IberSPEECH 2024
11-13 November 2024, Aveiro, Portugal

Direct Speech Synthesis from Non-audible Speech Biosignals: A Comparative Study

Javier Lobato Martín, José L. Pérez-Córdoba, Jose A. Gonzalez-Lopez

Dept. of Signal Theory, Telematics and Communications, University of Granada, Spain

javilm500@gmail.com, jlpc@ugr.es, joseangl@ugr.es

Abstract

This paper presents a speech restoration system that generates audible speech from articulatory movement data captured using Permanent Magnet Articulography (PMA). Several algorithms were explored for speech synthesis, including classical unit-selection and deep neural network (DNN) methods. A database containing simultaneous PMA and speech recordings from healthy subjects was used for training and validation. The system generates either direct waveforms or acoustic parameters, which are converted to audio via a vocoder. Results show intelligible speech synthesis is feasible, with Mel-Cepstral Distortion (MCD) values between 9.41 and 12.4 dB, and Short-Time Objective Intelligibility (STOI) scores ranging from 0.32 to 0.606, with a maximum near 0.9. Unit selection and recurrent neural network (RNN) methods performed best. Informal listening tests further confirmed the effectiveness of these algorithms in producing intelligible speech from PMA data.

Index Terms: Silent speech interfaces, direct speech synthesis, biosignals, unit selection, PMA, deep neural networks, CCA.

1. Introduction

Silent speech interfaces (SSIs) [1, 2, 3] are assistive technology designed to restore vocal communication for individuals with speech impairments. For this purpose, SSIs decode speech from non-acoustic biosignals produced in the human body during speech generation, bypassing the need for the actual acoustic signal. These biosignals can originate from various sources, such as neural activity in the brain's speech and language areas [4, 5, 6], electrical activity of facial muscles measured with electromyography (EMG) [7, 8], or the motion of speech articulators tracked through imaging techniques [9] or electromagnetic articulography [10, 11]. Since SSIs do not depend on the acoustic speech signal, they provide a novel method for restoring oral communication in speech-impaired individuals.

One approach to decoding speech from these biosignals is through direct speech synthesis [12, 13], which involves mapping the biosignals to a set of acoustic parameters suitable for speech synthesis. The most successful direct synthesis techniques employ a data-driven framework, where supervised machine learning techniques are used to model the mapping $\boldsymbol{y}_t = \boldsymbol{f}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t)$ between source feature vectors $\boldsymbol{x}$ extracted from the biosignals and target feature vectors $\boldsymbol{y}$ derived from speech signals. Training this function requires a dataset $\mathcal{D} = \{(\boldsymbol{X}_1, \boldsymbol{Y}_1), \ldots, (\boldsymbol{X}_M, \boldsymbol{Y}_M)\}$, consisting of pairs of sequences of feature vectors $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ from time-synchronous recordings, 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 main goal of this paper is to explore and compare various approaches for direct speech synthesis and spectral envelope generation from articulatory biosignals captured from the lips and tongue using a technique known as Permanent Magnet Articulography (PMA) [10, 11]. We implemented several machine learning techniques for PMA-to-speech conversion, ranging from classical unit selection algorithms to advanced deep neural network methods. These techniques were trained and evaluated on a dataset containing simultaneous PMA and speech recordings from healthy subjects engaged in different speech production tasks, from sequences of digits to phonetically-rich sentences.

The main novelty of this work lies in the introduction of new approaches to direct speech synthesis from PMA biosignals, including the evaluation of a wide range of algorithms on a specific PMA dataset to assess their performance. We highlight the strengths and weaknesses of each approach in various contexts. Additionally, we present hybrid methods, such as a direct PMA-to-Speech Unit Selection algorithm that eliminates the need for a vocoder, thereby simplifying the conversion process.

