Multi-view Temporal Alignment for Non-parallel Articulatory-to-Acoustic Speech Synthesis

Jose A. Gonzalez-Lopez[1], Miriam Gonzalez-Atienza[1], Alejandro Gomez-Alanis[1], José L. Pérez-Córdoba[1], and Phil D. Green[2]

[1]University of Granada, Granada, Spain
[2]University of Sheffield, Sheffield, U.K.

joseangl@ugr.es

Abstract

Articulatory-to-acoustic (A2A) synthesis refers to the generation of audible speech from captured movement of the speech articulators. This technique has numerous applications, such as restoring oral communication to people who cannot longer speak due to illness or injury. Most successful techniques so far adopt a supervised learning framework, in which time-synchronous articulatory-and-speech recordings are used to train a supervised machine learning algorithm that can be used later to map articulator movements to speech. This, however, prevents the application of A2A techniques in cases where parallel data is unavailable, e.g., a person has already lost her/his voice and only articulatory data can be captured. In this work, we propose a solution to this problem based on the theory of multi-view learning. The proposed algorithm attempts to find an optimal temporal alignment between pairs of non-aligned articulatory-and-acoustic sequences with the same phonetic content by projecting them into a common latent space where both views are maximally correlated and then applying dynamic time warping. Several variants of this idea are discussed and explored. We show that the quality of speech generated in the non-aligned scenario is comparable to that obtained in the parallel scenario.

Index Terms: multi-view learning, dynamic time warping, canonical correlation analysis, silent speech interface, speech restoration.

1. Introduction

A silent speech interface (SSI) is a type of assistive technology aimed at restoring normal, vocal communication to speech-impaired persons. It does so by decoding speech from biosignals, different from the actual acoustic signal, generated by the human body during speech production. These biosignals can range from the neural activity in the speech and language areas of the brain [1–3], electrical activity driving the face muscles captured by surface electrodes (i.e., electromyography (EMG)) [4–6], or motion capture of the speech articulators by means of imaging techniques [7] or electromagnetic articulography techniques [8–11].

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. Two alternative SSI approaches have been proposed to decode speech from speech-related biosignals [12]: silent speech recognition, which involves the use of automatic speech recognition (ASR) algorithms to transform the biosignals into text, and direct speech synthesis, which directly maps 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 in the literature when the biosignals encode information about the movements of the speech organs.

Most successful direct synthesis techniques so far adopt a data-driven framework in which supervised machine learning is used to model the mapping $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x})$ between source feature vectors $\boldsymbol{x}$ extracted from the biosignals and target feature vectors $\boldsymbol{y}$ computed from the speech signals. To train this function, a dataset $\mathcal{D} = \{(\boldsymbol{X}_1, \boldsymbol{Y}_1), \ldots, (\boldsymbol{X}_M, \boldsymbol{Y}_M)\}$ with pairs of sequences of feature vectors $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ extracted from time-synchronous recordings is used, where $\boldsymbol{X}_i \in \mathbb{R}^{d_x \times T_i}$ and $\boldsymbol{Y}_i \in \mathbb{R}^{d_y \times T_i}$. The need for parallel recordings, however, limits the application of direct synthesis techniques to only a few clinical scenarios, as described in [12]. For instance, people who have already lost their voices could not use this technology because of the impossibility of recording parallel data.

A solution to this problem, which we explore in this work, involves the use of speech recordings from voice donors (e.g., relatives 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 mouth along the speech recordings while articulatory data is captured. Even in this case, it is likely that the articulatory data will not be perfectly aligned with the speech recordings, having both modalities slightly different duration, thus preventing the direct application of standard supervised machine learning techniques.

To address this issue, in this work we propose an algorithm called multi-view temporal alignment by dependence maximisation in the latent space (TRANSIENCE)[1], which is based on the theory of multi-view learning [13, 14]. TRANSIENCE attempts to find the optimal temporal alignment between sequences of multiple views (e.g., audio and articulatory data) by first non-linearly projecting the data into a common, latent subspace where the views are maximally dependent and then aligning the resulting latent variables by means of the dynamic time warping (DTW) algorithm[2] [15]. We examine the performance of the proposed algorithm on a A2A task involving the conversion of articulatory data captured using permanent magnet articulography (PMA) [16, 17] to speech for multiple non-impaired

---

[1]Code is available at https://github.com/joseangl/transience.

