

Article

Non-Parallel Articulatory-to-Acoustic Conversion Using Multiview-based Time Warping

Jose A. Gonzalez-Lopez ^{1,*}, Alejandro Gomez-Alanis ¹, José L. Pérez-Córdoba ¹ and Phil D. Green ²

¹ Department of Signal Theory, Telematics and Communications, University of Granada, 18071, Granada, Spain; agomezalanis@ugr.es (A.G.-A.); jlpc@ugr.es (J.L.P.-C.);

² Department of Computer Science, University of Sheffield, S1 4DP, Sheffield, UK; p.green@sheffield.ac.uk

* Correspondence: joseangl@ugr.es

Abstract: In this paper, we propose a novel algorithm called multiview temporal alignment by dependence maximisation in the latent space (TRANSIENGE) for the alignment of time series consisting of sequences of feature vectors with different length and dimensionality of the feature vectors. The proposed algorithm, which is based on the theory of multiview learning, can be seen as an extension of the well-known dynamic time warping (DTW) algorithm but, as mentioned, it allows the sequences to have different dimensionalities. Our algorithm attempts to find an optimal temporal alignment between pairs of nonaligned sequences by first projecting their feature vectors into a common latent space where both views are maximally similar. To do this, powerful, nonlinear deep neural network (DNN) models are employed. Then, the resulting sequences of embedding vectors are aligned using DTW. Finally, the alignment paths obtained in the previous step are applied to the original sequences to align them. In the paper, we explore several variants of the algorithm that mainly differ in the way the DNNs are trained. We evaluated the proposed algorithm on a articulatory-to-acoustic (A2A) synthesis task involving the generation of audible speech from motion data captured from the lips and tongue of healthy speakers using a technique known as permanent magnet articulography (PMA). In this task, our algorithm is applied during the training stage to align pairs of nonaligned speech and PMA recordings that are later used to train DNNs able to synthesis speech from PMA data. Our results show the quality of speech generated in the nonaligned scenario is comparable to that obtained in the parallel scenario.

Keywords: deep learning; multiview learning; dynamic time warping; canonical correlation analysis; silent speech interface; latent embedding



Citation: Gonzalez-Lopez, J.A.; Gomez-Alanis, A.; Pérez-Córdoba, J.L.; Green, P.D. Non-Parallel Speech Synthesis from Articulatory Data using Multiview-based Time Warping. *Appl. Sci.* **2022**, *1*, 0. <https://doi.org/>

Academic Editor: Francesc Alías, Valentin Cardeñoso-Payo, David Escudero-Mancebo, César González-Ferreras and António Joaquim da Silva Teixeira.

Received:

Accepted:

Published:

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Silent speech interfaces (SSIs) are devices that enable speech communication in the absence of audible speech by decoding speech from other nonacoustic, speech-related biosignals generated by the human body during the process of speech production [1–3]. These biosignals can range from the neural activity in the speech and language areas of the brain [4–6], electrical activity driving the facial muscles captured by surface electrodes (i.e., electromyography (EMG)) [7–9], or motion capture of the speech articulators by means of imaging techniques [10] or electromagnetic articulography techniques [11–14].

SSIs have many potential applications, as summarized in [1]. For instance, an SSI could be used to enhance speech communication in noisy environments as the above-mentioned biosignals are mostly immune to acoustic noise. Also, SSIs might be used to enable covert communication to preserve privacy when speaking in public spaces. Yet another potential application, the one that motivates this work, is to restore oral communication to persons with speech impairments [3]. Thus, an SSI could be used to decode the words its user wants to say from the captured biosignals. Because SSIs do not rely on the acoustic speech signal, they offer a radically new form of restoring oral communication to people with speech impairments.

To decode speech from the biosignals, two alternative SSI approaches were proposed [3]: silent speech recognition, which involves the use of automatic speech recognition (ASR) algorithms to transform the biosignals into text, followed by text-to-speech synthesis and direct speech synthesis, which transforms the biosignals into a set of acoustic parameters amenable to speech synthesis. In this work, we focus on the latter approach, which is also known as articulatory-to-acoustic (A2A) synthesis when the biosignals encode information about the movements of the speech organs.

The most successful direct synthesis techniques so far adopted a data-driven framework, in which supervised machine learning is used to model the mapping $y = h(x)$ between source feature vectors x extracted from the biosignals and target feature vectors y computed from the speech signals, as shown in Figure 1. To train this function, a dataset $\mathcal{D} = \{(X_1, Y_1), \dots, (X_M, Y_M)\}$ with pairs of sequences of feature vectors (X_i, Y_i) extracted from time-synchronous recordings is used, where $X_i \in \mathbb{R}^{d_x \times T_i}$ and $Y_i \in \mathbb{R}^{d_y \times T_i}$ (M is the number of parallel sequences in the dataset, d_x and d_y are the dimensionality of the source and target vector, respectively, and T_i is the length of the i -th sequence pair). The need for parallel recordings, however, limits the application of direct synthesis techniques to only a few clinical scenarios, as described in [3]. For instance, people who already lost their voices could not use this technology because of the impossibility of recording parallel data to train the machine learning model, as shown in Figure 1.

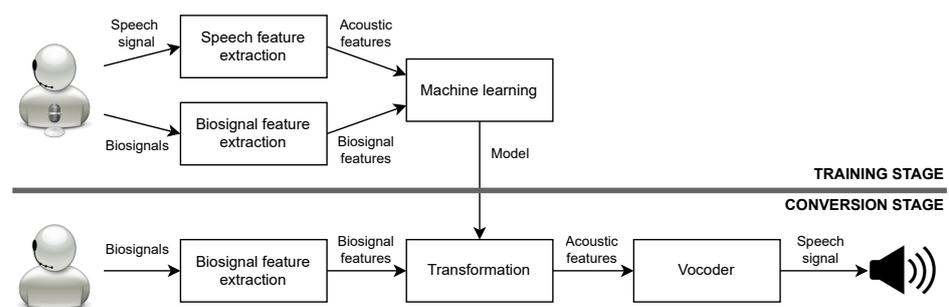


Figure 1. Block diagram of a typical (data-driven) direct synthesis technique. In training stage, a machine learning model is estimated, which represents relationship between feature vectors extracted from speech and nonspeech signals. This model is latter used in conversion stage to predict speech (acoustic) parameters from biosignals.

