applied sciences

MDPI

Editorial

Special Issue on IberSPEECH 2022: Speech and Language Technologies for Iberian Languages

José L. Pérez-Córdoba [1,*,†], Francesc Alías-Pujol [2,*,†] and Zoraida Callejas [3,*,†]

[1] SigMAT Research Group, Universidad de Granada, 18071 Granada, Spain
[2] Human-Environment Research (HER) Group, La Salle—Universitat Ramon Llull, 08022 Barcelona, Spain
[3] SISDIAL Research Group, Research Centre for Information and Communications Technologies (CITIC-UGR), Universidad de Granada, 18071 Granada, Spain
* Correspondence: jlpc@ugr.es (J.L.P.-C.); francesc.alias@salle.url.edu (F.-A.P.); zoraida@ugr.es (Z.C.)
† These authors contributed equally to this work.

Abstract: ThisSpecial Issue presents the latest advances in research and novel applications of speech and language technologies based on the works presented at the sixth edition of the IberSPEECH conference held in Granada in 2022, paying special attention to those focused on Iberian languages. IberSPEECH is the international conference of the Special Interest Group on Iberian Languages (SIG-IL) of the International Speech Communication Association (ISCA) and the Spanish Thematic Network on Speech Technologies (*Red Temática en Tecnologías del Habla*, or RTTH for short). Several researchers were invited to extend the contributions presented at IberSPEECH2022 due to their interest and quality. As a result, the Special Issue is composed of 11 papers that cover different research topics related to speech perception, speech analysis and enhancement, speaker verification and identification, speech production and synthesis, natural language processing, together with several applications and evaluation challenges.

Keywords: Iberianlanguages; speech production; speech synthesis; speech recognition; speaker identification; speech enhancement; summarization; semantic representations; natural language processing; neural networks; deep learning; evaluation challenge; audiovisual database

Citation: Pérez-Córdoba, J.L.; Alías-Pujol, F.; Callejas, Z. Special Issue on IberSPEECH 2022: Speech and Language Technologies for Iberian Languages. *Appl. Sci.* **2024**, *14*, 4505. https://doi.org/10.3390/app14114505

Received: 7 May 2024
Accepted: 23 May 2024
Published: 24 May 2024

1. Introduction

Research on speech technology for Iberian languages has a meeting point at IberSPEECH, a biannual conference originally conceived to bring together academia and companies from Spain and Portugal. However, in recent conferences IberSPEECH community has expanded to researchers and practitioners from other European countries with a shared interest in Iberian Languages. While research in speech and language technologies has the potential to yield very innovative outcomes, converting them into viable business ventures requires bridging the gap between academia and business. One way to achieve this goal is to create meeting points for industry and academia. With this aim, IberSPEECH2022, a three-day event that brought together the XII *Jornadas en Tecnologías del Habla* and the VIII Iberian SLTech Workshop, joined researchers, practitioners and entrepreneurs in speech and language technology to promote their interaction and discussion. In addition, and following the success of previous editions since 2006, the sixth event of the series also included a special session titled Albayzín Evaluations, which consisted of four challenges focused on evaluating different speech technologies over TV and radio broadcast data, with corpora provided by Radio Televisión Española (RTVE), Corporació Catalana de Mitjans Audiovisuals, and Corporación Aragonesa de Radio y Televisión.

This Special Issue presents the latest advances in research with novel applications of speech and language processing, paying special attention to those focused on Iberian languages, based on works presented at the 2022 edition of the conference held in Granada,

Spain [1]. The Special Issue is composed of ten high-quality papers containing extended versions of the conference contributions selected by the Technical Program Committee, together with a paper that summarizes the results of the 2022 Albayzín Evaluations. The main highlights of these works are described in this Editorial paper grouped in four main research topics.

2. Contributions

2.1. Natural Language Processing

Natural language processing (NLP) is a very dynamic and diverse field. The special issue showcases the breadth of research within NLP with papers covering important topics including summarizing, topic classification, chatbots, and datasets.