2. Related work

The topic of speech synthesis from non-audible biosignals has been extensively researched. Various machine learning techniques have been applied to model the mapping between biosignals and speech. These techniques include Gaussian mixture models (GMMs) [14, 13], hidden Markov models (HMMs) [15], support vector machines (SVMs) [16], concatenative unit-selection [17, 6, 18], and, more recently, different architectures of deep neural networks (DNNs) [19, 20, 8, 21]. Many of these approaches were initially developed for voice conversion (VC) [22], which aims to transform speech from one speaker to sound like that of another speaker. These methods have been utilized to enable speech synthesis from a variety of biosignals. This includes surface electromyography (EMG) [19, 8, 21], electromagnetic articulography (EMA) [13], radar signals captured from facial movements [16], ultrasound imaging of the tongue [9, 15], and brain signals captured using invasive electrodes [18, 6].

3. Method

3.1. Dataset

Four British participants (three men and one woman) took part in this study, where articulatory data and speech were recorded simultaneously as they completed two speech production tasks of varying phonetic complexity. The first task, based on the TIDigits database [23], involved sequences of up to seven English digits with a simple vocabulary of 11 words and 21 phones, aimed at demonstrating the system's ability to generate intelligible speech. The second task, from the CMU Arctic

10.21437/IberSPEECH.2024-18

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

Task	Subject	# of sentences	Minutes of data
TiDigits	S1	324	10.3
	S2	308	8.0
Arctic	S3	470	26.9
	S4	510	28.0

corpus [24], included phonetically rich sentences from public domain English texts, designed to test speech synthesis accuracy with a more complex vocabulary. Details of the recorded data for each subject are provided in Table 1.

Articulatory data was recorded using a Permanent Magnet Articulography (PMA) device [11], which tracked the movements of six small magnets attached to the lips and tongue using tissue adhesive. Magnetic field variations were measured by three tri-axial sensors at 100 Hz, while a fourth sensor compensated for the Earth's magnetic field. Speech was captured with a microphone placed 20 cm from the subjects and recorded at 48 kHz, later downsampled to 16 kHz. The study setup allowed for precise tracking of articulatory movements and high-quality speech recordings. More details about the recording protocol can be found in [11, 12, 13].

3.2. Signal processing

For speech synthesis, we evaluated two alternative approaches as discussed below: predicting short raw-audio segments from the PMA biosignals, or alternatively, predicting a low-dimensional acoustic representation of the speech signals. In the second approach, audio signals were parameterized using the WORLD vocoder [25] with a frame rate of 5 ms. The parameters included 25 mel-frequency cepstral coefficients (MFCCs), 1 band aperiodicity (BAP) value, 1 continuous F0 value in logarithmic scale, and 1 voiced/unvoiced (U/V) decision. PMA signals, on the other hand, were parameterized by applying principal component analysis (PCA) over contextual windows spanning several consecutive frames (more details provided below). Data partition was performed using k-fold cross validation, where in each iteration, 90% of the dataset was used for training and 10% for testing.

3.3. Speech synthesis algorithms

In this work, we evaluated five algorithms to synthesize audible speech from PMA biosignals. Figure 1 shows a block diagram illustrating the functioning of the implemented algorithms. During the training phase, a parallel database with simultaneous recordings of PMA and speech signals is processed to train a machine learning model that maps PMA feature vectors to acoustic vectors. At conversion time, the trained model is used to predict acoustic vectors from incoming PMA signals. Most of the algorithms we implemented predict MFCC parameters from the PMA biosignals. However, one algorithm, the unit selection algorithm, directly predicts short segments of speech waveforms, eliminating the need for a vocoder.