A solution to this problem, which we explore in this work, involves the use of previously available speech recordings from a voice donor (e.g., a relative or recordings of the patient's own voice made before the voice loss). Using these recordings, it would be possible (in principle) to obtain the necessary parallel data by asking the patient to repeat the speech recordings in silence while the necessary biosignals are captured, which is similar to karaoke. However, even in this case, it is likely that the captured biosignals will not be perfectly aligned with the speech recordings, since both modalities would have slightly different duration, thus preventing the application of standard, supervised machine learning techniques. Therefore, our problem now becomes concerned with aligning sequences obtained by sampling the same physical process with different sensors (e.g., speech and biosignal data). Once the sequences are aligned, standard machine learning techniques could be readily applied to model the relationship between the speech and biosignal data.

The problem stated above is similar to what happens in voice conversion (VC) [15–17], where the goal is to modify the spectral content (and possibly the prosody, too) of the speech of a source speaker in a way that it sounds as it was said by a different speaker (target speaker). To perform this transformation, a machine learning model (e.g., a deep neural network (DNN)) is trained with pairs of speech signals recorded from the source and target speakers while reading aloud a set of sentences. Again, a time-alignment procedure is

necessary to align the source and target speakers' signals because they might have different duration. To this end, the well-known DTW algorithm [18] is normally used.

The problem we focus on this work is, in a way, similar to the one described above for VC, that is, aligning feature sequences with the same phonetic content but possibly different duration. However, in contrast to VC, the direct application of DTW might not be possible in our case because the feature vectors extracted from the speech and bio- signals may have different dimensionality (i.e., $d_x \neq d_y$, in general), while the standard DTW algorithm requires that $d_x = d_y$. To address this issue, in this work we propose an extension of the DTW algorithm called multiview temporal alignment by dependence maximisation in the latent space (TRANSCIENCE) (Code is available at <https://github.com/joseangl/transience>; Accessed: 22-January-2022), which is based on the theory of multiview learning [19,20]. TRANSCIENCE attempts to find the optimal temporal alignment between sequences from different views (e.g., speech and biosignal data) by first projecting the feature vectors from the sequences into a common, latent subspace where the resulting embeddings have the same dimensionality and are maximally similar (we will define later what we mean by similarity). After that, the sequences of embedding vectors obtained for the different views are aligned by means of the DTW algorithm.

The remainder of this paper is organized as follows. First, in Section 2, the relevant related work is reviewed. The details of the proposed temporal alignment technique are presented in Section 3. Section 4 describes the experimental setup employed to evaluate the performance of the proposed technique. In particular, the proposed algorithm is evaluated on an A2A task involving the conversion of articulatory data captured using permanent magnet articulography (PMA) [21,22] to speech. Experimental results obtained on this task are shown in Section 5. We conclude in Section 6 and discuss potential future research directions.

2. Related Work

In the context of temporal alignment between sequences, there are several studies that have addressed this issue. Moreover, there are a number of real-world problems where these algorithms found application, e.g., image and video alignment, physiological signals (electroencephalogram (EEG) and EMG signals), video classification, etc. In the following paragraphs, we provide a review of previous works on the topic of sequence temporal alignment.

The most closely related work to our method is that of Trigeorgis et al. [23], in which an algorithm called deep canonical time warping (DCTW), which combines canonical correlation analysis (CCA) [24] with DTW, is proposed. The key differences in our approach with respect to DCTW can be summarized as follows. First, we evaluate different similarity metrics, not only CCA, to optimize the parameters of the DNNs used to map the feature vectors extracted from the multiple views into their common, latent subspace. Also, we introduce an autoencoder-based loss function, which helps to regularize the training and avoids naive solutions. Finally, inspired by the work in [25], we also propose the introduction of private latent variables for each view which aim at modeling the specific peculiarities within each view. Our technique also shares similarities with the generalized canonical time warping (GCTW) technique described in [26]. However, in contrast to this technique, TRANSCIENCE solves the optimal alignment problem with DTW rather than approximating the temporal warping with a set of predefined monotonic bases and optimizing the weights of these bases with a Gauss–Newton algorithm. Also, being based on CCA, GCTW computes the latent variables by applying a linear transformation to the feature vectors from the different views, while our method uses powerful autoencoders to nonlinearly transform the data, which could be expected to yield better alignments. Another technique similar to our proposal is the one presented in [27], which proposes an extension of CCA, an alignment-agnostic CCA (AA-CCA). AA-CCA models the uncertainty of alignments using a new data correlation term, and this allows to use not only decently aligned data (if available) when learning CCA, but also the unaligned ones. It is based on

the optimization of a constrained objective function that combines two terms: a correlation criterion and a context-based regularizer.

Contrary to now popular sequence-to-sequence (seq2seq) models (e.g., [28,29]), our method has different advantages. First, our aim is to align sequences of different lengths to obtain the necessary parallel data to train a machine learning technique, not to model a full-fledged mapping between sequences. In other words, our technique is only used to align the training data, while in test time the articulatory data is directly mapped to speech. Also, it is relatively straightforward to modify our method to process multiple views (more than 2), while the adaptation of seq2seq models is more involved.

3. multiview Temporal Alignment

The problem we address in this work can be formulated as finding the optimal alignment between two time series $\mathbf{X} = (x_1, \dots, x_{T_x})$ and $\mathbf{Y} = (y_1, \dots, y_{T_y})$, where the feature vectors x_i and y_j may have possibly different dimensionality. In our case, $\mathbf{X} \in \mathbb{R}^{d_x \times T_x}$ is a sequence of feature vectors extracted from the captured biosignal (whatever type the SSI is using) and $\mathbf{Y} \in \mathbb{R}^{d_y \times T_y}$ is the sequence of acoustic speech parameters. We further assume that both sequences encode the same phonetic content (i.e., same words in the same order) but have, possibly, different duration. Mathematically, this involves solving the following minimization problem,

$$\operatorname{argmin}_{\phi^x, \phi^y} \sum_{t=1}^T d(x_{\phi_t^x}, y_{\phi_t^y}), \quad (1)$$

where $T = \max(T_x, T_y)$, $d(x, y)$ is a distance function, $\phi^x \in \{1 : T_x\}^T$ and $\phi^y \in \{1 : T_y\}^T$ are warping functions that map the indices of the original time series to a common time axis where both series are aligned. This way, x_i and y_j will be aligned if $\phi_t^x = i$ and $\phi_t^y = j$ for a given t . The goal of the alignment algorithm is to find the warping paths $\langle \phi^x, \phi^y \rangle$ that minimize the total sum of distances among all possible alignments between both time series.