[2]It should be noted that the direct application of DTW to this problem is not possible because the views may have different dimensionality.

subjects.

The remainder of this paper is organised as follows. First, in Section 2, the relevant related work is reviewed and the main differences w.r.t. our technique are discussed. The details of the proposed technique are presented in Section 3. Section 4 describes the experimental setup, while the results are shown in Section 5. We conclude in Section 6 and discuss potential future research directions.

2. Related work

The most closely related work to our method is that of Trigeorgis *et al.* [18], in which an algorithm called deep canonical time warping (DCTW) combining canonical correlation analysis (CCA) [19] with DTW is proposed. The key differences of our approach w.r.t. DCTW can be summarised as follows. First, we evaluate different similarity metrics, not only CCA, to optimise the parameters of the deep neural networks (DNNs) used to map the multiple views into their common, latent subspace. Also, we introduce an autoenconder-based loss, which helps to regularise the training and avoids *naïve* solutions. Finally, inspired by the work in [20], we also propose the introduction of private latent variables for each view which aim at modelling the specific peculiarities within each view. Our technique also shares similarities with the generalized canonical time warping (GCTW) technique described in [21]. However, in contrast to this technique, TRANSIENCE solves the optimal alignment problem with DTW rather than approximating the temporal warping with a set of pre-defined monotonic bases and optimising the weights of these bases with a Gauss-Newton algorithm. Also, being based on CCA, GCTW computes the latent variables by *linearly projecting* the data from the different views, while our method uses powerful autoencoders to non-linearly transform the data, which we could be expected to yield better alignments.

Contrary to now popular sequence-to-sequence (seq2seq) models (e.g., [22, 23]), our method has different advantages. First, our aim is to align sequences of different lengths in order 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. Multi-view temporal alignment

The problem we address can be formulated as finding the optimal alignment between two time series $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{T_x})$ and $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{T_y})$ with possibly different dimensions. In our case, $\boldsymbol{X} \in \mathbb{R}^{d_x \times T_x}$ is a sequence of feature vectors extracted from an articulatory signal and $\boldsymbol{Y} \in \mathbb{R}^{d_y \times T_y}$ 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 minimisation problem,

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

where $T = \max(T_x, T_y)$, $d(\boldsymbol{x}, \boldsymbol{y})$ is a distance function, $\boldsymbol{\phi}^x \in \{1 : T_x\}^T$ and $\boldsymbol{\phi}^y \in \{1 : T_y\}^T$ are warping functions that map

the indices of the original time series to a common time axis. This way, $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ will be aligned if $\phi_t^x = i$ and $\phi_t^y = j$ for a given t.

To enable the alignment of sequences with different dimensionality, we assume that there exists a pair of transformation functions $\boldsymbol{f} : \mathbb{R}^{d_x} \to \mathbb{R}^{d_z}$ and $\boldsymbol{g} : \mathbb{R}^{d_y} \to \mathbb{R}^{d_z}$, modelled as DNNs in this work, that map the feature vectors into a common, latent space where the data from both views are maximally dependent (in the next section we will discuss how we measure this dependence). Thus, the problem in (1) now becomes,

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

For some fixed mapping functions $\boldsymbol{z}^x = \boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{z}^y = \boldsymbol{g}(\boldsymbol{y})$, the problem of temporal alignment of latent variable sequences $\boldsymbol{Z}^x = (\boldsymbol{z}_1^x, \ldots, \boldsymbol{z}_{T_x}^x)$ and $\boldsymbol{Z}^y = (\boldsymbol{z}_1^y, \ldots, \boldsymbol{z}_{T_y}^y)$ can be solved efficiently by means of the DTW algorithm [15]. Conversely, for fixed warping paths $\boldsymbol{\phi}^x$ and $\boldsymbol{\phi}^y$, the problem in (2) involves optimising the functions $\boldsymbol{f}(\cdot)$ and $\boldsymbol{g}(\cdot)$ in order to minimise $\sum_{t=1}^{T} d(\boldsymbol{f}(\boldsymbol{x}_{\phi_t^x}), \boldsymbol{g}(\boldsymbol{y}_{\phi_t^y}))$. If the mapping functions are modelled as DNNs, as in our case, this latter problem can be solved by back-propagation. Thus, TRANSIENCE algorithm solves (2) by alternating between the following two phases: (i) finding the optimum DNNs weights (Θ_f, Θ_g) by fixing the warping paths, and (2) applying DTW to compute the optimum warping paths (i.e., optimum alignments) $(\boldsymbol{\phi}^x, \boldsymbol{\phi}^y)$ by freezing the DNNs weights. The warping paths are initialised 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 $t = 1, \ldots, T$ and $T = \max(T_x, T_y)$.

