_{ch}$ no-propia del canal que un solo modelo genérico $\bar{w}_{ch}$ que las englobe a todas, agrupando por ello eventos con baja similitud entre sí, lo que acarrea una alta variabilidad y mayor complejidad en el modelado.

Existen otras posibles implementaciones de canales que pueden resultar interesantes. En concreto, los canales binarios que discriminan entre el ruido y cualquier otro tipo de evento, más conocidos en la literatura como *detectores de eventos*. Nótese que cualquier sistema STA/LTA, y, en general, cualquier sistema de detección de fase o *picking* puede ser caracterizado de esta forma.

Esquemas de configuración del sistema VSR-PSA. El proceso de construcción de un canal sigue básicamente los mismos pasos que el diseño del sistema base VSR-SSA (Capítulo 3), añadiendo dos etapas más encargadas de la extracción y filtrado en cada canal *ch* de las secuencias de probabilidades $\{p(\mathbf{x}, w_c)\}_{ch}$ para cada frame $\mathbf{x}$ y para cada modelo w_c. La especialización de los canales es posible gracias a que algunas variables que llamamos *variables PSA* o *variables del canal* pueden configurarse de forma específica en cada canal conforme a los posibles usos que queramos hacer del sistema. Acorde a estas funcionalidades, el diseño del sistema debe hacerse escogiendo entre uno de estos *criterios de configuración*:

max{PSA.joint}: Configurar el sistema general y los canales en particular para obtener los mejores resultados de reconocimiento para las clases originales $\{w_c\}$, obtenidos por el clasificador conjunto (tratando el sistema VSR-PSA como un sistema serie VSR-SSA).

max{PSA.ch}: Maximizar la eficiencia de reconocimiento de cada canal *ch* en su conjunto, lo que equivale a maximizar la eficiencia promedio para todas sus clases $\{w_c\}_{ch}$, medida por $\%cAcc$.

max{PSA.class}: Maximizar la eficiencia de reconocimiento en cada canal medida por $\%Acc(w_{ch})$ de solo una clase, w_{ch}, la denominada *clase propia* del canal.

El esquema de configuración $max\{PSA.joint\}$ es el adoptado por los sistemas clásicos en serie VSR-SSA, ya estudiados en la Sección 2.3. Los otros dos representan las nuevas posibilidades de análisis y flexibilidad que brinda el sistema VSR-PSA. En cualquier caso, en el Capítulo 6 también evaluaremos los resultados del clasificador conjunto, aunque el sistema no se configure específicamente para ello. Una vez escogido el criterio de configuración a seguir, los valores concretos de las variables PSA se asignan de una forma mixta:

- *Determinísticamente*: fijando valores concretos de ciertos parámetros conforme a análisis previos o a un conocimiento geofísico a priori de las señales.
- *Experimentalmente*: realizando pruebas supervisadas en abierto según la metodología indicada en la Sección 3.5.

5.1.1. Diseño de los canales

La construcción y configuración se estructura en las siguientes etapas:

1. **Adquisición de datos.** Como vemos en la Figura 5.1.1, el sistema PSA usa exactamente las mismas bases de datos que los sistemas serie.
2. **Extracción de características.** Parámetros de segmentación y filtrado tienen que ser configurados:
   a) *Filtrado espectral* en la banda $[f_L, f_H]_{ch}$ de las señales que entran al canal ch. Si configuramos el sistema mediante el criterio $max\{PSA.class\}$ para maximizar la eficiencia de la clase propia w_{ch} de cada canal, a priori, en un canal binario $PSA.bin$, prefiltraremos en la banda donde más energía espectral acumulen los eventos pertenecientes a w_c. Si nuestro objetivo es el diseño para optimizar los resultados de cada canal en general (usando el esquema de configuración $max\{PSA.ch\}$), no está asegurado que en un canal multiclase $PSA.mul$ donde se filtre atendiendo solo a su clase propia se mejore la eficiencia de reconocimiento del canal en conjunto.
   b) *Duración del segmento* o *frames* del canal. Se espera que para poder modelar correctamente la secuencialidad temporal de algunas clases de corta duración (LPs o VTs) el tamaño de los frames tenga que ser pequeño comparado con aquellos eventos de menor variabilidad temporal como LAHs, TRs o COLs. En todos los casos mantendremos un solapamiento entre segmentos del 50 %.
3. **Selección de características**. Para ello usamos en cada canal la técnica de reducción de dimensionalidad que ha resultado ser la más efectiva según el estudio hecho en el Capítulo 4: el algoritmo guiado por modelos DFS generalizado, de direccionalidad creciente y que usa $\%cAcc$ como medida discriminante, que abreviaremos como $DFS.cAcc$. Del conjunto original $C_{ch} = \{C_1, ..., C_K\}_{ch}$ de características de cada canal ch, el $DFS.cAcc$ ordena decrecientemente las

que son más eficientes conforme a uno de dos posibles criterios de configuración, $max\{PSA.class\}$ o $max\{PSA.ch\}$, para obtener el conjunto ordenado $S_{ch} = \{C_{ch,1}, ..., C_{ch.K}\}$ con $ch.K$ elementos, tal que $ch.K \leq K$. El número $ch.K$ de características que escojamos para describir las señales en cada canal también es un parámetro a configurar, y puede ser el mismo para todos los canales o distinto en cada uno de ellos, sugerido en cada caso por el algoritmo $DFS.cAcc$.

4. **Construcción de los modelos.** Se usan los HMMs como clasificadores en cada canal. Se configuran los siguientes parámetros:

   a) *Arquitectura del HMM:* se propone una arquitectura (definida por el nº de estados y los enlaces entre ellos) concreta para cada modelo de cada clase $\{w_c\}$, que se mantiene igual en todos los canales multiclase. En los canales binarios el modelo de la clase no-propia, $\bar{w}_{ch}$, promedia la arquitectura del resto de clases no-propias $\{w_c\} \neq w_{ch}$ en cada canal ch.

   b) *Componentes gausianas del HMM*: Se definen para cada canal ch de manera experimental conforme al criterio de configuración escogido; maximizar el $\%Acc$ para la clase w_{ch} o el $\%cAcc$ para el promedio de las clases $\{w_c\}_{ch}$.

   El aprendizaje tiene lugar de forma independiente para cada canal, tal y como se detalla en la Sección 3.3.2.1 y Subsección 3.7.2.

5. **Clasificación y evaluación del canal**. Cada canal se evalúa utilizando las mismas métricas del sistema de referencia VSR-SSA (Sección 3.4).

6. **Extracción de $\{p(\mathbf{x}, w_c)\}_{ch}$ para todas las clases $\{w_c\}_{ch}$ del canal**. La decodificación por Viterbi (Sección 3.3.2.1) realizada en el proceso de clasificación mediante HMMs no proporciona las secuencias $\{p(\mathbf{x}, w_c)\}$ para todos los vectores $\mathbf{x} = \{\mathbf{x}_t\}_{t=1:T}$ evaluados en cada uno de los modelos w_c. Estas secuencias se pueden aproximar de distintas formas descritas en la Sección A.7: directamente sobre la información dada por el software HTK o mediante una re-evaluación de eventos aislados (Cortés et al., 2009b).

7. **Filtros post-clasificación**. Una vez configurados los HMM y extraídas todas las probabilidades, el bloque del clasificador conjunto requiere que se normalicen cada secuencia $\{p(\mathbf{x}, w_{ch})\}_{ch}$ asociada a cada canal ch para poder compararlas a la hora de clasificar. El proceso de filtrado y normalización tiene una gran influencia en la eficiencia del reconocedor conjunto, siendo un punto clave en el diseño del sistema VSR-PSA. La mayoría de estos filtros pueden ser aplicados conjuntamente a todas las secuencias de cada canal o de forma independiente a cada una de las secuencias por separado. Se agrupan en:

   a) *Filtros de suavizado de probabilidades*: Un modelo asociado a una clase w_c que genere una secuencia $\{p(\mathbf{x}, w_c)\}$ con una alta variabilidad en un fichero dado, indica que, potencialmente, existe una alta incertidumbre

asociada a eventos de la clase w_c, lo que equivale a posibles eventos (o frames) w_c intercalados con eventos de otras clases. Con el objetivo de evitar demasiada variabilidad en las secuencias $\{p(\mathbf{x}, w_c)\}_{ch}$ dentro de cada canal ch y de ignorar probabilidades asociadas a frames o señales espúreos, se emplean filtros de suavizado que limitan el valor de $\{p(\mathbf{x}, w_c)\}$ a intervalos concretos.

b) *Normalización de probabilidades*: Se llevan todos los canales al mismo intervalo de valores, con el objeto de poder comparar las secuencias de las clases propias $\{w_{ch}\}$ en el decodificador conjunto. Un canal que tenga mucha variabilidad en $\{p(\mathbf{x}, w_c)\}_{ch}$ nos indica que es susceptible de contener muchos puntos de segmentación entre sus clases de canal $\{w_c\}_{ch}$.

c) *Filtros de equi-probabilidad*: Si fijamos un mismo valor de $sum(\{p\}_{ch}) = sum_{ch} \equiv \sum_c \sum_{\mathbf{x}} \{p(\mathbf{x}, w_c)\}_{ch}$ para todos los canales ch y suponemos que todos tienen el mismo número de vectores $\{\mathbf{x}\}$, estamos asumiendo que todos los canales modelan igual de bien los datos de entrada. Aún siendo una hipótesis razonable, puede no ser estrictamente cierta; si los canales se configuran para maximizar la eficiencia del modelo de la clase propia w_A de un canal A en detrimento del resto de modelos $\{w_c\} \neq w_A$ y en un registro de señal no hay eventos de la clase w_A, es muy probable que $sum_{ch(A)}$ tenga un valor menor que en otro canal B del que sí existan eventos en el fichero.

La Tabla 5.1.1 lista diferentes filtros susceptibles de ser aplicados junto con los objetivos e hipótesis que se asumen sobre los canales. El orden en el que se aplican es relevante y puede hacer cambiar el resultado final.

identificador	*implementación*	*objetivo*
CLIP_FILE_PCX	$\Delta\mathbf{p}_{ch} \in [pMIN, pMAX]$	evitar valores espúreos
CLIP_CLASS_PCX	$\Delta\mathbf{p}_c \in [\bar{\mathbf{p}}_c \pm N\sigma_{\mathbf{p}_c}]$	evitar valores espúreos
NORM_FILE_PCX	$\mathbf{\Delta p}_{ch} \equiv [pMIN, pMAX]$	fija incertidumbre en ch
STD_NORM_FILE_PCX	$\sigma_{\mathbf{p}_{ch}} \equiv 1$	canales equi-variables
FORCE_MEAN_FILE_PCX	$\bar{\mathbf{p}}_{ch} \equiv pMEAN$	canales equi-probables

Tabla 5.1.1.: $\{p(\mathbf{x}, w_{c=1:C})\}_{ch} \equiv \mathbf{p}_{ch}$ aplicados en cada canal ch para todas sus clases $\{w_c\}_{ch}$ en conjunto, o independientemente para cada una de sus clases w_c, $\{p(\mathbf{x}, w_c)\}_{ch} \equiv \mathbf{p}_c$.

Influencia del orden en el proceso de configuración. El orden en el que se desarrollan las etapas del diseño puede influir notablemente en los valores óptimos de las

variables, pues una configuración concreta en una etapa puede influir en otra. Sirvan como ejemplo lo siguientes casos:

- Una duración de los frames y una banda de filtrado concretas pueden aumentar el poder discriminatorio de unas características sobre otras, con lo que el algoritmo $DFS.cAcc$ variará su orden de selección para incorporar antes aquellas componentes más discriminativas. Si se aplica el $DFS.cAcc$ antes de fijar la duración de los frames y la banda de filtrado nos aseguramos que estos últimos se configurarán solo para el subconjunto S_{ch} de características finalmente seleccionadas y no para el conjunto C_{ch} inicial, del que luego se quitan las componentes menos discriminantes.
- Algunas características son más variables que otras y necesitan más componentes en el HMM para ser correctamente modeladas. Equivalentemente, el poder discriminatorio de una característica puede no ser el mismo con modelos simples que con otros más complejos, por lo que es muy probable que el algoritmo $DFS.cAc$c no seleccione el mismo conjunto de características para modelos de canal que usan solo 1 gausiana o para otros que usen 16 gausianas.

Este fenómeno ocurre tanto en sistemas serie VSR-SSA como en paralelo VSR-PSA y afecta a distintos niveles, variando el resultado según el orden en el que se aplican los distintos filtros, incluso, dentro de una misma etapa de configuración. La elección del orden *correcto* al configurar no es inmediata. La pregunta si existe el orden *correcto* debe ser formulada y respondida. Estas cuestiones caen fuera del ámbito de esta tesis; por lo que se adopta un criterio mixto para aumentar la independencia entre etapas de configuración y minimización del coste computacional. Partiendo de la configuración básica ya sugerida en el Capítulo 4 y siguiendo la metodología especificada en la Sección 3.5, el orden seguido es el siguiente:

1. Se configuran los HMMs: se fija la arquitectura y, experimentalmente, el nº de componentes gausianas
2. Se seleccionan las mejores características, aplicando el $DFS.cAcc$ sobre los modelos ya configurados
3. Se escogen los valores óptimos de del filtrado espectral y duración de los segmentos de señal de los que se extrae los vectores de características.

Diferencia entre los modelos de una misma clase propia de un canal múltiples y su homólogo binario: influencia de modelos *no propios* de un canal. Una cuestión interesante es si dada una clase propia w_{ch}, *¿el modelo de dicha clase propia* $w_{PSA.bin(ch)}$ *construido en un canal binario no debería ser igual a su homólogo* $w_{PSA.mul(ch)}$ *construido en un canal múltiple?*. Nótese que ambos tienen que describir el mismo subespacio de características... A priori: en cuanto el subespacio cambie (al seleccionar

distintas características o distinto tamaño de la ventana deslizante de parametrización) en el canal la respuesta clara es que no. Incluso, la respuesta puede seguir siendo negativa aún modelando el mismo subespacio: nótese que si nos basamos en pruebas experimentales para configurar un modelo, el resultado final de eficiencia de un modelo ($\%cAcc(w_{ch})$ o $\%cCorr(w_{ch})$) en un HMM depende del camino escogido en la red de búsqueda en el momento de su decodificación por Viterbi (Figura 3.3.3). La estructura de esta red de búsqueda o *macro-HMM* depende del resto de modelos del canal. Este hecho pone de manifiesto la influencia que el *resto de modelos* de un canal tiene en los resultados del canal y en la salida del decodificador conjunto.

5.1.2. Diseño del decodificador conjunto

El decodificador usado en la evaluación conjunta de los canales es una evolución del decodificador base diseñado para los GMM (Subsección 3.3.1) que permite la segmentación y clasificación de eventos registrados en el dominio continuo a partir de las secuencias $\{p(\mathbf{x}, w_{ch})\}_{ch}$ de probabilidades de las clases propias $\{w_{ch}\}$ de los canales. Se basa en un algoritmo sencillo que pretende aproximar al algoritmo de Viterbi mediante filtros semi-heurísticos sobre las secuencias $\{p(\mathbf{x}, w_{ch})\}_{ch}$. Esta nueva implementación se estructura en 2 fases:

1. Pre-filtrado de las secuencias de probabilidades $\{p(\mathbf{x}, w_{ch})\}_{ch}$
2. Segmentación secuencial basado en filtros de duración temporal y en probabilidades logarítmicas acumuladas

Una vez segmentados los eventos, la evaluación se realiza de la misma manera que en el sistema base VSR-SSA y que en los canales PSA. Nótese que este esquema de decodificación a pesar de estar aún en fase experimental ha obtenido buenos resultados preliminares como veremos en el Capítulo 6. Su sencilla implementación le confiere una baja complejidad computacional de $\mathcal{O}\{2 \cdot T\}$ frente a $\mathcal{O}\{T \cdot Q^2\}$ del algoritmo de Viterbi (siendo Q el número de estados de los HMM y T el número de observaciones de los datos de evaluación).

5.1.2.1. Pre-filtrado de las secuencias de probabilidades

Obtenidas las secuencias de probabilidades $\{p(\mathbf{x}, w_{ch})\}_{ch}$ en cada canal, se les somete a un pre-filtrado antes de pasar a la etapa de reconocimiento. La Tabla 5.1.2 detalla los dos tipos de filtros que se aplican:

- *Filtros para igualar el número de frames*: Dado que cada canal tiene su parametrización característica, sus correspondientes vectores de características pueden describir ventanas de señal de diferentes duraciones, y, por tanto, las secuencias $\{p(\mathbf{x}, w_{ch})\}_{ch}$ pueden tener distinto número de elementos en cada canal. Mediante funciones *spline* (curvas diferenciables definidas a trozos por

identificador	*implementación*	*objetivo*
PSA_FS_RESAMPLING	$Fs(\{p(\mathbf{x}, w_{ch})\}) \equiv Cte$	fija el mismo #frames $\forall$canales
SMOOTH_FILE_PCX	$LPF_{0,5}(\{p(\mathbf{x}, w_{ch})\})$	evita inserciones
CLASS_MIN_DUR	$dur_m(w_c) \leq dur(\mathbf{x}_c)$	ignora eventos $\mathbf{x}_c$ muy cortos
CLASS_MAX_DUR	$dur(\mathbf{x}_c) \leq dur_M(w_c)$	ignora eventos $\mathbf{x}_c$ muy largos
CLASS_DUR_PREFILTER	$dur(\mathbf{x}_c) \in [dur_m(w_c), dur_M(w_c)]$	limita la duración de $\mathbf{x}_c$

Tabla 5.1.2.: *Pre-filtrado de las secuencias de probabilidad* $\{p(\mathbf{x}, w_{ch})\}_{ch} \equiv \mathbf{p}_{ch}$ de las clases propias $\{w_{ch}\}$de cada canal ch.

polinomios -lineales en nuestro caso-) se interpolan todas las secuencias para tener el mismo número de vectores (el correspondiente a una duración de ventana de 2 segundos para Decepción y 4 para Colima). A continuación se filtra paso-baja para suavizar las probabilidades a 0.5 de la frecuencia normalizada.

- *Filtros de duración*: incluyen filtros de duración mínima y máxima para cada evento $\mathbf{x}_c$ que pertenezca a una clase concreta, los mismos que en la Sección 3.3.1.2. Además se implementa el llamado pre-filtrado de duración, aplicado sobre la duración total de la secuencia $\mathbf{x}$, que invalida como opción de clasificación en todo el fichero aquella clase que no cumpla las condiciones fijadas.

En la Figura 5.1.2 se muestran las secuencias $\{p(\mathbf{x}, w_{ch})\}_{ch}$ una vez filtradas con un intervalo de valores comparables entre sí, que constituyen la entrada al decodificador semi-heurístico.

5.1.2.2. Segmentación y clasificación semi-heurística de eventos

Nuestro decodificador adopta un sencillo esquema de pre-segmentación basado en probabilidades logarítmicas acumuladas del clasificador GMM (Subsubsección 3.3.1.2) al que se le añade un proceso semi-heurístico de *retro-segmentación* que re-ajusta las marcas de segmentación de los eventos. Dado un sismograma registrado en un fichero, secuencialmente se va pre-filtrando, segmentando y pre-clasificando eventos, tras lo cual se re-ajustando la duración del primer evento segmentado y posteriormente se continúa el proceso de filtrado, segmentación y pre-clasificación. La idea se esquematiza en el Algoritmo 5.1 y se compone en los siguientes pasos:

1. *Pre-filtrado temporal.* Se aplican filtros de tiempo mínimo para posteriormente ignorar las clases que no se ajustan a la duración de la secuencia de frames sin clasificar, representada por la secuencia $\{\mathbf{x} \setminus \mathbf{x}_{s-i}\}$. $\{\mathbf{x}_{s-i}\}$ son los segmentos que ya han sido clasificados.
2. *Pre-segmentación y pre-clasificación.* Se hallan los puntos de pre-segmentación $\{s, ..., s+j\}$, los cuales delimitan los eventos de una clasificación inicial. El proceso se realiza computando los puntos de *equi-probabilidad* mediante:

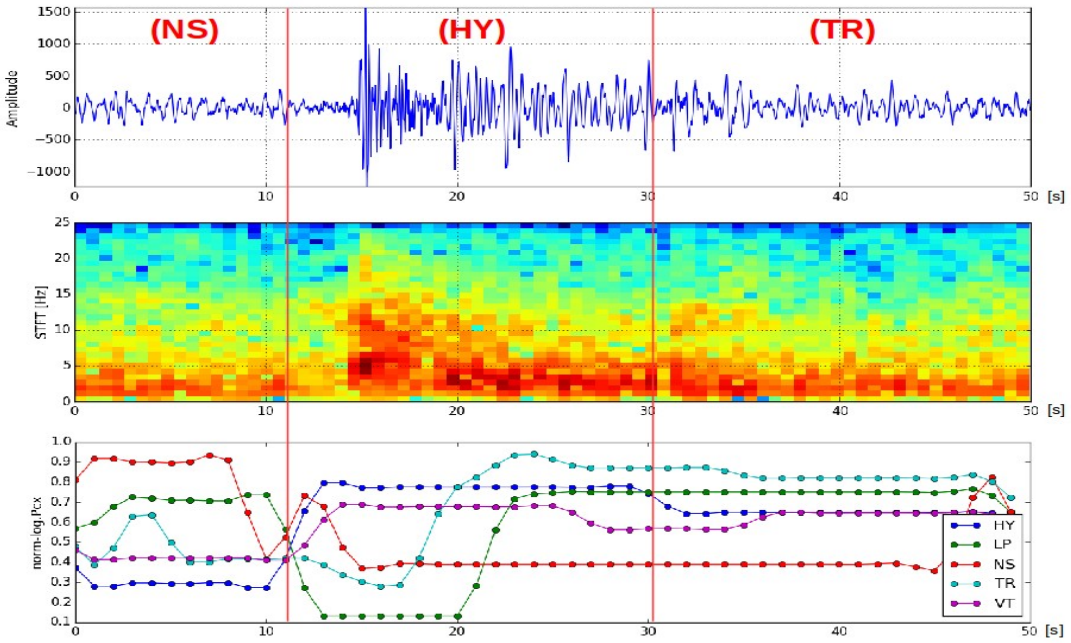


Figura 5.1.2.: VSR-PSA: probabilidades de los canales como entrada del decodificador conjunto. Cada canal ch genera una secuencia de probabilidad $\{p(\mathbf{x}, w_{ch})\}_{ch}$ de que cada vector $\mathbf{x}_t$ de características de la secuencia $\mathbf{x} = \{\mathbf{x}_t\}$ pertenezca al modelo w_{ch} de su clase propia. El decodificador conjunto PSA toma como entrada estas secuencias para segmentar el sismograma en 3 partes y clasificar cada uno de los eventos.

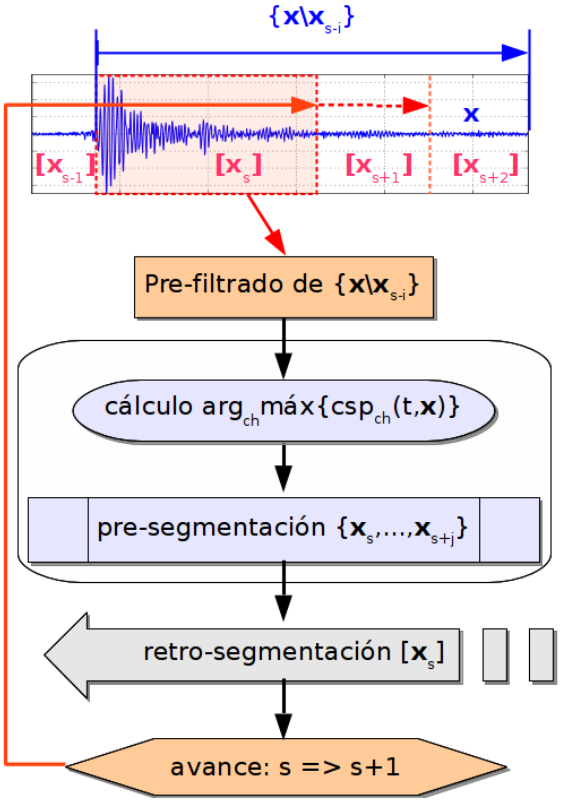
a) Para cada secuencia $\{p(\mathbf{x}, w_{ch})\}$ del canal ch se calcula su función de probabilidad logarítmica acumulada $csp_{ch}(t = T, \mathbf{x}) \equiv \sum_{t=1:T} \log p(\mathbf{x}_t, w_{ch})$.

b) Para cada frame del registro correspondiente al instante t, se calcula la clase $\hat{w}_t$ con mayor probabilidad acumulada, $\hat{w}_t \equiv \arg_{ch} \max\{csp_{ch}(t, \mathbf{x})\}$, asociando cada frame $\mathbf{x}_t$ con su clase más probable: $\mathbf{x}_t \rightarrow \hat{w}_t$, creando la secuencia de etiquetas $\{\hat{w}_t\}$.

c) Los puntos de equi-probabilidad vienen marcados por los instantes $t = \{s\}$ donde la secuencia $\{\hat{w}_t\}$ cambia de valor, $\hat{w}_{t=s} = w_A \neq w_B = \hat{w}_{t=s+1}$, lo que equivale a que en el instante $s{+}1$ el canal A, con mayor probabilidad acumulada $csp_{ch=A}(t, \mathbf{x})$, pasa a ser el de la clase B, $csp_{ch=B}(t, \mathbf{x})$.

3. *Retro-segmentación de* $[\mathbf{x}_s]$. Se ajusta la duración del evento primer evento $[\mathbf{x}_s]$ pre-clasificado. Nótese que en los frames $t \leq s$ se asocian a la clase w_A , mientras que en el instante $t = s + 1$ se asocia a w_B. Esto implica que en algún punto intermedio t_s entre estos instantes se puede interpolar que $csp_A(t_s, \mathbf{x}_s) = csp_B(t_s, \mathbf{x}_s)$, lo que equivale a decir que desde que el evento $\mathbf{x}_s$ comienza hasta el instante t_s tiene la misma probabilidad de asociarse a la clase w_A como a la w_B. Experimentalmente se traduce en que las marcas

Algoritmo 5.1 VSR-PSA: clasificación secuencial del decodificador. El flujo continuo de datos descrito por las probabilidades $\{p(\mathbf{x}, w_{ch})\}_{ch}$ prefiltradas se pre-segmenta en $\{\mathbf{x} \setminus \mathbf{x}_{s-i}\} = [\mathbf{x}_s, \mathbf{x}_{s+1}, ..., \mathbf{x}_{s+j}]$ eventos mediante los puntos $\{s\}$ de equi-probabilidad acumulada: instantes t donde un canal ch deja de tener la mayor probabilidad acumulada $csp_{ch}(t = s, \mathbf{x})\}$. En la retro-segmentación la duración del primer segmento $\mathbf{x}_s$ se acorta de forma semi-heurística. Posteriormente se continúa el proceso de pre-segmentación y ajuste para el resto de secuencia $\{\mathbf{x} \setminus \mathbf{x}_{s+1-i}\}$



que delimitan el final de los eventos están retardadas temporalmente (ver la Figura 5.1.3) y deben ajustarse a un punto anterior $(s - R)$ más adecuado. Para ello, nos desplazamos hacia atrás en el tiempo de manera semi-heurística, deslizando la marca temporal tantos frames R como sea posible siempre que se cumplan estas condiciones:

a) los frames recortados tienen más probabilidad de pertenecer a la clase B que a la A: $p(x_r, w_A) \leq p(x_r, w_B)\ \forall r = 1 : R$

b) el evento recortado sigue pasando los filtros de duración definidos en la Subsubsección 5.1.2.1

La Figura 5.1.3 nos muestra el resultado de la retro-segmentación sobre el sismograma. Observamos que las marcas que delimitan los eventos NS e HY quedan tras la etapa de pre-segmentación muy atrasadas temporalmente y que quedan aproximadamente bien ajustadas tras recortar su duración.

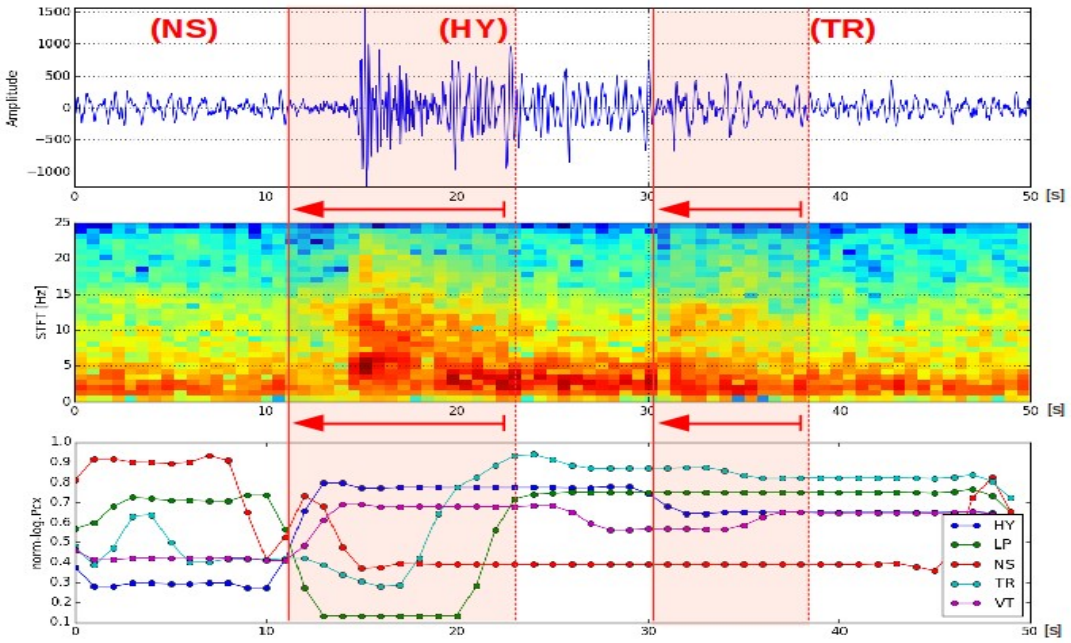


Figura 5.1.3.: VSR-PSA: retro-segmentación. Una vez pre-segmentado un fichero en los puntos de equi-probabilidad acumulada, las marcas de segmentación retornan progresivamente en el tiempo de forma semi-heurística.

4. *Avance secuencial*. Tras ajustar el tamaño del evento $\mathbf{x}_s$, el proceso se repite para el resto de frames que lo anteceden; la secuencia $\{\mathbf{x}_{t>(s-R)}\} = \{\mathbf{x} \backslash \mathbf{x}_{s+1-i}\}$.

5.1.3. Coste computacional de los sistemas VSR-PSA frente a los clásicos VSR-SSA

El proceso de paralelización del sistema lleva asociado un coste computacional extra inevitable respecto el sistema serie SSA. En principio:

- Se multiplica por 2 la complejidad en el sistema de canales binarios PSA.bin
- Se multiplica por C (el número de clases a considerar) en la arquitectura de canales múltiple PSA.mul

A esto habría que sumarle opcionalmente el coste del decodificador conjunto que aproximadamente es $\mathcal{O}\{2 \cdot T\}$, siendo T el número de observaciones (vectores de características) de la base de datos (el decodificado por Viterbi de los HMM tiene un orden de $\mathcal{O}\{T \cdot S^2\}$, con S el número de estados). Este incremento de complejidad es totalmente asumible en cualquier ordenador actual, dado la baja frecuencia de muestreo, en torno a $F_S = 100\,[Hz]$, de nuestras señales. Incluso el funcionamiento en tiempo real para 10 clases de la arquitectura PSA.mul de canales paralelo puede ejecutarse sin problemas en cualquier dispositivo portátil (móvil o tablet).

Nótese que para hace más de 10 años se podía ejecutar sin problemas un sistema de reconocimiento de voz en tiempo real en cualquier ordenador compatible con un microprocesador que funcionara a una frecuencia de 300 $[MHz]$, incluyendo la codificación de señal, que suele muestrearse con una $F_S = 16\,[kHz]$, 2 órdenes de magnitud mayor que nuestros sismos.

Hay que hacer sin embargo algunas puntualizaciones en torno al incremento de complejidad del sistema paralelo PSA respecto el serie SSA:

- Gracias al proceso de configuración específico de cada canal, tanto a nivel de topología (estados y componentes gausianas) como de selección de características, es bastante probable que la complejidad requerida por un canal PSA sea menor que la que requiera su respectivo sistema SSA para alcanzar tasas de eficiencia equivalentes.
- La flexibilidad inherente de la arquitectura PSA nos permite utilizar también otros reconocedores probabilísticos más sencillos que los HMM. En concreto, para clases con un patrón secuencial poco dependiente del tiempo (ruidos, tremores, colapsos y derrumbes) podríamos usar los GMM como modelos y una evolución del decodificador PSA.joint para reconocer en continuo.

5.2. Funcionalidades del sistema VSR-PSA

En esta sección presentaremos diversas utilidades con las que se ha dotado al sistema en paralelo propuesto. Las agruparemos en 2 categorías, en función de que bloque del esquema VSR-PSA se necesite para obtener los resultados: el análisis llevado a cabo en los canales o el clasificador conjunto. Los posibles usos derivados del sistema pueden, a su vez, dividirse en 2 grupos principales:

1. Como herramienta de *análisis*: que proporciona una información de tipo geofísico de las clases, destacando la extracción de las mejores características para describir eventos de una misma clase, así como estadísticos de tiempo, variabilidad y valores esperados de esas características.
2. Como *clasificador* de eventos: que engloba a detectores de eventos sísmicos (tipo STA/LTA), a discriminadores para una clase concreta, a reconocedores de eventos continuos, a detectores de eventos solapados y a sistemas de multi-etiquetado para un mismo registro sísmico.

5.2.1. Análisis por independiente de cada canal: detectores de clase

Dos tipos principales de estudios pueden llevarse a cabo por los canales: la selección de las características que mejor describen cada clase propia de cada canal y

la discriminación específica o reconocedor de eventos de un tipo concreto que sea especialmente interesante (explosiones, lahares, colapsos o tremores).

5.2.1.1. Selección de las mejores características para cada clase

Una de las funcionalidades más útiles del sistema VSR-PSA es la posibilidad de extraer las características más eficientes que permitan describir mejor a los eventos una misma clase. Este estudio es un primer paso al diseño de un sistema VSR no-supervisado universal, pues nos permite buscar parámetros que generalicen las clases independientemente del volcán donde hayan sido generadas.

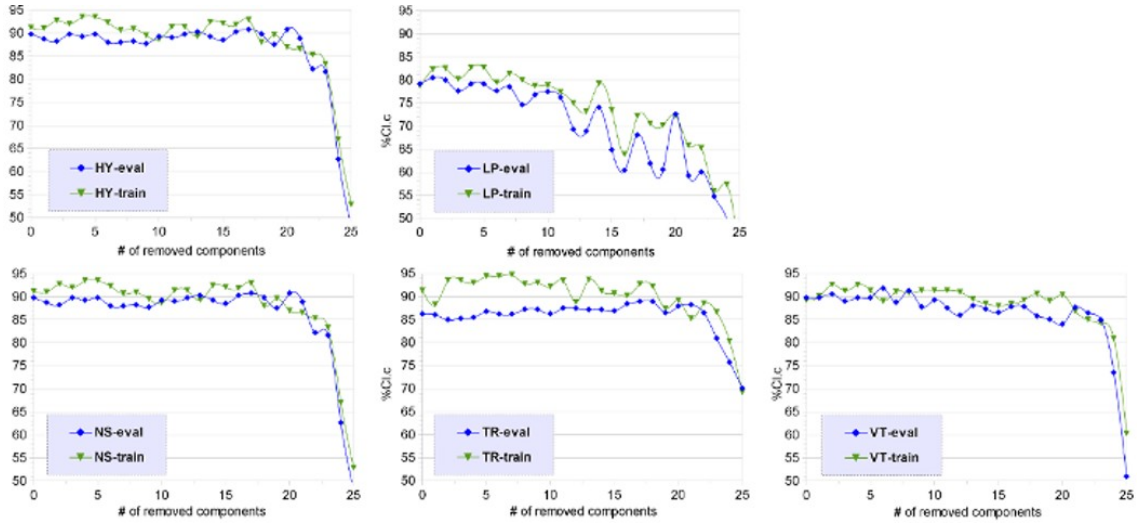


Figura 5.2.1.: Selección de características en cada canal. Se aplica el algoritmo DFS modificado en cada canal para 5 clases del volcán Decepción (Cortés et al., 2014). Se representa el % de clasificación en función del número de características quitadas del vector original, testeando tanto la DB de entrenamiento (verde) como la de evaluación (azul).

En el sistema VSR-PSA basado en GMMs propuesto por Cortés et al. (2014) se seleccionan las mejores características para cada una de las clases del volcán Decepción mediante la modificación del algoritmo DFS original propuesto por Álvarez et al. (2011) presentado en la Subsección 4.3.2 de esta tesis. La Figura 5.2.1 muestra la eficiencia del DFS: la tasa de clasificación en los canales se mantiene casi constante incluso después de quitar hasta el 90 % de las componentes originales del vector de descripción.

5.2.1.2. Detectores específicos de una clase concreta

La gran ventaja que tiene la arquitectura PSA de canales paralelos sobre el sistema serie SSA es que, más allá de los resultados de reconocimiento que al final son

evaluados conforme un etiquetado supervisado en el sistema SSA que puede ser poco fiable (Subsección 2.2.2), los canales PSA funcionando como detectores específicos para eventos de un tipo concreto tienen la capacidad de detectar eventos que el sistema serie ignora.