To enable the alignment of sequences with different dimensionality, we assume that there exists a pair of transformation functions $f : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_z}$ and $g : \mathbb{R}^{d_y} \rightarrow \mathbb{R}^{d_z}$, modeled as DNNs in this work, that project the original feature vectors into a common, latent space where the data from both views are maximally similar. Thus, the problem in (1) now becomes,

$$\operatorname{argmin}_{\phi^x, \phi^y} \sum_{t=1}^T d(f(x_{\phi_t^x}), g(y_{\phi_t^y})). \quad (2)$$

For some fixed mapping functions $z^x = f(x; \Theta_f)$ and $z^y = g(y; \Theta_g)$ (Θ_f and Θ_g denote the set of parameters for these functions), the problem of temporal alignment of the latent variable sequences $\mathbf{Z}^x = (z_1^x, \dots, z_{T_x}^x)$ and $\mathbf{Z}^y = (z_1^y, \dots, z_{T_y}^y)$ can be solved efficiently by means of the DTW algorithm. Conversely, for fixed warping paths ϕ^x and ϕ^y , the problem in (2) involves optimizing the functions $f(\cdot)$ and $g(\cdot)$ to minimize $\sum_{t=1}^T d(f(x_{\phi_t^x}), g(y_{\phi_t^y}))$. If the mapping functions are modeled as DNNs, as in our case, this latter problem can be solved by back-propagation. Thus, TRANSIENCE algorithm solves (2) by alternating between two phases: (i) finding the optimum set of DNN weights (Θ_f, Θ_g) by fixing the warping paths, and (ii) applying the DTW algorithm to compute the optimum alignments $\langle \phi^x, \phi^y \rangle$ between the sequences of latent vectors by freezing the DNN weights. The warping paths are initialized by uniformly aligning the sequences, i.e., $\phi_t^x = 1 + \left\lceil \frac{t-1}{T-1} (T_x - 1) \right\rceil$ and $\phi_t^y = 1 + \left\lceil \frac{t-1}{T-1} (T_y - 1) \right\rceil$ for $t = 1, \dots, T$ and $T = \max(T_x, T_y)$.

A block diagram of the proposed algorithm for temporal alignment of time series is shown in Figure 2. First, the feature vectors of both time series (denoted as $x_i; i = 1, \dots, T_x$ and $y_j; j = 1, \dots, T_y$ in the figure) are projected to a common latent space where the resulting embedding vectors (z_i^x and z_j^y) are maximally similar. To this end, two DNNs are

employed, one for each sequence. Next, both sequences of embedding vectors are aligned using the DTW algorithm, and the resulting warping paths $\langle \phi^x, \phi^y \rangle$ are employed to align the original sequences $X = (x_1, \dots, x_{T_x})$ and $Y = (y_1, \dots, y_{T_y})$. During the training stage, the DNN weights are optimized using a combination of different loss functions, as shown in Figure 2. These losses are described in detail in the following subsections.

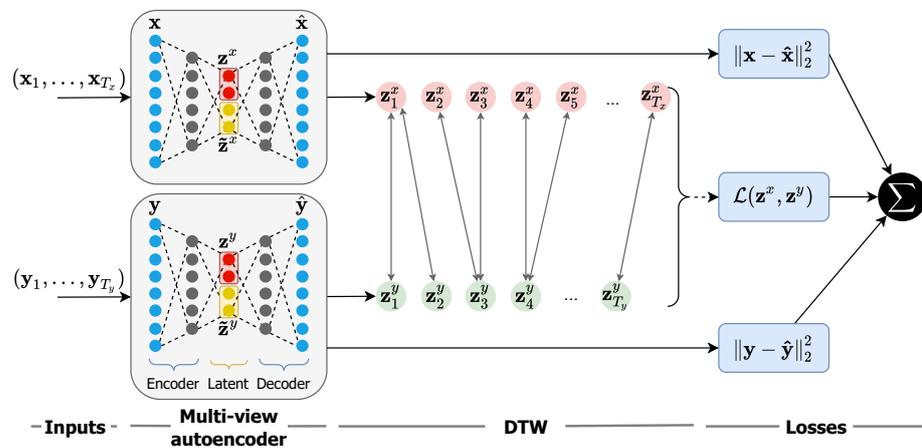


Figure 2. Block diagram of proposed TRANSIANCE algorithm for time series alignment. First, feature vectors of two sequences to be aligned are projected into a common, latent space (represented with red dots in figure) using DNNs trained for each view. Then, DTW algorithm is used to align resulting sequences of embedding vectors and, hence, original sequences. To train parameters of DNNs, different loss functions are employed (see text for more details).

3.1. Latent-Space Similarity Metrics

In this work, we evaluated three alternative loss functions $\mathcal{L}(z^x, z^y)$ used to optimize the weights of the DNNs shown in Figure 2 during the training stage. While all three losses attempt to optimize the similarity between pairs of embedding vectors computed by these DNNs, they differ in the specific metric used to compute this similarity. In particular, we propose to optimize the correlation, mutual information, and (minimize) a contrastive loss, respectively, between these vectors. These similarity metrics are described in the following sections.

3.1.1. Canonical Correlation Analysis

The first loss function we propose attempts to maximize the statistical correlation between pairs of aligned embedding vectors. Given a minibatch of N pairs of *aligned* observations $\mathcal{B} = \{(x_1, y_1), \dots, (x_N, y_N)\}$, the CCA loss function maximizes the correlation between the outputs of the DNNs, $f(x)$ and $g(y)$, as follows,

$$(\Theta_f^*, \Theta_g^*) = \operatorname{argmax}_{\Theta_f, \Theta_g} \sum_{i=1}^N \operatorname{corr}(f(x_i; \Theta_f), g(y_i; \Theta_g)). \tag{3}$$

As detailed in [19], this equals to maximizing the following loss function,

$$\mathcal{L}_{cca} = \sqrt{\operatorname{tr}(T^T T)}, \tag{4}$$

where $\operatorname{tr}(\cdot)$ is the trace operator and $T = \Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1/2}$. The covariance matrices $\Sigma_{xx} = \operatorname{cov}(f(x), f(x))$, $\Sigma_{xy} = \operatorname{cov}(f(x), g(y))$ and $\Sigma_{yy} = \operatorname{cov}(g(y), g(y))$ are estimated from the outputs of the DNNs. The DCTW algorithm described in [23] is a specific case of our TRANSIANCE algorithm when the CCA loss in (4) is used.