Here, we focus on predicting vocal-tract filter parameters from the PMA signals due to the limited ability of the PMA technique to capture information related to the glottal source [13]. This means that the vocoder parameters related to glottal source (i.e., BAP, F0, and U/V values) were not predicted by the our algorithms but were instead copied from the original audio

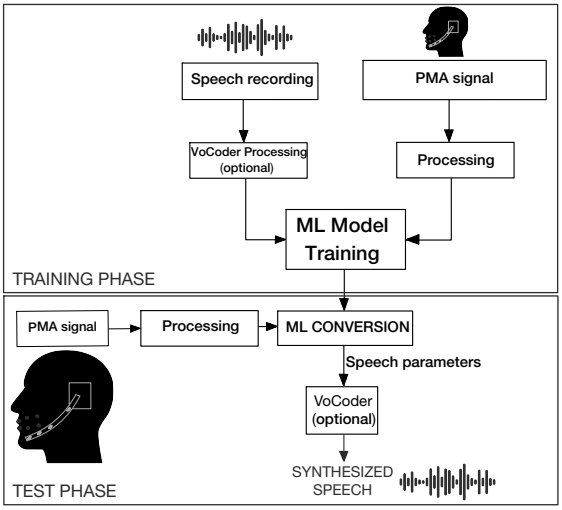

Figure 1: *Block diagram of the general approach used for mapping the PMA signals into audible speech.*

signals when synthesizing the final speech signals. More details about these algorithms are provided below.

3.3.1. Baseline linear methods

We evaluated two linear approaches as baseline methods for PMA-to-MFCC conversion: linear regression and canonical correlation analysis (CCA) [26, 27]. Both methods attempt to approximate the relationship $\boldsymbol{y} = f(\boldsymbol{x})$ between source PMA vectors $\boldsymbol{x}$ and MFCC feature vectors $\boldsymbol{y}$ as a linear regression. In particular, the linear regression approach models the relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$ using the equation

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}, \tag{1}$$

where $\boldsymbol{W}$ is the weight matrix and $\boldsymbol{b}$ is the bias vector. These parameters are estimated by least squares.

CCA-based linear regression, on the other hand, seeks to find linear transformations of $\boldsymbol{x}$ and $\boldsymbol{y}$ into a common latent space such that the correlation between the transformed variables in the latent space is maximized. CCA identifies projection matrices $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$ that maximize the correlation between $\boldsymbol{z}_x = \boldsymbol{W}_x\boldsymbol{x}$ and $\boldsymbol{z}_y = \boldsymbol{W}_y\boldsymbol{y}$, where $\dim(\boldsymbol{z}_x) = \dim(\boldsymbol{z}_y) \leq \min(\dim(\boldsymbol{x}), \dim(\boldsymbol{y}))$. Thus, the objective is to solve

$$\underset{\boldsymbol{W}_x, \boldsymbol{W}_y}{\operatorname{argmax}} \operatorname{corr}(\boldsymbol{W}_x\boldsymbol{x}, \boldsymbol{W}_y\boldsymbol{y}). \tag{2}$$

Once the projection matrices are determined, the predicted MFCCs $\boldsymbol{y}$ are computed using the following linear model:

$$\boldsymbol{y} = \boldsymbol{W}_y^{-1}\boldsymbol{W}_x\boldsymbol{x}. \tag{3}$$

In this approach, CCA effectively reduces the dimensionality of the data by projecting it onto correlated subspaces, which can lead to more accurate predictions by capturing the most relevant features when the original data contain noise or redundant information.

3.3.2. Deep neural networks

For comparison purposes, we also evaluated two deep neural network (DNN) based models for PMA-to-MFCC conversion: a standard feed-forward DNN and a gated recurrent unit (GRU)

recurrent neural network (RNN). These models have been successfully employed in previous silent speech interface (SSI) studies to model the relationship between silent-speech biosignals and acoustic parameters [8, 13, 19, 20].

The feed-forward DNN consists of 4 hidden layers, each containing 128 rectified linear units (ReLUs). In contrast, the RNN comprises 4 hidden layers with 128 GRUs in each layer. A total of 512 trainable parameters were used in both models. All models were implemented in PyTorch [28]. During training, the model weights were initialized randomly and optimized using the Adam algorithm and a learning rate of 1e-3. We used the Mean Squared Error (MSE) loss function to optimize the prediction of MFCCs parameters. The models were trained for 75 epochs.