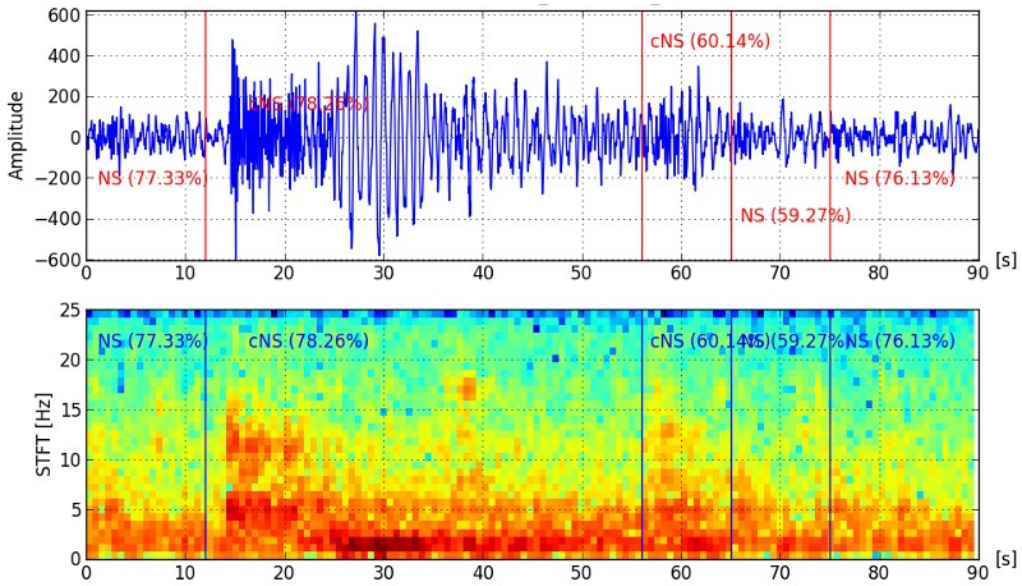


Figura 5.2.2.: Canal PSA binario *PSA.bin.NS* especializado en discriminar ruidos. Se distingue entre eventos ruidosos (NS) y el resto de eventos (cNS), incluyendo para cada evento valores de fiabilidad en su etiquetado. En la arquitectura serie SSA las 3 últimas etiquetas son: $\{NS(75{,}48\,\%) \mid HY(66{,}67\,\%) \mid NS(73{,}04\,\%)\}$. El canal especializado es capaz de distinguir en el último evento etiquetado 3 sub-eventos: $NS(73{,}04\,\%) \rightarrow \{cNS(60{,}14\,\%) \mid NS(59{,}27\,\%) \mid NS(76{,}13\,\%)\}$.

La Figura 5.2.2 muestra como el canal PSA.bin.NS es capaz de segmentar 3 eventos $\{cNS(60{,}14\,\%) \mid NS(59{,}27\,\%) \mid NS(76{,}13\,\%)\}$ donde el sistema SSA solo detecta uno $\{NS(73{,}04\,\%)\}$. Nótese que es muy probable que el segmento etiquetado como $\{cNS(60{,}14\,\%)\}$, a juzgar por su espectrograma, quizás debiera ser considerado como un evento híbrido al igual que su predecesor. De hecho, incluso el segmento $\{NS(59{,}27\,\%)\}$ pueda corresponderse con otro evento híbrido más pequeño. Los valores de robustez de etiquetado proporcionan una información muy valiosa: un valor por debajo del $70\,\%$ podrá ser considerado como el límite de fiabilidad para desechar el etiquetado dado por el sistema automático, considerar ignorar estas señales o guardar todos los segmentos que no alcanzan ese umbral para posteriormente estudiarlos y, si es apropiado, crear una nueva clase de *micro-híbridos*. Todos estos detalles dan a los expertos información adicional a la hora de analizar más en profundidad una base de datos o la sismicidad de una clase en concreto.

5.2.2. Análisis en conjunto de los canales

En este caso utilizamos la salida del decodificador PSA.joint como una salida de cualquier sistema SSA, más orientada al reconocimiento de sismos que al análisis. Sin embargo, obtenemos algunas ventajas sobre el sistema serie clásico:

- *Robustez* y *fiabilidad* que proporcionan los canales PSA, al estar específicamente diseñados para reconocer una clase.
- *Tasas de fiabilidad* o pertenencia a una clase para cada evento reconocido, promediadas a todo el evento o a nivel de cada uno de los segmentos del evento. Si bien estas pueden estimarse también en sistemas serie (Sección A.7), los resultados obtenidos a partir de los canales son mucho más fiables. Como veremos en los próximos apartados esta información detallada permite utilizar la salida del decodificador para el etiquetado múltiple, la detección de eventos solapados o la detección de eventos que no se ajustan a ninguna de las clases definidas en el sistema.

Utilizado exclusivamente como un reconocedor clásico, en la implementación práctica de la arquitectura PSA que detallamos en el Capítulo 6 comprobamos que los resultados del decodificador conjunto son comparables (incluso algo mejores, Tabla 6.6.1) que su homólogo en serie, incluso no habiendo sido configurado para ello. Es un buen punto de partida teniendo en cuenta que decodificador PSA tiene un sencillo diseño susceptible de ser mejorado.

En la Figura 5.2.3 observamos el comportamiento general de los canales (de tremor pulsante TP en este caso, $PSA.*.TP$) binarios y múltiples respecto a la salida conjunta ofrecida por el decodificador PSA ($PSA.*.JOIN$). Los canales, que usan como reconocedor los HMM y decodifican mediante el algoritmo e Viterbi (Sección 3.3.2.1), tienden a insertar más eventos que el decodificador PSA. Los canales múltiples y la salida conjunta tanto de los canales binarios como múltiples ofrecen más alternativas de cara a un etiquetado múltiple.

5.2.2.1. Etiquetado múltiple o alternativo

La opción de asignar distintas etiquetas por su tasa de confianza a un mismo evento aislado o presentar distintas secuencias de eventos como distintas alternativas en un reconocimiento continuo es lo que se entiende por *etiquetado múltiple*. Esta información puede resultar muy útil en sistemas de interacción hombre-máquina donde se ofrezcan distintas posibilidades de reconocimiento a un sistema experto que interactue con una persona para que pueda escoger entre una opción u otra en función de información adicional (distintos modelos de lenguaje o situación de contexto). En nuestro ámbito de interés esta funcionalidad es muy útil en el análisis de corpus de los que no se tiene mucha información geofísica a priori, como por ejemplo, un escenario de reconocimiento no supervisado en el que usan modelos construidos con eventos de un volcán distinto a donde se está reconociendo y el sistema presenta

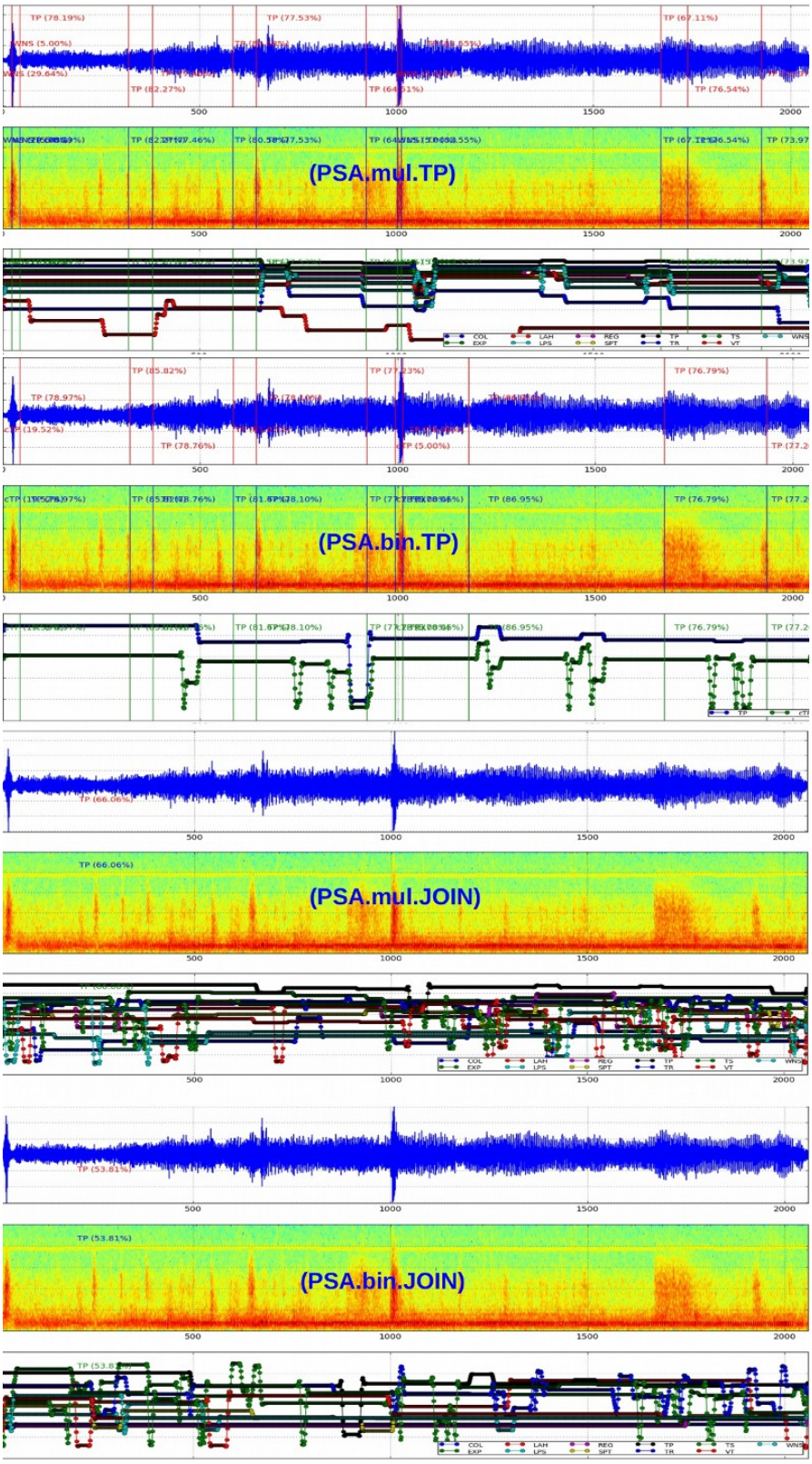


Figura 5.2.3.: Canales PSA vs. decodificador conjunto. Los decodificadores conjuntos (figs. inferiores, $PSA.*.JOIN$) tienden a insertar menos eventos que los decodificadores por Viterbi de cada canal (figs. superiores, $PSA.*.TP$). A su vez, los canales multiclase ($PSA.mul.*$) ofrecen más opciones de cara a un etiquetado múltiple que los binarios ($PSA.bin.*$).

varias *alternativas de etiquetado*, conforme a un criterio bien establecido basado en las probabilidades normalizadas de los canales.

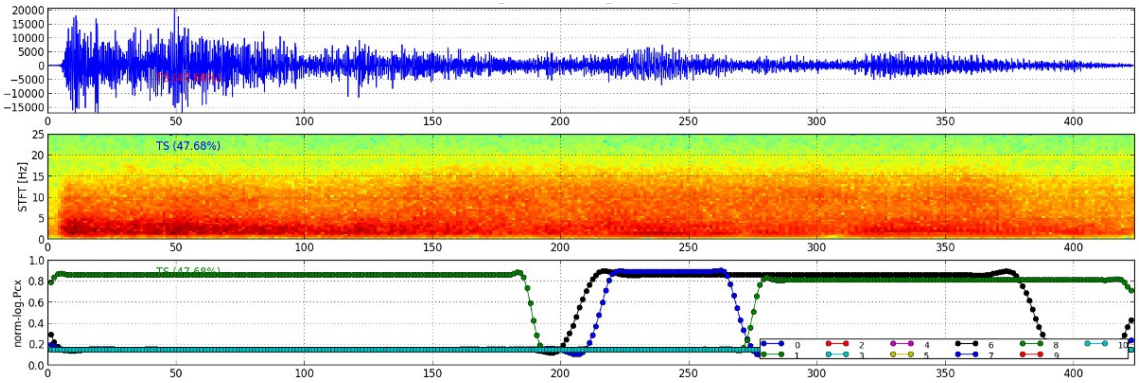


Figura 5.2.4.: Salida del codificador conjunto PSA.joint: etiquetado alternativo. El técnico experto etiqueta el evento como un tremor espasmódico $\{TS\}$. Sin embargo, el decodificador PSA tiende a segmentar el evento en $\{TS/TR/TP/TS\}$, proporcionando un etiquetado alternativo a considerar.

En la Figura 5.2.4 el sistema PSA sugiere un etiquetado alternativo al inicialmente realizado por el técnico experto. A partir de los valores de probabilidad normalizada o *tasas de confianza* de clase $\{p(\{\mathbf{x}\}, w_c)\}$ para cada modelo w_c que se generan en cada canal PSA específico, el decodificador conjunto segmenta en varios tremores $\{TS/TR/TP/TS\}$ mientras que en el etiquetado supervisado solo se asigna a uno $\{TS\}$, probablemente debido al. Un análisis detallado concuerda con el etiquetado automático del sistema.

Reconocimiento *N-best* frente al análisis por canales PSA. Los sistemas clásicos basados en modelos probabilísticos de clasificación de un evento aislado $\mathbf{x}$ ofrecen también esta posibilidad simplemente ordenando sus etiquetas de clase w_c decrecientemente conforme a los valores $\{p(\mathbf{x}, w_c)\}$ obtenidos para cada clase c. Esta funcionalidad se complica algo en el reconocimiento continuo de un registro con una secuencia$\{\mathbf{x}\}$ observaciones, donde los sistemas que reconocen en un espacio de estados pueden dar las N secuencias de estados o *caminos* que han obtenido más probabilidad en el proceso de decodificación (reconocimiento $N-best$, Young et al. (2006); Rabiner and Schafer (2007)). Nótese, que esto no equivale a tener los valores de probabilidad o confianza por independiente $\{p(\{\mathbf{x}\}, w_c)\}$ para cada modelo w_c como obtenemos con nuestro sistema PSA (Figura 5.2.4); en el caso de la red de búsqueda del algoritmo Viterbi en los HMM (Figura 3.3.3) una o varias de esas N secuencias de estados más probables pueden corresponder a distintos caminos del mismo modelo incluso en reconocimiento en aislado.

5.2.2.2. Detección de eventos solapados

Entendemos por un *evento solapado* aquel que se presenta al mismo tiempo en el sismograma que otro *evento de fondo*. Este fenómeno dificulta la efectividad de reconocimiento en los sistemas serie SSA, pero es compensado en los sistemas PSA gracias a que cada canal se diseñan de forma específica para discriminar un tipo de evento concreto, ganando en robustez al decodificar de forma conjunta. Un caso particular de eventos solapados es la detección de señales tal y como lo entendemos desde la perspectiva clásica: se detecta (o reconoce) cualquier evento que esté solapado al evento de fondo *ruido*. La Figura 5.2.5 muestra el ejemplo contrario: el decodificador PSA es capaz de detectar un evento solapado (en este caso un ruido instrumental) sobre un lahar.

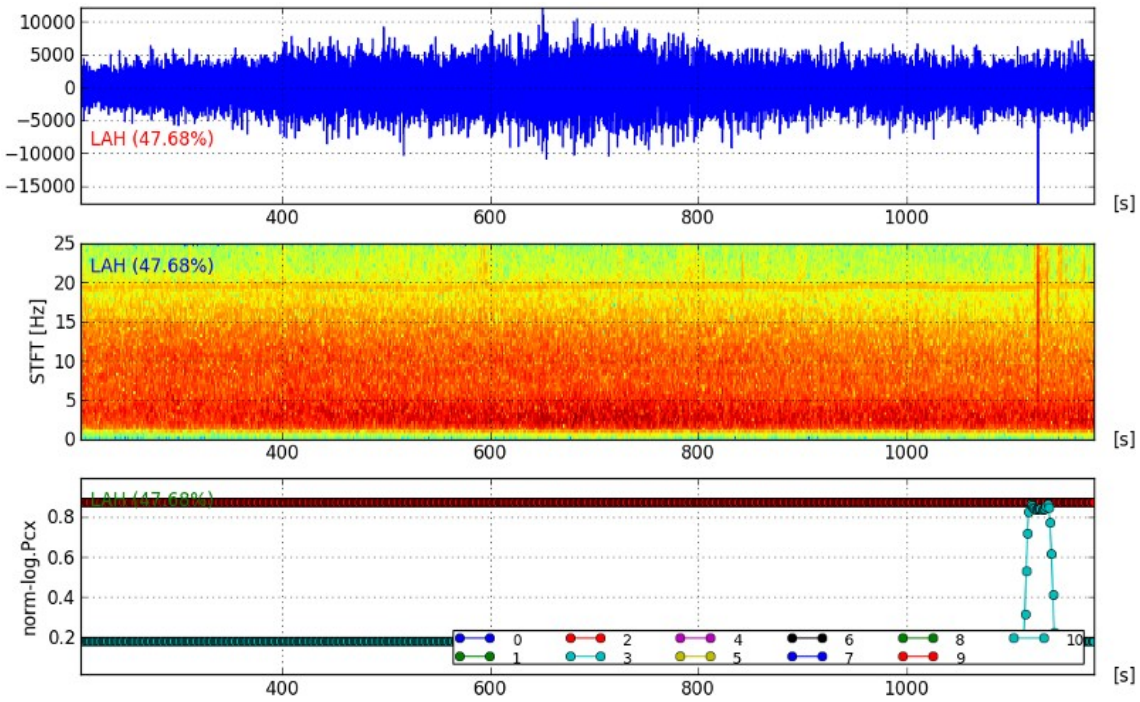


Figura 5.2.5.: Salida del codificador conjunto PSA.joint: eventos solapados. El decodificador PSA detecta un ruido {*WNS*} (probablemente instrumental) superpuesto al lahar {*LAH*}.

Si conocemos a priori el tipo de evento que queremos discriminar del evento de fondo la detección de eventos solapados puede también realizarse desde el canal específico de ese evento; por ejemplo, en el caso que queramos detectar enjambres de terremotos VTs solapados a un tremor. De hecho, sería la forma más correcta de proceder: usar el canal PSA.bin.VT para reconocer VTs solapados a cualquier otra clase. Si de forma general queremos estudiar la existencia del fenómeno de solapamiento en un corpus es más adecuado analizar la salida del decodificador conjunto PSA.

5.2.2.3. Definición de nuevas clases: análisis de eventos con baja confianza

Como ya comentamos en la Subsubsección 5.2.1.2, podemos definir un umbral en la tasa de fiabilidad asociada a cada evento reconocido por el sistema PSA. Esto puede hacerse indistintamente en la salida que proporcionan los canales o en la que da el decodificador conjunto, según busquemos sub-eventos de una clase concreta (distinto tipo de tremores o de eventos LP) o, en general, eventos que no pueden asignarse a las clases previamente definidas como ventiscas, ruido generado por el hombre o por instrumentos de registrado, aviones, tráfico, etc...

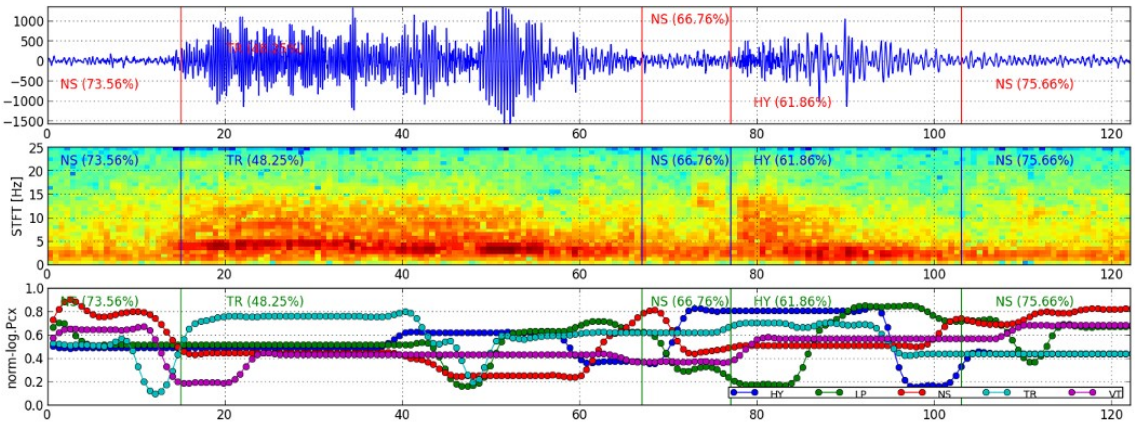


Figura 5.2.6.: Salida del codificador conjunto PSA.joint: clases no consideradas. Se muestran las etiquetas asignadas por el técnico experto y las probabilidades normalizadas $norm.log.Pcx$ de cada segmento $\mathbf{x}$ de pertenecer a cada una de las c clases. Observamos que en el evento segmentado como tremor espasmódico $\{TR\}$ el decodificador conjunto señala componentes de baja frecuencia $\{LP\}$ que puede interpretarse como un tremor armónico.

A modo de ejemplo, la Figura 5.2.6 detalla como la salida de reconocimiento del decodificador conjunto PSA muestra componentes de baja frecuencia (asignando segmentos a las clases LP o, incluso HY) en el evento etiquetado de forma supervisada por el técnico experto como tremor espasmódico (TR). Esta información pudiera ser considerada para definir una nueva clase en el corpus: el tremor armónico.

Parte III.

APLICACIONES DEL SISTEMA VSR-PSA

6. Sistema paralelo VSR-PSA como detector específico

En los próximos apartados vamos a construir desde cero un sistema de reconocimiento de sismos en paralelo (VSR-PSA) presentado teóricamente en el Capítulo 5. Con el objetivo de comparar resultados y extraer conclusiones generalizables a otros sistemas, al igual que hicimos con el sistema base en serie VSR-SSA (Capítulo 3), aplicaremos la misma técnica a dos volcanes muy distintos, el de la isla Decepción y el volcán de Fuego de Colima.

Empezaremos por definir la topología de los modelos HMM independientemente para cada tipo de evento a reconocer. Posteriormente vamos a seleccionar las características que mejor describan la clase propia de cada canal mediante el algoritmo discriminativo DFS generalizado detallado en el Capítulo 4. Analizaremos los resultados de esta selección entre las clases en común de los dos volcanes estudiando si pueden o no ser generalizables a otros volcanes. Una vez escogidos los 15 mejores descriptores para cada canal, los configuraremos por independiente para maximizar la eficiencia de reconocimiento de sus correspondientes clases propias y compararemos con los resultados del sistema base que usa 30 características.

6.1. Metodología y objetivos

En este capítulo estudiaremos experimentalmente cómo se configura y se comporta la arquitectura VSR-PSA paralela propuesta en el Capítulo 5 frente a la arquitectura VSR-SSA serie vista en el Capítulo 3. El sistema paralelo se configurará bajo el esquema $max\{PSA.class\}$ (Sección 5.1) que persigue optimizar cada canal para obtener la máxima eficiencia para reconocer los eventos pertenecientes a su clase propia funcionando como un conjunto de *detectores específicos* para cada tipo de señal. Estos detectores no solo serán capaces de aislar un evento del ruido de fondo (que es lo conseguido con los detectores clásicos), si no que están diseñados para discriminar la clase propia del resto de clases. Este esquema de configuración es especialmente útil en el análisis de los eventos, el etiquetado semi-supervisado y el estudio del solapamiento de eventos.

Adicionalmente estudiaremos las mejores características obtenidas para cada clase y compararemos la eficacia de reconocimiento obtenida en las dos arquitecturas, evaluando la mejora conseguida gracias al sistema VSR-PSA. Al igual que procedimos con la arquitectura serie (Capítulo 3) construiremos dos sistemas, uno para la base de datos del volcán de Decepción (*dec.95Mc*) y el otro para el volcán de Colima (*col.04Mc*). Ambas corresponden a la versión en continuo presentadas en Sección 3.1, con lo que los sistemas además de clasificar tendrán que detectar los eventos sobre un flujo continuo de señal.

Todo el proceso de configuración, construcción y evaluación se hará progresivamente en las siguientes etapas detalladas en Subsección 5.1.1:

1. **Topología (HMM)**. Construcción de modelos del canal: topología y componentes gausianas de los HMMs
2. **Descripción (DFS)**. Selección de características mediante el algoritmo *DFS.cAcc* generalizado
3. **Filtrado (BPF)**. Análisis de la banda óptima para el filtrado espectral BPF paso-banda
4. **Ventaneo** o *Windowing* **(WIN)**. Selección de la duración del segmento de parametrización

Nótese que los todos los pasos se realizan tanto en la arquitectura VSR-SSA serie como en la VSR-PSA en paralelo. La selección de la topología de los HMMs es común tanto para el sistema serie como para el paralelo. El filtrado y ventaneo ya se estudiaron para el sistema SSA en el Capítulo 3 al construir el sistema base y en el Capítulo 4 de reducción de dimensionalidad, escogiendo unos valores de $[1, 25]$ Hz para el filtrado y una duración de segmento de 2 segundos para Decepción y 4 para Colima. Todas las etapas se realizarán de forma independiente en cada uno de los canales VSR-PSA.

Escogeremos las mejores 15 características en cada canal proporcionadas por la selección mediante *DFS.cAcc* de entre las 30 componentes que finalmente se se-

leccionaron en la Sección 4.2 conformando el esquema mixto *geoLFCC.D*.30 de parametrización base. La banda espectral, la duración de la ventana de parametrización, y, en general, los valores óptimos de configuración y resultados dependen en gran medida de la naturaleza y número de características que se extraigan de cada segmento. Usamos 15 de las 30 características iniciales en base a elegir un tamaño de vector intermedio. Para un estudio más profundo del tema, nos remitimos a Cortés et al. (2014), que comparan distintos esquemas DFS de arquitecturas PSA con diversas parametrizaciones serie SSA.

En cuanto a las pruebas experimentales cabe destacar:

- Usaremos la misma metodología experimental descrita en la Subsección 4.1.6 de reducción de dimensionalidad. Todos los resultados están promediados entre 3 particiones de la base de datos, usando 2 de ellas para construir los modelos y la restante para evaluarlos. Se promedia además entre modelos HMMs construidos respectivamente para 1,2,4,8 y 16 componentes gausianas.
- Cada tipo de canal del sistema PSA tiene su propia configuración; distinguiremos entre esquemas de configuración para canales múltiples (*PSA.mul*) y binarios (*PSA.bin*).

6.2. Diseño de la topología de los HMMs

Dividiremos en dos secciones el estudio y evaluación de la estructura de los modelos; una, la *topología* propiamente, que se basa en el conocimiento a priori geofísico que tenemos de nuestras bases de datos y otra, la *selección de componentes gausianas* para cada HMM, en la que nos serviremos de una metodología experimental descrita en la sección anterior para escoger los mejores valores.

6.2.1. Topología de los HMMs

Basándonos en las propiedades generales de las señales sísmicas registradas en los volcanes vistas en el Capítulo 1 y más específicamente el extraído de la observación y análisis los eventos de las bases de datos de Colima y Decepción sobre las que hemos construido nuestros sistemas de referencia (Sección 3.1), vamos a definir el nº de estados y los enlaces que hay entre ellos considerando ciertas cuestiones relativas a los HMM (Subsección 3.3.2):

- **Estados**. Básicamente, el nº de *estados emisores* de vectores de características tiene que definirse acorde a los distintos sub-patrones que dentro de una misma clase de eventos podamos encontrar en el espacio de características Rabiner and Juang (1986). De forma más experimental, un criterio complementario usado por algunos autores (Ibáñez et al., 2009; Cortés et al., 2009b; Gutiérrez Espinoza, 2013) atiende a aumentar el nº de estados para disminuir el nº de

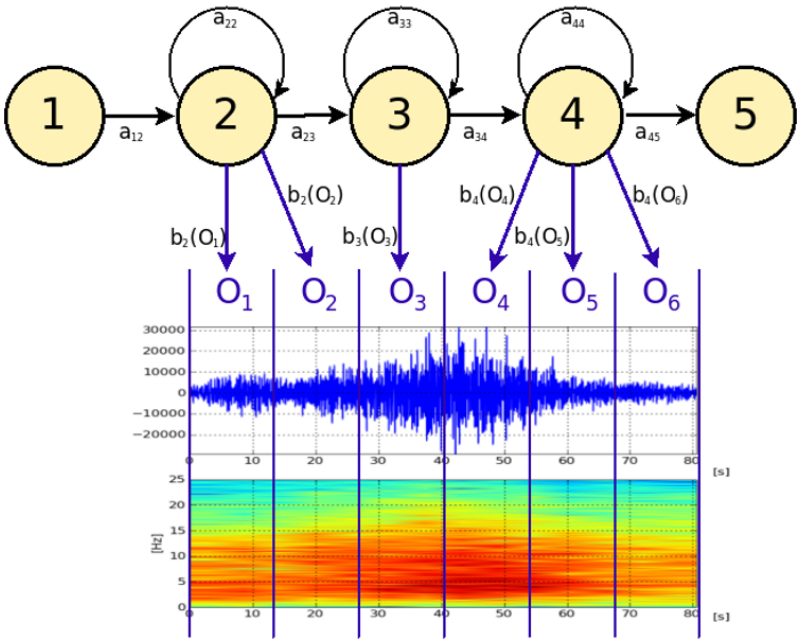


Figura 6.2.1.: Topología de los HMMs. en las clases secuenciales (VTs, LPs, REGs...) cada estado s está enlazado a sí mismo y al estado que le precede con una probabilidad de transición de evolucionar en cada instante t que se genera un vector de características O_t dada por $a_{s,s}$ y $a_{s,s+1}$ respectivamente. En las clases no secuenciales o cíclicas (NSs, TRs, COLs, LAHs...) existe una probabilidad $a_{4,2}$ de que desde el último estado emisor Q_4 se pase al primero Q_2, permitiendo un comportamiento cíclico del modelo.

inserciones. De forma general, puede afirmarse que el número óptimo de estados debe ser proporcional a la variabilidad en el espacio de características y variabilidad temporal de las clases, o, conjuntamente, proporcional a la variabilidad de una clase en su espacio de descripción. Siempre se añaden 2 estados más *no emisores* de interconexión para conectar los modelos HMM en la red de búsqueda y en el reconocimiento en continuo.

- **Enlaces**. De forma general, distinguiremos entre 2 tipos de clases:
  - *clases temporales* o *secuenciales*, que tienen una clara evolución de sus vectores de características de manera secuencial en el dominio temporal. Entre ellas las conforman casi todas las que tienen un origen estrictamente sísmico como los terremotos en los que puede distinguirse claramente la llegada secuencial del frente de ondas P, S, ondas superficiales, reflexiones y coda.
  - *clases no temporales* (*no secuenciales*): no poseen una clara evolución en sus características conforme evoluciona el tiempo. Ejemplos de ellas son los eventos ruidosos, los tremores y colapsos y lahares. Su reconocimiento

automático suele acarrear varios borrados e inserciones complicando la evaluación de resultados Figura 2.2.2.

Para adecuar el modelo a las propiedades de nuestras señales estudiadas en la Subsección 1.1.2 y en la Subsección 2.2.1, definiremos un HMM con una topología secuencial de izquierda a derecha indicada en la Figura 6.2.1 cuyos estados tras generar el vector de características puedan evolucionar al estado siguiente o a sí mismos. En el caso de las clases no secuenciales y con el objetivo de reducir inserciones permitiremos que el mismo modelo vuelva a generar vectores una vez que se ha llegado al final del mismo, existiendo una probabilidad de transición no nula de que desde el último estado emisor se pase al primero.

En un contexto general, la selección de estados debe asociarse al *modelado del dominio de características* mientras que la estructura de los enlaces debe adaptarse al *modelado temporal* de la clase. En función de ello, la Tabla 6.2.1 detalla en cada caso y para cada base de datos la topología propuesta en este capítulo. Se denota con el símbolo (->) las clases secuenciales y con (<>) las que incorporan un lazo recursivo desde el último estado emisor al primero. En cuanto a los estados, en todas las clases se incluyen los 2 no emisores y se añaden otros 2 de inicio y final de evento que asociamos como estados intermedios entre los eventos consecutivos a uno dado (en la mayoría de los casos el ruido), siendo el mínimo nº de estados según este criterio 5 (2 no emisores, 2 intermedios y 1 propio de cada clase). En clases de duración extensa como LAHs, COLs, EXPs y tremores se añaden más estados por la posibilidad de que contengan eventos solapados en ellas que compliquen su modelado en las características.

HMM.30: topología @ *dec.95Mc*					
	HY	*LP*	*NS*	*TR*	*VT*
estados	9	6	5	7	8
evolución	->	->	<>	<>	->

HMM.30: topología @ *col.04Mc*											
	COL	*EXP*	*LAH*	*LPS*	*REG*	*SPT*	*TP*	*TR*	*TS*	*VT*	*WNS*
estados	7	8	8	6	9	6	8	7	7	8	5
evolución	<>	->	<>	->	->	<>	<>	<>	<>	->	<>

Tabla 6.2.1.: *Topología de cada modelo HMM.* Se representa el nº de estados (incluidos los emisores) escogido para cada clase y si esta tiene una marcada evolución temporal (->) o no (<>). En caso negativo se añade un enlace desde el último estado al primero en el correspondiente HMM.

Nótese que la topología escogida para cada modelo HMM_c que represente a la clase c, es común en todas las arquitecturas SSA serie y PSA paralelo así como en todos

los canales. En caso del modelado de las clases $\{w_{ch}, \bar{w}_{ch}\}$ de un canal ch binario, la clase complementaria $\bar{w}_{ch}$ a la clase w_{ch} propia del canal promedia el nº de estados de las clases a las que representa y se define sin lazo recursivo.

6.2.2. Selección óptima del nº de gausianas

En esta etapa inicial de configuración $HMM.30$ (o $HMM\{geoLFCC.D.30\}$ de forma extendida), vamos a escoger el número de componentes gausianas de cada modelo HMM en función de la eficiencia $\%cAcc$ de reconocimiento resultados que arrojen las pruebas experimentales. Para ser consecuentes con la metodología detallada en la Sección 6.1, el número de gausianas óptimas se redondea a valores que sean potencia de 2. Este número se halla independientemente en las arquitecturas SSA y en cada canal PSA binario y múltiple, seleccionando en cada caso el valor experimentalmente mejor convenga a la clase propia de cada canal. Para el sistema SSA se escoge para el modelo de clase w_A el número de componentes que mayor $\%cAcc$ obtiene para dicho modelo.

La Tabla 6.2.2 muestra la selección realizada tras ejecutar los test de configuración. Es interesante constatar que la mayoría de las clases propias $\{w_{ch}\}$ necesitan ser modeladas con un mayor número de componentes en los canales binarios que en los múltiples. Esta complejidad añadida probablemente se deba a que al solo existir 2 clases, $\{w_{ch}, \bar{w}_{ch}\}$, en un canal binario PSA.bin(ch) sus correspondientes modelos tengan que ser más complejos para cubrir la variabilidad del espacio de características $\Omega_{\mathbf{x}}$ que los modelos múltiples que representan a las clases $\{w_c\}_{ch}$ en su homólogo canal múltiple PSA.mul(ch) al cubrir exactamente el mismo espacio $\Omega_{\mathbf{x}}$.

HMM.30: gausianas @ *dec.95Mc*						
	HY	*LP*	*NS*	*TR*	*VT*	media
SSA / PSA.mul(c)	2	16	1	2	8	*6*
PSA.bin(c)	8	8	2	1	2	*4*

HMM.30: gausianas @ *col.04Mc*												
	COL	*EXP*	*LAH*	*LPS*	*REG*	*SPT*	*TP*	*TR*	*TS*	*VT*	*WNS*	media
SSA / PSA.mul(c)	8	16	8	4	4	8	2	4	4	8	8	*7*
PSA.bin(c)	2	16	8	4	16	2	8	16	8	16	16	*10*

Tabla 6.2.2.: *Selección óptima de gausianas para cada clase.* Se muestran el número de componentes de cada modelo HMM para las clases $\{w_c\}$ del sistema serie SSA y las clases propias $\{w_{ch}\}$ de los canales $PSA.mul(c)$ múltiples y $PSA.bin(c)$ binarios.

En la Tabla 6.2.3 se recoge la eficiencia $\%cAcc$ de las clases propias $\{w_c\}$ en las arquitecturas serie SSA y paralelo $PSA.mul(c)$ múltiple y $PSA.bin(c)$ binario y se compara con la alcanzada por los mismos sistemas de referencia $BASE.30$ sin

optimizar las gausianas. En ambos casos se hace un promedio sobre los valores *%cAcc* asociados a cada uno de los 5 pasos de incremento de gausianas cuando se construyen los modelos (1,2,4,8 y 16 componentes en la etapa *BASE*.30 y de forma específica según indica la Tabla 6.2.2 para cada canal en la *HMM*.30). Como puede comprobarse en esta 1ª etapa de configuración, los resultados obtenidos para el sistema serie SSA son (y deben ser) exactamente los mismos que los obtenidos en cualquier canal múltiple de la arquitectura PSA pues al no haberse configurado nada de forma específica, los canales múltiples PSA son en realidad una repetición del sistema SSA. No ocurre así en los canales binarios que solo construyen dos modelos de clases por canal, en vez de un modelo para cada clase como los sistemas SSA y PSA.mul(c). Este es el mismo motivo por el cual las gausianas óptimas de los modelos coinciden en la Tabla 6.2.2 para las arquitecturas SSA y PSA.mul(c).