3.1.2. Maximum Mutual Information

As an alternative, we also consider maximizing the mutual information between the pairs of embedding vectors computed by the DNNs in Figure 2. This results in the following maximum mutual information (MMI) loss function,

$$\mathcal{L}_{mmi} = \sum_{i=1}^N p(\mathbf{f}(x_i), \mathbf{g}(y_i)) \log \frac{p(\mathbf{f}(x_i), \mathbf{g}(y_i))}{p(\mathbf{f}(x_i))p(\mathbf{g}(y_i))}. \quad (5)$$

The probability density functions (pdfs) in (5) are estimated using kernel density estimation (KDE) [30] as follows,

$$p(z_i) = \frac{1}{N-1} \sum_{j=1, j \neq i}^N K(z_i - z_j), \quad (6)$$

where an isotropic Gaussian kernel with trainable bandwidth σ_z is used in this work, i.e., $K(z) = \mathcal{N}(z; \mathbf{0}, \sigma_z \mathbf{I})$. Thus, three pdfs are estimated, the joint distribution $p(\mathbf{f}(x), \mathbf{g}(y))$ and the marginals $p(\mathbf{f}(x))$ and $p(\mathbf{g}(y))$, each one with its own trainable bandwidth.

3.1.3. Contrastive Loss

Finally, we also evaluate the contrastive loss (CL) function described in [25,31], which given a fixed latent variable from the first view $\mathbf{f}(x^+)$, takes an aligned positive example $\mathbf{g}(y^+)$ and an unaligned negative example $\mathbf{g}(y^-)$ from the second view and attempts to minimize the difference between the distances for the positive and negative examples:

$$\mathcal{L}_{cl} = \frac{1}{N} \sum_{i=1}^N \max(0, m + d(\mathbf{f}(x_i^+), \mathbf{g}(y_i^+)) - d(\mathbf{f}(x_i^+), \mathbf{g}(y_i^-))), \quad (7)$$

where m is a margin hyperparameter ($m = 0.5$ is used in this work) and $d(z^x, z^y) = 1 - \frac{z^x \cdot z^y}{\|z^x\| \|z^y\|}$ is the cosine similarity. The negative examples $\mathbf{g}(y^-)$ are generated by shuffling the outputs of the DNN for the second view before the loss is computed. Intuitively, the distances $d(\mathbf{f}(x^+), \mathbf{g}(y^+))$ in (7) should be small if both views are projected to similar (closer) representations in the common latent space, whereas the distances to the negative, unpaired examples $d(\mathbf{f}(x^+), \mathbf{g}(y^-))$ should be bigger because they are projected to different locations of that space.

3.2. multiview Autoencoder

to regularize the training of the DNNs and to avoid naive solutions (e.g., $\mathbf{f}(x) = \mathbf{g}(y) = c$, for all x and y , with c being a constant vector), we propose the introduction of an auxiliary, autoencoder-based reconstruction loss. As shown in Figure 2, this loss function minimizes the mean squared error (MSE) between the DNNs' inputs (x, y) and the reconstructed outputs ($\hat{x} = f^{-1}(f(x))$, $\hat{y} = g^{-1}(g(y))$), being f^{-1} and g^{-1} decoder networks that estimate the input feature vectors x and y from their latent projections z^x and z^y , respectively, as follows:

$$\mathcal{L}_{autoencoder} = \frac{1}{N} \left(\sum_{i=1}^N \left\| x_i - f^{-1}(f(x_i)) \right\|^2 + \sum_{i=1}^n \left\| y_i - g^{-1}(g(y_i)) \right\|^2 \right). \quad (8)$$

The weights of the DNNs in Figure 2 are now optimized by minimizing a weighted combination of the autoencoder-based loss in (8) and one of the similarity losses \mathcal{L}_{sim} described in Section 3.1, as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{sim} + \lambda \mathcal{L}_{autoencoder}, \quad (9)$$

where the hyperparameter λ is set to 1 in this work for simplicity.

3.3. Private Latent Variables

A key assumption of our temporal alignment algorithm TRANSIENCE is that the feature vectors of the sequence pair to be aligned share some common information. Although this is indeed our case, because the speech and bio-signals encode the same phone sequence, it is also true that each view may have its own unique characteristics, thus making the reconstruction loss in (8) difficult to optimize when only considering the shared latent variables. Therefore, it may be beneficial to model the unique characteristics of each view as well as the common characteristics shared among all the views. Inspired by the work in [25], we propose the introduction of private latent variables for each view \tilde{z}^x and \tilde{z}^y that aim at modeling the uniqueness of each view, as represented in Figure 2. The private variables are predicted from the inputs by a set of independent DNNs $\tilde{z}^x = \tilde{f}(x)$ and $\tilde{z}^y = \tilde{g}(y)$. These private latent variables are used in (8), in addition to the common, shared variables, for reconstructing the input data, i.e., $\hat{x} = f^{-1}(f(x), \tilde{f}(x))$ and $\hat{y} = g^{-1}(g(y), \tilde{g}(y))$.

When optimizing the weights of DNNs \tilde{f} and \tilde{g} , a standard Gaussian distribution $\tilde{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is chosen as the prior distribution for the private variables of each view. To enforce this distribution, we minimize the Kullback-Leibler (KL) divergence between the priors and the empirical distribution (modeled as a multivariate Gaussian distribution with diagonal covariance) estimated from the private variables as follows,

$$\mathcal{L}_{KL} = \frac{1}{2} \sum_{i=1}^{d_{\tilde{z}}} (\sigma_i^2 + \mu_i^2 - 1 - \log \sigma_i^2), \quad (10)$$

where the means and variances in (10) (μ_i, σ_i , respectively) are estimated for each private variable in each minibatch.

4. Experimental Setup

In this section, the experimental setup employed to evaluate the performance of the proposed technique is described. In particular, our algorithm is evaluated on an A2A task involving the conversion of articulatory data captured from the lips and tongue of healthy speakers to speech. More details about this can be found in the following subsections.