3.3.3. Unit selection

Finally, we also evaluated a non-parametric, unit selection (US) approach [29] for PMA-to-speech conversion. US is a classical text-to-speech (TTS) synthesis technique known for providing high-quality output even with limited training data. This techniques operates by selecting and concatenating small segments of speech, or 'units'. These units typically consist of small chunks of audio waveforms, such as single phonemes, or longer segments like syllables or words. The units are stored in a database (or codebook) containing time-aligned pairs of PMA and speech data, enabling the conversion of PMA signals to speech by selecting the pair with the highest similarity.

In the training phase, the database of unit pairs is created from time-synchronous recordings of PMA and speech signals. Our source and target units consist of overlapping PMA feature vectors and speech fragments, respectively, extracted from segments with a duration between [0.02 - 0.16] s. For the speech units, we evaluated two alternative approaches: either using audio waveform segments directly or MFCCs extracted with the WORLD vocoder [25]. Source units are normalized to the $[0, 1]$ range. Frame shift can be varied, so it is also incorporated into the analysis. When MFCCs were used, the overlap was kept at 5 ms as it was given by the vocoder. When using untreated voice, overlap is varied between 10% and 75% of each of the window length values, in 4 steps.

In the decoding phase, a sequence of acoustic units matching the input PMA signal is predicted. Firstly, the PMA signal is split into overlapping segments of source units using the same procedure as for the database creation in training time. In particular, we chose a variable window width (w_u) and a shift ($s_u = 0.05s$) for creating the source units. Then, given the sequence of target units computed from the input PMA signal $\boldsymbol{t}_{1:n} = \boldsymbol{t}_1, \ldots, \boldsymbol{t}_n$, a search procedure is implemented for recovering the closest sequence of units $\hat{\boldsymbol{u}}_{1:n} = \boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ to the target from the database. Mathematically, this procedure is governed by the following optimization process:

$$\hat{\boldsymbol{u}}_{1:n} = \underset{\boldsymbol{u}_{1:n}}{\operatorname{argmin}}\, C(\boldsymbol{t}_{1:n}, \boldsymbol{u}_{1:n}), \tag{4}$$

where $C(\boldsymbol{t}_{1:n}, \boldsymbol{u}_{1:n})$ denotes the total cost for a sequence of n units. This cost can be expanded as follows:

$$C(\boldsymbol{t}_{1:n}, \boldsymbol{u}_{1:n}) = \sum_{i=1}^{n} C^{(t)}(\boldsymbol{t}_i, \boldsymbol{u}_i) + \gamma \sum_{i=2}^{n} C^{(c)}(\boldsymbol{u}_{i-1}, \boldsymbol{u}_i). \tag{5}$$

Equation (5) is the sum of two terms involving target and concatenation costs. The target cost $C^{(t)}(\boldsymbol{t}_i, \boldsymbol{u}_i)$ measures the similarity between a database unit $\boldsymbol{u}_i$ and the target unit $\boldsymbol{t}_i$, computed as the Euclidean distance between the PMA source units (segments). The concatenation cost $C^{(c)}(\boldsymbol{u}_{i-1}, \boldsymbol{u}_i)$, on the other hand, measures the resemblance between two consecutive acoustic units, ensuring smooth transitions to avoid poor audio quality. This cost is calculated as the Cepstral distance between speech units at the point of concatenation. Finally, γ is an experimental weighting factor used to balance the importance of target and concatenation costs. In the evaluation, this weight is varied between 0 and 2 in 4 steps.

Solving (4) involves finding the optimal sequence of units from the database minimizing the total cost relative to the target units computed from the input PMA signal. This is efficiently achieved using the Viterbi algorithm [30]. To sped up this process, a pruning strategy was implemented, limiting the Viterbi search to a maximum of 10 active paths. This was facilitated by a *ball tree* data structure [31], which enabled efficient evaluation of the target cost for fewer database candidates. Our experiments showed that pruning had minimal effect on audio quality while significantly speeding up conversion time.

