Conversion from Facial Myoelectric Signals to Speech: A Unit Selection Approach

Marlene Zahner, Matthias Janke, Michael Wand, Tanja Schultz

Cognitive Systems Lab, Karlsruhe Institute of Technology, Germany

matthias.janke@kit.edu

Abstract

This paper reports on our recent research on surface electromyographic (EMG) speech synthesis: a direct conversion of the EMG signals of the articulatory muscle movements to the acoustic speech signal. In this work we introduce a *unit selection* approach which compares segments of the input EMG signal to a database of simultaneously recorded EMG/audio unit pairs and selects the best matching audio unit based on target and concatenation cost, which will be concatenated to synthesize an acoustic speech output. We show that this approach is feasible to generate a proper speech output from the input EMG signal. We evaluate different properties of the units and investigate what amount of data is necessary for an initial transformation. Prior work on EMG-to-speech conversion used a frame-based approach from the voice conversion domain, which struggles with the generation of a natural F_0 contour. This problem may also be tackled by our unit selection approach.

Index Terms: electromyography, silent speech interface, unit selection

1. Introduction

In the recent years novel speech processing approaches called *Silent Speech Interfaces* [1] became more and more popular. These systems enable speech communication or speech recognition even when the acoustic signal is not available, tackling mainly three application areas:

1. good performance in the presence of noise,
2. no disturbance of bystanders and preservation of privacy,
3. assistive devices for speech-disabled persons (e.g., laryngectomees).

Our method of processing silent speech relies on surface electromyography (EMG) [2], where the electrical action potentials of the articulatory muscles are recorded during speaking by electrodes on the skin surface. Solely these EMG signals are then used to trace back the original speech signal. We perform a direct conversion from EMG to the acoustic domain [3] so that a receiver, be it a computer or another human, can comprehend the intended message.

[3] introduced a direct mapping from EMG to speech using a frame-based voice conversion approach [4] that creates spectral acoustic features, but still uses the fundamental frequency (F_0) from the simultaneously recorded audible speech signal. [5] complemented this technique by generating F_0 from the EMG signal, but faced issues with the naturalness and prosody of the generated output. In this paper we introduce a direct EMG-to-speech mapping based on a *unit selection* approach [6]. In the training stage we build a database consisting of corresponding EMG and speech units, collected during simultaneous data recordings. In the conversion stage this database is used to search for the best matching speech units to a given input sequence of EMG segments.

We expect this new technique to have advantages compared to a frame-based approach (e.g. Gaussian Mixture Mapping, [3, 7, 8]). The main benefit is an improved naturalness in the synthesized speech. This is based on the fact that F0 variations are already contained in the speech units and no complex system for generating prosody on top of the speech frames is necessary.

The remainder of this paper is organized as follows: Sec. 2 presents the setup and describes the data corpus we used, followed by Sec. 3 which gives the details about our unit selection mapping approach. In Sec. 4, we present our experimental results, and Sec. 5 concludes the paper and outlines possible future work.

2. Data corpus information

For our unit selection approach we selected data recordings from our prior work [8, 9] which contain more than 500 utterances of EMG signals recorded during audible speech. The corpus consists of four recording sessions of two male speakers. While it is imaginable to use these sessions to build a speaker independent system, we only use it *session dependently*.

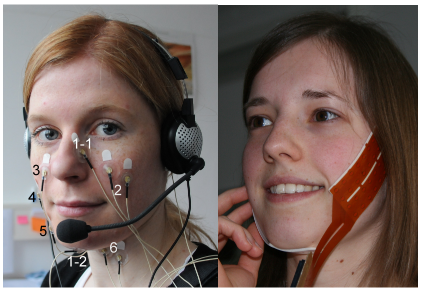

Figure 1: ***left****: Single electrode positioning, black numbers indicate unipolar derivation with reference electrodes behind ears (except channel 1), white numbers indicate bipolar derivation.* ***right****: Electrode array positioning, large array is positioned on the cheek, small array under the chin. See text for details.*

For the recording of the EMG signals, we used two different types of setups: a *single electrode* system and a novel *electrode array* approach. For the single electrode setup we used a computer-controlled 6-channel EMG data acquisition system (Varioport, Becker-Meditec, Germany). We captured signals from 1) the levator angulis oris, 2) the zygomaticus major, 3) the platysma, 4) the anterior belly of the digastric and 5) the tongue, see Fig. 1 (left) for the electrode positioning. All EMG signals were sampled at 600 Hz and filtered with an analog high-pass filter. The electrode positioning which yielded optimal results was adopted from [10].