In the age of information overload, the ability to identify key points in vast textual sources is crucial. The article by González et al. (contribution 1) focuses on extractive summarization, which involves selecting and extracting the most important information or key sentences from a given document or set of documents to create a concise summary. The authors present the *Attentional Extractive Summarization* framework, which exploits the attention mechanisms of Siamese hierarchical networks to learn the relations between documents and summary sentences. Following their innovative approach, it is no longer required to perform sentence labeling and oracle extraction, a laborious process involved in state-of-the-art approaches. The experimentation with two corpora of newspaper summarization shows it is possible to obtain high quality summaries.

Natural language understanding is also a key aspect in the interaction with chatbots. In particular, open-domain chatbots have the ability to maintain a conversation with users in a wide range of topics. This entails multiple challenges, as the chatbot must be able to manage an interaction in which topics vary over time, but the information provided must be consistent with the history of the dialogue and the conversation topic. In (contribution 8) Rodríguez-Cantelar et al. present novel Zero-shot approaches to accurately identify the topic and subtopic of the conversation at every moment, and the automatic estimation of inconsistent responses as a first step to endow chatbots with the ability to avoid contradictory responses to variations of semantically similar user inputs. The evaluation results show a good efficacy in both tasks with several dialogue datasets with up to 18 distinct topics.

Underlying the wide range of NLP topics, current advances in the area would not be possible without large high-quality resources for training and testing language models. Gutiérrez-Fandiño et al. (contribution 10) present the massive multilingual crawling corpus *esCorpius-m*. esCorpius-m contains data in 34 languages different from English, with a strong focus on Spanish. One of the most prominent characteristics of the dataset is that it is cleaner than other state-of-the art corpora thanks to a novel cleaning and de-duplication pipeline, and that it maintains document and paragraph boundaries, and incorporates traceability to the source documents. The deduplicated and cleaned corpus, shared in HuggingFace, has a size of 2.5 TB, with 645,772,362 paragraphs and 242,248,582,193 tokens.

2.2. Automatic Speech Recognition

Automatic Speech Recognition (ASR) is a technology that transforms human speech into readable text. This field has grown exponentially over the past decade, with ASR systems popping up in popular applications we use every day such Google Meet or Zoom for meeting transcriptions, and many more. Originally, ASR was focused solely on acoustic cues. Visual Speech Recognition (VSR) allows the development of complementary and/or alternative approaches to ASR to transcribe speech without the need for acoustic information, becoming instrumental to address those challenges posed by noisy communication environments as well as to develop silent speech interfaces for people with speech impairments, for instance. However, the interpretation of visual speech presents specific challenges such as visual ambiguities, the sensibility to changes on lighting conditions and the effect of inter- and intra-speaker variability, among others.

The article coauthored by Gimeno-Gómez and Martínez-Hinarejos (contribution 2) proposes the adaptation of end-to-end VSR systems to a specific speaker by introducing two adaptation techniques based on the so-called *Adapters* [2] to implement the fine-tuning module after speaker identification. The experiments conducted on the Spanish LIP-RTVE database [3] show that, on the one hand, both methods obtain similar recognition rates to those from state-of-the-art alternatives; a result especially relevant when a limited amount of speaker-dependent training data is available. On the other hand, the authors highlight the scalability of their approach with respect to the full-model fine-tuning techniques, reducing the training time and storage costs by up to 80% –thanks to the reduction of the number of learnable parameters–, but at the cost of a slight worsening their word error rate.

A study on classifying phones (speech sound) using electromyographic (EMG) signals obtained from the Spanish ReSSInt-EMG database is presented by Salomons et al. (contribution 3). This database is part of the ReSSInt project [4], which aims to restore speech for laryngectomees using an EMG-based Silent Speech Interface (SSI). Compared to previous studies on this topic, in this paper the authors revised the linear discriminant analysis (LDA) feature reduction procedure, which resulted in changing the number of LDA features from 28 to 21. This features reduction helps to reduce the training time and the complexity of the model, but the accuracy obtained remains similar.

The new study is conducted including new sessions and extending the experiments with different modalities regarding speaker and session dependency. Instead of a bagging classifier, a neural network is used as a classification method. The results obtained suggest that the development of an EMG-based SSI with sufficient performance for real-world applications requires a large and diverse database.

In the paper by Penagarikano et al. (contribution 5), a semisupervised speech data extraction method is presented and applied to create a new dataset designed for the development of fully bilingual Automatic Speech Recognition (ASR) systems for Basque and Spanish. The data collection are from plenary sessions of the Basque Parliament and is used to train domain-adapted models. Also, the main features of a fully bilingual ASR system for Basque and Spanish are presented. This system is based on the integration of acoustic, lexical and language models.