BASE.30 vs. *HMM.30* (*%cAcc*) @ *dec.95Mc*						
SSA	*HY*	*LP*	*NS*	*TR*	*VT*	media
BASE.30	56.83	54.78	21.03	45.32	56.92	*46.97*
HMM.30	66.32	40.96	40.10	52.78	55.79	*51.19*
PSA.mul(c)	*HY*	*LP*	*NS*	*TR*	*VT*	media
BASE.30	56.83	54.78	21.03	45.32	56.92	*46.97*
HMM.30	64.08	54.78	39.39	55.78	57.86	*54.38*
PSA.bin(c)	*HY*	*LP*	*NS*	*TR*	*VT*	media
BASE.30	49.02	60.39	46.24	53.74	56.13	*53.10*
HMM.30	52.70	59.81	55.77	68.77	54.91	*58.39*

BASE.30 vs. *HMM.30* (*%cAcc*) @ *col.04Mc*												
SSA	*COL*	*EXP*	*LAH*	*LPS*	*REG*	*SPT*	*TP*	*TR*	*TS*	*VT*	*WNS*	media
BASE.30	41.54	18.68	33.33	91.71	87.11	79.98	-34.57	3.41	3.27	85.96	13.92	*38.58*
HMM.30	43.56	-10.34	30.67	90.44	80.44	81.10	-10.77	-23.69	-7.64	87.02	-8.83	*32.00*
PSA.mul(c)	*COL*	*EXP*	*LAH*	*LPS*	*REG*	*SPT*	*TP*	*TR*	*TS*	*VT*	*WNS*	media
BASE.30	41.54	18.68	33.33	91.71	87.11	79.98	-34.57	3.41	3.27	85.96	13.92	*38.58*
HMM.30	43.56	18.68	33.33	91.40	85.33	79.73	-24.10	7.58	6.42	86.67	15.07	*40.33*
PSA.bin(c)	*COL*	*EXP*	*LAH*	*LPS*	*REG*	*SPT*	*TP*	*TR*	*TS*	*VT*	*WNS*	media
BASE.30	13.10	-160.87	30.67	81.76	-32.73	73.50	-57.82	-103.27	-42.24	35.09	14.71	*-13.46*
HMM.30	11.97	-160.87	30.67	81.13	24.00	75.83	-57.31	-103.27	-43.45	35.09	10.58	*-8.69*

Tabla 6.2.3.: *Etapa de configuración HMM.30: gausianas óptimas de los modelos HMM.* Se presenta el % de las clases $\{w_{ch}\}$ propias de los canales *PSA.mul(c)* múltiples y *PSA.bin(c)* binarios para las etapas de configuración *HMM*.30 frente a los sistemas de referencia *BASE*.30 (sin optimizar gausianas) en las arquitecturas *SSA* - serie y PSA - paralelo.

A partir de los resultados mostrados cabe puntualizar:

- En ambas bases de datos los resultados siempre mejoran respecto a los resultados de referencia $BASE.30$ en un promedio entorno al 10 % para la clase propia de los canales al optimizar sus componentes en la etapa $HMM.30$.
- Esta mejoría no tiene que darse necesariamente en los sistemas SSA; al evaluar la etapa $HMM.30$ cada modelo serie tiene su propio número optimizado de componentes en el mismo paso de evaluación, mientras que todos los modelos de un mismo canal PSA usan el mismo número de gausianas, el que más conviene a la clase propia w_{ch} del canal ch. Al igual que en la evaluación PSA, en las pruebas experimentales todos los modelos tienen el mismo número de componentes en cada paso. Teniendo en cuenta que la elección se hace en función de los experimentos de configuración, generalmente los resultados de reconocimiento para los modelos de las clases propias en la etapa $HMM.30$ suelen mejorar en promedio respecto la etapa $BASE.30$ mientras que los de los sistemas SSA no tienen a priori por qué, pues el test de configuración y el test de evaluación no se ejecutan en las mismas condiciones.
- Al realizar reconocimiento en continuo evaluado con el criterio de eficiencia promediada, $\%cAcc$, es muy común encontrarnos en Colima ($col.04Mc$) con valores negativos debido a las inserciones de eventos. A priori, este hecho necesariamente no significa desde un punto de vista geofísico que el reconocimiento sea alarmantemente pobre, cómo ya se indicó en la Figura 2.2.2
- Los resultados de los canales no siempre mejoran debido probablemente al redondeo (a la baja) realizado en el número óptimo de componentes gausianas para que coincida con potencias de 2. Este valor redondeado puede no ser óptimo en alguna partición, llegando incluso a bajar el promedio esperado en los test de configuración.

6.3. Selección de características

Una vez fijadas las gausianas óptimas para cada modelo estamos en disposición de ordenar la relevancia de las 30 características de la parametrización $geoLFCC.D.30$ usado en las etapas anteriores de configuración. El algoritmo de selección de características $DFS - recog = cAcc$ de direccionalidad creciente (Subsubsección 4.3.2.2), abreviado como $DFS.cAcc$, realizará esta tarea. La selección es independiente para la arquitectura SSA y para cada uno de los canales de los sistemas PSA.mul(c) múltiple y PSA.bin(c) binario. Los modelos HMM de un mismo canal son construidos con el número de gausianas óptimas definido en la etapa anterior de configuración $HMM.30$ (Tabla 6.2.2). Dividiremos esta sección en 2 partes; en la 1ª de ellas listamos y evaluamos la selección hecha por el DFS de las 15 mejores de las 30 características posibles, por lo que nos referiremos a esta etapa de configuración como $DFS.15$. En la 2ª analizaremos el efecto de la selección DFS comparando de forma general el sistema de referencia y todas las posibles arquitecturas PSA. Las tablas

de resultados completas para todas las componentes del vector *geoLFCC.D*.30 se presentan en el Apéndice B.

DFS.15: mejores 15 características de los canales *PSA.mul(c)* @ *col.04Mc*

COL	EXP	LAH	LPS	REG	SPT	TP	TR	TS	VT	WNS
lfcc1.D	f.80.D	f.nac	f.Slo	f.Slo	*f.Slo*	f.80	f.50.D	f.Slo	f.Slo	*htkE.D*
f.80.D	lfcc1	*lfcc1*	*lfcc1*	f.Max	f.Max	*f.Slo*	lfcc2.D	*lfcc1*	lfcc1	*fSlo.D*
lfcc2	f.80	f.Slo	f.nac	f.kur	*LPFerr*	f.kur	*f.80*	lfcc3.D	*f.nac*	f.50
f.nac	*LPFerr*	*lfcc1.D*	lfcc2	lfcc1.D	lfcc3.D	lfcc2	fSlo.D	f.skew	f.Max	lfcc1.D
f.skew	*f.50*	*f.kur*	*f.Max*	f.skew	fSlo.D	f.80.D	htkE.D	f.nac	LPFerr	lfcc2.D
f.50	f.nac	lfcc2.D	LPFerr	lfcc2	f.80.D	f.50.D	f.Slo	lfcc1.D	lfcc1.D	f.Max
f.Slo	f.50.D	f.80.D	f.kur	LPFerr	lfcc1.D	lfcc3	f.kur.D	f.80	lfcc2.D	f.Max.D
f.50.D	fSlo.D	LPFerr	f.80.D	lfcc5	htkE.D	f.50	f.20.D	lfcc2.D	lfcc2	lfcc2
f.kur	lfcc2	htkE	lfcc3.D	f.50	f.50.D	f.Max.D	f.Max.D	f.Max.D	htkE.D	lfcc3
lfcc4	f.kur	f.kur.D	lfcc4.D	f.50.D	f.skew	lfcc3.D	f.50	f.kur.D	lfcc5.D	f.50.D
htkE.D	f.skew	lfcc2	f.50	LPFerr.D	lfcc1	f.nac.D	f.skew	lfcc5.D	lfcc4.D	lfcc3.D
f.kur.D	f.nac.D	lfcc3.D	lfcc1.D	fSlo.D	f.kur	f.Max	LPFerr.D	lfcc2	lfcc4	f.Slo
lfcc4.D	f.Slo	htkE.D	f.20.D	lfcc1	lfcc4.D	f.20	f.nac.D	lfcc4	f.kur	f.nac.D
LPFerr	LPFerr.D	f.50.D	f.nac.D	lfcc4	f.20.D	f.skew	lfcc5	f.50.D	f.nac.D	LPFerr
f.Max	htkE.D	f.80	f.50.D	lfcc4.D	f.nac	htkE	LPFerr	f.nac.D	lfcc3.D	f.80

DFS.15: mejores 15 características de los canales *PSA.bin(c)* @ *col.04Mc*

COL	EXP	LAH	LPS	REG	SPT	TP	TR	TS	VT	WNS
f.kur	htkE.D	*lfcc1*	*f.Max*	f.80	*LPFerr*	f.50.D	f.nac	lfcc1.D	f.80	htkE.D
lfcc1.D	lfcc1.D	f.Max	*lfcc1*	f.50	htkE.D	f.Max.D	*f.80*	LPFerr	*f.nac*	fSlo.D
f.50	f.50.D	*f.kur*	LPFerr	f.50.D	*f.Slo*	lfcc4	LPFerr	fSlo.D	f.nac.D	f.Max.D
f.80.D	*f.50*	*lfcc1.D*	lfcc1.D	f.nac.D	lfcc1.D	*f.Slo*	f.80.D	*lfcc1*	f.50.D	f.Slo
f.nac	*LPFerr*	f.skew	f.50	f.nac	f.kur	f.50	f.Max.D	f.50	f.kur	LPFerr
htkE	f.80.D	f.Slo	lfcc3.D	f.80.D	f.Max	lfcc1.D	f.20	lfcc3	f.80.D	f.nac
f.kur.D	LPFerr.D	htkE.D	f.kur	f.Slo	f.nac	lfcc3.D	htkE.D	htkE.D	f.Slo	lfcc1.D
f.skew	f.80	f.nac	f.nac	f.kur	f.Max.D	htkE.D	lfcc4	f.80	f.50	f.80
htkE.D	f.nac	LPFerr	f.80.D	f.Max	f.80.D	f.80.D	lfcc3.D	f.Max	f.Max	f.Max
f.Slo	f.kur	f.80.D	f.kur.D	f.20	f.50.D	lfcc2.D	htkE	f.50.D	LPFerr	lfcc2.D
fSlo.D	lfcc2.D	lfcc2.D	fSlo.D	lfcc4	lfcc3.D	f.Max	f.skew.D	f.Max.D	lfcc2	lfcc2
LPFerr.D	fSlo.D	htkE	f.50.D	f.Max.D	f.nac.D	f.kur.D	lfcc1	lfcc5	lfcc2.D	f.50
lfcc3.D	f.Slo	fSlo.D	lfcc2	lfcc2	f.20	LPFerr	f.Max	f.80.D	f.20	f.skew
f.Max	f.nac.D	lfcc3	f.Slo	lfcc1.D	fSlo.D	f.kur	lfcc3	htkE	htkE.D	lfcc3
lfcc2	lfcc3	lfcc3.D	f.Max.D	fSlo.D	f.kur.D	f.nac	f.50.D	f.Slo	f.skew	lfcc1

Tabla 6.3.1.: *Etapa de configuración DFS.15: 15 mejores características para describir la clase propia de los canales PSA.mul(c) y PSA.bin(c) en la base col.04Mc.* La importancia de cada componente desciende según su posición en la columna.

6.3.1. Selección DFS de características en los canales PSA

Con el objetivo de convertir el sistema VSR-PSA en un conjunto de detectores especializados, la selección de características se hace conforme al criterio $max\{PSA.class\}$ con el que se optimizan los resultados de $\%cAcc$ para cada clase propia w_{ch} de cada canal ch en las arquitecturas PSA.mul(c) y PSA.bin(c) paralelo. Equivalentemente, en la arquitectura serie SSA se escogen aquellas características que consiguen la mejor eficiencia $\%cAcc$ de reconocimiento para todo el sistema en su conjunto.

DFS.15: mejores 15 características para las clases propias de cada canal PSA @ *dec.95Mc*

PSA.mul(c)					**PSA.bin(c)**				
HY	LP	NS	TR	VT	HY	LP	NS	TR	VT
f.nac	*f.kur*	f.Slo	fSlo.D	*f.Slo*	f.Slo	*f.Slo*	*f.80*	*f.nac*	*f.Slo*
LPFerr	*f.Slo*	*f.50*	f.20	lfcc2	*f.Max*	f.skew	*f.50*	f.kur	f.kur
fSlo.D	*f.nac*	*f.Max*	*f.Slo*	f.Max	f.50.D	*f.nac*	*f.Max*	LPFerr	f.nac
f.kur	*fSlo.D*	*f.80*	*f.nac*	*f.skew*	*f.nac*	*f.kur*	f.nac	f.50.D	htkE.D
f.Max	lfcc2	fSlo.D	f.Max	LPFerr	lfcc2.D	*fSlo.D*	f.50.D	*f.Slo*	*f.skew*
f.Slo	f.Max	f.kur	htkE.D	f.nac	f.kur	lfcc1.D	f.Slo	f.Max	lfcc2.D
htkE.D	LPFerr	lfcc2	lfcc2.D	f.50.D	lfcc1	f.80.D	f.20.D	htkE.D	lfcc2
lfcc2	f.80.D	LPFerr	lfcc2	f.kur	f.skew.D	f.50	lfcc2	f.20	LPFerr
f.nac.D	f.skew	f.nac	f.kur	lfcc1	f.20.D	f.80	fSlo.D	f.80	f.20.D
lfcc1	f.nac.D	f.20.D	f.80.D	fSlo.D	f.80.D	lfcc2.D	f.skew	f.50	lfcc4.D
f.skew.D	htkE.D	f.80.D	f.skew	lfcc2.D	f.kur.D	f.20	lfcc3	lfcc2.D	lfcc1
lfcc3	lfcc3	f.skew	LPFerr.D	f.kur.D	lfcc3.D	LPFerr.D	f.nac.D	fSlo.D	f.50.D
lfcc3.D	lfcc5.D	f.kur.D	f.50.D	lfcc1.D	lfcc4.D	f.skew.D	LPFerr	f.skew	fSlo.D
lfcc5.D	f.kur.D	f.Max.D	f.Max.D	htkE	lfcc1.D	f.Max	htkE.D	lfcc4	f.80.D
f.kur.D	f.50.D	LPFerr.D	f.kur.D	lfcc4	f.nac.D	f.50.D	f.kur.D	lfcc4.D	lfcc4

Tabla 6.3.2.: *DFS,15: 15 mejores características para describir la clase propia de cada canal múltiple $PSA.mul(c)$ y binario $PSA.bin(c)$ en la base $dec.95Mc$.*

Las 15 características seleccionadas por el DFS para las clases $\{w_{ch}\}$ propias de los canales PSA se listan en la Tabla 6.3.1 para el volcán de Colima y en la Tabla 6.3.2 para Decepción. La Tabla 6.3.3 muestra las características $DFS.15$ que aparecen conjuntamente seleccionadas en Colima y Decepción en sus canales PSA.mul(c), PSA.bin(c) y en ambos a la vez. Las clases que ambos corpus tienen comparten son los terremotos de largo periodo (LP/LPS), ruidos (NS/WNS), tremores espasmódicos (TR/TS) y sismos de origen volcano-tectónico (VT). Tras examinar los resultados listados es interesante recalcar:

- Al seleccionar las 5 características más importantes ($DFS.5$) en los canales múltiples y binarios, existe entre ellos una coincidencia de aproximadamente 50 % de Decepción y de tan solo el 33 % en Colima. Dichas características están enfatizadas en la Tabla 6.3.2 y la Tabla 6.3.2.

DFS.15: características de los canales PSA que son comunes en *dec.95Mc + col.04Mc*											
PSA.mul(c)				**PSA.bin(c)**				**PSA.mul(c)+PSA.bin(c)**			
LP	NS	TR	VT	LP	NS	TR	VT	LP	NS	TR	VT
f.kur	f.Slo	f.Slo	f.Slo	f.Slo	f.80	LPFerr	f.Slo	f.kur	f.Slo	f.Slo	f.Slo
f.Slo	f.50	f.skew	lfcc2	f.nac	f.50	f.50.D	f.kur	f.Slo	f.50	f.50.D	lfcc2
f.nac	f.Max	f.nac	f.Max	f.kur	f.Max	f.Slo	f.nac	f.nac	f.Max		LPFerr
lfcc2	f.80	lfcc2.D	LPFerr	f.Slo.D	f.nac	f.Max	f.skew	f.Max	f.80		f.nac
f.Max	f.Slo.D	f.Max.D	f.nac	lfcc.1.D	f.Slo	htkE.D	lfcc2.D	f.80.D	f.Slo.D		f.kur
LPFerr	lfcc2	f.kur.D	f.kur	f.80.D	lfcc2	f.80	lfcc2		lfcc2		lfcc2.D
f.80.D	LPFerr	lfcc2	lfcc1	f.50	f.Slo.D	f.50	LPFerr		LPFerr		
f.nac.D	f.Max.D	f.50.D	lfcc2.D	f.Max	f.skew	f.Slo.D	f.50.D				
f.50.D			lfcc1.D	f.50.D	LPFerr		f.80.D				
			lfcc4		htkE.D						

Tabla 6.3.3.: *DFS.15: características seleccionadas de los canales PSA.mul(c), PSA.bin(c) y PSA.mul(c) + PSA.bin(c) que son comunes en dec.95Mc y col.04Mc.* Las clases comunes en ambas bases de datos son (*dec.95Mc/col.04Mc*): LP/LPS, NS/WNS, TR/TS y VT/VT.

- Restringiéndonos a las mejores 5 componentes ($DFS.5$), la coincidencia entre ambas bases de datos es del 50 % en canales múltiples y de un 15 % en binarios.
- Las características de origen cepstral (vector $LFCC.D.10$, que está incluido dentro del vector base $geoLFCC.D.30$) solo representan el 6 % en Decepción y 23 % en Colima del vector $DFS.5$ cuando deberían abarcar el 33 %. En la selección $DFS.15$ las componentes $lfcc$ comunes a ambas bases representan el $9/35 \approx 26\,\%$ en canales PSA.mul(), un $4/36 \approx 11\,\%$ en binarios y el $3/19 \approx 16\,\%$ en la mezcla de ambos.
- En la selección $DFS.15$ existe una cierta concordancia entre características que aparecen en ambos canales PSA.mul(c) y PSA.bin(c) de ambas bases de datos: 5 para los LPs, 7 en NSs, 2 en TRs y 6 en VTs. La generalización de características es mucho mayor si nos limitamos a comparar por independiente en cada tipo de canal: un $35/60 \approx 58\,\%$ en canales múltiples y un $36/60 = 60\,\%$ en binarios.

Los valores de coincidencia para $DFS.5$ parecen bajos a priori, aunque en realidad no lo son tanto; el límite de aleatoriedad está en un 17 % (que es la probabilidad a priori de encontrar una característica en 1 de los 6 intervalos con 5 componentes cada uno en los que podemos dividir el vector $geoLFCC.D.30$). Ocurre igual en $DFS.15$; dada una componente seleccionada en un canal, la probabilidad de encontrarla a la vez en el resto de 3 selecciones $DFS.15$, cada una asociada a un canal PSA, es de $0{,}5 * 0{,}5 * 0{,}5 = 0{,}125 = 12{,}5\,\%$, lo que se traduce en que aleatoriamente, de media solo existirían $0{,}125 * 15 = 1{,}875 \approx 2$ componentes comunes en cada clase PSA que represente al mismo tipo de datos en Colima y Decepción. En este sentido,

cabe preguntarse si realmente la clase tremor espasmódico (TS) de Colima se puede equiparar a la clase tremor (TR) de Decepción .

HMM.30 vs. *DFS.15* (*%cAcc*) @ *dec.95Mc*						
SSA	*HY*	*LP*	*NS*	*TR*	*VT*	media
HMM.30	66.32	40.96	40.10	52.78	55.79	*51.19*
DFS.15	64.60	66.55	71.91	44.01	72.33	*63.88*
PSA.mul(c)	*HY*	*LP*	*NS*	*TR*	*VT*	media
HMM.30	64.08	54.78	39.39	55.78	57.86	*54.38*
DFS.15	66.33	66.55	52.42	34.71	66.35	*57.27*
PSA.bin(c)	*HY*	*LP*	*NS*	*TR*	*VT*	media
HMM.30	52.70	59.81	55.77	68.77	54.91	*58.39*
DFS.15	36.05	79.26	82.08	58.29	63.13	*63.76*

HMM.30 vs. *DFS.15* (*%cAcc*) @ *col.04Mc*												
SSA	*COL*	*EXP*	*LAH*	*LPS*	*REG*	*SPT*	*TP*	*TR*	*TS*	*VT*	*WNS*	media
HMM.30	43.56	-10.34	30.67	90.44	80.44	81.10	-10.77	-23.69	-7.64	87.02	-8.83	*32.00*
DFS.15	73.19	37.46	0.00	92.95	89.33	90.02	34.96	27.07	3.57	93.68	63.36	*55.05*
PSA.mul(c)	*COL*	*EXP*	*LAH*	*LPS*	*REG*	*SPT*	*TP*	*TR*	*TS*	*VT*	*WNS*	media
HMM.30	43.56	18.68	33.33	91.40	85.33	79.73	-24.10	7.58	6.42	86.67	15.07	*40.33*
DFS.15	64.01	44.04	31.33	91.05	68.99	85.17	-9.23	22.17	22.73	80.70	68.30	*51.75*
PSA.bin(c)	*COL*	*EXP*	*LAH*	*LPS*	*REG*	*SPT*	*TP*	*TR*	*TS*	*VT*	*WNS*	media
HMM.30	11.97	-160.87	30.67	81.13	24.00	75.83	-57.31	-103.27	-43.45	35.09	10.58	*-8.69*
DFS.15	62.01	-65.34	71.54	90.02	6.07	76.95	0.00	0.98	-12.42	58.25	20.63	*28.06*

Tabla 6.3.4.: *Etapa de configuración $DFS.15$: evaluación de las 15 mejores características en cada canal PSA.* % de eficiencia de los modelos $\{w_{ch}\}$ propios de los canales $PSA.mul(c)$ y $PSA.bin(c)$ cuando se construyen con 15 ($DFS.15$) y 30 características ($HMM.30$).

Una vez construido el vector $DFS.15\{geoLFCC.D.30\}$ específico para cada canal en Colima (mediante la Tabla 6.3.1) y en Decepción (Tabla 6.3.2), la Tabla 6.3.4 compara la eficiencia *%cAcc* en cada clase por independiente de los sistemas SSA, PSA.mul(c) y PSA.bin(c) al reducir las componentes del vector de características desde 30 (etapa $HMM.30$) a 15 (etapa $DFS.15$). Observamos que:

- Aún reduciendo el vector de características a la mitad, *los resultados de reconocimiento no solo se mantienen, si no que de media mejoran notablemente.* Dicha mejoría es más notable en Colima que en Decepción; y en general, es más amplia en sistemas SSA, seguidos de los canales PSA.bin(c) y por último los PSA.mul(c). La mejora sustancial gracias al DFS es algo ya constatado en el Capítulo 4 y en los trabajos de Álvarez et al. (2011), Cortés et al. (2014) y Cortés et al. (2015). En la Figura 4.5.1 del Capítulo 4 comprobamos también

como el DFS se muestra más eficiente en Colima que en Decepción estando formados ambos corpus por eventos aislados.

- *La mejora no es en todas las clases.* De hecho, en Decepción, la clase TR decrece una media de 13 puntos de $\%cAcc$ siempre en todas las arquitecturas y los terremotos HY empeoran hasta 16 puntos en los canales binarios. En Colima los terremotos lejanos, REG, pierden eficiencia al reducirse el vector. Probablemente este comportamiento sea una consecuencia de seleccionar un número fijo de componentes (15) para todos los modelos en vez de escoger el número de características que mejor se adecue a cada modelo de clase propia.
- *El comportamiento de los canales múltiples respecto a los binarios puede ser muy diferente.* En clases como los LAHs mientras pierde efectividad en PSA.mul(c) y gana más del doble en los PSA.bin(c). Los canales binarios parecen seguir la misma tendencia que el sistema SSA serie.
- *DFS se revela como una herramienta muy útil para reducir las inserciones.* Clases propicias a insertar eventos como tremores y ruidos, LAHs y EXPs multiplican su eficiencia al eliminar del vector componentes innecesarias.

6.3.2. Análisis y discusión de la selección discriminativa

Con el objetivo de contestar a la cuestión de si existen características que nos permitan describir eficientemente los mismos tipos de eventos acontecidos en distintos volcanes, en este apartado nos proponemos hacer un estudio general de la importancia de cada característica en las arquitecturas serie y paralelo y en los sistemas de reconocimiento en aislado y en continuo. Para ello vamos a contrastar los resultados aquí obtenidos con los presentados en el Capítulo 4 de reducción de dimensionalidad, donde se usan las versiones de eventos aislados $dec.95Ms$ y $col.04Ms$ de las bases de datos en continuo con las que trabajamos en este capítulo ($dec.95Mc$ y $col.04Mc$).

La Tabla 6.3.5 contiene información de la Tabla 4.2.6, que evalúa por independiente la eficiencia en $\%cCorr$ (equivalente en aislado al $\%cAcc$) de cada una de las 42 componentes geo-estadísticas del vector $geoSTATS.D.42$ en un sistema serie que denominamos $SSA.1$. Comparamos estos valores con los obtenidos en la sección anterior (Subsección 6.3.1) donde las columnas $PSA.mul(c)$ y $PSA.bin(c)$ representan un promedio del $\%cAcc$ obtenido por los modelos de las clases propias $\{w_{ch}\}$ de los canales PSA. Cabe destacar:

- *Escasa representación de componentes cepstrales.* Aunque de media debería haber un 33 %, solo hay un 13 % de características cepstrales (del vector $geoLFCC.\}D.10$) de la selección $DFS.15$ en el sistema serie SSA. El porcentaje se mantiene también en la arquitectura PSA.bin(c) y PSA.mul(c) en Decepción ($dec.95Mc$). En Colima ($col.04Mc$) sí alcanza justo el 33 %. Estos valores chocan con los que obtienen Cortés et al. (2015), en cuya selección de

15 mejores características de los vectores *geoLFCC.D.30 y geoSTATS.D.42*							
dec.95Mc			*dec.95Ms*	*col.04Mc*			*col.04Ms*
SSA	**PSA.mul(c)**	**PSA.bin(c)**	**SSA.1**	**SSA**	**PSA.mul(c)**	**PSA.bin(c)**	**SSA.1**
f.nac	*f.Slo*	*f.Slo*	*f.nac*	f.80	*f.Slo*	*f.nac*	*f.nac*
f.Slo	*f.Max*	*f.nac*	*f.50*	*f.kur.D*	lfcc1.D	lfcc1.D	*f.Max*
f.50.D	*f.nac*	*f.kur*	*f.skew*	*f.skew*	*f.nac*	*LPFerr*	*f.skew*
lfcc2	*fSlo.D*	*f.50.D*	f.80	*f.80.D*	*f.Max*	*f.Slo*	f.50
f.50	*f.kur*	*f.Max*	*f.Max*	*fSlo.D*	lfcc1	*f.Max*	*f.80*
LPFerr.D	lfcc2	*f.skew*	*LPFerr*	*f.50.D*	lfcc2	htkE.D	f.kur
f.Max	*LPFerr*	*lfcc2.D*	f.20.D	*f.Max*	*f.kur*	*f.80.D*	*f.Slo*
LPFerr	*f.skew*	*fSlo.D*	*f.Slo*	*f.Slo*	*LPFerr*	*f.kur*	htkE.D
htkE.D	*f.80.D*	f.80	f.nac.D	*htkE.D*	*f.50.D*	f.50	f.20
f.skew	f.kur.D	htkE.D	f.kur	lfcc4	*f.80.D*	*f.50.D*	A.nac
f.kur.D	*lfcc2.D*	f.20.D	*f.50.D*	*LPFerr.D*	lfcc3.D	*fSlo.D*	LPFerr
fSlo.D	*f.50.D*	lfcc2	A.kur	f.20.D	*lfcc2.D*	f.80	htkE
f.80.D	htkE.D	*LPFerr*	f.20	f.Max.D	f.80	f.Max.D	t.20
f.skew.D	f.50	f.50	*f.80.D*	lfcc2.D	*f.skew*	*lfcc2.D*	*f.50.D*
lfcc3.D	f.nac.D	*f.80.D*	A.std.D	*f.nac*	*fSlo.D*	*f.skew*	t.Max

Tabla 6.3.5.: *$DFS.15$: Vector mixto $geoLFCC.D.30$ frente al vector geo-estdístico $geoSTATS.D.15$.* Comparación de las 15 mejores características mixtas frente a las geo-estadísticas en las arquitecturas SSA, $PSA.mul(c)$ y $PSA.bin(c)$ en las bases $dec.95Mc$ y $col.04Mc$. Las componentes $geoSTATS.D.15$ se evalúan cada una por independiente en un sistema serie $SSA.1$ en las bases de eventos aislados $dec.95Ms$ y $col.04Ms$. Las características comunes en cada base de datos se enfatizan en *cursiva* en las columnas $SSA.1$. También se señalan las que lo son a la vez en $dec.95Mc$ y $col.04Mc$ en sistemas SSA y las comunes en los canales PSA. Las características en rojo no forman parte del vector base $geoLFCC.D.30$.

15 componentes del esquema $geoLFCC.D.30$ en el sistema SSA para Colima y Decepción en aislado, donde más del 80 % de las características seleccionadas son cepstrales.

- *Alta coincidencia de características geo-estadísticas en todas las arquitecturas.* La posibilidad de que una característica sea común entre los sistemas SSA, PSA.mul(c), PSA.bin(c) y SSA.1 es de 62 % en Decepción y un 50 % en Colima. La probabilidad de que dada una componente de la selección $DFS.15$ de esté además en las otras 3 arquitecturas es del $(13/30)^3 \approx 8\,\%$ en Decepción y del $(12/30)^3 \approx 6\,\%$ en Colima. Este hecho toma importancia si tenemos en cuenta que los sistemas SSA.1 usan eventos aislados y evaluación por independiente
- *Representación equilibrada de características dinámicas* ($.D$). De las componentes seleccionadas en los canales PSA, un 42 % corresponde a características contextuales o dinámicas (identificadas con $<caract.base>.D$, donde $<ca-$

rac.base> es el nombre original de la características base o estática). El porcentaje sube al 53 % en arquitecturas serie SSA. Para eventos aislados, sin embargo, estos número decrecen: tan solo el 20 % en SSA.1 y un 33 % en Cortés et al. (2015), lo que sugiere que las componentes dinámicas pueden estar relacionadas con la discriminación de eventos dentro de un flujo continuo de datos.

- *Coincidencia moderada en características comunes entre volcanes.* 10 de las 15 (67 %) mejores componentes del sistema serie SSA coinciden en ambos volcanes. También hay un 67 % de que la misma característica esté seleccionada conjuntamente en los canales múltiples y binarios de ambas bases de datos a la vez. Nótese que la probabilidad de que esto último ocurra aleatoriamente es del último $(15/30)^3 \approx 13\,\%$. Solo la característica $lfcc2.D$ de estas 10 comunes en todos los canales PSA es cepstral, y no hay ninguna cepstral común a los sistemas SSA.

El estudio de la información extraída durante esta sección de selección de características nos lleva a subrayar varias cuestiones:

- **Uso de otros esquemas de selección en el DFS.** La evaluación hecha en la Subsección 6.3.1, demuestran que el algoritmo DFS es en general un arma muy eficaz de cara a seleccionar características, logrando un efecto colateral de reducción inserciones. Dicha mejora no se da por igual en todos los canales, por lo que parece interesante que el número de componentes finalmente seleccionados en cada canal sea proporcional a la complejidad de su espacio de características a modelar.

- **Diferencia considerable entre canales PSA múltiples y binarios .** Esto hecho ya lo observamos en la fase anterior de configuración ($HMM.30$) al escoger en cada canal su número de componentes gausianas más adecuadas para su modelo de clase propia. En esta etapa las diferencias recaen no ya tanto en el conjunto de características seleccionado por el DFS si no en el incremento en $\%cAcc$ de este, que en promedio es mayor en canales binarios que en múltiples. Asimismo los modelos mejoran bastante más en Colima que en Decepción, gracias a la reducción de inserciones que se nota más en Colima.

- **Indicios de características *universales*, exportables entre volcanes.** Los resultados de la Tabla 6.3.3 y la Tabla 6.3.5 para sistemas $DFS.15$ indican una coincidencia al seleccionar las mejores 15 componentes del vector $geoLFCC.D.30$ de un 60 % entre canales del mismo tipo y un 67 % en el caso del sistema SSA serie. Si bien el estudio pormenorizado incluyendo distancias entre las características comunes y selecciones de distinto tamaño es demasiado extenso como para abarcarse en esta tesis, los resultados preliminares son prometedores siempre y cuando nos limitemos a canales PSA del mismo tipo. En el caso de evaluar características comunes para en canales múltiples y binarios a la vez esta concordancia baja notablemente a solo un 32 %

- **Modificación del vector de características** $geoLFCC.D.30$ **base.** La baja presencia de las componentes cepstrales en las características que tienen en común los canales PSA de Colima y Decepción (Tabla 6.3.3 y Tabla 6.3.5) indica que algunas características geo-estadísticas descartadas en el esquema $geoSTATS.D.42$ pueden ser relevantes en la discriminación de las clases propias de los canales PSA y no tanto en el proceso de selección realizado en el Sección 4.2 para el sistema serie SSA con eventos aislados. En concreto, aquellas componentes que específicamente están pensadas para discriminar una clase particular (como la correlación temporal $A.nac$ para detectar la correlación entre pulsos de clases cíclicas como TP o SPT). Por otro lado, el tamaño de nuestras bases de datos nos limita el número de componentes del vector inicial (Sección A.4).
- **Relevancia de características en continuo.** En el mismo contexto que el punto anterior, hay una alta probabilidad de que muchas de las características descartadas del vector geo-estadístico $geoSTATS.D.42$ sean apropiadas para detectar eventos en continuo, propiedad que no ha sido testeada en las pruebas de la Subsección 4.2.4 al hacerse con eventos aislados. Los resultados de la Tabla 6.3.5 nos revelan que probablemente haya una gran diferencia entre la selección DFS realizada en una base de datos en continuo y en su homóloga de eventos aislados que tiende a eliminar de la selección componentes dinámicas adecuadas para segmentar eventos.

6.4. Bandas óptimas para el filtrado espectral

En esta sección seleccionaremos el intervalo espectral $[f_L, f_H]_{ch}$ más conveniente para cada clase propia de un canal ch mediante pruebas de reconocimiento filtrando los datos en 4 bandas:

1. $[1, 10]\ Hz$ o *banda baja* , donde se acumula la energía espectral de eventos de largo periodo y tremores.
2. $[5, 15] Hz$ o *banda intermedia* , que elimina las bajas y altas frecuencias, adecuado para terremotos lejanos que hayan perdido su contenido de altas frecuencias y cuyas bajas frecuencias puedan confundirse con las de otros eventos.
3. $[10, 25] Hz$ o *banda alta*, que enfatiza el intervalo de altas frecuencias orientado a eventos *"ruidosos"* como explosiones, derrumbes y terremotos locales muy cercanos.
4. $[1, 25]\ Hz$ o *banda completa*, la que hemos estado usando hasta ahora, donde se registra la mayor parte de la energía espectral asociada de la sismicidad sentida en los volcanes.

La banda finalmente escogida para cada canal ch es aquella con la que se haya logrado mayor eficiencia $\%cAcc$ de reconocimiento para la clase propia w_{ch} de dicho

canal. Esta etapa ($BPF.15$) y la siguiente ($WIN.15$) no afectarán los resultados del sistema serie SSA, en los que se mantienen los valores de configuración base especificados en la Sección 6.1.

BPF.15: Bandas $[f_L, f_H]\,Hz$ @ *dec.95Mc*					
	HY	*LP*	*NS*	*TR*	*VT*
PSA.mul(c)	1-10	1-10	10-25	10-25	1-25
PSA.bin(c)	5-15	1-10	1-25	1-25	5-15

BPF.15: Bandas $[f_L, f_H]\,Hz$ @ *col.04Mc*											
	COL	*EXP*	*LAH*	*LPS*	*REG*	*SPT*	*TP*	*TR*	*TS*	*VT*	*WNS*
PSA.mul(c)	1-25	10-25	1-10	1-25	5-15	1-25	1-10	1-10	1-25	1-25	1-25
PSA.bin(c)	1-25	10-25	1-25	1-25	1-10	1-25	10-25	1-25	1-25	1-25	1-25

Tabla 6.4.1.: *$BPF.15$: Bandas espectrales óptimas para cada clase propia de los canales PSA.*

La Tabla 6.4.1 muestra los intervalos seleccionados mediante las pruebas experimentales para nuestros dos corpus de estudio en los canales binarios y múltiples. Destacamos:

- *Diferencia entre PSA.mul.(c) y PSA.bin(c) al escoger las bandas óptimas.* Los canales binarios tienden ligeramente a funcionar mejor en bandas más amplias que los múltiples. Puede estar relacionado con el hecho de que en canales binarios solo hay 2 modelos para describir todo el espacio $\Omega_{\mathbf{X}}$ de características frente a múltiples modelos en la arquitectura PSA.mul(c).
- *Las banda escogida para una clase propia no siempre está centrada en su intervalo de máxima energía.* Hay que recordar que el criterio de selección para la clase w_{ch} se fundamenta en la mejora de su eficiencia $\%cAcc(w_{ch})$ de reconocimiento, no necesariamente ligada a su descripción espectral. Por ello se escoge la banda que mejor discrimina a w_{ch}, y no necesariamente la que mejor permite su visualización. Por ejemplo: puede darse el caso de para los terremotos lejanos (REGs) la banda $[5, 15]\,Hz$ sea donde dinámicamente más varíen sus vectores de características, en oposición a la banda $[1, 10]\,Hz$ que no varíe tanto o, que pueda confundirse con otros eventos que también concentren ahí la mayor parte de su energía espectral. Si el modelo HMM_{REG} es capaz de captar esa dinámica de forma eficiente puede que el incremento de $\%cAcc(w_{REG})$ sea mayor en esta banda.