3.1. Latent-space dependence metrics

A key component of our algorithm is the distance function $d(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{y}))$ used in (2) to evaluate the dependence between pairs of aligned latent variables. Here, we evaluate three alternative loss functions that optimise this dependence by maximising the correlation, mutual information and (minimise) a contrastive loss, respectively, between the latent variables.

3.1.1. Canonical correlation analysis

Given a mini-batch of N pairs of *aligned* observations $\mathcal{B} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$, this loss function attempts to maximise the correlation between the outputs of the DNNs, $\boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{g}(\boldsymbol{y})$, as follows,

$$(\Theta_f^*, \Theta_g^*) = \underset{\Theta_f, \Theta_g}{\operatorname{argmax}} \sum_{i=1}^{N} \operatorname{corr}(\boldsymbol{f}(\boldsymbol{x}_i; \Theta_f), \boldsymbol{g}(\boldsymbol{y}_i; \Theta_g)). \quad (3)$$

As detailed in [13], this equals to maximising the following loss function,

$$\mathcal{L}_{cca} = \sqrt{\operatorname{tr}(\boldsymbol{T}'\boldsymbol{T})}, \quad (4)$$

where $\operatorname{tr}(\cdot)$ is the trace operator and $\boldsymbol{T} = \boldsymbol{\Sigma}_{xx}^{-1/2} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1/2}$. The covariance matrices $\boldsymbol{\Sigma}_{xx} = \operatorname{cov}(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x}))$, $\boldsymbol{\Sigma}_{xy} = \operatorname{cov}(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{y}))$ and $\boldsymbol{\Sigma}_{yy} = \operatorname{cov}(\boldsymbol{g}(\boldsymbol{y}), \boldsymbol{g}(\boldsymbol{y}))$ are estimated from the outputs of the DNNs. It should be noted that the DCTW algorithm described in [18] is a specific case of our TRANSIENCE algorithm when the CCA loss in (4) is used.

3.1.2. Maximum mutual information

As an alternative, we also consider maximising the mutual information between the latent variables as follows,

$$\mathcal{L}_{mmi} = \sum_{i=1}^{N} p(\boldsymbol{f}(\boldsymbol{x}_i), \boldsymbol{g}(\boldsymbol{y}_i)) \log \frac{p(\boldsymbol{f}(\boldsymbol{x}_i), \boldsymbol{g}(\boldsymbol{y}_i))}{p(\boldsymbol{f}(\boldsymbol{x}_i))p(\boldsymbol{g}(\boldsymbol{y}_i))}. \quad (5)$$

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

$$p(\boldsymbol{z}_i) = \frac{1}{N-1} \sum_{j=1, j \neq i}^{N} K(\boldsymbol{z}_i - \boldsymbol{z}_j), \quad (6)$$

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

3.1.3. Contrastive loss

Finally, we also evaluate the contrastive loss function described in [20, 24] which, given a fixed latent variable from the first view $\boldsymbol{f}(\boldsymbol{x}^+)$, takes an aligned positive example $\boldsymbol{g}(\boldsymbol{y}^+)$ and an unaligned negative example $\boldsymbol{g}(\boldsymbol{y}^-)$ from the second view and attempts to *minimise* the difference between the distances for the positive and negative examples:

$$\mathcal{L}_{contrastive} = \frac{1}{N} \sum_{i=1}^{N} \max(0, m + d(\boldsymbol{f}(\boldsymbol{x}_i^+), \boldsymbol{g}(\boldsymbol{y}_i^+))$$
$$- d(\boldsymbol{f}(\boldsymbol{x}_i^+), \boldsymbol{g}(\boldsymbol{y}_i^-))), \quad (7)$$