4.1. Dataset

The proposed alignment algorithm was evaluated on an A2A task involving the synthesis of speech from articulatory data captured using PMA. PMA is a technique for capturing the movements of the vocal apparatus during speech. The technique is based on sensing the changes in the magnetic field generated by the movement of small magnets attached to the speech articulators (lips and tongue, in our case) as the speaker 'mouths' words. In the current set-up, a total of six magnets were employed: four on the lips with dimensions of 1 mm (diameter) \times 5 mm (height), one on the tongue tip (2 mm \times 4 mm), and one on the middle of the tongue (5 mm \times 1 mm). The three spatial components (x, y, z) of the magnetic field generated by the magnets were then acquired by four triaxial magnetic sensors mounted on a rigid frame. Only three of the sensors were used for capturing articulatory movements, while the remaining sensor was used to compensate the effects of the earth's magnetic field on the captured articulatory data. More information about this technique can be found in [12,22].

Parallel data was recorded by four nonimpaired British subjects (2 males and 2 females) while reading aloud a subset of the CMU Arctic corpus [32]. Two alignment conditions

were evaluated: (i) intrasubject alignment, where PMA and speech signals recorded by the same subject in different sessions are aligned, and (ii) crosssubject alignment, where PMA signals recorded by a given subject are aligned with speech recorded by a different subject (with a possibly different gender as well). We also attempted to align PMA signals recorded from a female laryngectomy patient with a set of speech recordings made by her before losing the voice. The details of the dataset used for our experiments is summarized in Table 1.

Table 1. Details of dataset used for experiments.

Condition		# of Sentences	
		Training set	Evaluation set
Intra-subject	F → F	103 (11.3 min)	20 (1.2 min)
	M → M	134 (8.4 min)	20 (1.1 min)
Cross-subject	F → F	99 (6.0 min)	18 (0.9 min)
	F → M	332 (18.7 min)	20 (1.0 min)
	M → F	332 (21.9 min)	20 (1.2 min)
	M → M	414 (27.9 min)	20 (1.2 min)

Feature Extraction

The PMA and speech signals were parameterized as a series of feature vectors computed every 5 ms from 25 ms analysis windows. The speech signals were first downsampled from 48 kHz to 16 kHz and then converted to sequences of 28-dimension vectors using the WORLD vocoder [33]: 25 mel-generalised cepstral coefficients (MGCCs), 1 band aperiodicity (BAP) value, 1 continuous F_0 value in logarithmic scale, and 1 U/V decision. $\log F_0$ values in unvoiced frames were linearly interpolated from adjacent voiced frames. The 9-channel, background-canceled PMA signals were first oversampled from 100 Hz to 200 Hz to match the 5 ms frame rate. Then, segmented features were computed from the raw PMA signals by applying principal component analysis (PCA) over contextual windows with 11 frames. That is, for each frame of 25 ms, the preceding and succeeding 5 frames, along with the current one, were concatenated and PCA was applied for dimensionality reduction by retaining the 99% of the variance. Finally, the PMA and speech features were normalized to have zero mean and unit variance.

4.2. Implementation Details

Each DNN in TRANSIENGE was modeled as a 3-layer feed-forward neural network with $200 \times 100 \times 100$ hidden units and leaky rectified linear unit (LReLU) activations ($a = 0.03$) following [23]. The neural networks were trained as denoising autoencoders ($\sigma_{noise} = 0.5$) using the Adam algorithm [34] with a fixed learning rate of $\alpha = 1 \times 10^{-4}$ and a batch size of $N = 512$ samples. The dimensionality of the shared latent variables was set to $d_z = 20$ and fixed $d_{\hat{z}} = 10$ for the private variables. Finally, we used the cosine distance for the DTW algorithm in (2). For temporal alignment, only the MGCCs were used (augmented with delta and acceleration parameters).

4.3. PMA-to-Speech System

The aligned signals were used to train speaker-dependent A2A systems. We used the same setting as in our previous work [35]: DNNs with four hidden layers with 400 units in each layer, and rectified linear unit (ReLU) activations were used. The maximum-likelihood parameter generation (MLPG) algorithm [36,37] was applied over the DNN outputs to enhance the acoustic quality of the resynthesized waveforms.

4.4. Performance Evaluation

Because the ultimate goal of our work is the synthesis of speech signals from biosignals captured from the human body during speech production, we evaluated our alignment

technique by measuring the quality of the speech predicted from the PMA data. To this end, the following procedure was adopted. First, TRANSIENCE was applied to align the PMA and speech signals in the training dataset for each condition in Table 1. The aligned signals were then used to train PMA-to-speech systems, modeled as DNNs as described above. Finally, speech signals were synthesized from the PMA signals in the evaluation set in Table 1 using the DNNs trained in the previous step.

The speech signals resulting from this last step were used as the basis for our performance evaluation. In particular, we conducted two kinds of evaluation to assess the performance of our temporal alignment algorithm: objective and subjective. In the objective evaluation, we evaluated the quality of the resynthesized speech signals obtained from the test PMA signals by comparing them with that of the original speech recordings made by the subjects in our database. For this task, we used several objective metrics widely used in speech synthesis. Accuracy of spectral estimation was objectively evaluated using the mel-cepstral distortion (MCD) metric [38] between the MGCCs from the speech signals recorded by the subjects and those estimated from PMA data. For the excitation parameters, we computed the root mean squared error (RMSE) for measuring the estimation accuracy of the BAP parameter, the error rate for the voicing parameter, and the RMSE between the estimated and original $\log F_0$ contours in the voiced segments. For subjective evaluation, we conducted a set of listening tests whose details are provided below.

5. Results and Discussion

In this section, we show the performance results obtained by our technique on the A2A task described above.

5.1. Objective Evaluation

First, we evaluated the quality of speech predicted from the PMA signals in the evaluation set using the objective metrics described in Section 4.4. For TRANSIENCE, three variants were evaluated depending on the latent-space similarity loss function employed: \mathcal{L}_{cca} , \mathcal{L}_{mmi} , and \mathcal{L}_{cl} loss functions described in Section 3.1. Furthermore, for each of the three variants, we also evaluated the effect of the autoencoder-based loss described in Section 3.2 and the introduction of the private variables described in Section 3.3. For comparison purposes, we also evaluated the canonical time warping (CTW) technique described in [39], which is a particular case of TRANSIENCE combining standard (linear) CCA with DTW, thus not being able to model nonlinear latent mappings. We also provide the results obtained by an oracle system, in which the PMA and acoustic signals in the training dataset are aligned by using the *ideal* warping paths computed by applying DTW over the speech signals recorded by the same subjects. These signals are available for the healthy subjects in our database because the original recordings contained both (time-synchronous) speech and PMA signals.