El resultado del filtrado BPF óptimo se muestra en la Tabla 6.4.2. Aquellos canales que mantienen su banda en el intervalo $[1, 25]\,Hz$ por defecto no cambian su $\%cAcc$. Por ello el sistema SSA obtiene los mismos valores en esta etapa que en la anterior $DFS.15$. Observamos que:

BPF.15 vs. *DFS.15* (*%cAcc*) @ *dec.95Mc*						
SSA	*HY*	*LP*	*NS*	*TR*	*VT*	media
DFS/BPF.15	64.60	66.55	71.91	44.01	72.33	*63.88*
PSA.mul(c)	*HY*	*LP*	*NS*	*TR*	*VT*	media
DFS.15	66.33	66.55	52.42	34.71	66.35	*57.27*
BPF.15	69.41	69.07	61.09	49.35	66.35	*63.05*
PSA.bin(c)	*HY*	*LP*	*NS*	*TR*	*VT*	media
DFS.15	36.05	79.26	82.08	58.29	63.13	*63.76*
BPF.15	56.41	80.51	82.08	58.29	67.35	*68.93*

BPF.15 vs. *DFS.15* (*%cAcc*) @ *col.04Mc*												
SSA	*COL*	*EXP*	*LAH*	*LPS*	*REG*	*SPT*	*TP*	*TR*	*TS*	*VT*	*WNS*	media
DFS/BPF.15	73.19	37.46	0.00	92.95	89.33	90.02	34.96	27.07	3.57	93.68	63.36	*55.05*
PSA.mul(c)	*COL*	*EXP*	*LAH*	*LPS*	*REG*	*SPT*	*TP*	*TR*	*TS*	*VT*	*WNS*	media
DFS.15	64.01	44.04	31.33	91.05	68.99	85.17	-9.23	22.17	22.73	80.70	68.30	*51.75*
BPF.15	64.01	48.06	36.67	91.05	73.78	85.17	6.33	25.57	22.73	80.70	68.30	*54.76*
PSA.bin(c)	*COL*	*EXP*	*LAH*	*LPS*	*REG*	*SPT*	*TP*	*TR*	*TS*	*VT*	*WNS*	media
DFS.15	62.01	-65.34	71.54	90.02	6.07	76.95	0	0.98	-12.42	58.25	20.63	*28.06*
BPF.15	62.01	-40.64	71.54	90.02	42.67	76.95	8.76	0.98	-12.42	58.25	20.63	*34.43*

Tabla 6.4.2.: *Configuración BPF.15 vs. DFS.15: evaluación del filtrado espectral óptimo en cada canal PSA.* Se compara la mejora obtenida del filtrado paso banda (BPF) específico en los canales respecto a la etapa de configuración anterior *DFS.15*.

- *La mejora media en esta etapa ronda el en 4.5 puntos en la eficiencia %cAcc.* Se incrementa ligeramente y de manera equilibrada en todas las arquitecturas y en ambas bases de datos.
- *En los canales binarios se observan las mayores mejoras.* Los eventos REG (+36 puntos) y las EXPs (+15) en Colima junto con los +20 puntos de los eventos HY en Decepción, todos en canales binarios, son los ejemplos más significativos. En los canales múltiples destaca la mejora de la eficiencia en las clases NS (+9) y TR (+15) en Decepción y TP (+9) en Colima.

6.5. Análisis del tamaño óptimo de la ventana de parametrización

En esta última etapa (*WIN.15*) del diseño del sistema PSA probaremos experimentalmente en cada canal diversas duraciones de la ventana de parametrización que trocea (o *ventanea*) la señal continua en fragmentos (segmentos o *frames*) sobre los que se parametriza para generar los vectores de características. El tamaño por

defecto es de 2 segundos en Decepción y 4 para Colima. Teniendo en cuenta los resultados ya obtenidos en la Subsubsección 4.2.2.2 del Capítulo 4 de reducción de dimensionalidad y en la Sección A.1, omitiremos duraciones demasiado pequeñas, estudiando los siguientes tamaños:

- *[1,2,4 y 5]* segundos en Decepción
- *[2,4,8 y 10]* segundos en Colima

La elección de estos tamaños viene determinada por las propiedades generales de los eventos presentadas en la Subsección 1.1.2 y las características concretas de cada clase que existe en las bases de datos maestras de Decepción y Colima (Sección 3.1). Es conveniente recordar que teóricamente, aunque una ventana mayor se corresponde con una mayor resolución espectral, conviene que durante el tiempo que dura la ventana las características no cambien demasiado (queremos ventanas *cuasi-estacionarias*) para facilitar un modelado eficaz de los vectores asociados a cada estado HMM.

WIN.15: duración del vector [*s*] @ *dec.95Mc*					
	HY	*LP*	*NS*	*TR*	*VT*
PSA.mul(c)	4	4	4	4	4
PSA.bin(c)	2	2	2	5	4

WIN.15: duración del vector [*s*] @ *col.04Mc*											
	COL	*EXP*	*LAH*	*LPS*	*REG*	*SPT*	*TP*	*TR*	*TS*	*VT*	*WNS*
PSA.mul(c)	10	10	4	10	10	8	10	10	4	8	4
PSA.bin(c)	8	10	4	10	10	8	10	4	8	8	8

Tabla 6.5.1.: *$WIN.15$: Duración óptima del vector de características en canales $PSA.mul(c)$ múltiples y $PSA.bin(c)$ binarios.*

La Tabla 6.5.1 lista las duraciones de ventana óptimas para cada clase w_{ch} propia en cada canal *ch* PSA múltiple y binario. De ella extraemos la siguiente información:

- Hay una *tendencia general en todos los canales a usar duraciones largas*. De hecho, en todos los canales la duración experimental óptima es igual o mayor que la duración por defecto asignada al sistema SSA.
- *La selección en Colima no se adapta a lo esperado a priori*. La hipótesis de partida es que las clases secuenciales en las que hay una marcada evolución de los eventos con el tiempo se describan más eficazmente con ventanas no muy grandes que puedan describir dicha evolución en términos de patrones a modelar por los estados de los HMM, mientras que los eventos no secuenciales puedan ser descritos con ventanas de duración larga al no variar sensiblemente sus propiedades con el tiempo. Mientras que en Decepción se cumple esta hipótesis, en Colima no siempre; en concreto con los LAHs, LPSs, TRs y TSs.

WIN.15 vs. *BPF.15* (*%cAcc*) @ *dec.95Mc*						
SSA	*HY*	*LP*	*NS*	*TR*	*VT*	media
BPF/WIN.15	64.60	66.55	71.91	44.01	72.33	*63.88*
PSA.mul(c)	*HY*	*LP*	*NS*	*TR*	*VT*	media
BPF.15	69.41	69.07	61.09	49.35	66.35	*63.05*
WIN.15	77.21	77.26	61.52	65.27	64.81	*69.21*
PSA.bin(c)	*HY*	*LP*	*NS*	*TR*	*VT*	media
BPF.15	56.41	80.51	82.08	58.29	67.35	*68.93*
WIN.15	56.41	80.51	82.88	72.52	70.38	*72.38*

WIN.15 vs. *BPF.15* (*%cAcc*) @ *col.04Mc*												
SSA	*COL*	*EXP*	*LAH*	*LPS*	*REG*	*SPT*	*TP*	*TR*	*TS*	*VT*	*WNS*	media
BPF/WIN.15	73.19	37.46	0.00	92.95	89.33	90.02	34.96	27.07	3.57	93.68	63.36	*55.05*
PSA.mul(c)	*COL*	*EXP*	*LAH*	*LPS*	*REG*	*SPT*	*TP*	*TR*	*TS*	*VT*	*WNS*	media
BPF.15	64.01	48.06	36.67	91.05	73.78	85.17	6.33	25.57	22.73	80.70	68.30	*54.76*
WIN.15	71.11	81.48	36.67	96.19	91.55	90.24	32.52	54.06	22.73	90.18	68.30	*66.82*
PSA.bin(c)	*COL*	*EXP*	*LAH*	*LPS*	*REG*	*SPT*	*TP*	*TR*	*TS*	*VT*	*WNS*	media
BPF.15	62.01	-40.64	71.54	90.02	42.67	76.95	8.76	0.98	-12.42	58.25	20.63	*34.43*
WIN.15	72.65	56.51	71.54	97.46	84.44	90.53	46.03	0.98	-1.52	81.05	42.53	*58.38*

Tabla 6.5.2.: *Configuración WIN.15 vs. BPF.15: evaluación de la duración óptima del vector de características en cada canal PSA.* Se compara la mejora obtenida al usar una duración específica de la ventana de parametrización en los canales respecto a la etapa de configuración anterior *BPF.15*.

La evaluación de los canales tras fijar de manera experimental su duración de vector óptimo se presenta en la Tabla 6.5.2. Aquellos canales que no hayan cambiado su duración respecto a la etapa *BPF.15* anterior obtienen mantienen su valor de *%cAcc*. Se constata que:

- *Se incrementa considerablemente en promedio la eficiencia de reconocimiento en 11 puntos*, con lo que se comprueba que la metodología experimental al seleccionar la duración de ventana ha funcionado correctamente. Dicha mejora es muy llamativa en Colima, que alcanza de media los 18 puntos. La razón hay que buscarla en la reducción de las inserciones en clases como EXP (que eleva su *%cAcc* en más de 100 puntos...), WNS, TP o SPT, que sobre todo en los canales PSA.bin(c) mermaban con valores incluso negativos el promedio de *%cAcc*.
- *La arquitectura paralela en canales especializados PSA mejora de media 6 puntos los valores del sistema serie de referencia SSA*. Aún no siendo el objetivo principal de esta implementación PSA (planteado en la Sección 6.1), en ambas bases de datos los sistemas PSA mejoran en promedio al sistema serie. No

todas las clases aumentan su eficiencia: NSs y VTs de los canales PSA.mul(c) y HYs de PSA.bin(c) en Decepción se quedan a más de 10 puntos; TRs, VTs y WNSs en los canales múltiples de Colima también. En el caso de Decepción, NS y VT son las clases de mayor eficiencia de reconocimiento. En Colima, lo es VT, y TR y WNS son clases con tendencia a tener muchas inserciones, con lo que habrá que asumir que hace falta seguir mejorando sus canales binarios.

6.6. Discusión de resultados y conclusiones

Finalmente, en esta sección examinaremos la eficiencia de las parametrizaciones base de las que partíamos en capítulos anteriores y comprobaremos la evolución obtenida al construir el sistema VSR-PSA empleando las mejores 15 características seleccionadas por el algoritmo *DFS.cAcc*.

En la Tabla 6.6.1 resumimos los resultados obtenidos a lo largo de este capítulo y los comparamos con otros sistemas basados en la parametrización cepstral *MFCC.D.A* con 30 y 39 componentes utilizada en el Capítulo 3 al presentar el sistema de referencia SSA. Nótese que la arquitectura PSA ha sido configurada acorde al criterio $max\{PSA.class\}$ (Sección 5.1) para maximizar la eficacia de los canales al reconocer eventos pertenecientes a la clase propia de cada canal, lo que equivale a convertir la arquitectura VSR-PSA en detectores específicos en paralelo (Subsección 5.2.1), por lo que las columnas que promedian los resultados de cada clase propia en los canales múltiples PSA.mul(c) y binarios PSA.bin(c) serán enfatizadas en negrita. Con ánimo de facilitar las comparaciones, enfatizamos también los resultados correspondientes a las arquitecturas serie SSA y a los esquemas más relevantes: el sistema *BASE*.30 que usa el esquema base de parametrización *geoLFCC.D*.30 propuesto en la Sección 4.2 y *WIN*.15, que utiliza las mejores 15 componentes de dicho vector correspondiendo a la última etapa de configuración del sistema VSR-PSA. Observamos que:

- *La mejora tras configurar los canales PSA como detectores específicos (etapa WIN.15) respecto al sistema BASE.30 ronda un promedio de +30 puntos en el %cAcc*. Los modelos especializados de cada canal incrementan su eficiencia más de 20 puntos en Decepción. En Colima los resultados son aún mejores; casi 30 puntos en los canales múltiples PSA.mul(c) y más de 70 en los binarios PSA.bin(c).
- *Los canales PSA especializados mejoran un 12 % en Decepción y 14 % en Colima la eficiencia de reconocimiento respecto al sistema serie SSA*. Si comparamos con el sistema serie en la etapa de referencia *BASE*.30se alcanza un incremento hasta el 50 % en Decepción y más del 60 % en Colima, pero utilizando un vector solo con la mitad de componentes.
- *Los canales binarios PSA.bin(c) son los que más mejoran desde el sistema serie BASE.30*. Sin embargo, no puede extenderse esta afirmación a la arquitectura

en paralelo: el mayor incremento en eficiencia entre las etapas *BASE.30* y *DFS.15*se da PSA.bin(c) en Colima (con +72 puntos de *%cAcc* frente a los +28 de los múltiples) y en los canales múltiples PSA.mul(c) en Decepción (con +22 frente a 19 de los binarios).

SSA frente a PSA (*%cAcc*): sistemas BASE @ *dec.95Mc*							
vector	**SSA**	PSA.mul	**PSA.mul(c)**	JOINT.mul	PSA.bin	**PSA.bin(c)**	JOINT.bin
MFCC.D.A.39	**38.51**	-	-	59.76	-	**-**	49.95
MFCC.D.A.30	**48.58**	-	-	60.02	-	**-**	52.96
BASE.30	**46.97**	**-**	**46.97**	57.97	-	**53.10**	56.09
SSA frente a PSA (*%cAcc*): configuración de canales PSA @ *dec.95Mc*							
etapa	**SSA**	PSA.mul	**PSA.mul(c)**	JOINT.mul	PSA.bin	**PSA.bin(c)**	JOINT.bin
HMM.30	**51.19**	49.58	**54.38**	57.34	59.70	**58.39**	57.97
DFS.15	**63.88**	50.36	**57.27**	65.28	63.59	**63.76**	60.05
BPF.15	**63.88**	49.32	**63.05**	57.36	65.24	**68.93**	59.91
WIN.15	**63.88**	63.69	**69.21**	60.68	65.42	**72.34**	53.33

SSA frente a PSA (*%cAcc*): sistemas BASE @ *col.04Mc*							
vector	**SSA**	PSA.mul	**PSA.mul(c)**	JOINT.mul	PSA.bin	**PSA.bin(c)**	JOINT.bin
MFCC.D.A.39	**54.39**	-	**-**	86.12	-	**-**	62.71
MFCC.D.A.30	**56.98**	-	**-**	86.38	-	**-**	59.76
BASE.30	**38.58**	-	**38.58**	82.45	-	**-13.46**	68.15
SSA frente a PSA (*%cAcc*): configuración de canales PSA @ *col.04Mc*							
etapa	**SSA**	PSA.mul	**PSA.mul(c)**	JOINT.mul	PSA.bin	**PSA.bin(c)**	JOINT.bin
HMM.30	**32.00**	36.66	**40.33**	81.86	31.07	**-8.69**	65.68
DFS.15	**55.05**	37.52	**51.75**	84.53	51.26	**28.06**	59.34
BPF.15	**55.05**	28.50	**54.76**	84.57	52.41	**34.43**	61.78
WIN.15	**55.05**	51.47	**66.82**	82.54	66.50	**58.38**	66.25

Tabla 6.6.1.: *VSR-SSA vs. VSR-PSA: Resultados comparativos para distintas parametrizaciones y etapas de configuración.* Se muestra la eficiencia de reconocimiento *%cAcc* para las siguientes arquitecturas: serie (*SSA*), promedio de los resultados de los canales VSR-PSA para canales múltiples (*PSA.mul*) y binarios (*PSA.bin*), promedio de las c clases propias en canales múltiples (*PSA.mul(c)*) y binarios (*PSA.bin(c)*) y salida conjunta del sistema paralelo VSR-PSA para canales múltiples (*JOINT.mul*) y binarios (*JOINT.bin*).

- Aunque el sistema no ha sido configurado para ello, observamos como *la eficiencia %cAcc del decodificador conjunto PSA son equiparables a los del sistema serie SSA en Decepción y notablemente mejores en Colima.* Este comportamiento se mantiene en general en todas las etapas de configuración y en cualquier vector de parametrización. La salida del codificador conjunto (co-

lumnas *JOINT.mul* y *JOINT.bin*) junto con el promedio de *%cAcc* de los canales PSA (columnas *PSA.mul* y *PSA.bin*) no son directamente comparables con los resultados del resto de arquitecturas: para que lo fuese, los sistemas *PSA.mul* y *PSA.bin* deben ser configurados bajo el criterio $max\{PSA.ch\}$ y el decodificador conjunto optimizarse según el esquema $max\{PSA.joint\}$ (Sección 5.1).

- *La mejora de los modelos de las clases propias también repercute colateralmente en incrementar la eficiencia del canal.* Al igual que ocurre con el decodificador conjunto, el *%cAcc* de las columnas *PSA.mul* y *PSA.bin* mejora al mismo nivel que lo hace en PSA.mul(c) y PSA.bin(c). En los canales binarios incluso se supera al sistema SSA.

- *La parametrización cepstral MFCC.D.A obtiene resultados muy competitivos.* Tanto en el sistema serie SSA como en el decodificador conjunto PSA. La cuestión que se plantea es ¿cuánto más podría mejorar al configurarse el sistema PSA con este vector?. Lo especialmente interesante de los esquemas mixtos es que puedes incorporar nuevas características diseñadas especialmente para describir un tipo de eventos en concreto, algo que es imposible hacer con el esquema cepstral.

- En cada etapa de configuración se mejora un promedio de 5 puntos en la eficiencia *%cAcc* para los canales PSA.mul(c) múltiples y PSA.bin(c) binarios de Decepción. En Colima las etapas que más contribuyen al incremento de eficiencia son la selección de características (*DFS*.15) y la selección de óptima de la duración de la ventana de parametrización (*WIN*.15).

Para completar la información de la Tabla 6.6.1, vamos a desglosar la eficiencia para cada canal en cada etapa de configuración en la Tabla 6.6.2. El análisis de resultados revela:

- *Los resultados finales de la eficiencia para las clases propias de los canales PSA específicos son bastante buenos.* Tanto en Decepción como en Colima y en canales múltiples y binarios el proceso de configuración (desde la etapa inicial *BASE*,30 a la final *DFS*,15) incrementa satisfactoriamente el *%cAcc*. Todas las clases mejoran en cada canal tras configurarse. En Decepción solo PSA.mul(NS) y PSA.bin(HY) con 61,52 %*cAcc* y 56,41,52 %*cAcc* respectivamente están por debajo de la media de 63,88 %*cAcc* del sistema de referencia SSA - *WIN*.15. Los modelos de clases propias PSA alcanzan un promedio de 69,21 %*cAcc* en canales múltiples y 72,38 %*cAcc* en binarios. En Colima estos valores alcanzan los 66,82 % en los múltiples y los 58,38 % para binarios. Los resultados de algunas de sus clases están por debajo del 55,05 %*cAcc* de media del sistema SSA: LAH (36,67 %), TP (32,52 %), TR (54,06 %) y TS (22,73 %) en los canales múltiples y TP (46,03 %), TR (0,98 %), TS (−1,52 %) y WNS (42,53 %) en los binarios. Todas ellas son clases no secuenciales que tienden a insertar eventos.

Configuración $max\{PSA.class\}$ (%cAcc) @ *dec.95Mc*						
SSA	*HY*	*LP*	*NS*	*TR*	*VT*	media
BASE.30	56.83	54.78	21.03	45.32	56.92	*46.97*
HMM.30	66.32	40.96	40.10	52.78	55.79	*51.19*
DFS/BPF/WIN.15	64.60	66.55	71.91	44.01	72.33	*63.88*
PSA.mul(c)	*HY*	*LP*	*NS*	*TR*	*VT*	media
BASE.30	56.83	54.78	21.03	45.32	56.92	*46.97*
HMM.30	64.08	54.78	39.39	55.78	57.86	*54.38*
DFS.15	66.33	66.55	52.42	34.71	66.35	*57.27*
BPF.15	69.41	69.07	61.09	49.35	66.35	*63.05*
WIN.15	77.21	77.26	61.52	65.27	64.81	*69.21*
PSA.bin(c)	*HY*	*LP*	*NS*	*TR*	*VT*	media
BASE.30	49.02	60.39	46.24	53.74	56.13	*53.10*
HMM.30	52.70	59.81	55.77	68.77	54.91	*58.39*
DFS.15	36.05	79.26	82.08	58.29	63.13	*63.76*
BPF.15	56.41	80.51	82.08	58.29	67.35	*68.93*
WIN.15	56.41	80.51	82.08	72.52	70.38	*72.38*

Configuración $max\{PSA.class\}$ − %cAcc @ *col.04Mc*												
SSA	*COL*	*EXP*	*LAH*	*LPS*	*REG*	*SPT*	*TP*	*TR*	*TS*	*VT*	*WNS*	media
BASE.30	41.54	18.68	33.33	91.71	87.11	79.98	-34.57	3.41	3.27	85.96	13.92	*38.58*
HMM.30	43.56	-10.34	30.67	90.44	80.44	81.10	-10.77	-23.69	-7.64	87.02	-8.83	*32.00*
DFS/BPF/WIN.15	73.19	37.46	0.00	92.95	89.33	90.02	34.96	27.07	3.57	93.68	63.36	*55.05*
PSA.mul(c)	*COL*	*EXP*	*LAH*	*LPS*	*REG*	*SPT*	*TP*	*TR*	*TS*	*VT*	*WNS*	media
BASE.30	41.54	18.68	33.33	91.71	87.11	79.98	-34.57	3.41	3.27	85.96	13.92	*38.58*
HMM.30	43.56	18.68	33.33	91.40	85.33	79.73	-24.10	7.58	6.42	86.67	15.07	*40.33*
DFS.15	64.01	44.04	31.33	91.05	68.99	85.17	-9.23	22.17	22.73	80.70	68.30	*51.75*
BPF.15	64.01	48.06	36.67	91.05	73.78	85.17	6.33	25.57	22.73	80.70	68.30	*54.76*
WIN.15	71.11	81.48	36.67	96.19	91.55	90.24	32.52	54.06	22.73	90.18	68.3	*66.82*
PSA.bin(c)	*COL*	*EXP*	*LAH*	*LPS*	*REG*	*SPT*	*TP*	*TR*	*TS*	*VT*	*WNS*	media
BASE.30	13.10	-160.87	30.67	81.76	-32.73	73.50	-57.82	-103.27	-42.24	35.09	14.71	*-13.46*
HMM.30	11.97	-160.87	30.67	81.13	24.00	75.83	-57.31	-103.27	-43.45	35.09	10.58	*-8.69*
DFS.15	62.01	-65.34	71.54	90.02	6.07	76.95	0	0.98	-12.42	58.25	20.63	*28.06*
BPF.15	62.01	-40.64	71.54	90.02	42.67	76.95	8.76	0.98	-12.42	58.25	20.63	*34.43*
WIN.15	72.65	56.51	71.54	97.46	84.44	90.53	46.03	0.98	-1.52	81.05	42.53	*58.38*

Tabla 6.6.2.: *VSR-SSA vs. VSR-PSA: Mejora en cada etapa de configuración de los canales PSA frente al sistemas serie SSA.* Eficiencia de reconocimiento *%cAcc* para las arquitecturas: serie *SSA* y los modelos de las c clases propias en canales múltiples $PSA.mul(c)$ y binarios $PSA.bin(c)$ en las configuraciones $BASE.30$, $HMM.30$, $DFS.15$, $BPF.15$ y $WIN.15$.

- *Aún quedan algunas clases que se resisten a mejorar en los canales binarios.* Al comparar la eficiencia entre el sistema en serie SSA (SSA - $BASE$,30) y los canales ya configurados (PSA.mul(c) y PSA.bin(c) - WIN.15) nos encontramos con clases que empeoran ligeramente: HY (que baja su $\%cAcc$ en -0.5 puntos) en Decepción y REG (-3), TR (-2.5), TS (-4) y VT (-4) en Colima.

Para finalizar vamos a formular ciertas cuestiones surgidas resumir del estudio de estas tablas y de los resultados presentados en los apartados anteriores, cuyas respuestas pueden interpretarse como las conclusiones de este capítulo:

¿Qué hemos ganado con la arquitectura PSA paralelo respecto el sistema clásico SSA serie? La respuesta experimental obtenida en este capítulo es que tras haberse configurado ambos sistemas, es que la eficiencia de reconocimiento medida en $\%cAcc$ se incrementa de media un 13 % (desde 55.05 a 62.60) en Colima y un 11 % (desde 63.88 a 70.80). Más de un 10 % en ambas bases de datos es más que suficiente como para asumir los costes de aplicación del sistema paralelo PSA: una configuración más entretenida (Subsección 5.1.1) y un aumento en los recursos computacionales (Subsección 5.1.3).

Una respuesta más detallada nos lleva a remarcar ideas ya desarrolladas en el Capítulo 5:

- *Aumento de la robustez y fiabilidad.* Al estar cada canal PSA específicamente diseñado para reconocer un tipo de clase concreto
- *Aumento de la funcionalidad.* El sistema PSA configurado en este capítulo como detectores específicos en paralelo nos permite usarlo como detector de eventos solapados, detector de eventos potencialmente peligrosos para la población (lahares, explosiones, colapsos,...), etiquetado semi-automático y multi-etiquetado de bases de datos, etc...

Sin embargo, la respuesta más interesante es que *la arquitectura PSA en paralelo nos permite seguir mejorando el sistema de análisis y reconocimiento donde los sistemas serie SSA no pueden.* Mediante el diseño de características y la configuración específica en cada canal especializado en un solo tipo de evento, los valores de reconocimiento solo pueden seguir incrementando.

Canales PSA frente al decodificador conjunto El decodificador conjunto ($PSA.joint$) permite dar unos resultados únicos a partir de los obtenidos en los canales como si se tratase de un sistema en serie SSA (Subsección 5.1.2). El diseño de este decodificador es muy sencillo y puede considerarse aún que está en estado experimental: tiende a eliminar las inserciones pero parece tener problemas para segmentar correctamente los eventos. A pesar de ello y de que el sistema PSA implementado en este capítulo no ha sido configurado para optimizar la salida del decodificador conjunto los resultados preliminares son prometedores: son comparables (y ligeramente mejores) que en el sistema de referencia base SSA en Decepción y ampliamente mejores

en Colima. Resultados para ambos canales binarios y múltiples y con sistemas que usan distintas características.

Lo más llamativo del decodificador PSA.joint no es la tasa de eficiencia conseguida en esta implementación, si no que nuestro sistema PSA puede emplearse con cualquier tipo de datos (no solo sismos) y ser comparado con otras arquitecturas en paralelo ya existentes.

En cuanto a la funcionalidad, los canales PSA se optimizan con el objetivo de análisis y el etiquetado (semi) supervisado. Por el contrario, PSA.joint parece estar más indicado para el reconocimiento no supervisado o en tiempo real.

¿Que es mejor, los canales binarios o los múltiples? A partir de los resultados analizados en este capítulo no existe una respuesta clara: los canales binarios funcionan mejor en Decepción (72,38 %$cAcc$ frente a 69,21 % de los múltiples) que en Colima (58,38 % frente al 69,21 % de los múltiples), por lo que en promedio deberíamos decantarnos por la arquitectura múltiple.

Hay que destacar en favor de los canales binarios que son los que más han mejorado en el proceso de configuración, lo que puede ser indicativo (o no) de que potencialmente son más propensos a mejorar. En cualquier caso, hay una ventaja a priori en los sistemas binarios: su menor nivel de complejidad y tiempo requerido en su configuración.

A falta de más experimentación decidirse por una u otra arquitectura es precipitado. En este ámbito, sería muy interesante comprobar si la respuesta depende del número de clases y / o separabilidad de estas en el espacio de características: ¿funcionan mejor los sistemas binarios PSA.bin en corpus de datos con cuyo espacio de características sea fácilmente separable?.

¿Existe mucha diferencia entre los resultados en reconocimiento en continuo y los hallados en el Capítulo 4 para eventos aislados? Al examinar y comparar las tablas de resultados constatamos que existe más diferencia de la que pensábamos en un primer momento. Este hecho nos afecta a dos niveles: en la eficacia de reconocimiento y, especialmente, en la selección de las características del vector base de parametrización que hemos usado en este capítulo y en el capítulo Capítulo 4 de reducción de dimensionalidad.

En general es un error extrapolar resultados a partir de experimentos a otros escenarios cuando las condiciones de esos escenarios son distintas a las de experimentación. Los resultados extraídos en la Sección 6.3 demuestran que en los canales de la arquitectura PSA, o, incluso, más general en los sistemas en continuo, las características cepstrales (muy efectivas en los sistemas aislados SSA del Capítulo 4) aquí dejan de ser tan relevantes a favor de las características de origen geofísico. Igual ocurre con el carácter dinámico de las componentes: en continuo ganan eficacia respecto a las características estáticas. Los resultados indican que vale la pena seguir la linea de

investigación de Álvarez et al. (2009) en el desarrollo de nuevas características específicamente diseñadas para discriminar un tipo concreto de eventos, especialmente las clases no secuenciales como tremores y lahares. Estamos convencidos que con un vector base más efectivo los resultados obtenidos en los canales hubieran sido aún mejores.

¿Por qué no seleccionar sólo las 5 o 10 mejores características en el DFS? El hecho de que en la etapa de configuración $DFS.15$ (Subsección 6.3.1) existan modelos de clases propias que empeoren sus resultados al usar vectores de las mejores 15 características entre el grupo inicial de 30, indica que un esquema de configuración donde cada clase use el número de componentes que *necesite*, o, que los modelos de los canales tengan una complejidad independiente unos de otros conforme tengan que describir un subespacio de características más o menos variable, es una opción muy interesante tal y como se aprecia en Cortés et al. (2014). En este sentido, la elección de usar 15, 10 o 5 características es solo un criterio inicial de configuración.

¿Son las características selecionadas exportables de un volcán a otro? La experimentación hecha en la Subsección 6.3.2 de la etapa de configuración DFS,15 de selección de características, concluye que hay *indicios* de características que son seleccionadas a la vez en Decepción y en Colima para describir las clases que tienen en común: LPs, TSs y NSs (Tabla 6.3.3 y Tabla 6.3.5). Al menos si consideramos un tamaño del vector de selección lo suficientemente grande. Conforme reducimos el número de componentes en el estudio las coincidencias son menores. El porcentaje de coincidencia es del 60 % entre canales del mismo tipo y del 32 % si las características tienen que estar en canales binarios y múltiples de ambas bases de datos. Aunque los valores no parecen muy altos, dados que los eventos de Colima y Decepción son muy diferentes entre sí (ni siquiera son volcanes del mismo tipo) y que las características del vector base siempre son susceptibles de ser sustituidas por otras más eficientes y robustas, es un resultado alentador.

Definitivamente, se necesita bastante más trabajo de investigación para poder dar una respuesta fiable a la cuestión de la exportabilidad de características. Nuestra apuesta es que con una selección adecuada de componentes, bien diseñada y evaluada en varias bases de datos formadas por distintos volcanes, se pueden encontrar un conjunto (más o menos amplio) de características exportable a volcanes del mismo tipo. En realidad, ese desafío ya está ganado: los expertos geofísicos son capaces de distinguir un mismo tipo de evento en casi cualquier volcán, el verdadero reto está en poder exportar las características desde nuestro cerebro al sistema de parametrización del ordenador.

Parte IV.

CONCLUSIONES Y LÍNEAS DE INVESTIGACIÓN FUTURAS

7. Conclusiones

Para finalizar esta memoria del trabajo de investigación, a modo de conclusiones realizaremos un resumen de los temas tratados y de los resultados y enseñanzas que hemos obtenido en este proceso.

Seguiremos la misma estructura que se ha planteado en la tesis, repasando primero los retos relacionados con los sistemas de reconocimiento de sismos en la actualidad y la efectividad de algunas soluciones generales que proponemos para mejorar los sistemas VSR (Sección 7.1). En la Sección 7.2 examinaremos el tema de la reducción de dimensionalidad, así como la aportación hecha gracias al algoritmo discriminativo generalizado *DFS.cAcc*. Terminaremos evaluando la evolución tanto teórica como experimental que supone la arquitectura de reconocimiento en paralelo VSR-PSA frente a la clásica en serie SSA.

La motivación de este trabajo de investigación comienza con una simple pregunta: *¿Cómo se podría mejorar los sistemas VSR?*. A la luz del estudio realizado en el Capítulo 2, encontramos diversos elementos donde flaquean los sistemas actuales de reconocimiento de sismos:

- El *etiquetado* y *fiabilidad* de las bases de datos
- La *descripción* de los eventos en vectores de características
- El *modelado* de las clases
- La *evaluación* de los resultados

Para completar esta lista, también tenemos una serie de requerimientos por parte de los centros de monitorización que han de ser atendidos:

- El funcionamiento de los sistemas VSR de reconocimiento en *tiempo cuasi-real*
- La *portabilidad* de los sistemas, tanto en la integración con los sistemas de adquisición como la exportación de los modelos de un volcán a otro sin tener que construir de nuevo los modelos.

Todo ello nos conduce a buscar soluciones a estas cuestiones en esta tesis, conscientes de que algunas de ellas seguirán siendo retos y áreas de investigación abiertas en proyectos venideros. Para otras, en la Parte II hemos propuesto soluciones que funcionan relativamente bien: la creación de bases de datos maestras (Sección 3.6) para mejorar la fiabilidad de los datos, la descripción de eventos mediante la selección discriminativa de características dada por el algoritmo $DFS.cAcc$ generalizado (Subsubsección 4.3.2.2), el sistema en canales paralelos para realizar un modelado específico de cada clase (Capítulo 5 y Capítulo 6) y la evaluación % promediada por clase (Subsección 3.4.2) junto a la re-evalución de resultados de reconocimiento que no tenga en cuenta errores geofísicamente no relevantes (Subsección 3.4.3).

En cuanto a la fiabilidad y robustez de las bases de datos y la evaluación de resultados consideramos que no hay mucho margen para seguir mejorando. El trabajo de los expertos geofísicos que etiquetan eventos e interpretan resultados es fundamental. El sistema en paralelo VSR-PSA es una valiosa herramienta para realizar un etiquetado tentativo de eventos sugiriendo al experto distintas opciones con su respectiva tasa de fiabilidad en un reconocimiento semi-supervisado. Asimismo facilita la evaluación de los resultados gracias a la especialización de los canales en reconocer eventos bajo circunstancias ruidosas o cuando hay eventos solapados.

Por otra parte, tanto la mejora en el modelado como en la descripción de eventos son objetivos constantes en el aprendizaje automático y la inteligencia artificial (*IA*), y por ende, en VSR. Si bien la elección de un tipo u otro de modelos es más eficiente según el tipo de datos sobre los que se emplean (parece lógico utilizar modelos secuenciales para reconocer señales sismo-volcánicas secuenciales), como comprobamos en la Sección 2.3 la eficiencia alcanzada por la mayoría de ellos es más que satisfactoria y su uso o no vendrá marcado por otras razones relacionadas con

la implementación y escenario de sus funciones: complejidad, coste computacional y funcionalidad de reconocimiento en tiempo real.

Tanto en señales sismo-volcánicas como en el ámbito general de la IA, la descripción de los datos es el área más susceptible de ser mejorada. En VSR, técnicas de representación como los mapas auto-organizativos han gozado de una creciente popularidad en los últimos 10 años (Masiello et al., 2006; Esposito et al., 2008a; Köhler et al., 2009). En la Sección 6.3 mostramos como nuestra propuesta de usar características diseñadas específicamente para cada tipo de evento (Álvarez et al., 2009) en una arquitectura de canales paralelos (Cortés et al., 2014) logra reducir la complejidad del modelado a la mitad e incrementa la eficiencia de reconocimiento. Pensamos que la mejora puede ser aún mayor ideando características más específicas. La última gran tendencia en IA, y para grandes expertos el gran reto actual, es el aprendizaje profundo o *Deep Learning* que intenta aprender de forma automática formas más eficientes de representación de señales (Bengio (2009); Ngiam et al. (2011); Sutskever et al. (2013)). Hasta que en un futuro cercano los algoritmos automáticos de DL nos indiquen las mejores características, los expertos en geofísica tendrán que seguir diseñándolas específicamente.

En las próximas secciones vamos a enumerar de forma concisa las conclusiones obtenidas en cada una de las áreas principales donde nuestro trabajo de investigación ha presentado innovaciones: los sistemas VSR clásicos, la reducción de dimensionalidad en bases de datos de sismos y el diseño de la arquitectura VSR en paralelo.

7.1. Sistemas VSR actuales

1. **Los HMMs proporcionan un modelado consistente y apropiado para el reconocimiento VSR de sismos**. Después de analizar distintos enfoques y múltiples alternativas en la Subsección 2.4.1, nos decantamos por usar modelos capaces de describir patrones secuenciales dentro de un entorno probabilístico que nos facilite asignar valores de robustez a los eventos reconocidos. La mayor parte de los sistemas VSR alcanzan cotas de reconocimiento suficientemente altas. La decisión más que venir marcada por la tasa de reconocimiento, la cual es difícil de comparar al existir pocos trabajos que usen bases de datos comunes, viene determinada por los requerimientos que los sistemas VSR deben satisfacer de cara a ser implantados en un escenario de monitorización (Subsección 2.2.3):

    a) Reconocimiento sobre flujo continuo de datos en tiempo (cuasi)real

    b) Escalabilidad y robustez del sistema

    c) Capacidad de generalización de los modelos ante datos no etiquetados: reconocimiento no supervisado

La utilización de los HMMs en otras áreas como el reconocimiento del habla, donde han sido el estándar desde hace 20 años, junto con el respaldo de varios trabajos mostrando la efectividad y robustez de los sistemas VSR basados en HMMs (Ohrnberger, 2001; Alasonati et al., 2006; Benítez et al., 2007; Beyreuther and Wassermann, 2008; Avesani et al., 2012; Bicego et al., 2013; Gutiérrez Espinoza, 2013) valida sobradamente este modelado. El sistema VSR-SSA diseñado en el Capítulo 3 de este trabajo ha sido satisfactoriamente usado no solo para reconocer las señales de varios volcanes activos (Cortés et al., 2009a,b) y estudiar la reducción de dimensionalidad (Cortés et al., 2015), sino que además se ha integrado en sistemas de monitorización en tiempo real en Colima (Gonzalez-Amezcua et al., 2012) y en sistemas de predicción de erupciones (Boué et al., 2015).