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

3.2. Multi-view autoencoder

In order to regularise the training of the DNNs and to avoid naïve solutions (e.g., $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{y}) = \boldsymbol{c}$, for all $\boldsymbol{x}$ and $\boldsymbol{y}$, with $\boldsymbol{c}$ being a constant vector), we introduce an autoencoder-based reconstruction loss that minimises the mean squared error (MSE) between the DNNs' inputs $(\boldsymbol{x}, \boldsymbol{y})$ and the reconstructed outputs $(\hat{\boldsymbol{x}} = \boldsymbol{f}^{-1}(\boldsymbol{f}(\boldsymbol{x})), \hat{\boldsymbol{y}} = \boldsymbol{g}^{-1}(\boldsymbol{g}(\boldsymbol{y})))$, being $\boldsymbol{f}^{-1}$ and $\boldsymbol{g}^{-1}$ decoder networks that attempt to reconstruct $\boldsymbol{x}$ and $\boldsymbol{y}$ from their latent projections. The parameters of such networks, as well as those from the encoders $\boldsymbol{f}$ and $\boldsymbol{g}$, are trained by gradient-descend techniques using the following loss function,

$$\mathcal{L}_{autoencoder} = \frac{\lambda}{N} \left(\sum_{i=1}^{N} \left\| \boldsymbol{x}_i - \boldsymbol{f}^{-1}(\boldsymbol{f}(\boldsymbol{x}_i)) \right\|^2 \right.$$
$$\left. + \sum_{i=1}^{n} \left\| \boldsymbol{y}_i - \boldsymbol{g}^{-1}(\boldsymbol{g}(\boldsymbol{y}_i)) \right\|^2 \right), \quad (8)$$

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

Table 1: *Details of the dataset used for the experiments.*

Condition		# of sentences	
		Train	*Eval.*
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)

3.3. Private latent variables

A key assumption of TRANSIENCE is that the views share some common information. Although this is indeed the case for audio and articulatory data, where both views encode the same phonetic information, it is also true that each view may have its own unique characteristics, thus making the reconstruction loss in (8) difficult to optimise when only considering the shared latent variables. For instance, the PMA technique used in this work for articulator motion capture is known to model poorly the information about the speech prosody [25, 26], whereas this information can be easily decoded from the acoustic signal. Thus, 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 [20], we propose the introduction of private latent variables for each view $\tilde{\boldsymbol{z}}^x$ and $\tilde{\boldsymbol{z}}^y$ that aim at modelling the uniqueness of each view. The private variables are predicted from the inputs by a set of independent DNNs $\tilde{\boldsymbol{z}}^x = \tilde{\boldsymbol{f}}(\boldsymbol{x})$ and $\tilde{\boldsymbol{z}}^y = \tilde{\boldsymbol{g}}(\boldsymbol{y})$. These private latent variables are used in (8), in addition to the common variables, for reconstructing the input data, i.e., $\hat{\boldsymbol{x}} = \boldsymbol{f}^{-1}(\boldsymbol{f}(\boldsymbol{x}), \tilde{\boldsymbol{f}}(\boldsymbol{x}))$ and $\hat{\boldsymbol{y}} = \boldsymbol{g}^{-1}(\boldsymbol{g}(\boldsymbol{y}), \tilde{\boldsymbol{g}}(\boldsymbol{y}))$.

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

$$\mathcal{L}_{KL} = \frac{1}{2} \sum_{i=1}^{d_{\tilde{z}}} \left(\sigma_i^2 + \mu_i^2 - 1 - \log \sigma_i^2 \right), \quad (9)$$

where the mean and variances in (9) are estimated for each private variable in each mini-batch.

4. Experimental setup

4.1. Dataset

The proposed alignment algorithm was evaluated on a A2A task involving the synthesis of speech from articulatory data captured using PMA [9, 17]. Parallel data was recorded by four non-impaired British subjects (2 males and 2 females) while reading aloud a subset of the CMU Arctic corpus [27]. Two alignment conditions were evaluated: (i) intra-subject alignment, where PMA and speech signals recorded by the same subject in different sessions are aligned, and (ii) cross-subject alignment, where PMA signals recorded by a given subject are aligned with speech recorded by a different subject (with 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 summarised in Table 1.