The electrode array acquisition device (EMG-USB2, OT Bioelettronica, Italy) recorded the EMG signals using a large array of four rows of eight electrodes with 10mm inter-electrode distance (IED) and a second smaller array with one row of eight electrodes with 5mm IED. As illustrated in Fig. 1 (right) the large array was placed on the subject's cheek, while the smaller one was positioned under the chin. The array signals were sampled at 2048 Hz, using a bipolar derivation, where the activation differences between two adjacent channels in a row are calculated. We therefore obtain a total of 35 signal channels out of the 4×8 cheek electrodes and the 8 chin electrodes [9].

In addition to the EMG signal, we simultaneously recorded the audio signal with a standard close-talking microphone at a sampling rate of 16kHz. The audio signal is synchronized to the EMG signal with an analog marker. This is a prerequisite for building our database of corresponding EMG and speech segments.

The text corpus is based on [11] and consists of phonetically balanced English sentences which originated from the broadcast news domain. Each session was split into a *train* and *eval* set. The latter contains at least 10 different test sentences (plus repetitions), which are kept fixed across all sessions. For recording the data, the speaker read all prompted utterances in normal, audible speech in randomized order. This was supervised by a recording assistant to assure proper pronunciation and to guarantee a stable signal quality.

Table 1 gives detailed information about the durations of the recorded utterances.

Table 1: *Data corpus information for the recorded utterances, including speaker/session breakdown.*

Session	Accumulated data length, in (mm:ss)		# of train/eval utterances	
	Train	Eval	Train	Eval
Spk1-Single	27:10	01:19	500	20
Spk2-Single	26:54	00:49	496	13
Spk1-Array	31:01	00:47	500	10
Spk2-Array	25:44	01:10	500	20
Total	110:49	04:05	1996	63

3. Unit selection approach

The target of the proposed unit selection approach is to generate a natural-sounding waveform that resembles the original utterance. For this purpose we first build a database from simultaneously recorded EMG and audio signals. We extract segments of a fixed frame number, similar to Fig. 2. We refer to this number of frames as *unit width* w_u. To create a higher number of segments in the database, we shift the segments by 1 frame at a time, rather than shifting by the whole unit. Together, the EMG segment (*source*) and the associated speech segment (*target*) form one *unit*. Which features are used to represent speech and muscle movement is described in Sec. 4.1 and Sec. 4.2, respectively.

During the conversion phase, the EMG test sequence is also split up into overlapping segments of the same width, as shown in Fig. 2. We also vary the shift between two consecutive segments. The effect of this *unit shift* s_u is shown in Sec. 4.5. As default for our experiments we chose $w_u = 11$ and $s_u = 10$.

The desired units are then chosen from the database by means of two cost functions, the *target cost* c_t and the *concatenation cost* c_c [6]. The target cost measures the similarity between a test segment and a database unit. It is calculated as the mean Euclidean distance between the respective source segments $s_{test}^{(t)}$ and $s_{db}^{(t)}$:

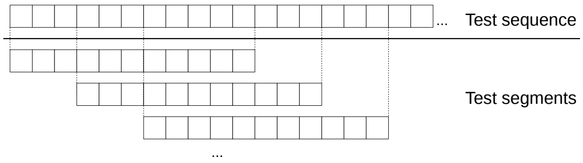

Figure 2: *Splitting of the test sequence, here with unit width $w_u = 11$ and unit shift $s_u = 3$.*

$$c_t = \frac{1}{w_u} \sum_{k=1}^{w_u} \sqrt{\sum_{d=1}^{D_E} (s_{test}^{(t)}(k,d) - s_{db}^{(t)}(k,d))^2}, \qquad (1)$$

where D_E denotes the dimensionality of the EMG features and $s^{(t)}(k,d)$ denotes the d-th dimension of the k-th frame of the segment at time index t.