2. **Los resultados preliminares indican que el sistema VSR basado en HMMs puede ser potencialmente exportable de un volcán a otro**. Aunque es necesario seguir investigando en esta linea, la experimentación realizada en Cortés et al. (2009b) donde se construyen modelos que describen eventos con 2 volcanes con una tasa de acierto considerable (una eficiencia del 70 % con casi 7000 eventos distribuidos en 10 clases), constata la robustez del sistema mixto y augura una linea de investigación interesante.

3. **Las bases de datos etiquetadas de forma manual por expertos geofísicos sufren de falta de fiabilidad**. La solución propuesta de construir bases de datos maestras viene a contrarrestar esta falta de robustez: trabajar sobre eventos que inequívocamente puedan ser asociados a una clase u otra es una excelente base de cara a construir modelos robustos y exportables. El diseño de estas bases de datos es, sin embargo, una tarea tediosa y delicada; se requiere un equilibrio entre la selección de eventos claramente representativos de una clase y la necesidad de un número mínimo de ellos para evitar modelos sobre-entrenados. Si es posible, es muy conveniente observar el mismo evento en distintas estaciones para minimizar los efectos de sitio y propagación.

4. **Una perspectiva geofísica es necesaria para evaluar correctamente la eficiencia de reconocimiento de un sistema VSR**. Las métricas clásicas usadas en el área del reconocimiento de patrones no son directamente aplicables en un entorno de monitorización de la actividad sísmica donde exista un posible riesgo para la población: hay inserciones de eventos que pueden ser correctas y no todos los borrados o sustituciones de eventos tienen la misma importancia geofísicamente hablando.

   Aún así, en la fase de construcción de los modelos del sistema, el objetivo debe seguir siendo modelar cada tipo de eventos de la forma más eficiente posible independientemente de su relevancia geofísica, que debe ser evaluada en una capa superior del sistema de vigilancia.

5. **Necesaria colaboración entre entidades y proyectos: bases de datos y tecnologías abiertas** . Siguiendo las directrices marcadas en otras áreas de

investigación, la creación de un centro donde se intercambien bases de datos etiquetadas de distintos volcanes y estén libremente disponibles tecnologías y modelos solo puede llevar a una mejora sustancial y un avance común del que toda la sociedad en general y las entidades científicas en particular se ven beneficiadas.

En este sentido, la celebración de conferencias y concursos internacionales para comparar efectividad de tecnologías usando corpus de datos comunes es muy interesante.

7.2. Descripción de eventos sísmicos y la reducción de dimensionalidad

1. **Una selección de características eficaz para describir los eventos juega un papel fundamental en el diseño de sistemas VSR**. En el área de reconocimiento de patrones es bien conocido que al reducir la dimensionalidad se simplifica los modelos y con ello el sistema en general. En el Capítulo 4 demostramos que la selección de características es incluso más efectiva que la transformación del espacio de descripción a la hora de reducir la complejidad (Sección 4.5). Escoger un subconjunto de componentes de un vector de características inicial en vez de transformarlas a otro tiene una ventaja añadida: el significado geofísico se mantiene, lo que favorece una posterior interpretación y análisis en el proceso de descripción de eventos.

2. **El algoritmo generalizado *DFS.cAcc* de selección de características es la técnica analizada más eficaz para reducir la dimensionalidad**. La generalización del algoritmo *DFS* de Álvarez et al. (2011) que se concreta en la evolución *DFS.cAcc* se revela como la mejor opción en el profundo estudio llevado a cabo en el Capítulo 4 (y respaldado en Cortés et al. (2015)). En dicho estudio se comparan diversos métodos clásicos y recientes de reducción de dimensionalidad a distintos niveles: filtros de selección, métodos guiados y transformaciones del espacio original, siendo el algoritmo *DFS.cAcc* la técnica más efectiva de manera global atendiendo al coste computacional, a la tasa de reconocimiento del vector de características y a las nulas suposiciones estadísticas que debe cumplir el espacio de descripción sobre el que es aplicado. (Subsección 4.5.1).

   La bondad del *DFS.cAcc* se ha comprobado tanto en sistemas SSA serie, donde se consigue superar los resultados del vector base *geoLFCC.D*.30 con solo 10 de las 30 componentes originales (Capítulo 4), como en los canales del sistema PSA en paralelo (Capítulo 6) donde que incrementan su *%cAcc* de efectividad de reconocimiento en 8 % en Decepción y más de un 20 % en Colima pero utilizando solo la mitad de las componentes del vector base.

3. **Resultados preliminares apuntan la posibilidad de diseñar características exportables**. La experimentación llevada a cabo con los canales PSA (Subsección 6.3.2) muestra que existe cierta coincidencia al seleccionar las mejores características para canales especializados en describir eventos de la *misma* clase en volcanes tan distintos como Decepción y Colima. Teniendo en cuenta que las componentes seleccionadas provienen del vector de descripción base *geoLFCC.D*.30 construido de forma general para describir eventos aislados, el margen de mejora es bastante prometedor: tal y como hace Álvarez et al. (2011) se pueden diseñar características específicas para cada tipo de evento, incrementar el número de componentes del vector base y dejar elegir las mejores al algoritmo *DFS.cAcc*. Para todo ello se necesario incrementar el tamaño de las bases de datos maestras de Decepción y Colima para evitar el sobre-entrenamiento de modelos.

7.3. Sistemas serie VSR-SSA frente arquitecturas en paralelo VSR-PSA

1. **La arquitectura de canales en paralelo PSA añade nuevas funcionalidades sobre el sistema clásico serie SSA**. El diseño de la arquitectura PSA orientada a canales especializados en reconocer eventos de un tipo concreto aporta robustez y fiabilidad respecto al sistema SSA. Adicionalmente, la estructura en paralelo proporciona nuevas capacidades de análisis como el multi-etiquetado de señales, la discriminación de eventos solapados y la especialización en la detección de eventos potencialmente peligrosos para la población. Puede además proporcionar una salida conjunta como el sistema SSA pero añadiendo tasas de fiabilidad de reconocimiento para cada evento gracias al decodificador PSA conjunto, que, aún encontrándose en fase de desarrollo ya obtiene resultados comparables a su homólogo SSA.
2. **Los canales PSA configurados como detectores específicos mejoran significativamente los resultados del sistema serie SSA**. Los test ejecutados en el Capítulo 6 con las bases de datos de Colima y Decepción demuestran que se incrementa la efectividad de reconocimiento a varios niveles:

   a) Más de 30 puntos en promedio de *%cAcc* en los canales *PSA*.15 que usan 15 componentes para describir sus eventos frente al sistema base SSA (*BASE*.30) en serie de 30 componentes.

   b) Más de un 12 % de los canales PSA respecto el sistema serie SSA cuando ambos usan 15 características. Comparado con el sistema en serie *BASE*.30 este incremento supone más del 50 %.

   No obstante, el éxito del sistema PSA no radica solo en esta mejora de eficiencia, si no en el incremento de la robustez del sistema y sobre todo, en

la potencialidad de los canales gracias al proceso de diseño especializado del que el sistema base SSA no puede beneficiarse. Los beneficios de adoptar un esquema paralelo frente al esquema clásico compesan el coste que hay que pagar por ello (Subsección 5.1.3): una complejidad computacional que dobla a los sistemas SSA en serie y es proporcional al número de clases en los canales múltiples.

8. Líneas de investigación futuras

Hay ciertos trabajos que son auto-conclusivos. Y este no es uno de ellos. En parte porque hemos desarrollado una idea sobre la que hay poco material anterior en el campo del reconocimiento de sismos, abriéndose todo un campo de nuevas ideas y experimentación que tiene aplicaciones inmediatas como hemos mostrado en la Parte III.

Tomando la inspiración en las tendencias actuales del área de inteligencia artificial, el reconocimiento de patrones y el aprendizaje automático, se abre un mundo de posibilidades de integración en el campo del VSR. Conjuntando estas lineas de investigación junto a las que nacen de forma natural del desarrollo de la tesis se convierte en un desafío explorar nuevas ideas, que agruparemos en retos principales (Sección 8.1) sobre los que esbozaremos lineas de actuación preliminares para abordarlos y otras propuestas presentadas en la Sección 8.2 de carácter más técnico que consideramos interesantes para mejorar el sistema de reconocimiento VSR con implantación prevista a medio o largo plazo.

8.1. Retos principales y su planteamiento inicial

8.1.1. Parametrización mejorada

La forma en la que describimos los datos juega un papel fundamental tanto en el proceso de interpretación y análisis de estos como en la eficacia del sistema (Subsección 4.1.1). Aunque en la Sección 4.1 se ha hecho un estudio exhaustivo para escoger un conjunto representativo de parámetros de naturaleza mixta que representen detalladamente a los eventos sísmicos, el área de extracción de características sigue generando interesantes temas de investigación:

- **Configuración específica para cada característica**. Tanto en una arquitectura serie (VSR-SSA) como en paralelo (VSR-PSA), el poder discriminatorio de las características podría mejorar si escogemos valores del proceso de ventaneo o de la segmentación en frames específicos para cada característica. Ya existen trabajos (Álvarez et al., 2009, 2011)que toman distintas duraciones de los segmentos en los que se trocea la señal continua en función de las características que se estén extrayendo a partir del sismograma.
- **Nuevas características**. Siguiendo la línea de investigación de Álvarez et al. (2009), merece el esfuerzo estudiar nuevas propuestas de descripción específicamente diseñadas para discriminar un evento respecto al resto.

8.1.2. Pre-segmentación de la señal en continuo

Teóricamente se muestra que cuanto más sencilla sea la red de búsqueda, el algoritmo de Viterbi en la decodificación de los HMM proporciona mejores resultados de reconocimiento (Sección 3.3.2.1). En la práctica esto se traduce en que conviene tener ficheros de datos que contengan el menor número posible de eventos en ellos. Experimentalmente ya hemos comprobado este hecho cuando comparamos los resultados de una base de datos en continuo con la de su homóloga conteniendo solo eventos aislados, por ejemplo, al contrastar la eficiencia de reconocimiento lograda por los corpus *dec.95Ms* y *col.04Ms* en aislado del tema Capítulo 4 de reducción de dimensionalidad con la de las bases *dec.95Mc* y *col.04Mc* usadas en el sistema base (Sección 3.7) o en los canales PSA en paralelo (Capítulo 6).

Por todo ello, pensamos que la opción de pre-segmentar el flujo continuo de datos en un módulo de pre-procesado puede mejorar los resultados del reconocimiento VSR en continuo. Proponemos dos lineas de actuación complementarias:

- **Pre-segmentación de eventos mediante técnicas de detección del frente de ondas**. Los sistemas de *picking* que son capaces de detectar la llegada del frente de ondas primarias se han estado usando como discriminador de eventos sísmicos desde hace más de 40 años (Leprettre et al., 1998; Withers et al., 1998). En la actualidad han evolucionado hasta convertirse en sistemas

muy efectivos para detectar eventos (Beyreuther et al., 2012; Alvarez et al., 2013)

- **Pre-segmentación de eventos usando discriminantes de ruido.** En Beyreuther and Wassermann (2008) la segmentación de los sismos se efectúa cuando se cruza un umbral de mínima probabilidad marcado por el triple de la probabilidad asociada al ruido. Se promedia la probabilidad en cada clase usando 3 D.HMMs con distinto nº de estados. En una fase de post-procesado se utilizan filtros de mínima duración que pretenden solucionar errores de clasificación. En una evaluación sobre un mes entero de registro, este sistema de segmentación consigue segmentar un 81 % de los sismos frente a un 90 % de un detector basado en una implementación del algoritmo STA/LTA.

Como una aplicación más del sistema VSR-PSA, puede ser muy interesante comparar los anteriores detectores con la segmentación de eventos proporcionada a la salida de un canal PSA específicamente diseñado para detectar el ruido (del que partimos con un resultado inicial de un 82 % para la arquitectura PSA.mul(NS) en Decepción, Tabla 6.6.2)

8.1.3. Reconocimiento con las 3 componentes de la señal

En la mayoría de los sistemas VSR estudiados en la Sección 2.3, se utiliza solo la componente vertical (Z) de los sismogramas, bien por cuestiones de simplicidad o por no estar disponibles las otras 2 (N y E). El mayor inconveniente al usar 3D es simple: el sistema triplicaría su coste computacional. Nosotros proponemos dos técnicas complementarias:

- Sistemas que en el proceso de descripción extraigan las características de cada eje Z, N y E, las multiplexen en un vector de características de una sola dimensión para luego dejar que el algoritmo *DFS.cAcc* seleccione las mejores componentes y reduzca la dimensionalidad del vector a la 3ª parte, para obtener un vector final de igual tamaño que el utilizado inicialmente en la dirección Z.
- Un vector de parametrización que extraiga características mixtas obtenidas a partir de las 3 componentes direccionales, como por ejemplo el azimut, la polarización, la energía o el retardo entre frentes de ondas PS. Hay que tener en cuenta que muchas de estas características pueden no ser exportables al ser demasiado locales o estar afectadas por efectos de sitio.

Nótese uno de los objetivos principales de estudiar las 3 componentes del sismograma desde el punto de vista de la descripción de eventos es la posibilidad de encontrar características que nos permitan diferenciar entre eventos con una mayor energía asociada a ondas superficiales que a ondas volumétricas. Otra ventaja es poder discriminar más fácilmente entre sismos que tengan una clara llegada de ondas S frente a aquellos que no.

8.2. Lineas secundarias de investigación

8.2.1. Reconocimiento a nivel de sub-evento

A pesar de ser una evolución básica, inspirada en el reconocimiento de voz, existen pocas investigaciones encontradas en torno a este enfoque. Cabe destacar el trabajo pionero de Liu and Fu (1983) que de forma automática divide sus eventos en 13 primitivas y consigue muy buenos resultados de reconocimiento (91 % discriminando entre dos clases) con un sencillo clasificador basado en vecindades.

8.2.2. Uso de otros modelos en la arquitectura PSA

Como alternativa a los modelos GMM y HMM propuestos, la estructura del sistema VSR-PSA permite el uso distintos tipos de clasificadores para distintos canales. La mejora del clasificador conjunto es otra linea de investigación abierta.

8.2.2.1. Uso de modelos probabilísticos alternativos

La flexibilidad el sistema en paralelo VSR-PSA permite usar clasificadores específicos para cada canal con la única condición de que sean capaces de proporcionar las probabilidades $\{p(\mathbf{x}, w_c)\}$ de cada frame de la secuencia $\mathbf{x} = \{\mathbf{x}_t\}$ para cada uno de los modelos de las clases $\{w_c\}$. Las siguientes lineas de investigación son prometedoras:

- *Usar clasificadores no-estructurados para clases sin evolución temporal marcada.*
- *Implementar clasificadores que específicamente proporcionen* $\{p(\mathbf{x}, w_c)\}$. La alternativa inmediata es usar los GMM. Sin embargo, las pruebas iniciales realizadas no han logrado igualar los resultados de los HMMs, probablemente por su falta de modelado estructurado. En este sentido, otros clasificadores probabilísticos basados en redes bayesianas pueden resultar interesantes. O, incluso, usar librerías nativas en Python de modelos HMM que sean capaces de dar $\{p(\mathbf{x}, w_c)\}$, dado que, como vimos en Sección A.7, mediante HTK solo podemos estimar estas probabilidades.

8.2.2.2. Mejora del clasificador conjunto

El clasificador conjunto del sistema VSR-PSA se ha implementado de forma que permita incluir la funcionalidad de reconocimiento de un sistema serie VSR-SSA. Con el objetivo de incrementar la eficiencia de clasificación pueden diseñarse esquemas más completos del reconocedor o incorporar ciertas mejoras básicas como la integración explícita del algoritmo de Viterbi en el clasificador conjunto.

8.2.3. Modelado explícito de la gramática y del lenguaje

De nuevo, inspirándonos en los sistemas de reconocimiento del habla, en el caso de que tengamos un corpus suficientemente grande de datos en continuo, podría ser interesante modelar explícitamente las secuencias de los eventos detectados o *lenguaje* en un doble nivel (Subsubsección 2.1.4.3):

1. *Modelado gramatical*: que hasta ahora hemos usado tan solo para distinguir entre reconocimiento en continuo o clasificación de eventos aislados, puede extenderse para forzar que un evento EV esté delimitado entre señales de modelos NS de ruido ($< NS - EV - NS >$), para definir explícitamente un evento híbrido HY como un terremoto volcano-tectónico VT seguido de uno de largo periodo LP ($HY = VT - LP$;) o que un tremor pulsante TP es una secuencia de pequeños sismos de baja frecuencia sobre un tremor espasmódico TS ($TP =< TS - LP >$).
2. *Modelado estadístico del lenguaje*: Puede ser interesante combinar el modelado de la señal con un modelado estructural a nivel de secuencias tal y como se hace en voz para corpus de gran tamaño. Estadísticamente puede construirse un modelo de predicción del evento en el instante actual t si conocemos los N eventos que le preceden. Este tipo de modelado ha demostrado su eficacia en bases de datos con muchos eventos en el ámbito del reconocimiento del habla. Incluso, desde el punto de vista de evaluación del riesgo y prevención de desastres, es muy interesante utilizar distintos modelos en función de la sismicidad histórica de cada volcán o de la fase eruptiva en la que se encuentre.

Bibliografía

Acernese, F., Ciaramella, A., De Martino, S., De Rosa, R., Falanga, M., Tagliaferri, R., 2003. Neural networks for blind-source separation of Stromboli explosion quakes. Neural Networks, IEEE Transactions on 14 (1), 167–175.

Aizerman, A., Braverman, E. M., Rozoner, L., 1964. Theoretical foundations of the potential function method in pattern recognition learning. Automation and remote control 25, 821–837.

Aki, K., 1957. Space and time spectra of stationary stochastic waves, with special reference to microtremors. Bulletin of the Earthquake Research Institute 35, 415–456.

Aki, K., Chouet, B., Fehler, M., Zandt, G., Koyanagi, R., Colp, J., Hay, R. G., 1978. Seismic properties of a shallow magma reservoir in Kilauea Iki by active and passive experiments. Journal of Geophysical Research: Solid Earth (1978–2012) 83 (B5), 2273–2282.

Aki, K., Christoffersson, A., Husebye, E. S., 1977. Determination of the three-dimensional seismic structure of the lithosphere. Journal of Geophysical Research 82 (2), 277–296.

Alasonati, P., Wassermann, J., Ohrnberger, M., 2006. Signal classification by wavelet based hidden Markov models: application to seismic signals of volcanic origin. Statistics in Volcanology (1), 161–174.

Allan, J., Carmichael, I., 1984. Lamprophyric lavas in the Colima graben, SW Mexico. Contributions to Mineralogy and Petrology 88 (3), 203–216.

Allen, R. V., 1978. Automatic earthquake recognition and timing from single traces. Bulletin of the Seismological Society of America 68 (5), 1521–1532.

Almendros, J., 1999. Análisis de señales sismo-volcánicas mediante técnicas de array. Ph.D. thesis, Universidad de Granada.

Almendros, J., Alguacil, G., Del Pezzo, E., Ortiz, R., 1997. Array tracking of the volcanic tremor source at Deception Island, Antarctica. Geophysical Research Letters 24 (23), 3069–3072.

Almendros, J., Ibáñez, J. M., Alguacil, G., Del Pezzo, E., Jan. 1999. Array analysis using circular-wave-front geometry: an application to locate the nearby seismo-volcanic source. Geophysical Journal International 136 (1), 159–170.

Álvarez, I., Cortés, G., De la Torre, A., Benítez, C., García, L., Lesage, P., Arambula, R., González, M., 2009. Improving feature extraction in the automatic classification of seismic events. Application to Colima and Arenal volcanoes. In: Geoscience and Remote Sensing Symposium, 2009 IEEE International, IGARSS 2009. Vol. 4. IEEE, p. 526.

Álvarez, I., García, L., Cortés, G., Benítez, C., De la Torre, A., 2011. Discriminative Feature Selection for Automatic Classification of Volcano-Seismic Signals. Geoscience and Remote Sensing Letters, IEEE (99), 1–5.

Álvarez, I., García, L., Cortés, G., Benítez, C., de la Torre, A., Ibáñez, J., 2010. Comparison of volcano-seismic signals from different volcanoes: characterization of different sources os events and strategies for the evaluation of automatic recognizers of volcano-seismic events. In: CITIES ON VOLCANOES, 6TH EDITION. TENERIFE 2010 . pp. 97–97.

Alvarez, I., Garcia, L., Mota, S., Cortes, G., Benitez, C., De la Torre, A., 2013. An Automatic P-Phase Picking Algorithm Based on Adaptive Multiband Processing. Geoscience and Remote Sensing Letters, IEEE 10 (6), 1488–1492.

Anderson, K. R., 1978. Automatic analysis of microearthquake network data. Geoexploration 16 (1–2), 159–175.

Arámbula, R., 2011. Clasificación automática de eventos sísmicos volcánicos y análisis de la actividad sísmica reciente en el Volcán de Colima. Ph.D. thesis, Universidad Nacional Autónoma de México.

Arámbula-Mendoza, R., Lesage, P., Valdés-González, C., Varley, N., Reyes-Dávila, G., Navarro, C., 2011. Seismic activity that accompanied the effusive and explosive eruptions during the 2004–2005 period at Volcán de Colima, Mexico. Journal of Volcanology and Geothermal Research 205 (1–2), 30–46.

Arauzo-Azofra, A., Aznarte, J. L., Benítez, J. M., 2011. Empirical study of feature selection methods based on individual feature evaluation for classification problems. Expert Systems with Applications 38 (7), 8170–8177.

Arciniega-Ceballos, A., Chouet, B. A., Dawson, P., 1999. Very long-period signals associated with vulcanian explosions at Popocatepetl Volcano, Mexico. Geophysical Research Letters 26 (19), 3013–3016.

Aspinall, W., Carniel, R., Jaquet, O., Woo, G., Hincks, T., 2006. Using hidden multistate Markov models with multi-parameter volcanic data to provide empirical evidence for alert level decision-support. Journal of volcanology and geothermal research 153 (1), 112–124.

Avesani, R., Azzoni, A., Bicego, M., Orozco-Alzate, M., 2012. Automatic Classification of Volcanic Earthquakes in HMM-Induced Vector Spaces. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Springer, pp. 640–647.

Avossa, C., Giudicepietro, F., Marinaro, M., Scarpetta, S., 2003. Supervised and Unsupervised Analysis Applied to Strombolian E.Q. In: Apolloni, B., Marinaro, M., Tagliaferri, R. (Eds.), Neural Nets. Vol. 2859 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 173–178.

Baker, P. E., McReath, I., Harvey, M., Roobol, M., Davies, T., 1975. The geology of the South Shetland Islands: V. Volcanic evolution of Deception Island. Vol. 78. British Antarctic Survey.

Bamler, R., Eineder, M., Adam, N., Zhu, X., Gernhardt, S., 2009. Interferometric potential of high resolution spaceborne SAR. Photogrammetrie-Fernerkundung-Geoinformation 2009 (5), 407–419.

Barberi, F., Bertagnini, A., Landi, P., Principe, C., 1992. A review on phreatic eruptions and their precursors. Journal of volcanology and geothermal research 52 (4), 231–246.

Bartlett, M. S., 2001. Face Image Analysis by Unsupervised Learning. Kluwer Academic Publishers, Norwell, MA, USA.

Baum, L. E., Eagon, J., et al., 1967. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. Bull. Amer. Math. Soc 73 (3), 360–363.

Baum, L. E., Petrie, T., 1966. Statistical inference for probabilistic functions of finite state Markov chains. The annals of mathematical statistics, 1554–1563.

Baum, L. E., Petrie, T., Soules, G., Weiss, N., 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The annals of mathematical statistics, 164–171.

Baum, L. E., Sell, G., 1968. Growth transformations for functions on manifolds. Pacific Journal of Mathematics 27 (2), 211–227.

Bengio, Y., 2009. Learning deep architectures for AI. Foundations and trends® in Machine Learning 2 (1), 1–127.

Benítez, C., Ibáñez, J., García, L., Cortés, G., Álvarez, I., March 2009. Analysis of volcanic seismicity at Deception Island, Stromboli volcano and Mt. Etna using an automatic CHMM-based recognition method. In: Bean, C. B. A., 6th Framework, E. C. P. (Eds.), VOLUME Project: VOLcanoes, Understanding Subsurface Mass MoveMEnt. Bean, C.J., Braiden, A.K., Lokmer, I., Martini, F. and O'Brien, G.S, School of Geological Sciences, University College Dublin, pp. 140–149.

Benítez, C., Ramírez, J., Segura, J. C., Ibáñez, J., Almendros, J., García-Yeguas, A., Cortés, G., 2007. Continuous HMM-based seismic-event classification at Deception Island, Antarctica. Geoscience and Remote Sensing, IEEE Transactions on 45 (1), 138–146.

Bermejo, P., Gámez, J. A., Puerta, J. M., 2011. A {GRASP} algorithm for fast hybrid (filter-wrapper) feature subset selection in high-dimensional datasets. Pattern Recognition Letters 32 (5), 701–711.

Beyreuther, M., Carniel, R., Wassermann, J., 2008. Continuous Hidden Markov Models: Application to automatic earthquake detection and classification at Las Cañadas caldera, Tenerife. Journal of Volcanology and Geothermal Research 176 (4), 513–518.

Beyreuther, M., Hammer, C., Wassermann, J., Ohrnberger, M., Megies, T., 2012. Constructing a Hidden Markov Model based earthquake detector: application to induced seismicity. Geophysical Journal International 189 (1), 602–610.

Beyreuther, M., Wassermann, J., 2008. Continuous earthquake detection and classification using discrete Hidden Markov Models. Geophysical Journal International 175 (3), 1055–1066.

Beyreuther, M., Wassermann, J., 2011. Hidden semi-Markov Model based earthquake classification system using Weighted Finite-State Transducers. Nonlinear Processes in Geophysics 18, 81–89.

Bezdek, J. C., Pal, S. K., 1992. Fuzzy models for pattern recognition. Vol. 56. IEEE Press, New York.

Bhanu, B., Lin, Y., 2003. Genetic algorithm based feature selection for target detection in SAR images. Image and Vision Computing 21 (7), 591–608.

Bicego, M., Acosta-Muñoz, C., Orozco-Alzate, M., 2013. Classification of Seismic Volcanic Signals Using Hidden-Markov-Model-Based Generative Embeddings. Geoscience and Remote Sensing, IEEE Transactions on 51 (6), 3400–3409.

Biesiada, J., Duch, W., 2007. Feature selection for high-dimensional data—a Pearson redundancy based filter. In: Computer Recognition Systems 2. Springer, pp. 242–249.

Bishop, C. M., 1995. Neural networks for pattern recognition. Oxford university press.

Bishop, C. M., 2007. Pattern recognition and machine learning (information science and statistics). Springer.

Blaschke, T., Berkes, P., Wiskott, L., 2006. What is the relation between slow feature analysis and independent component analysis? Neural Computation 18 (10), 2495–2508.

Blaschke, T., Wiskott, L., 2004. CuBICA: independent component analysis by simultaneous third- and fourth-order cumulant diagonalization. Signal Processing, IEEE Transactions on 52 (5), 1250–1256.

Bolstad, W. M., 2004. Introduction to Bayesian statistics. John Wiley & Sons.

Boser, B. E., Guyon, I. M., Vapnik, V. N., 1992. A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory. ACM, pp. 144–152.

Bouchard, G., Triggs, B., et al., 2004. The tradeoff between generative and discriminative classifiers. In: 16th IASC International Symposium on Computational Statistics (COMPSTAT'04). pp. 721–728.

Boué, A., Lesage, P., Cortés, G., Valette, B., Reyes-Dávila, G., 2015. Real-time eruption forecasting using the material Failure Forecast Method with a Bayesian approach. Journal of Geophysical Research: Solid Earth 120 (4), 2143–2161, 2014JB011637.
URL http://dx.doi.org/10.1002/2014JB011637

Bouvier, R., 1972. Seismic Pattern Recognition. Master's thesis, Ohio School of Engineering.

Bowman, B. C., Dowla, F., May 1992. Real-time classification of signals from three-component seismic sensors using neural nets. Presented at the IEEE Transactions on Instrumentation and Measurement and 4th Institut Industriel De Transfert De Technologie (IITT) International Conference on Artificial Intelligence and Expert Systems, Paris (France), 21-22 Oct. 1992, 21–22.

Bretón, M., Ramírez, J., Navarro, C., 2002. Summary of the historical eruptive activity of Volcán de Colima, Mexico 1519–2000. Journal of Volcanology and Geothermal Research 117 (1), 21–46.

Cabras, G., 2011. Advanced component analysis techniques for signal decomposition and their applications to audio restoration and volcanic seismology. Ph.D. thesis, Università degli studi di Udine.

Cabras, G., Carniel, R., Wassermann, J., 2008. Blind source separation: An application to the mt. Merapi volcano, Indonesia. Fluctuation and Noise Letters 8 (03n04), L249–L260.

Cabras, G., Carniel, R., Wassermann, J., 2010. Signal enhancement with generalized ICA applied to Mt. Etna volcano, Italy. Bollettino di Geofisica Teorica ed Applicata 51 (1), 57–73.

Cárdenas-Peña, D., Orozco-Alzate, M., Castellanos-Dominguez, G., 2013. Selection of time-variant features for earthquake classification at the Nevado-del-Ruiz volcano. Computers & Geosciences 51 (0), 293–304.

Cardoso, J.-F., 1997. Infomax and maximum likelihood for blind source separation. Signal Processing Letters, IEEE 4 (4), 112–114.

Carmona, E., Almendros, J., Serrano, I., Stich, D., Ibáñez, J., 2012. Results of seismic monitoring surveys of Deception Island volcano, Antarctica, from 1999–2011. Antarctic Science 24 (05), 485–499.

Carniel, R., Di Cecca, M., Jaquet, O., 2006. A user-friendly, dynamic web environment for remote data browsing and analysis of multiparametric geophysical data within the MULTIMO project. Journal of volcanology and geothermal research 153 (1), 80–96.

Caruana, R., 1996. Algorithms and Applications for Multitask Learning. In: In Proceedings of the Thirteenth International Conference on Machine Learning. Morgan Kaufmann, pp. 87–95.

Chen, C., 1977. COMPARISON OF SEISMIC FEATURES EXTRACTED BY DIGITAL SIGNAL PROCESSING TECHNIQUES. 2, 148–150.

Chen, C., Apr 1978a. On digital signal modelling and classification with the teleseismic data. In: Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '78. Vol. 3. pp. 529–531.

Chen, C., 1978b. Seismic pattern recognition. Geoexploration 16 (1), 133–146.

Chen, C., Wang, Z., 2006. ICA and factor analysis application in seismic profiling. In: Geoscience and Remote Sensing Symposium, 2006. IGARSS 2006. IEEE International Conference on. IEEE, pp. 1560–1563.

Chouet, B., 1981. Ground motion in the near field of a fluid-driven crack and its interpretation in the study of shallow volcanic tremor. Journal of Geophysical Research: Solid Earth (1978–2012) 86 (B7), 5985–6016.

Chouet, B., 1996a. New methods and future trends in seismological volcano monitoring. In: Scarpa, R., Tilling, R. (Eds.), Monitoring and mitigation of volcano hazards. Springer-Verlag, pp. 23–97.

Chouet, B., 2003. Volcano seismology. Pure and Applied Geophysics 160 (3-4), 739–788.

Chouet, B. A., 1996b. Long-period volcano seismicity: its source and use in eruption forecasting. Nature 380 (6572), 309–316.

Chouet, B. A., Matoza, R. S., 2013. A multi-decadal view of seismic methods for detecting precursors of magma movement and eruption. Journal of Volcanology and Geothermal Research 252, 108–175.

Chouet, B. A., Page, R. A., Stephens, C. D., Lahr, J. C., Power, J. A., 1994. Precursory swarms of long-period events at Redoubt Volcano (1989–1990), Alaska: Their origin and use as a forecasting tool. Journal of Volcanology and Geothermal Research 62 (1), 95–135.

Cichocki, A., Georgiev, P., 2003. Blind source separation algorithms with matrix constraints. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences 86 (3), 522–531.

Comon, P., 1994a. Independent component analysis, a new concept? Signal processing 36 (3), 287–314.

Comon, P., 1994b. Tensor diagonalization, a useful tool in signal processing. In: 10th IFAC Symposium on System Identification (IFAC-SYSID). Vol. 1. pp. 77–82.

Cortés, A., Garduño Monroy, V., Navarro-Ochoa, C., Komorowski, J., Saucedo, R., Macías, J., Gavilanes, J., 2005. Cartas Geológicas y Mineras 10. Carta Geológica del Complejo Volcánico de Colima, con Geología del Complejo Volcánico

de Colima: México DF, Universidad Nacional Autónoma de México, Instituto de Geología, escala 1 (10,000).

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine learning 20 (3), 273–297.

Cortés, G., Arámbula, R., Álvarez, I., Benítez, C., Ibáñez, J., Lesage, P., González-Amézcua, M., Reyes-Dávila, G., March 2009a. Analysis of Colima, Popocatépetl and Arenal volcanic seismicity using an automatic Continuous Hidden Markov Models-based recognition system. In: Bean, C. B. A., 6th Framework, E. C. P. (Eds.), VOLUME Project: VOLcanoes, Understanding Subsurface Mass MovemENt. Bean, C.J., Braiden, A.K., Lokmer, I., Martini, F. and O'Brien, G.S, School of Geological Sciences, University College Dublin, pp. 150–160.

Cortés, G., Arambula, R., Gutiérrez, L., Benítez, C., Ibáñez, J., Lesage, P., Álvarez, I., García, L., 2009b. Evaluating robustness of a HMM-based classification system of volcano-seismic events at colima and popocatepetl volcanoes. In: Geoscience and Remote Sensing Symposium, 2009 IEEE International, IGARSS 2009. Vol. 2. IEEE, p. 1012.

Cortés, G., Benítez, M., García, L., A. I., Ibáñez, J., 2015. A Comparative Study of Dimensionality Reduction Algorithms Applied to Volcano-Seismic Signals. Accepted for publication.

Cortés, G., García, L., Álvarez, I., Benítez, C., de la Torre, A., Ibáñez, J., 2014. Parallel System Architecture (PSA): An efficient approach for automatic recognition of volcano-seismic events. Journal of Volcanology and Geothermal Research 271 (0), 1–10.

Curilem, G., Vergara, J., Fuentealba, G., Acuña, G., Chacón, M., 2009. Classification of seismic signals at Villarrica Volcano (Chile) using neural networks and genetic algorithms. Journal of Volcanology and Geothermal Research 180 (1), 1–8.

Curilem, M., Huenupan, F., San Martin, C., Fuentealba, G., Cardona, C., Franco, L., Acuña, G., Chacón, M., 2014a. Feature Analysis for the Classification of Volcanic Seismic Events Using Support Vector Machines. In: Nature-Inspired Computation and Machine Learning. Springer, pp. 160–171.

Curilem, M., Vergara, J., San Martin, C., Fuentealba, G., Cardona, C., Huenupan, F., Chacón, M., Khan, M. S., Hussein, W., Yoma, N. B., 2014b. Pattern recognition applied to seismic signals of the Llaima volcano (Chile): An analysis of the events' features. Journal of Volcanology and Geothermal Research 282, 134–147.

De la Torre, A., Peinado, A., Rubio, A., Sánchez, V., 1997. A DFE-based algorithm for feature selection in speech recognition. In: Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on. Vol. 2. IEEE, pp. 1519–1522.

De Lauro, E., De Martino, S., Falanga, M., Palo, M., 2009. Decomposition of high-frequency seismic wavefield of the Strombolian-like explosions at Erebus volcano

by independent component analysis. Geophysical Journal International 177 (3), 1399–1406.

Del Pezzo, E., Esposito, A., Giudicepietro, F., Marinaro, M., Martini, M., Scarpetta, S., Feb. 2003. Discrimination of earthquakes and underwater explosions using neural networks. Bulletin of the Seismological Society of America 93 (1), 215–223.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 1–38.

Dowla, F. U., Taylor, S. R., Anderson, R. W., 1990. Seismic discrimination with artificial neural networks: preliminary results with regional spectral data. Bulletin of the Seismological Society of America 80 (5), 1346–1373.

Draper, B. A., Baek, K., Bartlett, M. S., Beveridge, J. R., 2003. Recognizing faces with PCA and ICA. In: Computer Vision And Image Understanding, special issue on face recognition. pp. 115–137.

Drugman, T., Gurban, M., Thiran, J., 2007. Relevant Feature Selection for Audio-Visual Speech Recognition. In: Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on. IEEE, pp. 179–182.

Duin, R. P., Orozco-Alzate, M., Londono-Bonilla, J. M., 2010. Classification of volcano events observed by multiple seismic stations. In: Pattern Recognition (ICPR), 2010 20th International Conference on. IEEE, pp. 1052–1055.

Dupont, S., Luettin, J., 2000. Audio-visual speech modeling for continuous speech recognition. Multimedia, IEEE Transactions on 2 (3), 141–151.