The presented ASR system is able to deal with code switching [5] events (in bilingual communities, speakers sometimes mix languages and jump spontaneously from one language to another) in a natural and computationally efficient fashion. Performance results, Word Error Rates (WER), show a considerable reduction compared with the used baseline system.

2.3. Speech Synthesis

Voice cloning techniques aim to generate synthetic speech that resembles the voice of a particular speaker. The article by González-Docasal et al. (contribution 4) reflects on the relevance of considering the most appropriate input speech corpora from the target speaker as a key factor of the voice cloning process when applying deep learning. The approach focuses on selecting the most-suitable corpus subset considering the lower heterogeneity in terms of phonetic coverage and utterance speed, as well as the lower signal-to-noise ratio, as a means of obtaining cloned speech with higher synthetic quality. This quality is measured by computing two estimators of the well-known Mean Opinion Score (MOS), named NISQA and MOSnet, avoiding the need of including human perceptual evaluations in the selection process. The paper also presents an algorithm to calculate the sentence alignment of the synthesized audio at the character level in the process. The experiments consider both low and high-quality speech corpora, being the latter used as a reference, for evaluating the proposal on the Tacotron-2 text-to-speech voice-cloning framework. The results suggest that introducing a robust pre-training step when developing voice-cloning based on deep learning can significantly reduce the amount of data needed for training. It is worth mentioning that the authors found that training a voice-cloning model with only 3 h of speech data (even in a different language) from a pre-trained model resulted in similar synthetic quality to training the voice-cloning approach from scratch

with larger datasets, thus, making the training process of this kind of techniques based on deep learning more efficient.

From a different perspective of those approaches considering deep learning techniques, speech can be also generated from articulatory-based models. Recent advances on the production of voice through 3D-based vocal tract geometries using numerical simulations have been able to generate speech signals such as vowels and diphtongs, among other short utterances. For techniques based on the source-filter model, the estimation of both the glottal source and the vocal tract transfer function becomes instrumental, being particularly challenging when trying to improve expressiveness of current approaches.

In the work conducted by Freixes et al. (contribution 7), the authors compare the performance of current state-of-the-art glottal inverse filtering (GIF) techniques, devoted to decompose the glottal source from the vocal tract, on a common reference dataset. The performance of different variants of iterative adaptive inverse filtering (IAIF) and quasi closed phase (QCP) approaches are implemented and evaluated on OPENGLOT's Repository I [6] extended with female vowels. Several standard GIF error metrics are computed to evaluate the obtained glottal flow signals. As a result of the conducted benchmarking on both male and female synthetic vowels with different phonation types and F0s, the authors argue that QCP-based approaches outperform their IAIF-based alternatives for almost all error metrics and evaluated scenarios, besides behaving more stable across sex (male-female), phonation type (from lax to tense: whispery, breathy, normal, and creaky), F0 values (from 100 to 360 Hz with steps of 20 Hz), and vowels. Moreover, the results show that applying GIF on female vowels is more challenging than on male vowels, presenting a general decrease in main errors when moving from tense to lax phonations.

2.4. Affective Computing and Applications

Speech Emotion Recognition (SER) plays a crucial role in applications involving human-machine interaction. These systems still struggle to accurately discern the emotional state of their users, which is a fundamental aspect of human communication. However, the scarcity of suitable emotional speech datasets presents a major challenge for accurate SER systems. Most datasets include artificially simulated emotions and feature a small number of audio samples and speakers. To enhance the overall performance and generalization capabilities of subsequent systems, researchers have employed a technique known as cross-corpus or joint training. This approach involves amalgamating multiple datasets during the training phase. Cross-corpus or joint training has been utilized in certain SER systems as a means to overcome the challenges posed by the aforementioned dataset limitations, for example see [7].

In the paper by Pastor et al. (contribution 9), authors explore the cross-corpus strategy as an extension of the training set and investigates the cross-corpus strategy in greater detail. They explore additional language variability by incorporating a dataset in a different language, and they evaluate the system's performance by gradually increasing the amount of matched data in the training set. Furthermore, another self-supervised (SS) representation, WavLM [8], is assessed, considering its favorable performance in previous studies. The obtained results show that combining databases, even when they encompass different languages, can enhance the system's performance. Moreover, The study presented in this paper reveals that the WavLM representation outperforms other representations in the SER task.