Table 2 shows the objective results for the different systems. TRANSIENCE using the CL loss yields the best objective results, outperforming the rest of the similarity metrics and, even, the oracle system. In particular, relative gains of 18.36% and 21.46% are achieved in the MCD metric w.r.t. using the CCA and MMI losses. Unfortunately, it seems that the introduction of the autoencoder-based loss and the private latent variables do not improve the results, which may be due to the dimensionality of the private variables not being enough to capture the peculiarities of each view. It is also surprising that the simple (linear) CTW technique outperforms its nonlinear version TRANSIENCE-CCA. In future work, we should look at this issue.

Table 2. Summary of the objective results obtained by each alignment technique on an A2A task where speech was synthesized from PMA data captured from the lips and tongue using DNNs. The following metrics were computed between the predicted speech signals and the original speech signals recorded by the speakers: (a) speech spectral envelope distortion (MCD); (b) aperiodic component distortion (RMSE); (c) speech fundamental frequency (F_0) distortion (RMSE); and (d) voicing error rate (%). Best result for each set of parameters is highlighted in bold face.

Method	MGCC MCD (dB)	BAP RMSE (dB)	F_0 RMSE (Hz)	Voicing Err. Rate (%)
Oracle	7.81	0.43	14.75	23.79
CTW	8.55	0.59	15.98	23.08
+autoenc.	9.20	0.88	16.70	25.30
+priv. vars.	8.83	0.58	15.74	21.47
TRANSIENCE-CCA	9.37	0.85	16.40	27.95
+autoenc.	10.02	1.46	15.95	34.08
+priv. vars.	10.48	1.24	15.79	31.88
TRANSIENCE-MMI	9.74	0.69	16.43	22.25
+autoenc.	9.97	1.09	16.92	23.72
+priv. vars.	9.86	1.70	16.41	21.75
TRANSIENCE-CL	7.65	0.12	15.28	24.10
+autoenc.	7.76	0.20	14.98	24.06
+priv. vars.	7.82	0.30	14.58	23.68

5.2. Subjective Evaluation

We also conducted an ABX test to subjectively evaluate the quality of the resynthesized speech signals. To implement the test, we used the web-based BeagleJS framework [40]. Twenty-seven listeners (recruited among students enrolled in our courses) participated in the test, who had to judge which of two versions of the same signal produced by any combination of two of the five systems in Table 2 was more similar to a reference (one of the signals recorded by the subjects). Each listener evaluated 10 sample pairs for each of the 10 possible system combinations (i.e., 100 pairs evaluated in total by each listener). For this task, only the “basic” systems in Table 2 were evaluated (i.e., without the use of the autoencoder-loss nor the private variables), because this setting produced the best overall objective results. The test was conducted in a quiet room while listeners wore good-quality headphones. Listeners were instructed to set the volume at a comfortable level at the start of the test. The total duration of the test was approximately 20 min.

Figure 3 shows the results of the listening test. The most preferred system by a large margin was our algorithm TRANSIENCE using the CL function, being this system on par with the Oracle system. Similar to the triplet loss used in applications such as face or voice biometrics [41,42], we speculate that the CL loss function allows for a more precise way to measure the similarity between pairs of latent vectors sampled from different views and, hence, to more accurately train the DNNs used by our algorithm.

Interestingly, the CTW system obtained higher preference scores than its nonlinear version (TRANSIENCE+CCA) and that of the MMI system. It may be because the optimization process became stuck in poor local-minima for the latter systems. However, more research is needed to shed some light into this problem.

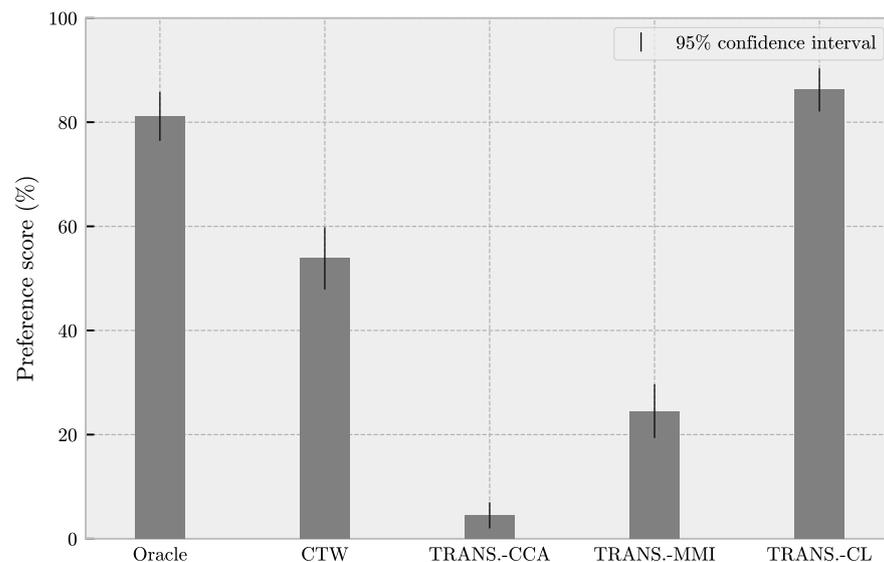


Figure 3. Results of the ABX listening test on speech quality.

6. Conclusions

In this paper we proposed a new procedure for the alignment of times series. The proposed method, called TRANSIENGE, can be seen as an extension to the well-known DTW algorithm, but allowing the alignment of pairs of sequences consisting of feature vectors with different dimensionality. To enable this feature, powerful DNNs are used to non-linearly map the feature vectors from the sequences into a common, latent space where the resulting embedding vectors have the same dimensionality and are maximally similar. Once these embedding vectors are obtained, the DTW algorithm is used to compute the optimum alignment between them. We proposed different versions of this algorithm, mainly affecting the way the DNNs are trained.

To evaluate the proposed technique, an A2A task involving the synthesis of audible speech from articulatory signals was employed. To this end, articulatory and speech data recorded from healthy speakers for multiple sessions were used. Our algorithm was then used to align the articulatory and speech signals and, once aligned, the aligned signals were employed for training DNNs able to estimate the speech parameters from the articulatory signals; thus, enabling people with speech impairments to communicate again. Our results showed that it is feasible to deploy direct synthesis techniques in nonparallel scenarios. In particular, both the objective and subjective evaluation conducted showed us that the quality of speech signals obtained by our algorithm approaches (when not outperforming) the results achieved when using an oracle alignment technique.