Endo, E. T., Murray, T., 1991. Real-time seismic amplitude measurement (RSAM): a volcano monitoring and prediction tool. Bulletin of Volcanology 53 (7), 533–545.

Escalante, B., Alberto, N., Wiskott, L., 2012. Slow feature analysis: Perspectives for technical applications of a versatile learning algorithm. KI-Künstliche Intelligenz 26 (4), 341–348.

Esposito, A., Giudicepietro, F., D'Auria, L., Scarpetta, S., Martini, M., Coltelli, M., Marinaro, M., 2008a. Unsupervised neural analysis of very-long-period events at Stromboli volcano using the self-organizing maps. Bulletin of the Seismological Society of America 98 (5), 2449–2459.

Esposito, A., Giudicepietro, F., D'Auria, L., Scarpetta, S., Martini, M., Coltelli, M., Marinaro, M., 2008b. Unsupervised neural analysis of very-long-period events at Stromboli volcano using the self-organizing maps. Bulletin of the Seismological Society of America 98 (5), 2449–2459.

Esposito, A. M., Scarpetta, S., Giudicepietro, F., Masiello, S., Pugliese, L., Esposito, A., 2006. Nonlinear exploratory data analysis applied to seismic signals. In: Neural Nets. Springer, pp. 70–77.

Falsaperla, S., Fortuna, L., Graziani, S., Nunnari, G., 1992. Automatic classification of seismic events by neural networks. In: IGARSS '92; Proceedings of the 12th Annual International Geoscience and Remote Sensing Symposium. pp. 224–226.

Faure, C., Soldano, H., Van Der Pyl, T., 1984. Descriptive methods and processing of seismic signals. Geoexploration 23 (1), 17–34.

Fehler, M., Chouet, B., 1982. Operation of a digital seismic network on Mount St. Helens volcano and observations of long period seismic events that originate under the volcano. Geophysical Research Letters 9 (9), 1017–1020.

Feig, E., Winograd, S., 1992. Fast algorithms for the discrete cosine transform. Signal Processing, IEEE Transactions on 40 (9), 2174–2193.

Filzmoser, P., Hron, K., Reimann, C., Garrett, R., 2009. Robust factor analysis for compositional data. Computers & Geosciences 35 (9), 1854–1861.

Fischer, T. P., Morrissey, M. M., Calvache, V. M. L., Gomez, M. D., Torres, C. R., Stix, J., Williams, S. N., 1994. Correlations between SO2 flux and long-period seismicity at Galeras volcano. Nature 368 (6467), 135–137.

Fisher, R. A., 1922. On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 309–368.

Fisher, R. A., 1936. The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics 7 (2), 179–188.

Fogel, D. B., 2006. Evolutionary computation: toward a new philosophy of machine intelligence. Vol. 1. John Wiley & Sons.

Fortuna, L., Graziani, S., Lo Presti, M., Nunnari, G., Jun 1991. A Neural Network For Seismic Events Classification. In: Geoscience and Remote Sensing Symposium, 1991. IGARSS '91. Remote Sensing: Global Monitoring for Earth Management., International. Vol. 3. pp. 1663–1666.

Franzius, M., Wilbert, N., Wiskott, L., 2008. Invariant Object Recognition with Slow Feature Analysis. In: Kůrková, V., Neruda, R., Koutník, J. (Eds.), Artificial Neural Networks - ICANN 2008. Vol. 5163 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 961–970.

Friedman, J. H., 1997. On bias, variance, 0/1—loss, and the curse-of-dimensionality. Data mining and knowledge discovery 1 (1), 55–77.

Friedman, N., Geiger, D., Goldszmidt, M., 1997. Bayesian network classifiers. Machine learning 29 (2-3), 131–163.

Fukunaga, K., 1990. Introduction to statistical pattern recognition. Access Online via Elsevier.

Gaby, J. E., Anderson, K. R., 1984. Hierarchical segmentation of seismic waveforms using affinity. Geoexploration 23 (1), 1–16, seismic Signal Analysis and Discrimination {III}.

Gales, M., Young, S., 2008. The application of hidden Markov models in speech recognition. Foundations and Trends in Signal Processing 1 (3), 195–304.

Galli, L., Castellani, C., Page, G., Saccorotti, G., March 2009. Wavelet decomposition and advanced denoising techniques for analysis and classification of seismic signals. In: Bean, C. B. A., 6th Framework, E. C. P. (Eds.), VOLUME Project: VOLcanoes, Understanding Subsurface Mass MoveMEnt. Bean, C.J., Braiden, A.K., Lokmer, I., Martini, F. and O'Brien, G.S, School of Geological Sciences, University College Dublin, pp. 118–129.

Gammerman, A., Vovk, V., Vapnik, V., 1998. Learning by transduction. In: Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., pp. 148–155.

García, L., Ibáñez, J., Cortés, G., Álvarez, I., de la Torre, A., 2010. Automatic and semi-automatic tools for segmentation and labelling large databases of volcano-seismic signals. In: CITIES ON VOLCANOES, 6TH EDITION. TENERIFE 2010 . pp. 97–97.

García-Yeguas, A., Koulakov, I., Ibáñez, J. M., Rietbrock, A., 2012. High resolution 3D P wave velocity structure beneath Tenerife Island (Canary Islands, Spain) based on tomographic inversion of active-source data. Journal of Geophysical Research: Solid Earth (1978–2012) 117 (B9).

Garduño Monroy, V. H., Saucedo-Girón, R., Jiménez, Z., Gavilanes-Ruiz, J. C., Cortes-Cortés, A., Uribe-Cifuentes, R. M., 1998. La Falla Tamazula, limite suroriental del bloque Jalisco, y sus relaciones con el complejo volcánico de Colima, Mexico. Revista Mexicana de Ciencias Geológicas 15 (2), 132–144.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., 1993. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA STI/Recon Technical Report N 93, 27403.

Gerbrands, J. J., 1981. On the relationships between SVD, KLT and PCA. Pattern recognition 14 (1), 375–381.

Ghahramani, Z., Hinton, G. E., et al., 1996. The EM algorithm for mixtures of factor analyzers. Tech. rep., Technical Report CRG-TR-96-1, University of Toronto.

Giacinto, G., Paolucci, R., Roli, F., 1997. Application of neural networks and statistical pattern recognition algorithms to earthquake risk evaluation. Pattern Recognition Letters 18 (11-13), 1353–1362.

González, M. B., 2011. El volcán de Fuego de Colima: seis siglos de actividad eruptiva, 1523-2010. Universidad de Colima.

Gonzalez-Amezcua, M., Arambula-Mendoza, R., Reyes-Davila, G., Cortes, G., Lesage, P., Benitez, C., Ibáñez, J., Valdes-Gonzalez, C., 2012. Automated classification of volcanic seismic signals using Hidden Markov Models (HMMs) in quasi realtime at Volcan de Colima. In: Cities on Volcanos, 7th Edition.

González-Ferrán, O., et al., 1995. Volcanes de Chile. Instituto Geográfico Militar Chileno.

Gutiérrez Espinoza, L. A., Octubre 2013. Sistema de detección y clasificación de señales sísmico-volcánicas utilizando Modelos Ocultos de Markov (HMMs): Aplicación a volcanes activos de Nicaragua e Italia. Ph.D. thesis, Universidad de Granada.

Guyon, I., 2008. Practical feature selection: from correlation to causality. Mining massive data sets for security: advances in data mining, search, social networks and text mining, and their applications to security, 27–43.

Guyon, I., Elisseeff, A., Mar. 2003. An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182.

Haindl, M., Somol, P., Ververidis, D., Kotropoulos, C., 2006. Feature selection based on mutual correlation. In: Progress in Pattern Recognition, Image Analysis and Applications. Springer, pp. 569–577.

Hall, M. A., 1999. Correlation-based feature selection for machine learning. Ph.D. thesis, The University of Waikato.

Ham, F. M., Iyengar, I., Hambebo, B. M., Garces, M., Deaton, J., Perttu, A., Williams, B., 2012. A Neurocomputing Approach for Monitoring Plinian Volcanic Eruptions Using Infrasound. Procedia Computer Science 13 (0), 7–17, proceedings of the International Neural Network Society Winter Conference (INNS-WC2012).

Hardoon, D. R., Szedmak, S., Shawe-Taylor, J., 2004. Canonical correlation analysis: An overview with application to learning methods. Neural Computation 16 (12), 2639–2664.

Harman, H. H., 1967. Modern factor analysis.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., Tibshirani, R., 2009. The elements of statistical learning. Vol. 2. Springer.

Havskov, J., Alguacil, G., 2010. Instrumentation in earthquake seismology. Vol. 22. Springer Science & Business Media.

Helz, R. T., 1993. Drilling report and core logs for the 1988 drilling of Kilauea Iki lava lake, Kilauea Volcano, Hawaii, with summary descriptions of the occurrence of foundered crust and fractures in the drill core. Tech. rep., US Geological Survey,.

Hloupis, G. P., 2009. Seismological data acquisition and signal processing using wavelets. Ph.D. thesis, School of Engineering and Design PhD Theses.

Hoogenboezem, R., 2010. Automatic classification of segmented seismic recordings at the nevado del ruiz volcano, colombia. Master's thesis, Master's thesis, Delft University of Technology.

Hsu, H.-H., Lu, M.-D., 2008. Feature Selection for Cancer Classification on Microarray Expression Data. In: Intelligent Systems Design and Applications, 2008. ISDA'08. Eighth International Conference on. Vol. 3. IEEE, pp. 153–158.

Hyvarinen, A., 1999. Fast and robust fixed-point algorithms for independent component analysis. Neural Networks, IEEE Transactions on 10 (3), 626–634.

Hyvarinen, A., Karhunen, J., Oja, E., 2001. Independent Component Analysis. Wiley, New York.

Hyvärinen, A., Oja, E., 2000. Independent component analysis: algorithms and applications. Neural networks 13 (4), 411–430.

Ibáñez, J., Benítez, C., Gutiérrez, L., Cortés, G., García-Yeguas, A., Alguacil, G., 2009. The classification of seismo-volcanic signals using Hidden Markov Models as applied to the Stromboli and Etna volcanoes. Journal of Volcanology and Geothermal Research 187 (3-4), 218–226.

Ibáñez, J., Carmona, E., Almendros, J., Saccorotti, G., Del Pezzo, E., Abril, M., Ortiz, R., 2003. The 1998–1999 seismic series at Deception Island volcano, Antarctica. Journal of volcanology and geothermal research 128 (1), 65–88.

Ibáñez, J., Del Pezzo, E., Almendros, J., La Rocca, M., Alguacil, G., Ortíz, R., García, A., 2000. Seismovolcanic signals at Deception Island volcano, Antarctica: Wave field analysis and source modeling. Journal of Geophysical Research 105 (B6).

Ives, R., 1975. Dynamic spectral ratios as features in seismological pattern recognition. In: Proc. Conf. Computer Graphics, Pattern Recognition Data Structure. pp. 211–213.

Jain, A. K., Duin, R. P. W., Mao, J., 2000. Statistical pattern recognition: A review. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE 22 (1), 4–37.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning. Springer.

Jiang, H., Li, X., Liu, C., 2006. Large margin hidden Markov models for speech recognition. Audio, Speech, and Language Processing, IEEE Transactions on 14 (5), 1584–1595.

Jiménez, A., 2003. Una visión unificada de las redes neuronales y la estadística multivariante. Ph.D. thesis, Cádiz: Universidad de Cádiz.

John, G. H., Kohavi, R., Pfleger, K., et al., 1994. Irrelevant Features and the Subset Selection Problem. In: ICML. Vol. 94. pp. 121–129.

John, H., 1992. Holland, Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence.

Johnson, C. E., Bittenbinder, A., Bogaert, B., Dietz, L., Kohler, W., 1995. Earthworm: A flexible approach to seismic network processing. Iris newsletter 14 (2), 1–4.

Jolliffe, I. T., 2002. Principal Component Analysis. Springer Verlag, New York.

Jones, E., Oliphant, T., Peterson, P., et al., 2001. SciPy: Open source scientific tools for Python.
URL http://www.scipy.org/

Joswig, M., 1990. Pattern recognition for earthquake detection. Bulletin of the Seismological Society of America 80 (1), 170–186.

Kawakatsu, H., Kaneshima, S., Matsubayashi, H., Ohminato, T., Sudo, Y., Tsutsui, T., Uhira, K., Yamasato, H., Ito, H., Legrand, D., 2000. Aso94: Aso seismic observation with broadband instruments. Journal of Volcanology and Geothermal Research 101 (1), 129–154.

Kilburn, C. R., 2003. Multiscale fracturing as a key to forecasting volcanic eruptions. Journal of Volcanology and Geothermal Research 125 (3), 271–289.

King, D. E., 2009. Dlib-ml: A Machine Learning Toolkit. Journal of Machine Learning Research 10, 1755–1758.

Kira, K., Rendell, L. A., 1992. A Practical Approach to Feature Selection. In: Proceedings of the Ninth International Workshop on Machine Learning. ML '92. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 249–256.

Köhler, A., Ohrnberger, M., Scherbaum, F., Sep. 2009. Unsupervised feature selection and general pattern discovery using Self-Organizing Maps for gaining insights into the nature of seismic wavefields. Comput. Geosci. 35 (9), 1757–1767.

Kohonen, T., 1988. Self-organization and associative memory. Self-Organization and Associative Memory, 100 figs. XV, 312 pages.. Springer-Verlag Berlin Heidelberg New York. Also Springer Series in Information Sciences, volume 8 1.

Kohonen, T., 2001. Self-organizing maps. Vol. 30. Springer Science & Business Media.

Koller, D., Friedman, N., 2009. Probabilistic graphical models: principles and techniques. MIT press.

Kompella, V. R., Luciw, M., Schmidhuber, J., 2011. Incremental slow feature analysis. In: Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Two. IJCAI'11. AAAI Press, pp. 1354–1359.

Konstantinou, K. I., Schlindwein, V., 2003. Nature, wavefield properties and source mechanism of volcanic tremor: a review. Journal of Volcanology and Geothermal Research 119 (1), 161–187.

Kotsiantis, S., 2007. Supervised Machine Learning: A Review of Classification Techniques. Informatica 31, 249–268.

Kubichek, R. F., Quincy, E., 1985. Statistical modeling and feature selection for seismic pattern recognition. Pattern Recognition 18 (6), 441–448.

Kuhnl, T., Kummert, F., Fritsch, J., 2011. Monocular road segmentation using slow feature analysis. In: Intelligent Vehicles Symposium (IV), 2011 IEEE. pp. 800–806.

Kuncheva, L. I., 1996. On the equivalence between fuzzy and statistical classifiers. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 4 (03), 245–253.

Kvaerna, T., Ringdal, F., 1992. Integrated array and three-component processing using a seismic microarray. Bulletin of the Seismological Society of America 82 (2), 870–882.

Ladd, M. D., Alam, M. K., Sleefe, G. E., Nguyen, H. D., 2000. Seismic and acoustic signal identification algorithms. In: AeroSense 2000. International Society for Optics and Photonics, pp. 106–120.

Lahr, J., Chouet, B., Stephens, C., Power, J., Page, R., 1994. Earthquake classification, location, and error analysis in a volcanic environment: Implications for the magmatic system of the 1989–1990 eruptions at Redoubt Volcano, Alaska. Journal of Volcanology and Geothermal Research 62 (1), 137–151.

Landis, J. R., Koch, G. G., 1977. The Measurement of Observer Agreement for Categorical Data. Biometrics 33 (1), pp. 159–174.

Langer, H., Falsaperla, S., Masotti, M., Campanini, R., Spampinato, S., Messina, A., 2009. Synopsis of supervised and unsupervised pattern classification techniques applied to volcanic tremor data at Mt Etna, Italy. Geophysical Journal International 178 (2), 1132–1144.

Langer, H., Falsaperla, S., Messina, A., Spampinato, S., Behncke, B., 2011. Detecting imminent eruptive activity at Mt Etna, Italy, in 2007–2008 through pattern classification of volcanic tremor data. Journal of Volcanology and Geothermal Research 200 (1), 1–17.

Langer, H., Falsaperla, S., Powell, T., Thompson, G., 2006. Automatic classification and a-posteriori analysis of seismic event identification at Soufrière Hills volcano, Montserrat. Journal of Volcanology and Geothermal Research 153 (1–2), 1–10, mULTIMO: Multi-Parameter Monitoring, Modelling and Forecasting of Volcanic Hazard. Results from an European Project.

Lasserre, J., Bishop, C. M., 2007. Generative or Discriminative? Getting the Best of Both Worlds. BAYESIAN STATISTICS 8, 3–24.

Lee, H., Choi, S., 2003. Pca+ hmm+ svm for eeg pattern classification. In: Signal Processing and Its Applications, 2003. Proceedings. Seventh International Symposium on. Vol. 1. IEEE, pp. 541–544.

Leet, R. C., 1988. Saturated and subcooled hydrothermal boiling in groundwater flow channels as a source of harmonic tremor. Journal of Geophysical Research: Solid Earth (1978–2012) 93 (B5), 4835–4849.

Leiva, J. M., 2007. Extracción de características mediante criterios basados en teoría de la información. Ph.D. thesis, Universidad Carlos III de Madrid, España.

Leprettre, B., Martin, N., Glangeaud, F., Navarre, J.-P., 1998. Three-component signal recognition using time, time-frequency, and polarization information-application to seismic detection of avalanches. Signal Processing, IEEE Transactions on 46 (1), 83–102.

Lesage, P., 2009. Interactive Matlab software for the analysis of seismic volcanic signals. Computers & Geosciences 35 (10), 2137–2144.

Levinson, S. E., Rabiner, L. R., Sondhi, M. M., 1983. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. Bell System Technical Journal, The 62 (4), 1035–1074.

Li, H., Jiang, T., Zhang, K., 2006. Efficient and robust feature extraction by maximum margin criterion. Neural Networks, IEEE Transactions on 17 (1), 157–165.

Liu, H.-H., Fu, K.-S., 1983. An application of syntactic pattern recognition to seismic discrimination. Geoscience and Remote Sensing, IEEE Transactions on 21 (2), 125–132.

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., Arnaldi, B., 2007. A review of classification algorithms for EEG-based brain–computer interfaces. Journal of Neural Engineering 4.

Lutu, P. E., Engelbrecht, A. P., 2010. A decision rule-based method for feature selection in predictive data mining. Expert Systems with Applications 37 (1), 602–609.

Makhoul, J., 1975. Linear prediction: A tutorial review. Proceedings of the IEEE 63 (4), 561–580.

Masiello, S., Esposito, A., Scarpetta, S., Giudicepietro, F., Esposito, A., Marinaro, M., 2006. Application of self organized maps and curvilinear component analysis to the discrimination of the Vesuvius seismic signals. In: 5th Workshop On Self-Organized Maps.

Masotti, M., Campanini, R., Mazzacurati, L., Falsaperla, S., Langer, H., Spampinato, S., 2008. TREMOrEC: a software utility for automatic classification of volcanic tremor. Geochemistry, Geophysics, Geosystems 9 (4).

Masotti, M., Falsaperla, S., Langer, H., Spampinato, S., Campanini, R., 2006. Application of Support Vector Machine to the classification of volcanic tremor at Etna, Italy. Geophysical Research Letters 33.

Maurice, R., Stacey, D., Wiens, D. A., Shore, P. J., Vera, E., Dorman, L. M., 2003. Seismicity and tectonics of the South Shetland Islands and Bransfield Strait from a regional broadband seismograph deployment. Journal of Geophysical Research: Solid Earth (1978–2012) 108 (B10).

McCulloch, W. S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics 5 (4), 115–133.

McNutt, S. R., 1986. Observations and analysis of B-type earthquakes, explosions, and volcanic tremor at Pavlof Volcano, Alaska. Bulletin of the Seismological Society of America 76 (1), 153–175.

McNutt, S. R., 1996. Seismic monitoring and eruption forecasting of volcanoes: a review of the state-of-the-art and case histories. In: Monitoring and mitigation of volcano hazards. Springer, pp. 99–146.

McNutt, S. R., 2005. Volcanic seismology. Annu. Rev. Earth planet. Sci. 32, 461–491.

Mendoza-Rosas, A. T., De la Cruz-Reyna, S., 2008. A statistical method linking geological and historical eruption time series for volcanic hazard estimations: applications to active polygenetic volcanoes. Journal of Volcanology and Geothermal Research 176 (2), 277–290.

Merialdo, B., Apr 1988. Phonetic recognition using hidden Markov models and maximum mutual information training. In: Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on. pp. 111–114 vol.1.

Messina, A., Langer, H., 2011. Pattern recognition of volcanic tremor data on Mt. Etna (Italy) with KKAnalysis—A software program for unsupervised classification. Computers & Geosciences 37 (7), 953–961.

Miller, A., Stewart, R., White, R., Luckett, R., Baptie, B., Aspinall, W., Latchman, J., Lynch, L., Voight, B., 1998. Seismicity associated with dome growth and collapse at the Soufriere Hills Volcano, Montserrat. Geophysical Research Letters 25 (18), 3401–3404.

Mitchell, T. M., 1997. Machine learning. Burr Ridge, IL: McGraw Hill.

Mohri, M., Pereira, F., Riley, M., 2002. Weighted finite-state transducers in speech recognition. Computer Speech & Language 16 (1), 69–88.

Mora, M. M., Lesage, P., Dorel, J., Bard, P.-Y., Métaxian, J.-P., Alvarado, G. E., Leandro, C., 2001. Study of seismic site effects using H/V spectral ratios at Arenal Volcano, Costa Rica. Geophysical research letters 28 (15), 2991–2994.

Murphy, K. P., 2002. Dynamic bayesian networks: representation, inference and learning. Ph.D. thesis, University of California, Berkeley.

Murphy, M., Cercone, J., Mar 1993. Neural network techniques applied to seismic event classification. In: System Theory, 1993. Proceedings SSST '93., Twenty-Fifth Southeastern Symposium on. pp. 343–347.

Murthy, S. K., 1998. Automatic construction of decision trees from data: A multi-disciplinary survey. Data mining and knowledge discovery 2 (4), 345–389.

Narendra, P. M., Fukunaga, K., Sept 1977. A Branch and Bound Algorithm for Feature Subset Selection. Computers, IEEE Transactions on C-26 (9), 917–922.

Neuberg, J., O'Gorman, C., 2002. A model of the seismic wavefield in gas-charged magma: application to Soufriere Hills Volcano, Montserrat. Geological Society, London, Memoirs 21 (1), 603–609.

Ng, A. Y., Jordan, M. I., 2001. On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. In: NIPS. pp. 841–848.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A. Y., 2011. Multimodal deep learning. In: Proceedings of the 28th international conference on machine learning (ICML-11). pp. 689–696.

Noether, G. E., 1981. Why kendall tau. Teaching Statistics 3 (2), 41–43.

Normandin, Y., 1996. Maximum Mutual Information Estimation of Hidden Markov Models. In: Lee, C.-H., Soong, F., Paliwal, K. (Eds.), Automatic Speech and Speaker Recognition. Vol. 355 of The Kluwer International Series in Engineering and Computer Science. Springer US, pp. 57–81.

Ohrnberger, M., 2001. Continuous automatic classification of seismic signals of volcanic origin at Mt. Merapi, Java, Indonesia. Ph.D. thesis, Universität Potsdam, Germany.

Oliver, N., Garg, A., 2002. MMIHMM: Maximum Mutual Information Hidden Markov Models.

Oliveros, A., Carniel, R., Tárraga, M., Aspinall, W., 2008. On the application of hidden markov model and bayesian belief network to seismic noise at Las Cañadas caldera, Tenerife, Spain. Chaos, Solitons & Fractals 37 (3), 849–857.

Omori, F., 1911. The Usu-san eruption and earthquake and elevation phenomena. Bull. Imp. Earthq. Inv. Com. 5, 1–38.

Orlic, N., Loncaric, S., 2010. Earthquake—explosion discrimination using genetic algorithm-based boosting approach. Computers |& Geosciences 36 (2), 179–185.

Orozco-Alzate, M., Acosta-Muñoz, C., Londoño Bonilla, J. M., 2012. The Automated Identification of Volcanic Earthquakes: Concepts, Applications and Challenges. Earthquake Research and Analysis-Seismology, Seismotectonic and Earthquake Geology, 345–370.

Oura, K., Zen, H., Nankaku, Y., Lee, A., Tokuda, K., 2006. Hidden semi-Markov model based speech recognition system using weighted finite-state transducer. In: Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on. Vol. 1. IEEE, pp. I–I.

Pal, S. K., Mitra, S., 1999. Neuro-fuzzy pattern recognition: methods in soft computing. John Wiley & Sons, Inc.

Parpoula, C., Drosou, K., Koukouvinos, C., 2013. Large-Scale Statistical Modelling via Machine Learning Classifiers. Journal of Statistics Applications & Probability 2 (3), 203–222.

Patterson, E. K., Gurbuz, S., Tufekci, Z., Gowdy, J., 2002. CUAVE: A new audio-visual database for multimodal human-computer interface research. In: Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on. Vol. 2. IEEE, pp. II–2017.

Pearson, K., 1901. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2 (11), 559–572.

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Analysis and Machine Intelligence, IEEE Transactions on 27 (8), 1226–1238.

Pennebaker, W. B., 1992. JPEG: Still image data compression standard. Springer.

Pernkopf, F., Bilmes, J., 2005. Discriminative versus generative parameter and structure learning of Bayesian network classifiers. In: Proceedings of the 22nd international conference on Machine learning. ACM, pp. 657–664.

Pinnegar, C., 2006. Polarization analysis and polarization filtering of three-component signals with the time–frequency S transform. Geophysical Journal International 165 (2), 596–606.

Potamianos, G., Neti, C., Luettin, J., Matthews, I., 2004. Audio-visual automatic speech recognition: An overview. Issues in visual and audio-visual speech processing 22, 23.

Prudencio, J., De Siena, L., Ibáñez, J., Del Pezzo, E., García-Yeguas, A., Díaz-Moreno, A., 2015. The 3D Attenuation Structure of Deception Island (Antarctica). Surveys in Geophysics 36 (3), 371–390.

Rabiner, L., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77 (2), 257–286.

Rabiner, L., Juang, B.-H., 1986. An introduction to hidden Markov models. ASSP Magazine, IEEE 3 (1), 4–16.

Rabiner, L. R., Juang, B.-H., 1993. Fundamentals of speech recognition. Prentice Hall signal processing series. Prentice Hall.

Rabiner, L. R., Schafer, R. W., 2007. Introduction to digital speech processing. Foundations and trends in signal processing 1 (1), 1–194.

Raina, R., Shen, Y., Mccallum, A., Ng, A. Y., 2003. Classification with hybrid generative/discriminative models. In: Advances in neural information processing systems.

Rey, J., Somoza, L., Martínez-Frías, J., 1995. Tectonic, volcanic, and hydrothermal event sequence on Deception Island (Antarctica). Geo-Marine Letters 15 (1), 1–8.

Reznik, Y. A., Hinds, A. T., Zhang, C., Yu, L., Ni, Z., 2007. Efficient fixed-point approximations of the 8x8 inverse discrete cosine transform. In: Proc. SPIE. Vol. 6696. p. 669617.

Riggelsen, C., Ohrnberger, M., Scherbaum, F., 2007. Dynamic bayesian networks for real-time classification of seismic signals. In: Knowledge Discovery in Databases: PKDD 2007. Springer, pp. 565–572.

Ripepe, M., Ciliberto, S., Della Schiava, M., 2001. Time constraints for modeling source dynamics of volcanic explosions at Stromboli. Journal of Geophysical Research: Solid Earth (1978–2012) 106 (B5), 8713–8727.

Rogers, J., Stephens, C., 1991. SSAM: a PC-based seismic spectral amplitude measurement system for volcano monitoring. Seism. Res. Lett 62, 22.

Rosenblatt, F., 1962. Principles of neurodynamics.

Rowe, C., Aster, R., Kyle, P., Schlue, J., Dibble, R., 1998. Broadband recording of Strombolian explosions and associated very-long-period seismic signals on Mount Erebus Volcano, Ross Island, Antarctica. Geophysical research letters 25 (13), 2297–2300.

Roweis, S. T., Saul, L. K., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290 (5500), 2323–2326.

Saeys, Y., Inza, I. n., Larrañaga, P., 2007. A review of feature selection techniques in bioinformatics. bioinformatics 23 (19), 2507–2517.

San-Martin, C., Melgarejo, C., Gallegos, C., Soto, G., Curilem, M., Fuentealba, G., 2010. Feature extraction using circular statistics applied to volcano monitoring. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Springer, pp. 458–466.

Sassa, K., 1936. Micro-seismometric study on eruptions of the volcano Aso. Mem. Coll. Sci., Kyoto Imp. Univ., Ser. A 19, 11–56.

Saucedo, R., Macıas, J., Bursik, M., Mora, J., Gavilanes, J., Cortes, A., 2002. Emplacement of pyroclastic flows during the 1998–1999 eruption of Volcan de Colima, Mexico. Journal of Volcanology and Geothermal Research 117 (1), 129–153.

Scarpetta, S., Giudicepietro, F., Ezin, E., Petrosino, S., Del Pezzo, E., Martini, M., Marinaro, M., 2005. Automatic classification of seismic signals at Mt. Vesuvius volcano, Italy, using neural networks. Bulletin of the Seismological Society of America 95 (1), 185–196.

Scholz, M., Fraunholz, M., Selbig, J., 2008. Nonlinear principal component analysis: neural network models and applications. In: Principal manifolds for data visualization and dimension reduction. Springer, pp. 44–67.

Schuldt, C., Laptev, I., Caputo, B., Aug 2004. Recognizing human actions: a local SVM approach. In: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. Vol. 3. pp. 32–36 Vol.3.

Seidl, D., Schick, R., Riuscetti, M., 1981. Volcanic tremors at Etna: a model for hydraulic origin. Bulletin volcanologique 44 (1), 43–56.

Sha, F., Saul, L. K., 2006. Large margin hidden Markov models for automatic speech recognition. In: Advances in neural information processing systems. pp. 1249–1256.

Shalizi, C., Sep. 2009. The Truth about Principal Components and Factor Analysis. Tech. rep., CMU.

Silverman, B. W., Jones, M. C., 1989. E. Fix and JL Hodges (1951): An important contribution to nonparametric discriminant analysis and density estimation: Commentary on Fix and Hodges (1951). International Statistical Review/Revue Internationale de Statistique, 233–238.

Sprekeler, H., Wiskott, D. L., August 2008. Understanding Slow Feature Analysis: A Mathematical Framework.

Stevens, S. S., Volkmann, J., Newman, E. B., 1937. A Scale for the Measurement of the Psychological Magnitude Pitch. The Journal of the Acoustical Society of America 8 (3), 185–190.

Stich, D., Almendros, J., Jiménez, V., Mancilla, F., Carmona, E., 2011. Ocean noise triggering of rhythmic long period events at Deception Island volcano. Geophysical Research Letters 38 (22).

Sturton, S., Neuberg, J., 2003. The effects of a decompression on seismic parameter profiles in a gas-charged magma. Journal of volcanology and geothermal research 128 (1), 187–199.

Sutskever, I., Martens, J., Dahl, G., Hinton, G., 2013. On the importance of initialization and momentum in deep learning. In: Proceedings of the 30th international conference on machine learning (ICML-13). pp. 1139–1147.

Sutton, R. S., Barto, A. G., 1998. Reinforcement learning: An introduction. MIT press.

Takahashi, T., Satofuka, Y., 2002. Generalized theory of stony and turbulent muddy debris-flow and its practical model. Journal of Japan Society of Erosion Control Engineering 55 (3), 33–42.

Takahashi, T., Tsujimoto, H., 2000. A mechanical model for Merapi-type pyroclastic flow. Journal of Volcanology and Geothermal Research 98 (1), 91–115.

Tarbuck, E. J., Lutgens, F. K., Tasa, D., 2012. Earth science. Prentice Hall.

Tekalp, A. M., Tekalp, A. M., 1995. Digital video processing. Vol. 1. Prentice Hall PTR Upper Saddle river, NJ.

Tenenbaum, J. B., De Silva, V., Langford, J. C., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290 (5500), 2319–2323.

Theodoridis, S., Koutroumbas, K., 2009. Pattern recognition. Academic Press.

Tobias Blaschke, T. Z. L. W., 2007. Independent Slow Feature Analysis and Nonlinear Blind Source Separation.

Tuncer, Y., Tanik, M. M., Allison, D. B., 2008. An overview of statistical decomposition techniques applied to complex systems. Computational Statistics & Data Analysis 52 (5), 2292–2310.

Van Hulse, J., Khoshgoftaar, T., Napolitano, A., Wald, R., 2009. Feature Selection with High-Dimensional Imbalanced Data. In: Data Mining Workshops, 2009. ICDMW '09. IEEE International Conference on. pp. 507–514.

Vapnik, V., 2000. The nature of statistical learning theory. springer.

Varley, N., Arámbula-Mendoza, R., Reyes-Dávila, G., Sanderson, R., Stevenson, J., 2010a. Generation of Vulcanian activity and long-period seismicity at Volcán de Colima, Mexico. Journal of Volcanology and Geothermal Research 198 (1), 45–56.

Varley, N. R., Arámbula-Mendoza, R., Reyes-Dávila, G., Stevenson, J., Harwood, R., 2010b. Long-period seismicity during magma movement at Volcán de Colima. Bulletin of volcanology 72 (9), 1093–1107.

Veksler, O., 2004. Pattern Recognition. University course.

Vergniolle, S., Jaupart, C., 1990. Dynamics of degassing at Kilauea volcano, Hawaii. Journal of Geophysical Research: Solid Earth (1978–2012) 95 (B3), 2793–2809.

Viterbi, A. J., 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. Information Theory, IEEE Transactions on 13 (2), 260–269.

Vogel, A., Zhu, Y., Brandes, K., 1992. Earthquake prognostics-Expert Systems and application of Artificial Intelligence techniques as part of logistical approach to earthquake disaster mitigation. In: Proceedings of the workshop: Application of artificial intelligence techniques in seismology and engineering seismology. Vol. 23. pp. 43–52.

Voight, B., 1989. A relation to describe rate-dependent material failure. Science 243 (4888), 200–203.

Wagner, G. S., Owens, T. J., 1996. Signal detection using multi-channel seismic data. Bulletin of the Seismological Society of America 86 (1A), 221–231.

Wassermann, J., 2012. Volcano Seismology, IASPEI New manual of seismological observatory practice 2 (NMSOP-2), 2nd Edition. Potsdam : Deutsches GeoForschungsZentrum GFZ, Potsdam, Ch. 13, pp. 1–77, dOI:10.2312/GFZ.NMSOP-2_ch13.

Wassermann, J., Beyreuther, M., Ohrnberger, M., Carniel, R., 2007. The usability of hidden markov modelling in seismic detection and automatic warning level estimation of volcanic activity. Seismol. Res. Lett. 78, 249.

Wiskott, L., 1999. Learning invariance manifolds. Neurocomputing 26, 925–932.

Wiskott, L., Berkes, P., Franzius, M., Sprekeler, H., Wilbert, N., 2011. Slow feature analysis. Scholarpedia 6 (4), 5282.

Wiskott, L., Sejnowski, T. J., 2002. Slow feature analysis: Unsupervised learning of invariances. Neural computation 14 (4), 715–770.

Withers, M., Aster, R., Young, C., Beiriger, J., Harris, M., Moore, S., Trujillo, J., 1998. A comparison of select trigger algorithms for automated global seismic phase and event detection. Bulletin of the Seismological Society of America 88 (1), 95–106.

Xue, J.-H., Titterington, D., 2008. Comment on "On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes". Neural Processing Letters 28 (3), 169–187.

Young, S., 1994. The HTK Hidden Markov Model Toolkit: Design and Philosophy. Tech. Rep. TR152, Cambridge University Engineering Dept.

Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P. C., 2006. The HTK Book, version 3.4. Cambridge University Engineering Department, Cambridge, UK.

Yu, L., Ye, J., Liu, H., 2007. Dimensionality Reduction for Data Mining - Techniques, Applications and Trends. In: SIAM International Conference on Data Mining Proceedings. SIAM.

Yueqing, Z., Vogel, A., Ping, H., 1996. Expert System for Earthquake Hazard Assessment: ESEHA. In: Schenk, V. (Ed.), Earthquake Hazard and Risk. Vol. 6 of Advances in Natural and Technological Hazards Research. Springer Netherlands, pp. 199–209.

Zamalloa, M., Rodrigues-Fuentes, L., Penagarikano, M., Bordel, G., Uribe, J., 2008. Feature dimensionality reduction through genetic algorithms for faster speaker recognition. In: 16th European Signal Processing Conference.

Zhang, Z., Tao, D., 2012. Slow feature analysis for human action recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34 (3), 436–450.

Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., Liu, H., 2010. Advancing feature selection research. ASU Feature Selection Repository.

Zito, T., Wilbert, N., Wiskott, L., Berkes, P., 2008. Modular toolkit for Data Processing (MDP): a Python data processing framework. Front. Neuroinform. 2.

Zobin, V. M., Arámbula, R., Bretón, M., Reyes, G., Plascencia, I., Navarro, C., Téllez, A., Campos, A., González, M., León, Z., et al., 2015. Dynamics of the January 2013–June 2014 explosive-effusive episode in the eruption of Volcán de Colima, México: insights from seismic and video monitoring. Bulletin of Volcanology 77 (4), 1–13.

Zobin, V. M., Navarro-Ochoa, C. J., Reyes-Dávila, G. A., 2006. Seismic quantification of the explosions that destroyed the dome of Volcán de Colima, Mexico, in July–August 2003. Bulletin of volcanology 69 (2), 141–147.