Depending on the type of acoustic units, two procedures are used to generate the final audio waveform from the optimal unit sequence. If the acoustic units are MFCCs, they are fed into the WORLD vocoder to produce the final speech waveform, with BAP, log F0 and U/V parameters obtained from the original signals, as stated above. Alternatively, if raw audio segments are used, the final audio output is generated using the *overlap-and-add* method, where multiple overlapping units are combined to create a single, coherent audio waveform. Specifically, the overlapping audio units are smoothed using the weight function $w[\,]$ proposed in [32]:

$$w[i] = \frac{\exp(-0.2 \cdot a[i])}{\hat{w}}, \quad i = 1 \ldots m \tag{6}$$

where m is the number of overlapping units for a given frame and $a[i]$ is given by,

$$a[i] = \begin{cases} \left[\frac{n}{2}, \frac{n}{2} - 1, \ldots, 1, 1, \ldots, \frac{n}{2} - 1, \frac{n}{2}\right], & \text{if } n \text{ is even} \\ \left[\frac{n}{2}\right], \left[\frac{n}{2}\right] - 1, \ldots, 1, \ldots, \left[\frac{n}{2}\right] - 1, \left[\frac{n}{2}\right], & \text{if } n \text{ is odd} \end{cases} \tag{7}$$

where $\hat{w} = \sum_{i=1}^{n} exp(-0.2)a[i]$ normalizes the weights such that their sum equals 1. This process is repeated for each frame.

3.4. Evaluation metrics

We employed speech-quality and intelligibility metrics from the literature to evaluate our proposed methods. First, Mel-cepstral distortion (MCD) [33] in decibels (dB), was employed to measure the overall quality of the synthesised signals compared to the original audios. Lower MCD values indicates higher quality synthesis. Short-term objective intelligibility (STOI) [34] was used to assess the intelligibility of the synthesized audios. This metric allows for reliable intelligibility evaluations without the need for listening tests, with higher values indicating better intelligibility. For the evaluation, we applied a 10-fold cross-validation strategy.

Table 2: *Results for the TiDigits task.*

Algorithm	**MCD**		**STOI**	
	Avg.	*Std.*	*Avg.*	*Std.*
Linear regress.	10.95	0.05	0.50	0.02
CCA linear regress.	12.33	0.04	0.43	0.01
DNN	11.19	0.00	0.51	0.00
RNN	**9.41**	0.00	0.56	0.00
US w/ vocoder	10.96	0.05	0.54	0.03
US w/o vocoder	11.39	0.07	**0.59**	0.09

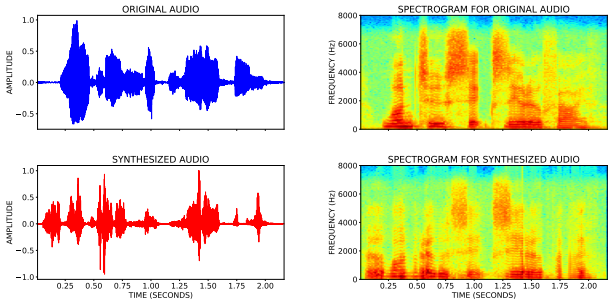

Figure 2: *Comparison of speech waveforms and spectrograms for the digit sequence "One Two Six Zero Five" for natural speech (first row) and RNN-predicted speech (second row).*

Table 3: *Results for the Arctic task.*

Algorithm	**MCD**		**STOI**	
	Avg.	*Std.*	*Avg.*	*Std.*
Linear regress.	11.15	0.04	0.43	0.01
CCA linear regress.	12.51	0.03	0.39	0.02
DNN	11.83	0.00	0.42	0.00
RNN	**10.44**	0.00	**0.49**	0.00
US w/ vocoder	12.33	0.05	0.42	0.02
US w/o vocoder	12.95	0.07	0.47	0.08