Prosody is one of the essential characteristics of human voice communication, with prosody we can communicate everything from emotions to the purpose of a sentence. Through automatic speech processing, it is possible to analyze prosody and extract prosody characteristics that can be used in a variety of speech-related applications. In particular, by analyzing prosody characteristics, it is possible to identify differences between the voices of typical speakers and individuals with Down syndrome (DS) [9].

A study to obtain prosody features to detect problems in individuals with Down syndrome is carried out in contribution 11. By means of these features it is possible to help therapists to design training therapies adapted to the particular problems of each

user. These prosodic features have been extracted from a speech corpus from individuals with Down syndrome. The corpus was obtained through the use of a tool (a video game) designed to train prosody and pragmatics in individuals with Down syndrome.

2.5. Albayzin Evaluations

Following other similar evaluation initiatives to share and compare recent advances on speech technologies, the Albayzin evaluation campaign was launched in 2006 supported by the RTTH. The Albayzin evaluation series have become a well-known and established framework for the Iberian speech research community, especially for Spanish.

As reported by Lleida et al. (contribution 6), the challenges considered for the 2022 edition were the following ones: speech-to-text transcription, speaker diarization and identity assignment, text and speech alignment, and search on speech. When compared to previous editions, it is to note that the IberSpeech2022 related edition included two main novelties. On the one hand, a text and speech alignment challenge was included in the Albayzin evaluation series. And, on the other, new and more challenging databases from broadcast media content were released for all the evaluation tasks, being some of them such as the RTVE (Radio Televisión Española) and the Basque Parliament databases specifically created for the corresponding challenges. The paper summarizes the main characteristics of the evaluation databases, as well as the evaluation tasks, detailing the metrics considered. Moreover, the paper describes the main features and discusses the results obtained by the approaches of the different research teams that participated in each one of the challenges and sub-challenges, improving the results obtained with respect to those reported in previous challenges.

Author Contributions: Writing—original draft preparation, J.L.P.-C., F.A.-P., Z.C.; writing—review and editing, J.L.P.-C., F.A.-P., Z.C.; visualization, J.L.P.-C., F.A.-P., Z.C.; supervision, J.L.P.-C., F.A.-P., Z.C. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASR	Automatic Speech Recognition
EMG	Electromyography
GIF	Glottal Inverse Filtering
IAIF	Iterative Adaptive Inverse Filtering
ISCA	International Speech Communication Association
LDA	Linear Discriminant Analysis
MOS	Mean Opinion Score
NLP	Natural Language Processing
QCP	Quasi Closed Phase
RTTH	Red Temática en Tecnologías del Habla
RTVE	Radio Televisión Española
SER	Speech Emotion Recognition
SIG-IL	Special Interest Group on Iberian Languages
SS	Self-Supervised
SSI	Silent Speech Interface
VSR	Visual Speech Recognition
WER	Word Error Rate

List of Contributions

1. González, J.; Segarra, E.; García-Granada, F.; Sanchis, E.; Hurtado, L.F. Attentional Extractive Summarization. *Appl. Sci.* **2023**, *13*, 1458. https://doi.org/10.3390/app13031458.
2. Gimeno-Gómez, D.; Martínez-Hinarejos, C.-D. Comparing Speaker Adaptation Methods for Visual Speech Recognition for Continuous Spanish. *Appl. Sci.* **2023**, *13*, 6521. https://doi.org/10.3390/app13116521.