Zobin, V. M., Orozco-Rojas, J., Reyes-Dávila, G. A., Navarro, C., 2005. Seismicity of an andesitic volcano during block-lava effusion: Volcan de Colima, Mexico, November 1998–January 1999. Bulletin of volcanology 67 (7), 679–688.

Índice de Tablas

Índice de figuras

Algoritmos

Parte V.

APÉNDICE

A. Cuestiones prácticas

A.1. ¿Cuál es la duración del segmento óptima para cada característica geo-estadística?

Vamos a analizar el efecto que tiene la duración de los segmentos en los que troceamos los eventos sobre la tasa de reconocimiento que obtenemos independientemente para cada característica del vector mixto *geoSTATS.D.42* propuesto en la Subsubsección 4.2.4.2. El significado de cada componente y la propiedad de los datos que pretende describir se explican en la Sección 4.2. Para el test usaremos las señales de las bases de datos maestras *dec.95Ms* y *col.04Ms* definidas en la Sección 3.6. Seguiremos la metodología experimental detallada en la Subsección 4.1.6.

La Figura A.1.1 representa para diferentes tamaños de ventana la tasa de reconocimiento cuando los datos se describen con solo una característica. A priori, analizándola no parece existir ningún patrón predeterminado como en el caso de las parametrizaciones basadas en transformadas (Figura 4.2.5), lo que sugiere que la elección de un tamaño concreto está más relacionada con el tipo de evento que con la característica en sí. Este hecho es un precursor del estudio de sistemas en paralelo y es analizado en profundidad en el Capítulo 6.

Llama la atención la variabilidad de hasta 20 puntos en la tasa de reconocimiento para algunas componentes como *A.skew_D* o *htk.E_D*. El estudio de distintas duraciones de ventana para cada componente queda fuera del alcance de esta tesis, si

	dec.95Ms			*col.04Ms*		
duración [s]	GMM.4g	HMM.5e.4g	*media*	GMM.4g	HMM.7e.4g	*media*
0.2	42.60	44.73	*43.67*	32.94	41.03	*36.99*
0.5	39.87	46.76	*43.32*	33.11	44.58	*38.85*
1	39.83	47.38	*43.61*	32.29	46.02	*39.16*
2	40.91	48.38	*44.65*			
5	41.81	49.33	*45.57*	34.69	52.69	*43.69*
10				35.26	53.96	*44.61*
media	41.00	47.32	*44.16*	33.66	47.65	*40.66*

Tabla A.1.1.: *Configuración de la duración de segmento para el vector geo-estadístico geoSTATS.D.42.* Promedio del *%cCorr* obtenido para cada componente.

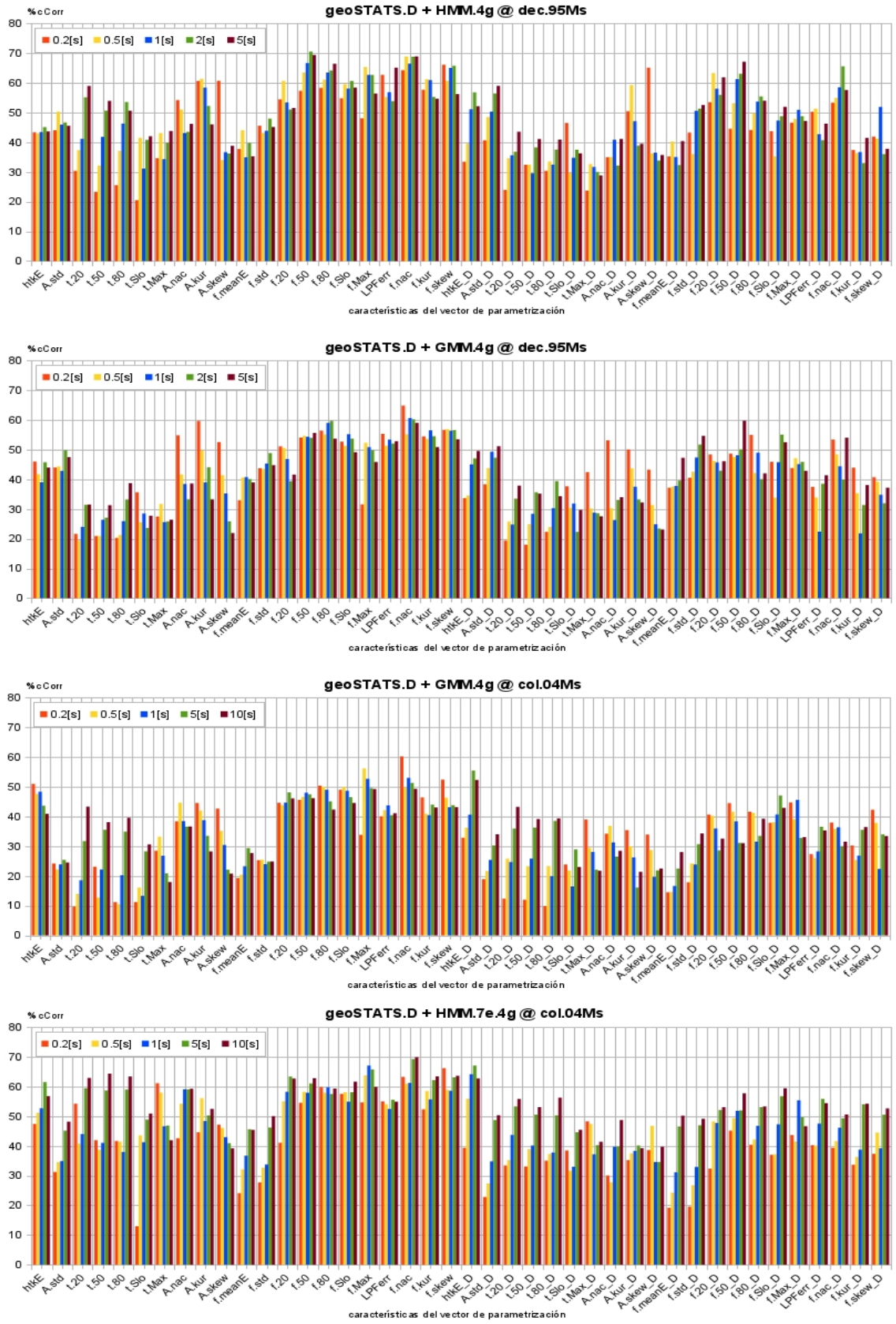


Figura A.1.1.: Eficiencia de reconocimiento de cada característica geoestadística de *geoSTATS.D.42* para distintos tamaños de ventana.

bien otros autores (Álvarez et al., 2009, 2011) han utilizado tamaños distintos en sus trabajos. Scarpetta et al. (2005) cambian la duración del segmento acorde con la característica a extraer y la estación donde se registran los datos. Promediando entre las componentes del vector en la Tabla A.1.1, comprobamos que, en general, hay una ligera mejora al usar ventanas grandes (de más de 1 segundo) en *dec.95Ms.* Dicha mejora es considerablemente mayor en *col.04Ms.*

A.2. ¿Influye la *direccionalidad* en la selección secuencial de características?

Por *direccionalidad* entendemos la acción que respecto al conjunto de características puede tomar el algoritmo de selección en cada bucle de reconocimiento:

- *direccionalidad creciente* o *Sequential Forward Selection - SFS* (Jain et al., 2000): añade la característica más discriminativa según una medida dada al conjunto de características ya seleccionadas
- *direccionalidad decreciente* (*Sequential Backward Selection - SBS*): elimina la componente menos discriminativa.

Típicamente, un algoritmo SC_RI creciente empezará seleccionando la característica más discriminativa de entre todas las posibles en el 1er bucle de análisis y terminará escogiendo la mejor de entre las 2 restantes en el último bucle para añadirla al conjunto de características ya seleccionadas. Si la direccionalidad es decreciente (tal y como fue diseñado el algoritmo DFS original por De la Torre et al., 1997) empezará por eliminar la menos discriminativa de entre el conjunto total de características en el 1er bucle y acabará seleccionando entre las 2 restantes en el último bucle. Aunque la intuición nos lleva a pensar que el algoritmo debería llegar al mismo resultado independientemente de su direccionalidad, un análisis esquemático nos demuestra que esto no es así...

Sea C un conjunto de características con F elementos, tal que $C=\{C_1,...,C_F\}$. Un algoritmo SC_RI evaluado en una partición de datos BD necesita F-1 bucles de reconocimiento para ordenar C en una secuencia S usando como medida discriminante D (cAcc %). Sea D_A la medida discriminante obtenida cuando los datos son descritos por A, subconjunto de análisis de C, y sean S_i y $S \backslash C_i$ los conjuntos S y S\C en el bucle$_i$. Nótese que en cada bucle se cumple que $S \backslash C_i \bigcup S_i = C$.

Tal como se muestra en el Algoritmo A.1, el análisis creciente obtiene la secuencia $S_{crec}=\{C_3,C_1,C_4,C_2\}$ de las características más discriminantes en orden decreciente, mientras que usando una direccionalidad decreciente hallamos $S_{dec}=\{C_1,C_2,C_4,C_3\}$ como la secuencia de las menos discriminativas de forma decreciente. Un algoritmo independiente de la direccionalidad implica que $S_{crec}(i)=S_{dec}(F+1-i)$ para cualquier i en $\{1..F\}$, condición que no se cumple en este ejemplo. Como puede observarse,

Algoritmo A.1 Comparación entre direccionalidad creciente y decreciente de algoritmos *SC_RI*

En este ejemplo C={htkE, kur@t, f30, fSlo}, D=cAcc %) y la partición de datos usada es BD=dec.95M_1_2.

- **Direccionalidad creciente**: el algoritmo comienza con S={} para ir añadiendo en el bucle$_i$ la característica más discriminativa de S\C$_{i-1}$, esto es, la que obtiene el mayor D$_A$ cuando es añadida al conjunto S$_{i-1}$de características previamente seleccionadas en el bucle anterior:

	inicio=bucle$_0$	**bucle$_1$**	**bucle$_2$**	**bucle$_3$**	**bucle$_4$=fin**
C$_1$=htkE		D$_1$=34.90	*D$_{13}$=61.63*		
C$_2$=kur@t		D$_2$=31.51	D$_{23}$=52.48	D$_{123}$=63.92	
C$_3$=f30		*D$_3$=60.43*			
C$_4$=fSlo		D$_4$=32.08	D$_{34}$=61.32	*D$_{134}$=67.10*	
S	{}	{C$_3$}	{C$_1$,C$_3$}	{C$_1$,C$_3$,C$_4$}	{C$_3$,C$_1$,C$_4$,C$_2$}
S\C	{C$_1$,C$_2$,C$_3$,C$_4$}	{C$_1$,C$_2$,C$_4$}	{C$_2$,C$_4$}	{C$_2$}	{}

- **Direccionalidad decreciente:** se va eliminando en el bucle$_i$ la característica C$_j$ menos discriminativa de S\C$_{i-1}$, es decir, aquella que consigue el mayor D$_A$ cuando C$_j$ no forma parte de A:

	inicio=bucle$_0$	**bucle$_1$**	**bucle$_2$**	**bucle$_3$**	**bucle$_4$=fin**
C$_1$=htkE		*D$_{234}$=70.95*			
C$_2$=kur@t		D$_{134}$=67.10	*D$_{34}$=61.32*		
C$_3$=f30		D$_{124}$=61.32	D$_{24}$=51.24	D$_4$=32.08	
C$_4$=fSlo		D$_{123}$=63.92	D$_{23}$=52.48	*D$_3$=60.43*	
S	{}	{C$_1$}	{C$_1$,C$_2$}	{C$_1$,C$_2$,C$_4$}	{C$_1$,C$_2$,C$_4$,C$_3$}
S\C	{C$_1$,C$_2$,C$_3$,C$_4$}	{C$_2$,C$_3$,C$_4$}	{C$_3$,C$_4$}	{C$_3$}	{}

por el diseño de los algoritmos SC_RI, se analizan distintos subconjuntos de características A en los bucles *simétricos* (bucles *i* en direccionalidad creciente $\leftrightarrow$ *F+1-i* en decreciente) por lo que las características seleccionadas tras terminar el bucle de análisis no tienen por qué coincidir, aunque sí el valor de D$_A$para los mismos subconjuntos A.

Demostrada la *no-direccionalidad* de los algoritmos SC_RI, cabe preguntarse otras 2 cuestiones adicionales:

i) ¿Bajo que condiciones se alcanzaría la simetría en la direccionalidad? Experimentalmente apreciamos que la no simetría parece estar relacionada con la dependencia estadística existente entre las características. En las pruebas con algoritmos SC mixtos se puede comprobar la tendencia a la simetría cuando el espacio de características se decorrela en subespacios independientes (mediante PCA o ICA)

previamente a la selección mediante SC_RI. También se aprecia la tendencia a la no simetría con conjuntos de características (linealmente) dependientes entre sí como {f30, fSlo=f70-f30, tSlo, t70,...}. En este sentido, es interesante estudiar el comportamiento de un algoritmo SC_RI ante un conjunto que contenga características repetidas {f30, fSlo, kur@t, fSlo, f90}.

ii) ¿Cuál es mejor elección, una direccionalidad creciente, decreciente o alguna otra alternativa? Existe algunos motivos que inclinan la balanza hacia una dirección creciente:

1. *Coste computacional*. Los SC_RI crecientes requieren menos tiempo de ejecución pues los subconjuntos de análisis a examinar en los primeros bucles contienen menos elementos, lo que implica modelos más sencillos y evaluación más rápida. Los modelos más complicados (con más características) se evalúan en los últimos bucles, cuando hay que obtener menos valores de D al ser el conjunto S\C más pequeño.

2. *Fiabilidad*. Teóricamente es más sencillo (Drugman et al., 2007) seleccionar las características más discriminantes en un conjunto pequeño que en uno mayor debido a la mayor probabilidad de que exista correlación entre características dentro un subconjunto A cuantos más elementos contenga y, a que las características no discriminantes enmascaren a las discriminantes, degradando la medida D (De la Torre et al., 1997).

3. *Información obtenida*. La direccionalidad creciente nos proporciona tras finalizar el 1er bucle la medida discriminante D_f para cada característica por sí sola, información que puede ser muy útil para evaluar dicha característica independientemente de las demás. Asimismo, en el 2º bucle nos proporciona D_{fg} para todas (menos una) combinaciones posibles de pares de características, con lo que podemos evaluar la correlación que hay entre ellas.

No obstante, la discusión no está cerrada; algunos autores prefieren la direccionalidad decreciente argumentando que la evaluación de una característica hecha en conjunto con todas las demás ayuda a hallar una posible complementariedad entre ellas tal y como hemos mostrado en la Subsección 4.1.4. Aún así, existen muchas estrategias y algoritmos de *búsqueda* para incluir o eliminar una característica en un subconjunto seleccionado, agrupadas en estrategias de búsqueda completa, heurísticas y no-deterministas (Yu et al., 2007), algunas de ellas como la técnica *Branch-and-Bound* (Narendra and Fukunaga, 1977), aparte del estudio de todas las posibles combinaciones, aseguran hallar el conjunto óptimo (acorde con la medida D) a costa de aumentar el coste computacional (Jain et al., 2000).

A.3. ¿Cuál es la mejor configuración para el algoritmo *DFS-rsv* de selección de características?

El proceso de configuración del algoritmo *DFS-rsv* (Subsubsección 4.3.2.2) equivale a hallar los valores óptimos de los parámetros $(rsv, smooth, \alpha)$ definidas en la Subsubsección 4.3.2.2. Para ello realizamos distintas pruebas y escogeremos los valores que mejor resultados globales proporcionen en las bases de datos maestras *dec.95Ms* y *col.04Ms* (Sección 3.6). En este caso usaremos el vector mixto *geoLFCC.D.26* dado por Cortés et al. (2014).

La Figura A.3.1 muestra las gráficas de configuración para *dec.95Ms*. La Tabla A.3.1 promedia sus resultados respecto al número de características seleccionadas. Al igual que en el caso de la selección mediante filtros (Subsubsección 4.3.1.4) cuando se alcanza el tamaño medio del vector las curvas toman una forma similar, dentro del intervalo [85,90] *%cCorr*. A priori es difícil decantarse por una configuración con-

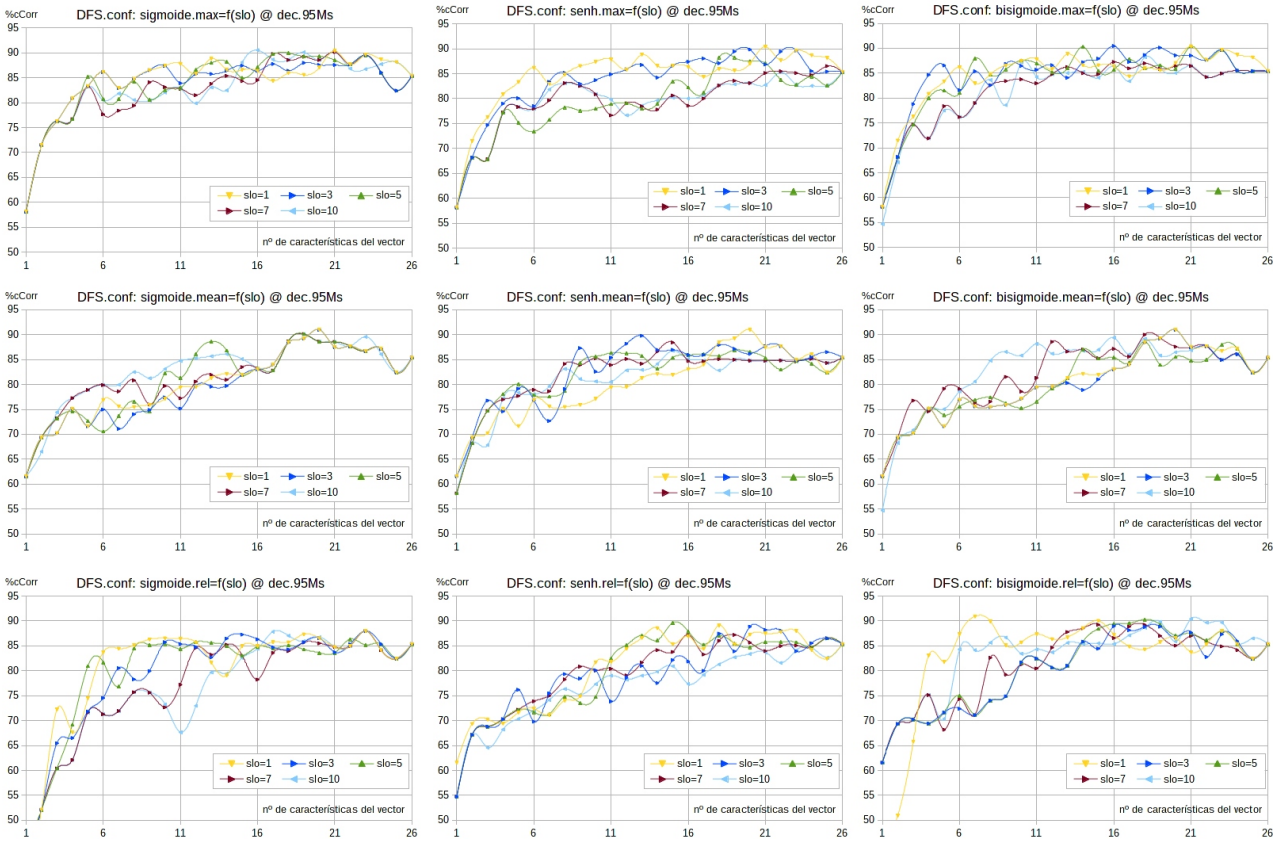


Figura A.3.1.: Configuración del algoritmo *DFS-rsv* para *dec.95Ms*. Las curvas muestran el $\%cCorr = \%cCorr(rsv, smooth, \alpha = slo)$ para la parametrización *geoLFCC.D.26*.

rsv=max						
smooth	$\alpha = 1$	$\alpha = 3$	$\alpha = 5$	$\alpha = 7$	$\alpha = 10$	*media*
sigmoide	84.34	83.92	83.55	82.62	83.03	*83.49*
senh	84.34	83.31	79.27	79.48	79.28	81.13
bisigmoide	84.34	*84.53*	83.81	81.70	81.33	83.14
media	*84.34*	83.92	82.21	81.27	81.22	82.59
rsv=mean						
smooth	$\alpha = 1$	$\alpha = 3$	$\alpha = 5$	$\alpha = 7$	$\alpha = 10$	*media*
sigmoide	80.20	79.56	80.95	81.17	82.49	80.87
senh	80.09	*82.60*	81.76	81.83	80.78	*81.41*
bisigmoide	80.20	79.90	80.03	82.13	82.50	80.95
media	80.16	80.68	80.91	81.71	*81.92*	81.08
rsv=rel						
smooth	$\alpha = 1$	$\alpha = 3$	$\alpha = 5$	$\alpha = 7$	$\alpha = 10$	*media*
sigmoide	80.54	79.35	79.77	76.81	76.22	78.54
senh	80.41	78.95	79.74	79.61	77.02	79.15
bisigmoide	82.39	80.12	80.66	81.14	*83.03*	*81.47*
media	*81.11*	79.47	80.06	79.19	78.76	79.72

Tabla A.3.1.: *Configuración del algoritmo DFS-rsv: valores promedio respecto el tamaño del vector geoLFCC.D.26 para la BD dec.95Ms.*

creta, aunque la forma de las gráficas indica que el binomio $(rsv = max, smooth = sigmoide)$ tiende a estabilizarse antes en un rango alto de $\%cCorr$.

Según la Tabla A.3.1, el uso de un suavizado frente a otro no varía demasiado los resultados dentro de una misma función *rsv* discriminante. Tampoco es posible afirmar categóricamente que grado de suavizado es el óptimo. La opción $(rsv = max, smooth = sigmoide)$ se confirma como el mejor esquema promediado, y $(rsv = max, smooth = bisigmoide, \alpha = 3)$ los valores óptimos de configuración para *dec.95Ms*.

Analizando la Figura A.3.2 de configuración para *col.04Ms* se observa que $(rsv = mean, smooth = sigmoide)$ es el esquema que antes alcanza valores altos (sobre [80,85] $\%cCorr$). La función de discriminación $(rsv = rel)$ es claramente inferior a las otras. De nuevo, el análisis de las gráficas no arroja luz sobre la mejor elección de la curva o el grado de suavizado, existiendo más diferencias entre ellas que las que encontrábamos en *dec.95Ms*.

La Tabla A.3.2 revela que la mejor configuración es $(rsv = max, smooth = sigmoide, \alpha = 10)$, coincidiendo además con el mejor binomio $(rsv = max, smooth = sigmoide)$ que consigue un 77.94 $\%cCorr$, similar al 77.80 $\%cCorr$ de la opción $(rsv = mean, smooth = sigmoide)$.

Figura A.3.2.: Configuración del algoritmo *DFS-rsv* para *col.04Ms* $\%cCorr = \%cCorr(rsv, smooth, \alpha = slo)$

Finalmente promediamos los resultados entre las dos bases de datos. En ambos casos la opción $(rsv = max, smooth = sigmoide)$ indica una mejor selección de las características, pero no hay un consenso en cuanto al nivel de suavizado. La Tabla A.3.3 sintetiza las curvas de configuración de la Figura A.3.1 y la Figura A.3.2. Los valores óptimos se obtienen para $(rsv = max, smooth = sigmoide, \alpha = 5)$.

Fiabilidad del proceso de configuración para el algoritmo DFS-rsv. Nótese, que a pesar de todos los experimentos realizados no existe la certeza de que los valores obtenidos en la configuración de $(rsv, smooth, \alpha)$ sean la mejor elección para cualquier base de datos y parametrización usada. De hecho, ni siquiera a pesar de las múltiples pruebas realizadas podemos garantizar que los valores hallados sean robustos. Sí que parece existir una tendencia hacia el la función discriminante $(rsv = max)$ y un suavizado dado por las funciones (bi)sigmoide con un grado bajo. Aún así, el nivel de incertidumbre es demasiado alto como para afirmar que esta tendencia sea general.

Llegados a este punto, solo restan dos opciones posibles:

rsv=max						
smooth	$\alpha = 1$	$\alpha = 3$	$\alpha = 5$	$\alpha = 7$	$\alpha = 10$	*media*
sigmoide	77.02	77.64	78.20	78.41	*78.42*	*77.94*
senh	76.23	76.26	67.74	66.95	66.63	70.76
bisigmoide	76.81	76.03	75.51	76.20	75.21	75.95
media	*76.69*	76.65	73.82	73.85	73.42	74.88
rsv=mean						
smooth	$\alpha = 1$	$\alpha = 3$	$\alpha = 5$	$\alpha = 7$	$\alpha = 10$	*media*
sigmoide	77.70	77.80	77.93	*78.16*	77.40	*77.80*
senh	77.75	77.31	73.41	75.67	75.64	75.95
bisigmoide	77.75	76.13	73.66	77.05	71.98	75.32
media	*77.73*	77.08	75.00	76.96	75.01	76.36
rsv=rel						
smooth	$\alpha = 1$	$\alpha = 3$	$\alpha = 5$	$\alpha = 7$	$\alpha = 10$	*media*
sigmoide	68.16	69.08	70.98	70.38	70.35	*69.79*
senh	69.49	69.55	69.00	69.41	*71.48*	*69.79*
bisigmoide	68.65	69.49	68.55	69.58	69.11	69.08
media	68.77	69.37	69.51	69.79	*70.31*	69.55

Tabla A.3.2.: *Configuración del algoritmo DFS-rsv: valores promedio respecto el tamaño del vector geoLFCC.D.26 para col.04Ms.*

1. *Se efectúa el proceso de configuración del algoritmo DFS-rsv para cada BD y parametrización a usar.* Lo que conlleva una costosa etapa de inicialización y configuración del sistema.
2. *Se usan valores promedio de suavizado.* Teniendo en cuenta que los resultados para $(rsv = max, rel)$ no son muy diferentes y que las curvas de suavizado sigmoide y bisigmoide tampoco, se puede optar por esquemas de configuración no muy extremos. De igual manera, parece razonable esperar que la diferencia en el conjunto de características finalmente seleccionadas sea también pequeña.

A.4. ¿Cuál es el tamaño mínimo de una BD para evitar el sobre- entrenamiento de los modelos?

Unos modelos sobre-entrenados sobre un corpus de datos pierden capacidad de generalización y, por tanto, de representar correctamente a otros datos, al ser construidos con un margen de variabilidad muy pequeño, adaptado a los datos utilizados en la fase de entrenamiento. Este hecho se traduce en una pérdida de eficacia de reconocimiento cuando los modelos se usan para clasificar datos con una estadística

rsv=max						
smooth	$\alpha = 1$	$\alpha = 3$	$\alpha = 5$	$\alpha = 7$	$\alpha = 10$	*media*
sigmoide	80.68	80.78	*80.87*	80.51	80.73	*80.71*
senh	80.28	79.79	73.51	73.21	72.95	75.95
bisigmoide	80.57	80.28	79.66	78.95	78.27	79.55
media	*80.51*	80.28	78.01	77.56	77.32	78.74
rsv=mean						
smooth	$\alpha = 1$	$\alpha = 3$	$\alpha = 5$	$\alpha = 7$	$\alpha = 10$	*media*
sigmoide	78.95	78.68	79.44	79.66	79.94	*79.34*
senh	78.92	*79.95*	77.58	78.75	78.21	78.68
bisigmoide	78.98	78.01	76.84	79.59	77.24	78.13
media	78.95	78.88	77.96	*79.34*	78.46	78.72
rsv=rel						
smooth	$\alpha = 1$	$\alpha = 3$	$\alpha = 5$	$\alpha = 7$	$\alpha = 10$	*media*
sigmoide	74.35	74.22	75.38	73.60	73.28	74.16
senh	74.95	74.25	74.37	74.51	74.25	74.47
bisigmoide	75.52	74.81	74.60	75.36	*76.07*	*75.27*
media	*74.94*	74.42	74.78	74.49	74.53	74.63

Tabla A.3.3.: *Configuración del algoritmo DFS-rsv: valores promedio respecto el vector geoLFCC.D.26 y las bases de datos maestras, dec.95Ms y col.04Ms.*

ligeramente diferente al corpus original de entrenamiento y puede evitarse usando una mayor cantidad de datos de entrenamiento para aumentar la variabilidad entre eventos de una misma clase.

El límite inferior de eventos de la misma clase necesario para modelarla depende de la *complejidad* de su modelo, esto es, del número de parámetros necesarios para definirlo y del tipo de modelo, en nuestro caso, los GMMs. Asumiendo un evento sismo-volcánico como una sucesión de segmentos (Cortés et al., 2014) y, teniendo en cuenta que la complejidad de un HMM con topología lineal, con E estados emisores y que usa gausianas para modelar el espacio de características es equiparable a la complejidad de una combinación lineal de GMMs con E elementos, podemos restringir el estudio únicamente a los GMMs.

El mínimo número de observaciones necesarias para definir un GMM viene dado por la mitad de los componentes de su matriz de covarianza: $0{,}5C^2$, siendo C el tamaño del vector de características. En el caso de usar matrices diagonales (como en esta tesis) se reduce de $0{,}5C^2$ a C (el tamaño de la diagonal). Si los GMMs usan G componentes gausianas y son diagonales necesitamos CG por cada GMM. Por tanto, identificando cada estado emisor de los HMMs como un GMM diagonal se necesitaría ECG observaciones mínimas para definir un HMM de EE estados emisores con G componentes gausianas diagonales cada uno.

BDs	clase	datos[s]	ventana[s]	%train	%over	observaciones	*EE*	*G*	$C_{\max}$
dec.95*M*	VT	608	2	2/3	50	405	3	4	*33*
col.04*M*	VT	2369	4	2/3	50	789	5	4	*39*

Tabla A.4.1.: *Tamaño máximo del vector de características*, C_{max}, para evitar el sobre-entrenamiento de modelos HMM definidos por EE estados emisores y G gausianas diagonales en función del $\%train$ de eventos usados en el entrenamiento y del $\%over$ del solapamiento ed la ventana de características. Se muestra como ejemplo la clase VT en la base de datos de Colima ($col.04M$) y Decepción ($dec.95M$).

En la Tabla A.4.1 se resume para una configuración usual la complejidad máxima (representada por el máximo nº de características, C_{max}, del vector de parametrización) permitida al construir los modelos para la clase con menos datos de cada BD. Dada una duración total de los datos ($datos[s]$) de una clase, de los cuales una parte ($\%train$) están dedicados a entrenamiento, obtenemos el nº de observaciones usadas para modelar dicha clase tras el proceso de ventaneo definido por la duración de la ventana deslizante ($ventana[s]$) y su % de solapamiento ($\%over$). El nº de estas observaciones define el tamaño del vector de parametrización mediante la Ecuación A.4.1:

$$ECG \leqslant observaciones \tag{A.4.1}$$

Programas que permiten el sobre-entrenamiento de modelos.
Con el objeto de evitar errores en el proceso de modelado y minimizar el efecto del sobre-entrenamiento, muchos programas automáticamente fijan un mínimo de variabilidad que deben tener las componentes gausianas de los GMM/HMM inicializando las matrices de covarianza. En el caso de HTK, este viene definido por la variable *varFloor* (Young et al., 2006). Esta automatización puede ser a veces contraproducente, pues permite construir modelos que no cumplen la Ecuación A.4.1. Aún así, una variabilidad mínima NO es garantía de que los modelos no se vean degradados por la falta de datos. En caso de sobre-entrenamiento, el software desarrollado en este trabajo automáticamente impone una variabilidad mínima en la definición de los modelos, previo aviso al usuario.

¿Hasta cuando hay que aumentar el tamaño de la base de datos de entrenamiento? No es necesario seguir incluyendo datos de entrenamiento si se cumplen estos supuestos:

- Tenemos un conjunto de características lo suficientemente bueno, tal que, un humano experto en la materia sea capaz de clasificar correctamente usando la descripción de las señales que da esa parametrización
- Tenemos un modelo clasificador lo suficientemente complejo para describir los datos de entrenamiento (equivalentemente: tenemos bajo error de predicción

en el test cerrado) o, lo que es lo mismo, nuestro clasificador es de *pequeño bias*

- los datos de entrenamiento son mucho más numerosos que el número de parámetros que nuestro modelo debe aprender. Equivalentemente: existe poca probabilidad de cometer sobre-entrenamiento (nuestros modelos tienen *baja varianza*)

A.5. ¿Se deben normalizar los registros sísmicos antes del proceso de extracción de características?

A favor:

- Aumenta la exportabilidad de las características al llevar los valores del sismograma dentro de un rango de amplitud predeterminado ([-1,1]) antes del proceso de extracción. Con ello, contribuye a construir modelos exportables.

En contra:

- Elimina información de la energía absoluta de la señal, propiedad importante para caracterizar geofísicamente a los eventos.
- Disminuye la capacidad de generalización de los modelos.

En el caso de que la información de la energía absoluta sea relevante para el proceso de reconocimiento podemos adoptar algunas de estas soluciones intermedias:

- Incluir en el set de características únicamente un componente que de la energía absoluta, y posteriormente normalizar la señal.
- Acondicionamiento o pre-procesamiento de la señal: eliminar su media en cada registro y filtrar en un rango de frecuencias de interés ([1-25]Hz ó [1-50]Hz)
- Hacer un escalado no lineal previo de la señal (usando logaritmos, por ejemplo) para disminuir el rango dinámico del sismograma, previo la extracción de características, fomentando así la exportabilidad pero sin perder la información de energía absoluta.

A.6. ¿Cómo influye la variabilidad de una característica en su capacidad para diferenciar entre clases de eventos?

La desviación estándar de una variable es una medida ampliamente usada para asignarle más peso en el área de codificación. En cuanto a clasificación, que una característica varíe a lo largo de un flujo de datos no es necesariamente indicador de su poder discriminativo. Pueden darse los siguientes casos:

1. Que varíe cuando se detecta un evento y/o varíe en distintas partes de un evento
2. Que varíe aleatoriamente, sin relación directa con la llegada de un evento
3. Que tenga un valor parecido para eventos del mismo tipo.

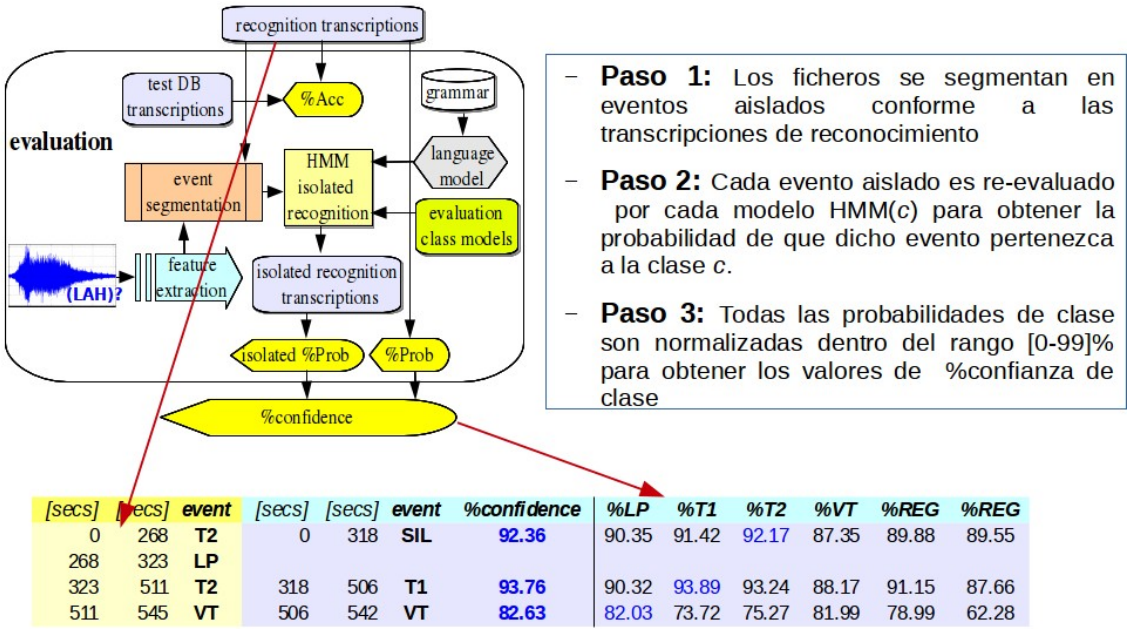
Nótese que excepto en el caso 2, en general podría afirmarse que la variabilidad de una característica es una propiedad interesante desde el punto de vista de la clasificación. Sin embargo, no siempre es cierto: el caso 3 es el más útil desde el punto de vista de clasificación e implicaría poca variabilidad entre eventos de la misma clase, pero cierta variabilidad entre eventos de clases distintas.

En cuanto al caso 1; por sí mismo no implica una separabilidad entre clases. La diferencia de la distribución de esa variabilidad según las clases es lo que nos da el grado de separabilidad entre ellas. Estas distribuciones es precisamente lo que intentan describir los modelos de clasificación.

A.7. Evaluación de las probabilidades $\{p(\mathbf{x}, w_C)\}$ dada la secuencia $\mathbf{x} = \{\mathbf{x}_t\}$ por cada HMM asociado a las clases $\{w_c\}$

Dado un registro de datos en continuo descrito por la secuencia de vectores $\mathbf{x} = \{\mathbf{x}_t\}_{t=1:T}$ y un conjunto, $\{HMM_C\}$ de HMMs cada uno asociado a una clase w_c, una vez construida la red de búsqueda o macro-HMM simbolizado por M_{RB} (Figura 3.3.3) el algoritmo de Viterbi (Sección 3.3.2.1) usado en la decodificación halla la secuencia de estados $\mathbf{q_x} = \{q_t\}_{t=1:T}$ o *camino* que maximice la probabilidad $p(\mathbf{x}, \mathbf{q}; M_{RB})$. Hallada $\mathbf{q_x}$ es inmediato extraer la secuencia de etiquetas $\{\hat{w}_1, ..., \hat{w}_N\}$ como resultado del proceso de reconocimiento en continuo.