The concatenation cost is based on the cepstral distance at the point of concatenation to ensure a proper smoothness of the acoustic output. It is calculated as the mean Euclidian distance between the overlapping frames of the target segments of two database units $t_{db}^{(t)}$ and $t_{db}^{(t+1)}$:

$$c_c = \frac{1}{o_u} \sum_{k=1}^{o_u} \sqrt{\sum_{d=1}^{D_A} (t_{db}^{(t+1)}(k,d) - t_{db}^{(t)}(k+s_u,d))^2}, \qquad (2)$$

where D_A denotes the dimensionality of the audio features and $o_u = w_u - s_u$ is the number of overlapping frames of two units. If two units have a natural transition, meaning they originated from the same utterance with starting points i and $i + s_u$, they are favored because the overlapping frames are the same, resulting in a concatenation cost of 0.

Additionally, a weight for the two cost functions is needed to balance naturalness and distinctiveness.

The search for the optimal unit sequence is implemented as a Viterbi search through a fully connected network [12]. The goal is to minimize the overall cost of the chosen sequence. In our experiments, the Viterbi search was limited to a maximum of 100 active paths. This constraint decreased the output quality only marginally while considerably reducing the computation time.

After determining the optimal unit sequence, the overlapping audio segments are smoothed using a weight function as proposed by [13]. We define n as the number of units which share a frame, illustrated by the hatched frames of the chosen segments in Fig. 3. Since this number of units n varies depending on the unit shift, the weight w for each unit's affected frame is calculated as follows:

$$w[i] = \frac{exp(-0.2 \cdot a[i])}{\hat{w}}, \; i = 1 \ldots n, \qquad (3)$$

with

$$a[i] = \begin{cases} [\frac{n}{2}, \frac{n}{2} - 1, \ldots, 1, 1, \ldots, \frac{n}{2} - 1, \frac{n}{2}], & n \text{ even} \\ [\lceil \frac{n}{2} \rceil, \lceil \frac{n}{2} \rceil - 1, \ldots, 1, \ldots, \lceil \frac{n}{2} \rceil - 1, \lceil \frac{n}{2} \rceil], & n \text{ odd,} \end{cases} \qquad (4)$$

where $\hat{w} = \sum_{i=1}^{n} exp(-0.2 \cdot a[i])$ is used for normalizing the weights to sum to 1. Fig. 3 shows an example of the smoothing process: The hatched frames of the chosen segments are weighted and added up to create one output frame. This process is repeated at each frame.

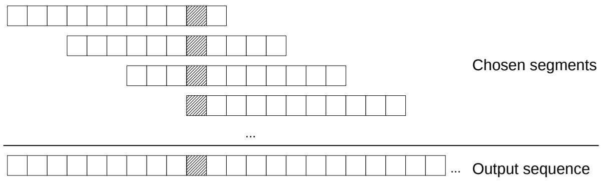

Figure 3: *Creating the output sequence from the chosen audio segments.*

4. Experiment setup

4.1. Acoustic features

In the acoustic signal domain, an excitation-filter model of speech is considered. 25 Mel-cepstral coefficients (MCEPs) [14] are extracted as filter parameters and fundamental frequency (F_0) estimates are derived as excitation features for every 10 ms in a frame with 27 ms length. These features represent the acoustic speech unit. To finally synthesize the converted speech signal we use two different methods:

- The Mel Log Spectrum Approximation (MLSA) filter method [15], which takes the generated F_0 and MCEPs as input.
- Choosing the optimal unit sequence as described in Sec. 3 and using the obtained time stamps to slice and concatenate the corresponding original waveforms of the training utterances. This direct concatenation method uses the speech segments directly and is therefore done without further signal processing.

4.2. Electromyographic features

We evaluate a feature which is based on a composition of *time-domain features* [16]. For a given feature $\mathbf{f}$, $\bar{\mathbf{f}}$ is its frame-based time-domain mean. $\mathbf{P_f}$ is the corresponding frame-based power, and $\mathbf{z_f}$ is the frame-based zero-crossing rate. $S(\mathbf{f}, n)$ is the stacking of adjacent frames of the feature $\mathbf{f}$ in the size of $2n + 1$ ($-n$ to n) frames, which is used in order to account for time-context information.