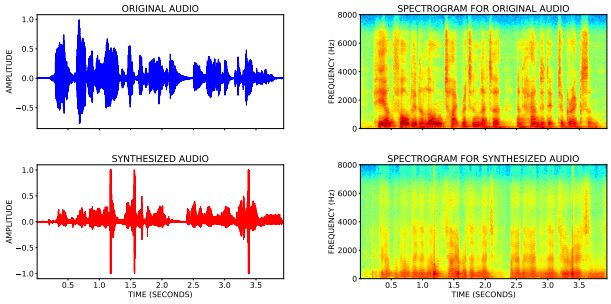

Figure 3: *Comparison of speech waveforms and spectrograms for the Arctic sentence "He unfolded a long typewritten letter and handed it to Gregson" for natural speech (first row) and RNN-predicted speech (second row).*

4. Results

4.1. TiDigits results

Table 2 shows the MCD and STOI results for each method in the TiDigits task. For each method, the grand average (Avg.) and standard deviation (Std.) of each metric is provided. The RNN method achieves the highest speech quality with an average MCD of 9.41 dB, suggesting its superior ability to model long-term speech correlations. This is a significant improvement over the other methods, which may produce less smooth audio due to frame-by-frame mapping. In terms of STOI, the unit selection method without a vocoder (US w/o vocoder) scores highest at 0.59, albeit with higher variability. The RNN method also performs well with a STOI of 0.56, effectively balancing spectral distortion and intelligibility. Overall, the RNN approach is the most balanced and effective, achieving the best MCD and a competitive STOI.

Fig. 2 shows an example of speech predicted by the best-performing method, the RNN, compared with the natural speech signal recorded by participant S1. This technique effectively predicts the global structure of the waveform, including silence periods and speech formants.

4.2. Arctic results

Table 3 shows the results obtained in the Arctic task, involving the production of phonetically-rich sentences. Overall, performance metrics are worse for the Arctic task compared to the TiDigits task (Table 2), due to the increased phonetic complexity. Again, RNN is the method achieving the lowest MCD values (average 10.44), indicating the highest speech quality and effectiveness in handling complex phonetic variations. The RNN method also performs best in terms of intelligibility with a STOI of 0.49, demonstrating robustness in producing clear speech under demanding conditions. The US w/o vocoder method achieves a relatively high STOI of 0.47, but with significant variability, indicating inconsistent performance. These results highlight the RNN's superior capability in managing both spectral distortion and speech intelligibility in phonetically-rich tasks.

Fig. 3 shows an example of speech predicted by the RNN method in comparison with the corresponding original speech signal recorded by participant S3. Again, we see that the general structure of the speech signal is well predicted from the PMA data.

5. Conclusions

The main conclusion derived from this work is that it is possible to synthesize speech using PMA biosignals using various proposed methods. The PMA signals provide sufficient information correlated with voice to generate intelligible and high-quality speech. Our results show that, while the MCD values obtained are higher than those commonly reported in the literature which, in turn, can be due to the fact that we are using a different vocoder, the synthesized speech remains clear and understandable. The highest performance was observed in the simplest TiDigits task. Among the evaluated methods, the vocoder-free unit selection and GRU-RNN methods stood out, achieving the best MCD and STOI metrics. Informal listening tests further confirmed that these algorithms provided the highest intelligibility. Consequently, these two methods are the most promising for future research and development in PMA-to-speech synthesis. Future work will aim to enhance the performance of these methods for more complex speech synthesis tasks. Additionally, we plan to explore the application of these methods to other types of biosignals.

6. Acknowledgements

This work was supported by grant PID2022-141378OB-C22 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU.

7. References

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[2] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-based spoken communication: A survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2257–2271, 2017.

[3] J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. M. M. 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.