Table 2: *Summary of the objective results.*

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**
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
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
Contrastive	**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

4.2. Implementation details

Each DNN in TRANSIENCE was modelled 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 [18]. The neural networks were trained as denoising autoencoders ($\sigma_{noise} = 0.5$) using the Adam algorithm [28] with a fixed learning rate of $1e-4$ and a batch size of $N = 512$ samples. The dimensionality of the shared latent variables was set to $d_z = 20$ and fixed $d_{\tilde{z}} = 10$ for the private variables. The PMA signals were parameterised by applying principal component analysis (PCA) over contextual windows with 11 frames. The acoustic signals, on the other hand, were parameterised using the WORLD vocoder [29] with a frame rate of 5 ms as 25 mel-generalised cepstral coefficients (MGCCs), 1 band aperiodicity (BAP) value, 1 continuous F_0 value in logarithmic scale and 1 U/V decision. For temporal alignment, only the MGCCs were used (augmented with delta and acceleration parameters). Finally, the cosine distance was used for DTW.

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 [26]: DNNs with 4 hidden layers with 400 units in each layer and rectified linear unit (ReLU) activations were used. The maximum-likelihood parameter generation (MLPG) algorithm [30,31] was applied over the DNN outputs to enhance the acoustic quality of the re-synthesised waveforms.

5. Results

5.1. Objective evaluation

First, we evaluated the quality of the re-synthesised speech signals obtained from the test PMA signals by comparing them with the original speech recordings made by the non-impaired subjects. For this task, we used several objective metrics widely used in speech synthesis: mel-cepstral distortion (MCD), root mean squared error (RMSE) of the predicted BAP and F_0 values and the error rate for the voicing parameter. For TRANSIENCE, three variants were evaluated depending on the latent-space similarity loss function employed: CCA, maximum mutual information (MMI), and Contrastive losses 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 [32], which is a particular case of TRANSIENCE combin-
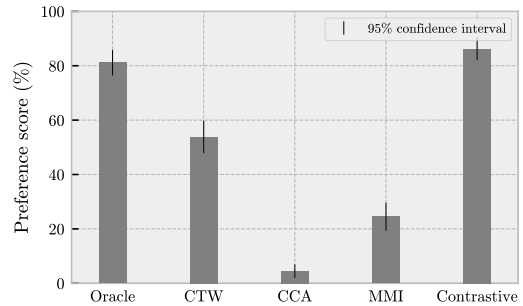

Figure 1: *Results of the ABX test on speech quality.*

ing standard (linear) CCA with DTW, thus not being able to model non-linear 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.

Table 2 shows the objective results for the different systems. TRANSIENCE using the contrastive loss yields the best alignments and, hence, the best objective results, outperforming the other 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 non-linear version TRANSIENCE+CCA. In future work, we should look at this issue.

5.2. Subjective evaluation

We also conducted an ABX test to subjectively evaluate the quality of the resynthesised speech signals. 27 listeners participated in the test, who have to judge which of two versions of the same signal produced by any combination of two of the 5 systems in Table 2 was more similar to a reference (one of the signals recorded by the subjects). Each listeners evaluated 10 sample pairs for each of the 10 possible system combinations (i.e., 100 pairs evaluated in total). For this task, only the "basic" systems in Table 2 were evaluated (i.e., without autoencoder-loss and private variables), because this setting produced the best objective results. Fig. 1 shows the results of the listening test. The most preferred system by a large margin was Contrastive, being on par with the Oracle system. Interestingly, the CTW system obtained higher preference scores than its non-linear version (TRANSIENCE+CCA) and that the MMI system. It may be due that the optimisation process get stuck in poor local-minima for the latter systems. However, more research is needed to shed some light into this problem.

6. Conclusions

We have proposed a new method for the alignment of times series with different dimensionality. Our evaluation on a A2A task involving the synthesis of audible speech from articulatory signals has shown that it is feasible to deploy direct synthesis techniques in non-parallel scenarios. Future work include evaluating our technique using more data from clinical population and introducing more constraints for the alignment (e.g., phonetic constraints).

7. References

[1] F. H. Guenther, J. S. Brumberg, E. J. Wright, A. Nieto-Castanon, J. A. Tourville, M. Panko, R. Law, S. A. Siebert, J. L. Bartels, D. S. Andreasen, P. Ehirim, H. Mao, and P. R. Kennedy, "A wireless brain-machine interface for real-time speech synthesis," *PLoS ONE*, vol. 4, no. 12, pp. e8218–e8218, 2009.