La evaluación de cada uno de los vectores $\mathbf{x}_t$ de la secuencia $\mathbf{x}$ por cada uno de los modelos $\{w_c\}$ no es posible hacerla directamente por Viterbi, que solo proporciona una secuencia de estados. A esto se une el inconveniente técnico de que el software usado, HTK (versión 3.4) de Young et al. (2006) no da directamente el valor de $p(\mathbf{x}_t, q_t)$, con q_t el estado que genera el vector $\mathbf{x}_t$ de la secuencia $\mathbf{x}$ en un instante



[secs]	[secs]	event	[secs]	[secs]	event	%confidence	%LP	%T1	%T2	%VT	%REG	%REG
0	268	T2	0	318	SIL	92.36	90.35	91.42	92.17	87.35	89.88	89.55
268	323	LP										
323	511	T2	318	506	T1	93.76	90.32	93.89	93.24	88.17	91.15	87.66
511	545	VT	506	542	VT	82.63	82.03	73.72	75.27	81.99	78.99	62.28

Figura A.7.1.: Extracción de las probabilidades $\{p(\mathbf{x}, w_c)\}$ por re-evaluación en reconocimiento aislado. Usando las transcripciones de reconocimiento se segmentan los ficheros en eventos aislados y se vuelven a evaluar por cada modelo HMM asociado a cada clase w_c (Cortés et al., 2009b).

t concreto, sino un promedio de probabilidad correspondiente al intervalo $[t, t+U]$ en el que la secuencia $\{q_t, q_{t+1}, ..., q_{t+U}\}$ permanece en el mismo estado lógico del HMM correspondiente. Aún así, es posible estimar con cierto grado de confianza las probabilidades $\{p(\mathbf{x}, w_c)\}$ para cada modelo de diversas maneras:

1. *Interpolado a partir de las secuencias promediadas de HTK*, asociando el valor promediado de probabilidad $p(\mathbf{x}_{A,t}, q_t)$ a los vectores $\mathbf{x}_t$ de un evento $\mathbf{x}_A$ que son generados por el mismo estado lógico del HMM_A podemos estimar $\{p(\mathbf{x}_A, w_A)\}$ si el sistema ha asignado $\mathbf{x}_A \rightarrow w_A$. El problema es hallar $\{p(\mathbf{x}_A, w_c)\}$ para un evento $\mathbf{x}_A$ para las clases $w_c \neq w_A$. Una solución inicial es asociar es $\{p(\mathbf{x}_A, w_{c\neq A})\} \equiv m_c$, siendo $m_c = \text{mín}_t\{p(\mathbf{x}_{t,c}, w_{c\neq A})\}$ y $\mathbf{x}_{t,c}$ todos los vectores de la secuencia $\mathbf{x}$ asociados a la clase w_c en un fichero de datos evaluado.

2. *Re-evaluación de eventos aislados.* Cortés et al. (2009b) logran obtener las probabilidades $\{p(\mathbf{x}, w_c)\}$ para cada clase w_c y para cada vector $\mathbf{x}_t$ de la secuencia $\mathbf{x} = \{\mathbf{x}_t\}_{t=1:T}$ re-evaluando en aislado los eventos de un fichero continuo que hayan sido reconocidos. El proceso se ilustra en la Figura A.7.1:

   a) Segmentación de los eventos de un fichero a partir de las marcas temporales contenidas en las transcripciones de reconocimiento

   b) Evaluación en aislado de cada evento segmentado $\mathbf{x}_S$ por cada modelo HMM perteneciente a cada una de las clases w_c a considerar: en cada

evaluación solo se permite crear una red de búsqueda a través de un único HMM_c asociado a la clase w_c. Esto posibilita que HTK extraiga las probabilidades $p(\mathbf{x}_{S,t}, w_c)$ para cada vector $\mathbf{x}_{S,t}$ de $\mathbf{x}_S$ en cada instante t.

c) Normalización en el intervalo [0,1] de todas las probabilidades de clase $\{p(\mathbf{x}, w_c)\}$. Estas probabilidades normalizadas se denominan *tasa de confianza (robustez) de clase* de que el evento segmentado $\mathbf{x}_S$ pertenezca a la clase w_c.

Como vemos en la Figura A.7.1, la extracción de las tasas de robustez de clase posibilita el *re-etiquetado en aislado* de todos los ficheros en vez de usar el etiquetado original dado por el reconocimiento en continuo: un evento se asigna a la clase que obtiene la mayor tasa de robustez. Cortés et al. (2009b) obtienen valores similares de $\%cAcc$ con este re-etiquetado comparado con el etiquetado en continuo en un test abierto y continuo con una base de datos compuesta por eventos de Colima y Popocatépetl.

Se observa que los resultados al aplicar uno u otro método son bastante similares, así que utilizamos el interpolado al ser bastante más rápido.

B. Tablas de selección de características

Estos son las tablas de selección de características correspondientes al Capítulo 6, donde se implementa una arquitectura PSA en paralelo en la que cada canal se configura para funcionar como un detector y clasificador especializado en un tipo de evento concreto representado por el modelo de clase propia del canal. Se construyen dos sistemas, uno para cada base de datos en continuo, la del volcán de Colima (*col.04Mc*) y la del volcán Decepción (*dec.95Mc*), cuyos eventos son descritos por secuencias de vectores *geoLFCC.D.30* de 30 características mixtas detalladas en la Sección 4.2.

Estas tablas son una extensión de las que se encuentran en el proceso de configuración de los sistemas PSA en la Subsección 6.3.1 y muestran la efectividad que cada característica posee para describir los eventos de la clase propia de cada canal PSA, dada por su *posición* en la columna. Dicha posición es asignada por el algoritmo discriminatorio generalizado *DFS.cAcc* (Subsubsección 4.3.2.2).

Relevancia de las características *geoLFCC.D.30* en los canales PSA.mul(c) @ col.04Mc										
COL	EXP	LAH	LPS	REG	SPT	TP	TR	TS	VT	WNS
lfcc1.D	f.80.D	f.nac	f.Slo	f.Slo	f.Slo	f.80	f.50.D	f.Slo	f.Slo	htkE.D
f.80.D	lfcc1	lfcc1	lfcc1	f.Max	f.Max	f.Slo	lfcc2.D	lfcc1	lfcc1	fSlo.D
lfcc2	f.80	f.Slo	f.nac	f.kur	LPFerr	f.kur	f.80	lfcc3.D	f.nac	f.50
f.nac	LPFerr	lfcc1.D	lfcc2	lfcc1.D	lfcc3.D	lfcc2	fSlo.D	f.skew	f.Max	lfcc1.D
f.skew	f.50	f.kur	f.Max	f.skew	fSlo.D	f.80.D	htkE.D	f.nac	LPFerr	lfcc2.D
f.50	f.nac	lfcc2.D	LPFerr	lfcc2	f.80.D	f.50.D	f.Slo	lfcc1.D	lfcc1.D	f.Max
f.Slo	f.50.D	f.80.D	f.kur	LPFerr	lfcc1.D	lfcc3	f.kur.D	f.80	lfcc2.D	f.Max.D
f.50.D	fSlo.D	LPFerr	f.80.D	lfcc5	htkE.D	f.50	f.20.D	lfcc2.D	lfcc2	lfcc2
f.kur	lfcc2	htkE	lfcc3.D	f.50	f.50.D	f.Max.D	f.Max.D	f.Max.D	htkE.D	lfcc3
lfcc4	f.kur	f.kur.D	lfcc4.D	f.50.D	f.skew	lfcc3.D	f.50	f.kur.D	lfcc5.D	f.50.D
htkE.D	f.skew	lfcc2	f.50	LPFerr.D	lfcc1	f.nac.D	f.skew	lfcc5.D	lfcc4.D	lfcc3.D
f.kur.D	f.nac.D	lfcc3.D	lfcc1.D	fSlo.D	f.kur	f.Max	LPFerr.D	lfcc2	lfcc4	f.Slo
lfcc4.D	f.Slo	htkE.D	f.20.D	lfcc1	lfcc4.D	f.20	f.nac.D	lfcc4	f.kur	f.nac.D
LPFerr	LPFerr.D	f.50.D	f.nac.D	lfcc4	f.20.D	f.skew	lfcc5	f.50.D	f.nac.D	LPFerr
f.Max	htkE.D	f.80	f.50.D	lfcc4.D	f.nac	htkE	LPFerr	f.nac.D	lfcc3.D	f.80
lfcc2.D	f.Max	f.skew.D	LPFerr.D	lfcc5.D	LPFerr.D	lfcc1	f.nac	htkE	LPFerr.D	f.skew.D
lfcc3.D	lfcc3.D	lfcc4.D	f.kur.D	f.nac.D	f.80	f.nac	lfcc1.D	f.Max	f.skew	f.20
fSlo.D	lfcc2.D	lfcc5.D	lfcc5.D	lfcc2.D	lfcc2	htkE.D	lfcc3.D	f.80.D	fSlo.D	f.skew
f.nac.D	lfcc3	f.20.D	f.80	f.20.D	f.nac.D	f.20.D	f.80.D	lfcc5	f.80.D	lfcc1
f.80	f.20	f.Max.D	fSlo.D	f.80.D	lfcc5.D	f.kur.D	f.20	f.kur	f.kur.D	f.20.D
lfcc1	lfcc1.D	f.skew	lfcc2.D	htkE.D	f.skew.D	LPFerr.D	lfcc3	lfcc4.D	f.80	f.kur
lfcc5.D	lfcc5	LPFerr.D	f.Max.D	lfcc3.D	lfcc2.D	f.skew.D	f.kur	fSlo.D	f.20	lfcc4
f.skew.D	lfcc4.D	fSlo.D	f.skew.D	f.Max.D	f.20	LPFerr	f.skew.D	f.20.D	lfcc5	f.nac
f.20.D	lfcc4	f.nac.D	htkE.D	f.20	f.50	lfcc5	f.Max	f.skew.D	f.Max.D	f.80.D
LPFerr.D	f.Max.D	lfcc3	lfcc5	f.nac	f.kur.D	lfcc1.D	lfcc4.D	LPFerr.D	f.20.D	lfcc5.D
f.20	f.kur.D	lfcc5	lfcc4	f.skew.D	f.Max.D	lfcc4.D	lfcc5.D	f.50	f.skew.D	lfcc4.D
htkE	htkE	lfcc4	htkE	lfcc3	lfcc3	lfcc5.D	lfcc4	f.20	f.50	f.kur.D
lfcc5	f.20.D	f.Max	f.skew	htkE	lfcc4	fSlo.D	lfcc1	LPFerr	f.50.D	LPFerr.D
f.Max.D	f.skew.D	f.50	f.20	f.80	htkE	lfcc2.D	htkE	lfcc3	htkE	lfcc5
lfcc3	lfcc5.D	f.20	lfcc3	f.kur.D	lfcc5	lfcc4	lfcc2	htkE.D	lfcc3	htkE

Tabla B.0.1.: *Relevancia de las 30 componentes del vector mixto geoLFCC.D.30 para describir la clase propia de cada canal múltiple PSA.mul(c) en la base col.04Mc.*. En cada columna las características se presentan en orden descendente de importancia.

Relevancia de las características *geoLFCC.D.30* en los canales PSA.bin(c) @ col.04Mc										
COL	EXP	LAH	LPS	REG	SPT	TP	TR	TS	VT	WNS
f.kur	htkE.D	lfcc1	f.Max	f.80	LPFerr	f.50.D	f.nac	lfcc1.D	f.80	htkE.D
lfcc1.D	lfcc1.D	f.Max	lfcc1	f.50	htkE.D	f.Max.D	f.80	LPFerr	f.nac	fSlo.D
f.50	f.50.D	f.kur	LPFerr	f.50.D	f.Slo	lfcc4	LPFerr	fSlo.D	f.nac.D	f.Max.D
f.80.D	f.50	lfcc1.D	lfcc1.D	f.nac.D	lfcc1.D	f.Slo	f.80.D	lfcc1	f.50.D	f.Slo
f.nac	LPFerr	f.skew	f.50	f.nac	f.kur	f.50	f.Max.D	f.50	f.kur	LPFerr
htkE	f.80.D	f.Slo	lfcc3.D	f.80.D	f.Max	lfcc1.D	f.20	lfcc3	f.80.D	f.nac
f.kur.D	LPFerr.D	htkE.D	f.kur	f.Slo	f.nac	lfcc3.D	htkE.D	htkE.D	f.Slo	lfcc1.D
f.skew	f.80	f.nac	f.nac	f.kur	f.Max.D	htkE.D	lfcc4	f.80	f.50	f.80
htkE.D	f.nac	LPFerr	f.80.D	f.Max	f.80.D	f.80.D	lfcc3.D	f.Max	f.Max	f.Max
f.Slo	f.kur	f.80.D	f.kur.D	f.20	f.50.D	lfcc2.D	htkE	f.50.D	LPFerr	lfcc2.D
fSlo.D	lfcc2.D	lfcc2.D	fSlo.D	lfcc4	lfcc3.D	f.Max	f.skew.D	f.Max.D	lfcc2	lfcc2
LPFerr.D	fSlo.D	htkE	f.50.D	f.Max.D	f.nac.D	f.kur.D	lfcc1	lfcc5	lfcc2.D	f.50
lfcc3.D	f.Slo	fSlo.D	lfcc2	lfcc2	f.20	LPFerr	f.Max	f.80.D	f.20	f.skew
f.Max	f.nac.D	lfcc3	f.Slo	lfcc1.D	fSlo.D	f.kur	lfcc3	htkE	htkE.D	lfcc3
lfcc2	lfcc3	lfcc3.D	f.Max.D	fSlo.D	f.kur.D	f.nac	f.50.D	f.Slo	f.skew	lfcc1
lfcc2.D	f.kur.D	lfcc4	lfcc4.D	f.20.D	f.50	fSlo.D	f.Slo	f.nac.D	lfcc1.D	f.20.D
f.skew.D	f.20.D	lfcc2	f.skew	lfcc3	LPFerr.D	lfcc4.D	fSlo.D	f.skew	fSlo.D	f.20
f.50.D	htkE	f.50.D	f.80	lfcc1	lfcc2.D	lfcc3	f.kur.D	LPFerr.D	f.Max.D	f.kur.D
LPFerr	f.skew.D	f.Max.D	lfcc5	htkE	lfcc2	f.nac.D	lfcc2.D	f.kur.D	LPFerr.D	lfcc3.D
lfcc3	lfcc4	lfcc5	lfcc5.D	LPFerr	f.80	f.80	f.nac.D	lfcc2	f.20.D	f.kur
lfcc5	f.Max.D	f.20.D	lfcc3	htkE.D	f.skew	f.20	lfcc5	lfcc2.D	lfcc3.D	f.80.D
f.20.D	lfcc4.D	LPFerr.D	lfcc2.D	f.skew	lfcc4.D	LPFerr.D	lfcc2	f.kur	f.skew.D	f.nac.D
lfcc4.D	f.20	f.50	htkE.D	lfcc5.D	lfcc4	lfcc1	f.kur	f.skew.D	lfcc4	f.50.D
f.20	f.Max	lfcc4.D	f.20	lfcc5	htkE	lfcc2	lfcc4.D	f.nac	lfcc3	lfcc4
lfcc5.D	lfcc2	f.kur.D	f.skew.D	f.kur.D	f.20.D	lfcc5	lfcc5.D	lfcc5.D	htkE	lfcc5
f.Max.D	f.skew	lfcc5.D	LPFerr.D	lfcc4.D	f.skew.D	f.skew	lfcc1.D	f.20.D	lfcc5.D	LPFerr.D
lfcc4	lfcc1	f.20	f.20.D	LPFerr.D	lfcc3	f.skew.D	f.50	lfcc3.D	lfcc1	lfcc5.D
f.80	lfcc5	f.nac.D	htkE	f.skew.D	lfcc5	f.20.D	f.skew	lfcc4.D	f.kur.D	lfcc4.D
f.nac.D	lfcc3.D	f.skew.D	lfcc4	lfcc2.D	lfcc5.D	htkE	f.20.D	f.20	lfcc4.D	f.skew.D
lfcc1	lfcc5.D	f.80	f.nac.D	lfcc3.D	lfcc1	lfcc5.D	LPFerr.D	lfcc4	lfcc5	htkE

Tabla B.0.2.: *Relevancia de las 30 componentes del vector mixto geoLFCC.D.30 para describir la clase propia de cada canal binario PSA.bin(c) en la base col.04Mc.* En cada columna las características se presentan en orden descendente de importancia.

Relevancia de las características *geoLFCC.D.30* en los canales PSA.mul(c) y PSA.bin(c) @ dec.95Mc									
PSA.mul(c)					**PSA.bin(c)**				
HY	LP	NS	TR	VT	HY	LP	NS	TR	VT
f.nac	f.kur	f.Slo	fSlo.D	f.Slo	f.Slo	f.Slo	f.80	f.nac	f.Slo
LPFerr	f.Slo	f.50	f.20	lfcc2	f.Max	f.skew	f.50	f.kur	f.kur
fSlo.D	f.nac	f.Max	f.Slo	f.Max	f.50.D	f.nac	f.Max	LPFerr	f.nac
f.kur	fSlo.D	f.80	f.nac	f.skew	f.nac	f.kur	f.nac	f.50.D	htkE.D
f.Max	lfcc2	fSlo.D	f.Max	LPFerr	lfcc2.D	fSlo.D	f.50.D	f.Slo	f.skew
f.Slo	f.Max	f.kur	htkE.D	f.nac	f.kur	lfcc1.D	f.Slo	f.Max	lfcc2.D
htkE.D	LPFerr	lfcc2	lfcc2.D	f.50.D	lfcc1	f.80.D	f.20.D	htkE.D	lfcc2
lfcc2	f.80.D	LPFerr	lfcc2	f.kur	f.skew.D	f.50	lfcc2	f.20	LPFerr
f.nac.D	f.skew	f.nac	f.kur	lfcc1	f.20.D	f.80	fSlo.D	f.80	f.20.D
lfcc1	f.nac.D	f.20.D	f.80.D	fSlo.D	f.80.D	lfcc2.D	f.skew	f.50	lfcc4.D
f.skew.D	htkE.D	f.80.D	f.skew	lfcc2.D	f.kur.D	f.20	lfcc3	lfcc2.D	lfcc1
lfcc3	lfcc3	f.skew	LPFerr.D	f.kur.D	lfcc3.D	LPFerr.D	f.nac.D	fSlo.D	f.50.D
lfcc3.D	lfcc5.D	f.kur.D	f.50.D	lfcc1.D	lfcc4.D	f.skew.D	LPFerr	f.skew	fSlo.D
lfcc5.D	f.kur.D	f.Max.D	f.Max.D	htkE	lfcc1.D	f.Max	htkE.D	lfcc4	f.80.D
f.kur.D	f.50.D	LPFerr.D	f.kur.D	lfcc4	f.nac.D	f.50.D	f.kur.D	lfcc4.D	lfcc4
lfcc2.D	lfcc4.D	lfcc2.D	LPFerr	lfcc5.D	lfcc2	lfcc1	f.skew.D	f.nac.D	f.50
f.Max.D	f.20	f.skew.D	lfcc1.D	lfcc3.D	htkE.D	lfcc2	f.kur	f.20.D	LPFerr.D
f.80.D	f.80	f.50.D	htkE	f.80.D	f.80	f.nac.D	f.80.D	f.kur.D	f.nac.D
LPFerr.D	f.skew.D	lfcc3.D	f.nac.D	f.50	lfcc4	lfcc4.D	LPFerr.D	f.Max.D	lfcc3.D
lfcc4.D	lfcc3.D	lfcc5.D	f.50	f.20.D	LPFerr.D	lfcc3.D	lfcc1.D	f.skew.D	lfcc5.D
f.20.D	lfcc1	htkE.D	f.20.D	LPFerr.D	lfcc3	lfcc5.D	lfcc4	lfcc3.D	f.Max.D
f.50	lfcc1.D	f.20	lfcc3.D	f.nac.D	fSlo.D	LPFerr	lfcc3.D	LPFerr.D	f.20
f.skew	f.Max.D	lfcc5	f.80	lfcc4.D	lfcc5.D	f.kur.D	lfcc2.D	lfcc1.D	f.80
lfcc1.D	lfcc2.D	lfcc4.D	f.skew.D	f.Max.D	f.skew	f.Max.D	f.Max.D	lfcc2	f.Max
f.50.D	f.50	f.nac.D	lfcc4	f.80	LPFerr	f.20.D	lfcc4.D	lfcc5.D	lfcc5
lfcc5	f.20.D	lfcc3	lfcc5.D	lfcc3	f.Max.D	lfcc4	lfcc5	lfcc1	f.kur.D
lfcc4	lfcc5	lfcc4	lfcc4.D	lfcc5	htkE	htkE.D	f.20	lfcc3	lfcc3
f.80	LPFerr.D	lfcc1.D	lfcc3	f.20	f.20	lfcc5	lfcc5.D	htkE	f.skew.D
f.20	lfcc4	lfcc1	lfcc1	f.skew.D	lfcc5	htkE	lfcc1	f.80.D	htkE
htkE	htkE	htkE	lfcc5	htkE.D	f.50	lfcc3	htkE	lfcc5	lfcc1.D

Tabla B.0.3.: *Relevancia de las 30 mejores características del vector mixto geoLFCC.D.30 para describir la clase propia de cada canal múltiple PSA.mul(c) y binario PSA.bin(c) en la base dec.95Mc.* La importancia de cada componente viene dada por su posición en la columna.

Relevancia de las características del vector mixto *geoLFCC.D.30*						
	dec.95Mc			*col.04Mc*		
posición	**SSA**	**PSA.mul(c)**	**PSA.bin(c)**	**SSA**	**PSA.mul(c)**	**PSA.bin(c)**
1	f.nac	f.Slo	f.Slo	f.80	f.Slo	f.nac
2	f.Slo	f.Max	f.nac	f.kur.D	lfcc1.D	lfcc1.D
3	f.50.D	f.nac	f.kur	f.skew	f.nac	LPFerr
4	lfcc2	fSlo.D	f.50.D	f.80.D	f.Max	f.Slo
5	f.50	f.kur	f.Max	fSlo.D	lfcc1	f.Max
6	LPFerr.D	lfcc2	f.skew	f.50.D	lfcc2	htkE.D
7	f.Max	LPFerr	lfcc2.D	f.Max	f.kur	f.80.D
8	LPFerr	f.skew	fSlo.D	f.Slo	LPFerr	f.kur
9	htkE.D	f.80.D	f.80	htkE.D	f.50.D	f.50
10	f.skew	f.kur.D	htkE.D	lfcc4	f.80.D	f.50.D
11	f.kur.D	lfcc2.D	f.20.D	LPFerr.D	lfcc3.D	fSlo.D
12	fSlo.D	f.50.D	lfcc2	f.20.D	lfcc2.D	f.80
13	f.80.D	htkE.D	LPFerr	f.Max.D	f.80	f.Max.D
14	f.skew.D	f.50	f.50	lfcc2.D	f.skew	lfcc2.D
15	lfcc3.D	f.nac.D	f.80.D	f.nac	fSlo.D	f.skew
16	f.20.D	lfcc5.D	f.nac.D	lfcc3.D	htkE.D	lfcc1
17	lfcc2.D	f.Max.D	f.skew.D	f.nac.D	f.50	f.kur.D
18	lfcc3	lfcc3.D	lfcc4.D	f.kur	f.nac.D	lfcc2
19	lfcc5	LPFerr.D	LPFerr.D	lfcc1.D	lfcc4.D	lfcc3.D
20	f.kur	f.20.D	lfcc3.D	f.20	f.kur.D	f.nac.D
21	lfcc5.D	f.80	f.kur.D	f.50	f.20.D	lfcc3
22	lfcc4	f.skew.D	lfcc4	LPFerr	LPFerr.D	f.20
23	f.20	f.20	lfcc1	lfcc5.D	lfcc5.D	LPFerr.D
24	f.80	lfcc3	f.20	lfcc1	f.Max.D	htkE
25	lfcc1.D	lfcc1.D	lfcc1.D	f.skew.D	lfcc4	lfcc4
26	lfcc4.D	lfcc4.D	f.Max.D	lfcc2	f.skew.D	f.20.D
27	f.nac.D	lfcc1	lfcc5.D	htkE	lfcc5	f.skew.D
28	f.Max.D	lfcc4	lfcc3	lfcc5	f.20	lfcc5
29	lfcc1	lfcc5	lfcc5	lfcc4.D	lfcc3	lfcc4.D
30	htkE	htkE	htkE	lfcc3	htkE	lfcc5.D

Tabla B.0.4.: *Comparación de la relevancia de las características del vector geoLFCC.D.30 para describir cada clase en las arquitecturas SSA, PSA.mul(c) y PSA.bin(c) en las bases dec.95Mc y col.04Mc.* Se presentan valores promedio entre los canales de los sistemas paralelo PSA. La importancia de cada componente viene dada por su posición en la columna.

C. Divulgación científica

La realización de esta memoria es solo una parte de la actividad investigadora y profesional llevada a cabo desde el año 2005 dentro del grupo interdisciplinar de vulcanología a cargo del Instituto Andaluz de Geofísica y el Dpto. de Teoría de la Señal, Telemática y Comunicaciones de la Universidad de Granada. Durante ese periodo, diversos trabajos han complementado la realización de esta tesis.

C.1. Artículos en revistas especializadas

- Cortés, G.; Benítez, C.; García, L.; Álvarez, I.; Ibañéz, J.M. *"A Comparative Study of Dimensionality Reduction Algorithms Applied to Volcano-Seismic signals"*. in Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Jorunal of, vol.PP, no.99, pp.1-11, accepted for publication Sept. 2015, doi:10.1109/JSTARS.2015.2479300.
- Boué, A..; Lesage, P.; Cortés, G.; Valette, B.; Reyes-Dávila, G. *"Real-time eruption forecasting using the material Failure Forecast Method with a Bayesian approach"*. Journal of Geophysical Research: Solid Earth, vol.120, no.4, pp.2143-2161, 2015.
- Cortés, G.; García, L.; Álvarez, I.; Benítez, C.; De la Torre, A.; Ibáñez, J.M. *"Parallel System Architecture (PSA): An efficient approach for automatic recognition of volcano-seismic events"*. Journal of Volcanology and Geothermal Research, vol.271, no.0, pp.1-10, 2014.
- Carmona, E.; Almendros, J.; Martín, R.; Cortés, G.; Alguacil, G.; Moreno, J.; Martín, B.; Martos, A; Serrano, I.; Stich, D.; Ibáñez, J.M. *"Advances in seismic monitoring at Deception Island volcano (Antarctica) since the International Polar Year"*. Annals of Geophysics, vol.57, no.3, 2014
- Alvarez, I.; Garcia, L.; Mota, S.; Cortes, G.; Benitez, C.; De la Torre, A. *"An Automatic P-Phase Picking Algorithm Based on Adaptive Multiband Processing"*. Geoscience and Remote Sensing Letters, IEEE , vol.10, no.6, pp.1488,1492, Nov. 2013.
- Alvarez, I.; Garcia, L.; Cortes, G.; Benitez, C.; De la Torre, A. *"Discriminative Feature Selection for Automatic Classification of Volcano-Seismic Signals"*. Geoscience and Remote Sensing Letters, IEEE, vol.9, no.2, pp.151-155, 2012.

- Ibañez, J.M.; Benítez, C.; Gutiérrez, L.A.; Cortés, G., García-Yeguas, A.; Alguacil, G. *"The xlassification of seismo-volcanic signals using Hidden Markov Models as applied to Stromboli and Etna volcanoes"*. Journal of Volcanology and Geothermal Research, vol.187, no.3-4, pp.218-227, 2009.
- Benítez, C.; Ramírez, J.; Segura, J.C.; Ibáñez, J.M.; Almendros, J.; García-Yeguas, A.; Cortés, G. *"Continuous HMM-Based Seismic-Event Classification at Deception Island, Antarctica"*. In Geoscience and Remote Sensing, IEEE Trans.on, vol.45, pp.138-146, 2007.

C.2. Capítulos de libro

- Cortés, G.; Arámbula, R.; Álvarez, I.; Benítez, C.; Ibáñez, J.M.; Lesage, P.; González, M.; Reyes, G. *"Analysis of Colima, Popocatépetl and Arenal volcanic seismicity using an automatic Continuous Hidden Markov Models based recognition"*. VOLUME Project: VOLcanoes, Understanding Subsurface Mass MoveMEnt. ISBN: 978-1-905254-39-2. C.J. Bean and European Commission. 6th Framework Programme. 2: 150-160. 2009.
- Benítez, C.; Lesage, P.; Cortés, G.; Segura, J.C.; Ibáñez, J.M.; De la Torre, A. *"Automatic recognition of volcano-seismic events based on Continuous Hidden Markov Models based recognition"*. VOLUME Project: VOLcanoes, Understanding Subsurface Mass MoveMEnt. ISBN: 978-1-905254-39-2. C.J. Bean and European Commission. 6th Framework Programme. 2: 130-140. 2009.
- Benítez, C.; Ibáñez, J.M.; García, L.; Gutiérrez, L.A.; Cortés, G.; Álvarez, I. *"Analysis of volcano seismicity at Deception Island, Stromboli volcano and Mt Etna using an automatic CHMM based recognition method"*. VOLUME Project: VOLcanoes, Understanding Subsurface Mass MoveMEnt. ISBN: 978-1-905254-39-2. C.J. Bean and European Commission. 6th Framework Programme. 2: 140-150. 2009.

C.3. Ponencias en congresos internacionales

- Cortés, G.; Benítez, C.; Ibáñez, J.M.; González, E.; Arámbula, R.; Lesage, P.; Orozco, J. *"Towards an Unsupervised HMM-based Automatic Classification System: Application to a Joint Database Built from Colima, Popocatépetl and Deception Seismo-Volcanic Events"*. Cities on Volcanoes CoV 2010.
- Cortés, G.; Arámbula-Mendoza, R.; Gutiérrez, L.A.; Benítez, C.; Ibañez, J.M.; Lesage, P. *"Evaluating robustness of a hmm-based classification system of volcano-seismic events at Colima and Popocatépetl volcanoes"*. IEEE Transactions of Geoscience and Remoting Sensing Symposium TGARSS 2009.

- Cortés, G.; García, L.; Benítez, C.; Segura, J.C. *"HMM-Based Continuous Sign Language Recogniton using a Fast Optical Flow Parameterization of Visual Information"*. Artículo presentado en la sesión oral del congreso internacional INTERSPEECH-ICSLP 2006, Pittsburgh, U.S.A, Septiembre 2006.

C.4. Comunicaciones en congresos internacionales

- Boué, A.; Lesage, P.; Cortés, G.; Valette B.; Spica, Z.; Reyes-Dávila, G.; Arámbula-Mendoza, R.; Budi-Santoso, A. *"Performance of the 'material failure forecasting Method' in real-time situation: a Bayesian approach applied on effusive and explosive eruptions"*. In International Union of Geodesy and Geophysics – 26th General Assembly, Prague, Czech Republic, IUGG 2015.
- Díaz, A.; Álvarez, I.; De la Torre, A., García, L,; Benítez, C.; Cortés, G. *"Application of a cross correlation-based picking algorithm to an active seismic experiment in Sicily and Aeolian Islands"*. In EGU General Assembly Conference Abstracts, 2014.
- Boué, A.; Lesage, P.; Cortés, G.; Benítez, C.; Ibáñez, J.; Alvarez, I.; De La Torre, A.; Gutierrez, L.A.; Arámbula-Mendoza, R..; González-Amézcua, M.; Reyes-Dávila, G. *"Automatic classification of seismo-volcanic signals as a tool to improve eruption forecasting"*. IAVCEI Scientific Assembly, Kagoshima, Japan, IAVCEI 2013.
- Carmona, E.; Almendros, J.; Martín, R.; Cortés, G.; Alguacil, G.; Moreno, J.; Martín B. et al. *"Seismic monitoring at Deception Island volcano (Antarctica): Recent advances"*. In EGU General Assembly Conference Abstracts, vol. 14, p. 12717. 2012.
- Boué, A.; Lesage, P.; Cortés, G.; Benítez, C.; Ibáñez, J.M.; Alvarez, I.; De La Torre, A.; Gutierrez, L.A., Arámbula-Mendoza, R.; González-Amézcua, M.; Reyes-Dávila, G. *"Improving the Material Failure Forecast Method (FFM) for eruption prediction by automatic classification of volcano-seismic signals."* Cities on Volcanoes 7, Colima, Mexico. 2012.
- Gonzalez-Amezcua, M.; Arámbula-Mendoza, R.; Reyes-Dávila, G.; Cortés, G.; Lesage, P.; Benítez, C.; Ibáñez, J.,M.; Valdés-González, C. *"Automated classification of volcanic seismic signals using Hidden Markov Models (HMMs) in quasi realtime at Volcán de Colima*". Cities on Volcanoes 7, Colima, Mexico. 2012.
- Cortés, G.; García, L.; Álvarez, I.; González, E.; De la Torre, A.; Ibáñez, J.M. *"Insufficient Amount of Data in Automatic Recognition of Volcano-Seismic Events: Strategies for Dealing with this Problem"*. Cities Of Volcanoes CoV 2010.

- Álvarez, I.; Cortés, G.; De la Torre, A.; Benítez, C.; Ibañez, J.M.; Lesage, P.; Arámbula-Mendoza, R.; González-Amezcua, M. *"Improving feature extraction in the automatic classification of seismic events. Application to Colima and Arenal volcanoes"*. IEEE Transactions of Geoscience and Remoting Sensing Symposium TGARSS 2009.
- Gutiérrez, L.A.; Ibañez, J.M.; Ramírez, J.; Benítez, C.; Tenorio, V. "Volcano-seismic signal detection and classification processing using Hidden Markov Models. Application to San Cristóbal volcano". IEEE Transactions of Geoscience and Remoting Sensing Symposium TGARSS 2009.

C.5. Proyectos de investigación

Proyectos internacionales

- (2013-2015) MED-SUV: MEDiterranean SUpersite Volcanoes (MED-SUV, EC-FP7).
- (2005-2009) VOLUME: VOLcanoes: Understanding sub-surface mass move-MEnt. EC-FP6-018471.

Proyectos nacionales

- (2009-2011) HISS: Título: Modelos Sísmicos De Alta Resolución De Volúmenes Sismogénicos De Volcanes Activos, Islas De Tenerife Y Decepción, Y Su Impacto En La Valoración Del Peligro Volcánico. CGL2008-01660.
- (2012-2014) EPHESTOS: Desarrollo de modelos de propagación de ondas sísmicas en medios altamente heterogéneos y sus efectos: aplicación a regiones volcánicas activas. (Una mejora de los protocolos de alerta temprana y de los modelos de riesgo volcánico). CGL2011-29499-C02-01.

C.6. Docencia en congresos internacionales

- Lesage, P.; Boue, A.; Arambula-Mendoza, R.; Cortés, G. Taller sobre clasificación automática de señales sismo-volcánicas. Ponente en el congreso internacional Cities On Volcanoes, 2012, CoV.7, Colima, Mexico, 24-27 de Noviembre 2012.

D. Nomenclatura

BW - Baum-Wellch algorithm / **BW** - algoritmo de Baum-Wellch

DB - Database / **BD** - Base de Datos

DCT - Discrete Cosine Transform / **TDC** - Transformada Discreta del Coseno

DFS - Discriminant Feature Selection / **DSC** - algoritmo Discriminante de Selección de Características

EM - Expectation - Maximization algorithm / **EM** - algoritmo de Expectación-Maximización

FA - Factor Analysis / **AF** - Análisis Factorial

FDA - Fisher Discriminant Analysis / **ADF** - Análisis Discriminante de Fisher

FFT - Fast Fourier Transform / **TRF** - Transformada Rápida de Fourier

GMM - Gaussian Multivariate Models / **MGM** - Modelos de Gaussianas Multivariadas

KDE - Kernel Density Estimation / **EDN** - Estimadores de Densidad basada en Núcleos

HMM - Hidden Markov Models / **MOM** - Modelos Ocultos de Markov

LDA - Linear Discrimant Analysis / **ADL** - Análisis Discriminante Lineal

LPC - Linear Prediction Coding / **CPL** - Codificación mediante Predicción Lineal

MAP - Maximum A Posteriori / **MAP** - Máxima probabilidad A Posteriori

MCE - Minimum Classification Error / **MEC** - Mínimo Error de Clasificación

MLE - Maximum Likelihood Estimate / **EMV** - Estimación de Máxima Verosimilitud

MFCC - Mel-Frequency Cepstral Coeficcients / **CCFM** - Coeficientes Cepstrales de la Frecuencias Mel

NN - Nearest Neighbors / **VC** - Vécinos más Cercanos

PCA - Principal Component Analysis / **ACP** - Análisis de Componentes Principales

PDF - Probability Density Function / **FDP** - Función Densidad de Probabilidad

SC - Selección de Características

SC_F - Selección de Características mediante Filtros en vez de modelos de clases

SC_I - Selección de Características analizando Independientemente unas de otras

SC_M - Selección de Características guiada mediante Modelos

SC_S - Selección de Características analizando conjuntamente Subgrupos

STFT - Short Time Fourier Transform / **TFTC** - Transformada de Fourier de Tiempo Corto

RC - Reducción de Características

RD - Reducción de Dimensionalidad

rsv - *r*eliability *s*core *v*alue / **vmf** - *v*alor de la *m*edida de la *f*iabilidad

VSR - Volcano-Seismic Recognition / **RSV** - Reconocimiento de Sismos registrados en Volcanes

WFST - Weighted Finite State Transducers / **TPEF** - Transductores Ponderados de Estados Finitos