3. Salomons, I.; del Blanco, E.; Navas, E.; Hernáez, I.; de Zuazo, X. Frame-Based Phone Classification Using EMG Signals. *Appl. Sci.* **2023**, *13*, 7746. https://doi.org/10.3390/app13137746.
4. González-Docasal, A.; Álvarez, A. Enhancing Voice Cloning Quality through Data Selection and Alignment-Based Metrics. *Appl. Sci.* **2023**, *13*, 8049. https://doi.org/10.3390/app13148049.
5. Penagarikano, M.; Varona, A.; Bordel, G.; Rodriguez-Fuentes, L.J. Semisupervised Speech Data Extraction from Basque Parliament Sessions and Validation on Fully Bilingual Basque-Spanish ASR. *Appl. Sci.* **2023**, *13*, 8492. https://doi.org/10.3390/app13148492.
6. Lleida, E.; Rodriguez-Fuentes, L.J.; Tejedor, J.; Ortega, A.; Miguel, A.; Bazán, V.; Pérez, C.; de Prada, A.; Penagarikano, M.; Varona, A.; et al. An Overview of the IberSpeech-RTVE 2022 Challenges on Speech Technologies. *Appl. Sci.* **2023**, *13*, 8577. https://doi.org/10.3390/app13158577.
7. Freixes, M.; Joglar-Ongay, L.; Socoró, J.C.; Alías-Pujol, F. Evaluation of Glottal Inverse Filtering Techniques on OPENGLOT Synthetic Male and Female Vowels. *Appl. Sci.* **2023**, *13*, 8775. https://doi.org/10.3390/app13158775.
8. Rodríguez-Cantelar, M.; Estecha-Garitagoitia, M.; D'Haro, L.F.; Matía, F.; Córdoba, R. Automatic Detection of Inconsistencies and Hierarchical Topic Classification for Open-Domain Chatbots. *Appl. Sci.* **2023**, *13*, 9055. https://doi.org/10.3390/app13169055.
9. Pastor, M.A.; Ribas, D.; Ortega, A.; Miguel, A.; Lleida, E. Cross-Corpus Training Strategy for Speech Emotion Recognition Using Self-Supervised Representations. *Appl. Sci.* **2023**, *13*, 9062. https://doi.org/10.3390/app13169062.
10. Gutiérrez-Fandiño, A.; Pérez-Fernández, D.; Armengol-Estapé, J.; Griol, D.; Kharitonova, K.; Callejas, Z. esCorpius-m: A Massive Multilingual Crawling Corpus with a Focus on Spanish. *Appl. Sci.* **2023**, *13*, 12155. https://doi.org/10.3390/app132212155.
11. Corrales-Astorgano, M.; González-Ferreras, C.; Escudero-Mancebo, D.; Cardeñoso-Payo, V. Prosodic Feature Analysis for Automatic Speech Assessment and Individual Report Generation in People with Down Syndrome. *Appl. Sci.* **2024**, *14*, 293. https://doi.org/10.3390/app14010293.

References

1. IberSPEECH 2002 Conference. Available online: http://iberspeech2022.ugr.es/ (accessed on 23 April 2024).
2. Bapna, A.; Arivazhagan, N.; Firat, O. Simple, scalable adaptation for neural machine translation. In Proceedings of the EMNLP-IJCNLP., ACL, Hong Kong, China, 3–7 November 2019; pp. 1538–1548.
3. Gimeno-Gómez, D.; Martínez-Hinarejos, C.D. LIP-RTVE: An Audiovisual Database for Continuous Spanish in the Wild. In Proceedings of the LREC, Marseille, France, 20–25 June 2022; pp. 2750–2758.
4. Hernaez, I.; Gonzalez Lopez, J.A.; Navas, E.; Pérez Córdoba, J.L.; Saratxaga, I.; Olivares, G.; Sanchez de la Fuente, J.; Galdón, A.; Garcia, V.; del Castillo, J.; et al. ReSSInt project: Voice restoration using Silent Speech Interfaces. *Proc. IberSPEECH 2022* **2022**, 226–230. [CrossRef]
5. Gardner-Chloros, P. *Code-Switching*; Cambridge University Press: Cambridge, UK, 2009. [CrossRef]
6. Alku, P.; Murtola, T.; Malinen, J.; Kuortti, J.; Story, B.; Airaksinen, M.; Salmi, M.; Vilkman, E.; Geneid, A. OPENGLOT—An Open Environment for the Evaluation of Glottal Inverse Filtering. *Speech Commun.* **2019**, *107*, 38–47. [CrossRef]
7. Zong, Y.; Zheng, W.; Zhang, T.; Huang, X. Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression. *IEEE Signal Process. Lett.* **2016**, *23*, 585–589. [CrossRef]
8. Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1505–1518. [CrossRef]
9. Corrales-Astorgano, M.; Escudero-Mancebo, D.; González-Ferreras, C. Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome. *Speech Commun.* **2018**, *99*, 90–100. [CrossRef]