Regarding future work, we would like to evaluate our technique using different types of biosignals (e.g., EEG) and data obtained from clinical population. Furthermore, even better alignment could be achieved by introducing additional constraints (e.g., phonetic constraints) or, alternatively, using state-of-the-art neural network architectures, such as the ResNET [43] or DenseNET [44] architectures.

Author Contributions: The individual contributions are provided as follows. Contributions: conceptualization, J.A.G.-L., A.G.-A., J.L.P.-C. and P.D.G.; methodology, J.A.G.-L., A.G.-A., J.L.P.-C. and P.D.G.; software, J.A.G.-L., A.G.-A. and J.L.P.-C.; validation, J.A.G.-L. and A.G.-A.; formal analysis, J.A.G.-L., A.G.-A. and P.D.G.; investigation, J.A.G.-L., A.G.-A., J.L.P.-C. and P.D.G.; resources, J.A.G.-L., J.L.P.-C. and P.D.G.; data curation, J.A.G.-L., A.G.-A. and J.L.P.-C.; writing—original draft preparation, J.A.G.-L.; writing—review and editing, J.A.G.-L., A.G.-A., J.L.P.-C. and P.D.G.; visualization, J.A.G.-L., A.G.-A. and J.L.P.-C.; supervision, J.A.G.-L.; project administration, J.A.G.-L. and J.L.P.-C.; funding acquisition, J.A.G.-L., and J.L.P.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Spanish State Research Agency (SRA) grant number PID2019-108040RB-C22/SRA/10.13039/501100011033, and the FEDER/Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades project no. B-SEJ-570-UGR20.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the UK National Research Ethics Service.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to ethical reasons.

Acknowledgments: We thank Prof. James M. Gilbert and Dr. Lam A. Cheah from the University of Hull (UK) for their help during data collection.

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

Abbreviations

The following abbreviations are used in this manuscript:

A2A	Articulatory-to-Acoustic
AA-CCA	Alignment-Agnostic CCA
ASR	Automatic Speech Recognition
BAP	Band APeriodicity
CCA	Canonical Correlation Analysis
CL	Contrastive Loss
CTW	Canonical Time Warping
DCTW	Deep Canonical Time Warping
DNN	Deep Neural Network
DTW	Dynamic Time Warping
EEG	Electroencephalogram
EMG	Electromyography
GCTW	Generalized Canonical Time Warping
KDE	Kernel Density Estimation
KL	Kullback-Leibler
LReLU	Leaky Rectified Linear Unit
MCD	Mel-Cepstral Distortion
MGCC	Mel-Generalised Cepstral Coefficient
MLPG	Maximum-Likelihood Parameter Generation
MMI	Maximum Mutual Information
MSE	Mean Squared Error
PCA	Principal Component Analysis
pdf	Probability Density Function
PMA	Permanent Magnet Articulography
ReLU	Rectified Linear Unit
RMSE	Root Mean Squared Error
seq2seq	Sequence-to-Sequence
SSI	Silent Speech Interface
TRANSIENCE	multiview Temporal Alignment by Dependence Maximisation in the Latent Space

References

1. Denby, B.; Schultz, T.; Honda, K.; Hueber, T.; Gilbert, J.M.; Brumberg, J.S. Silent speech interfaces. *Speech Commun.* **2010**, *52*, 270–287. <https://doi.org/10.1016/j.specom.2009.08.002>.
2. Schultz, T.; Wand, M.; Hueber, T.; Krusienski, D.J.; Herff, C.; Brumberg, J.S. Biosignal-Based Spoken Communication: A Survey. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 2257–2271. <https://doi.org/10.1109/TASLP.2017.2752365>.
3. Gonzalez-Lopez, J.A.; Gomez-Alanis, A.; Martín-Doñas, J.M.; Pérez-Córdoba, J.L.; Gomez, A.M. Silent speech interfaces for speech restoration: A review. *IEEE Access* **2020**, *8*, 177995–178021. <https://doi.org/10.1109/ACCESS.2020.3026579>.