[4] C. Herff, D. Heger, A. De Pesters, D. Telaar, P. Brunner, G. Schalk, and T. Schultz, "Brain-to-text: decoding spoken phrases from phone representations in the brain," *Frontiers in neuroscience*, vol. 9, p. 217, 2015.

[5] 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.

[6] M. Angrick, S. Luo, Q. Rabbani, D. N. Candrea, S. Shah, G. W. Milsap, W. S. Anderson, C. R. Gordon, K. R. Rosenblatt, L. Clawson *et al.*, "Online speech synthesis using a chronically implanted brain–computer interface in an individual with als," *Scientific reports*, vol. 14, no. 1, p. 9617, 2024.

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

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

[9] 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 Communication*, vol. 52, no. 4, pp. 288–300, 2010.

[10] 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," *Medical engineering & physics*, vol. 30, no. 4, pp. 419–425, 2008.

[11] L. A. Cheah, J. Bai, J. A. Gonzalez, S. R. Ell, J. M. Gilbert, R. K. Moore, and P. D. Green, "A user-centric design of permanent magnetic articulography based assistive speech technology," in *International Conference on Bio-inspired Systems and Signal Processing*, vol. 2. SCITEPRESS, 2015, pp. 109–116.

[12] J. Gonzalez Lopez, L. Cheah, J. Gilbert, J. Bai, S. Ell, P. Green, and R. Moore, "A silent speech system based on permanent magnet articulography and direct synthesis," *Computer Speech & Language*, vol. 39, 03 2016.

[13] 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 Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2362–2374, 2017.

[14] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech," *Speech communication*, vol. 54, no. 1, pp. 134–146, 2012.

[15] T. Hueber and G. Bailly, "Statistical conversion of silent articulation into audible speech using full-covariance HMM," *Computer Speech & Language*, vol. 36, pp. 274–293, 2016.

[16] A. R. Toth, K. Kalgaonkar, B. Raj, and T. Ezzat, "Synthesizing speech from Doppler signals," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4638–4641.

[17] M. Zahner, M. Janke, M. Wand, and T. Schultz, "Conversion from facial myoelectric signals to speech: a unit selection approach," in *Proc. Interspeech 2014*, 2014, pp. 1184–1188.

[18] C. Herff, L. Diener, M. Angrick, E. Mugler, M. C. Tate, M. A. Goldrick, D. J. Krusienski, M. W. Slutzky, and T. Schultz, "Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices," *Frontiers in Neuroscience*, vol. 13, 2019.

[19] L. Diener, M. Janke, and T. Schultz, "Direct conversion from facial myoelectric signals to speech using deep neural networks," in *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–7.

[20] 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*, 2017, pp. 3986–3990.

[21] R. Song, X. Zhang, X. Chen, X. Chen, X. Chen, S. Yang, and E. Yin, "Decoding silent speech from high-density surface electromyographic data using transformer," *Biomedical Signal Processing and Control*, vol. 80, p. 104298, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809422007522

[22] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[23] R. Leonard, "A database for speaker-independent digit recognition," in *ICASSP'84. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 9. IEEE, 1984, pp. 328–331.

[24] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004.

[25] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[26] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3-4, pp. 321–377, 12 1936. [Online]. Available: https://doi.org/10.1093/biomet/28.3-4.321

[27] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[28] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[29] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings*, vol. 1. IEEE, 1996, pp. 373–376.

[30] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[31] M. Dolatshah, A. Hadian, and B. Minaei-Bidgoli, "Ball*-tree: Efficient spatial indexing for constrained nearest-neighbor search in metric spaces," *arXiv preprint arXiv:1511.00628*, 2015.

[32] Z. Wu, T. Virtanen, T. Kinnunen, E. Chng, and H. Li, "Exemplar-based unit selection for voice conversion utilizing temporal information." in *INTERSPEECH*. Lyon, 2013, pp. 3057–3061.

[33] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, vol. 1. IEEE, 1993, pp. 125–128.

[34] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.