[2] H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani, "Towards reconstructing intelligible speech from the human auditory cortex," *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[3] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, 2019.

[4] T. Schultz and M. Wand, "Modeling coarticulation in EMG-based continuous speech recognition," *Speech Commun.*, vol. 52, no. 4, pp. 341—-353, 2010.

[5] M. Wand, M. Janke, and T. Schultz, "Tackling speaking mode varieties in EMG-based speech recognition," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 10, pp. 2515—2526, 2014.

[6] M. Janke and L. Diener, "EMG-to-speech: Direct generation of speech from facial electromyographic signals," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2375–2385, 2017.

[7] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Commun.*, vol. 52, no. 4, pp. 288–300, 2010.

[8] P. W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader, and B. Conrad, "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain Lang.*, vol. 31, no. 1, pp. 26–35, 1987.

[9] M. J. Fagan, S. R. Ell, J. M. Gilbert, E. Sarrazin, and P. M. Chapman, "Development of a (silent) speech recognition system for patients following laryngectomy," *Med. Eng. Phys.*, vol. 30, no. 4, pp. 419–425, 2008.

[10] J. A. Gonzalez, L. A. Cheah, J. M. Gilbert, J. Bai, S. R. Ell, P. D. Green, and R. K. Moore, "A silent speech system based on permanent magnet articulography and direct synthesis," *Comput. Speech. Lang.*, vol. 39, pp. 67–87, 2016.

[11] J. A. Gonzalez, L. A. Cheah, A. M. Gomez, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth, "Direct speech reconstruction from articulatory sensor data by machine learning," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2362–2374, 2017.

[12] J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. M. Martín-Doñas, J. L. Pérez-Córdoba, and A. M. Gomez, "Silent speech interfaces for speech restoration: A review," *IEEE Access*, vol. 8, pp. 177 995–178 021, 2020.

[13] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. ICML*, 2013, pp. 1247–1255.

[14] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multiview representation learning," in *Proc. ICML*, 2015, pp. 1083–1092.

[15] L. R. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.

[16] F. Fang, J. Yamagishi, I. Echizen, and J. Lorenzo-Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," in *Proc. ICASSP*, 2018, pp. 5279–5283.

[17] J. M. Gilbert, S. I. Rybchenko, R. Hofe, S. R. Ell, M. J. Fagan, R. K. Moore, and P. Green, "Isolated word recognition of silent speech using magnetic implants and sensors," *Med. Eng. Phys.*, vol. 32, no. 10, pp. 1189–1197, 2010.

[18] G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "Deep canonical time warping for simultaneous alignment and representation learning of sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1128–1138, 2017.

[19] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3-4, pp. 321–377, 1936.

[20] W. Wang, X. Yan, H. Lee, and K. Livescu, "Deep variational canonical correlation analysis," *arXiv preprint arXiv:1610.03454*, 2016.

[21] F. Zhou and F. De la Torre, "Generalized canonical time warping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 279–294, 2015.

[22] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. NIPS*, 2014, pp. 3104–3112.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[24] K. M. Hermann and P. Blunsom, "Multilingual models for compositional distributed semantics," in *Proc. Annual Meeting of the Association for Computational Linguistics*, 2014, pp. 58–68.

[25] J. A. Gonzalez, L. A. Cheah, J. Bai, S. R. Ell, J. M. Gilbert, R. K. Moore, and P. D. Green, "Analysis of phonetic similarity in a silent speech interface based on permanent magnetic articulography," in *Proc. Interspeech*, 2014, pp. 1018–1022.

[26] J. A. Gonzalez, L. A. Cheah, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth, "Evaluation of a silent speech interface based on magnetic sensing and deep learning for a phonetically rich vocabulary," in *Proc. Interspeech*, 2017, pp. 3986–3990.

[27] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *5th ISCA Workshop on Speech Synthesis*, 2004, pp. 223–224.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] M. Morise, F. Yokomuri, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE T. Inf. Syst.*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[30] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, 2000, pp. 1315–1318.

[31] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, 2007.

[32] F. Zhou and F. Torre, "Canonical time warping for alignment of human behavior," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 22, 2009, pp. 2286–2294.