4. Guenther, F.H.; Brumberg, J.S.; Wright, E.J.; Nieto-Castanon, A.; Tourville, J.A.; Panko, M.; Law, R.; Siebert, S.A.; Bartels, J.L.; Andreasen, D.S.; et al. A wireless brain-machine interface for real-time speech synthesis. *PLoS ONE* **2009**, *4*, e8218. <https://doi.org/10.1371/journal.pone.0008218>.
5. Akbari, H.; Khalighinejad, B.; Herrero, J.L.; Mehta, A.D.; Mesgarani, N. Towards reconstructing intelligible speech from the human auditory cortex. *Sci. Rep.* **2019**, *9*, 1–12. <https://doi.org/10.1038/s41598-018-37359-z>.
6. Anumanchipalli, G.K.; Chartier, J.; Chang, E.F. Speech synthesis from neural decoding of spoken sentences. *Nature* **2019**, *568*, 493–498. <https://doi.org/10.1038/s41586-019-1119-1>.
7. Schultz, T.; Wand, M. Modeling coarticulation in EMG-based continuous speech recognition. *Speech Commun.* **2010**, *52*, 341–353. <https://doi.org/10.1016/j.specom.2009.12.002>.
8. Wand, M.; Janke, M.; Schultz, T. Tackling speaking mode varieties in EMG-based speech recognition. *IEEE Trans. Biomed. Eng.* **2014**, *61*, 2515–2526. <https://doi.org/10.1109/TBME.2014.2319000>.
9. Janke, M.; Diener, L. EMG-to-speech: Direct generation of speech from facial electromyographic signals. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **2017**, *25*, 2375–2385. <https://doi.org/10.1109/TASLP.2017.2738568>.
10. Hueber, T.; Benaroya, E.L.; Chollet, G.; Denby, B.; Dreyfus, G.; Stone, M. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Commun.* **2010**, *52*, 288–300. <https://doi.org/10.1016/j.specom.2009.11.004>.
11. Schönle, P.W.; Gräbe, K.; Wenig, P.; Höhne, J.; Schrader, J.; Conrad, B. Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain Lang.* **1987**, *31*, 26–35. [https://doi.org/10.1016/0093-934X\(87\)90058-7](https://doi.org/10.1016/0093-934X(87)90058-7).
12. Fagan, M.J.; Ell, S.R.; Gilbert, J.M.; Sarrazin, E.; Chapman, P.M. Development of a (silent) speech recognition system for patients following laryngectomy. *Med. Eng. Phys.* **2008**, *30*, 419–425. <https://doi.org/10.1016/j.medengphy.2007.05.003>.
13. Gonzalez, J.A.; Cheah, L.A.; Gilbert, J.M.; Bai, J.; Ell, S.R.; Green, P.D.; Moore, R.K. A silent speech system based on permanent magnet articulography and direct synthesis. *Comput. Speech, Lang.* **2016**, *39*, 67–87. <https://doi.org/10.1016/j.csl.2016.02.002>.
14. Gonzalez, J.A.; Cheah, L.A.; Gomez, A.M.; Green, P.D.; Gilbert, J.M.; Ell, S.R.; Moore, R.K.; Holdsworth, E. Direct speech reconstruction from articulatory sensor data by machine learning. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **2017**, *25*, 2362–2374. <https://doi.org/10.1109/TASLP.2017.2757263>.
15. Kain, A.; Macon, M. Spectral voice conversion for text-to-speech synthesis. In Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 1998), Seattle, WA, USA, 15 May 1998; Volume 1, pp. 285–288. <https://doi.org/10.1109/ICASSP.1998.674423>.
16. Stylianou, Y.; Cappe, O.; Moulines, E. Continuous probabilistic transform for voice conversion. *IEEE Trans. Audio Speech Lang. Process.* **1998**, *6*, 131–142. <https://doi.org/10.1109/89.661472>.
17. Mohammadi, S.H.; Kain, A. An overview of voice conversion systems. *Speech Commun.* **2017**, *88*, 65–82. <https://doi.org/10.1016/j.specom.2017.01.008>.
18. Rabiner, L.R.; Juang, B.H. *Fundamentals of Speech Recognition*; Prentice-Hall: Upper Saddle River, NJ, USA, 1993.
19. Andrew, G.; Arora, R.; Bilmes, J.; Livescu, K. Deep canonical correlation analysis. In Proceedings of the International Conference on Machine Learning (ICML 2013), Atlanta, GA, USA, 16–21 June 2013; pp. 1247–1255.
20. Wang, W.; Arora, R.; Livescu, K.; Bilmes, J. On deep multiview representation learning. In Proceedings of the International Conference on Machine Learning (ICML 2015), Lille, France, 6–11 July 2015; pp. 1083–1092.
21. Fang, F.; Yamagishi, J.; Echizen, I.; Lorenzo-Trueba, J. High-quality nonparallel voice conversion based on cycle-consistent adversarial network. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5279–5283.
22. Gilbert, J.M.; Rybchenko, S.I.; Hofe, R.; Ell, S.R.; Fagan, M.J.; Moore, R.K.; Green, P. Isolated word recognition of silent speech using magnetic implants and sensors. *Med. Eng. Phys.* **2010**, *32*, 1189–1197. <https://doi.org/10.1016/j.medengphy.2010.08.011>.
23. Trigeorgis, G.; Nicolaou, M.A.; Schuller, B.W.; Zafeiriou, S. Deep canonical time warping for simultaneous alignment and representation learning of sequences. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1128–1138. <https://doi.org/10.1109/TPAMI.2017.2710047>.
24. Hotelling, H. Relations between two sets of variates. *Biometrika* **1936**, *28*, 321–377. <https://doi.org/10.1093/biomet/28.3-4.321>.
25. Wang, W.; Yan, X.; Lee, H.; Livescu, K. Deep variational canonical correlation analysis. *arXiv* **2016**, arXiv:1610.03454.
26. Zhou, F.; De la Torre, F. Generalized canonical time warping. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 279–294. <https://doi.org/10.1109/TPAMI.2015.2414429>.
27. Sahbi, H. Learning CCA Representations for Misaligned Data. In Proceedings of the European Conference on Computer Vision, ECCV—Workshop, Munich, Germany, 8–14 September 2018; Springer: Munich, Germany, 2018; Volume 11132, pp. 468–485. https://doi.org/10.1007/978-3-030-11018-5_39.
28. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 8–13 December 2014; pp. 3104–3112.
29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
30. Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* **1962**, *33*, 1065–1076.

31. Hermann, K.M.; Blunsom, P. Multilingual models for compositional distributed semantics. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 22–27 June 2014; pp. 58–68. <https://doi.org/10.3115/v1/P14-1006>.
32. Kominek, J.; Black, A.W. The CMU Arctic speech databases. In Proceedings of the 5th ISCA Workshop on Speech Synthesis, Pittsburgh, PA, USA, 14–16 June 2004; pp. 223–224.
33. Morise, M.; Yokomuri, F.; Ozawa, K. WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Trans. Inf. Syst.* **2016**, *99*, 1877–1884. <https://doi.org/10.1587/transinf.2015EDP7457>.
34. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
35. Gonzalez, J.A.; Cheah, L.A.; Green, P.D.; Gilbert, J.M.; Ell, S.R.; Moore, R.K.; Holdsworth, E. Evaluation of a silent speech interface based on magnetic sensing and deep learning for a phonetically rich vocabulary. In Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 3986–3990.
36. Tokuda, K.; Yoshimura, T.; Masuko, T.; Kobayashi, T.; Kitamura, T. Speech parameter generation algorithms for HMM-based speech synthesis. In Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, 5–9 June 2000; pp. 1315–1318. <https://doi.org/10.1109/ICASSP.2000.861820>.
37. Toda, T.; Black, A.W.; Tokuda, K. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 2222–2235. <https://doi.org/10.1109/TASL.2007.907344>.
38. Kubichek, R. Mel-cepstral distance measure for objective speech quality assessment. In Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, Victoria, BC, Canada, 19–21 May 1993; pp. 125–128.
39. Zhou, F.; Torre, F. Canonical time warping for alignment of human behavior. In Proceedings of the International Conference Neural Information Processing Systems, Vancouver, BC, Canada, 7–10 December 2009; Volume 22, pp. 2286–2294.
40. Kraft, S.; Zölzer, U. BeagleJS: HTML5 and JavaScript based framework for the subjective evaluation of audio quality. In Proceedings of the Linux Audio Conference, Karlsruhe, Germany, 1–4 May 2014.
41. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
42. Gomez-Alanis, A.; Gonzalez-Lopez, J.A.; Peinado, A.M. A kernel density estimation based loss function and its application to asv-spoofing detection. *IEEE Access* **2020**, *8*, 108530–108543.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
44. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.