For an EMG signal with normalized mean $x[n]$, the nine-point double-averaged signal $w[n]$ is defined as

$$w[n] = \frac{1}{81} \sum_{k=-4}^{4} \sum_{l=-4}^{4} x[n + k + l] \qquad (5)$$

The high-frequency signal is $p[n] = x[n] - w[n]$, and the rectified high-frequency signal is $r[n] = |p[n]|$. The final feature $\mathbf{TDn}$ is defined as follows:

$$\mathbf{TDn} = S(\mathbf{TD0}, n), \text{where } \mathbf{TD0} = [\bar{\mathbf{w}}, \mathbf{P_w}, \mathbf{P_r}, \mathbf{z_p}, \bar{\mathbf{r}}] \qquad (6)$$

This $\mathbf{TDn}$ feature is computed from each of the EMG channels, then a fused vector is formed by stacking each channel vector, and a Linear Discriminant Analysis (LDA) is applied to reduce the dimensionality of the final feature vector to 32. The LDA matrix is computed on the train data of each session on the 45 English phones, plus one silence phone. We use this feature because it performs better than simple features tested in the preparation phase and also to be able to compare the results to a frame-based approach from our prior work [8].

It is known (e.g., [17, 18]) that there exists an anticipatory effect of the EMG signals compared to the simultaneously recorded speech signals. We model this anticipatory effect by adding a time delay of 50ms to the EMG signals.

Note that the electrode array recordings provide 35 channels, instead of six EMG channels given by the single electrodes. We visually inspected the EMG signals and discarded the noise channels to finally use a set of 18 channels. Further details about the positioning and processing of the electrode array signals can be found in [9].

4.3. Experimental results

For evaluation of the proposed EMG-to-speech conversion we use the Mel Cepstral Distortion (MCD) [19]. The MCD is a scaled Euclidean distance between the spectral features of the target audible speech and the spectral features of the synthesized EMG speech in decibel. Smaller numbers implicate better results.

First, the MCD is computed for each frame, then it is averaged over all frames of an utterance. Note that the source EMG signal and the target audio signal are recorded synchronously, hence the converted audio signal and the target audio signal are automatically aligned as well and we do not need to perform any alignment here.

The F_0 estimation accuracy is evaluated by a voiced/unvoiced decision rate. U/U and V/V give the amount of all frames that are correctly recognized as unvoiced or voiced, respectively.

4.4. Audible/audible units

To evaluate the general feasibility of our approach and to investigate whether the amount of training data is sufficient, we performed an oracle experiment, where the audible features are taken as source **and** target features. The source segment of each unit consists of the audible MCEPs and the target segment contains the MCEPs plus the F_0 feature. We refer to this method as *Aud2Aud*. The Unit Selection system is trained as described in Sec. 3 and tested on unseen data to investigate if enough units are available to synthesize a proper acoustic output. We obtain an intelligible speech output, indicated by a mean MCD of 4.14. 90.6% of the voiced and 89.1% of the unvoiced frames are classified correctly. The results on all 4 sessions are shown in Table 2.

Table 2: *Mel cepstral distortions and Voiced/Unvoiced (V/U) accuracies with Aud2Aud mapping approach for the single electrode (S) and array (A) recording sessions.*

	Spk1-S	Spk2-S	Spk1-A	Spk2-A
MCD	4.08	4.14	3.95	4.40
V-Acc	91.25%	89.75%	92.16%	89.11%
U-Acc	85.71%	88.78%	91.36%	90.74%

The top of Fig. 6 shows the exemplary spectrograms of a synthesized output, generated with the MLSA-filter method, the direct wave-segment concatenation method, as well as the target audible utterance.

4.5. EMG/audible units

Our observations from the Aud2Aud mapping show that the amount of training data is sufficient to generate a proper speech output and that an F_0 accuracy of around 90% can be achieved. We therefore applied our framework to an electromyographic feature input. We evaluated different frame sizes and stack-

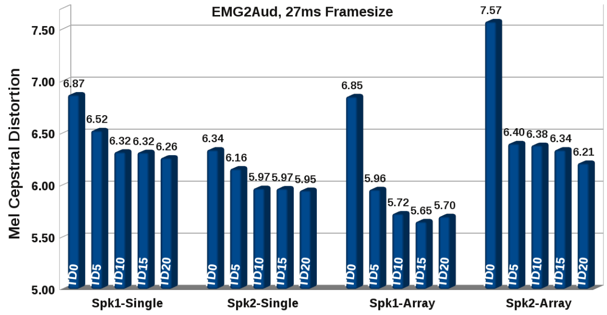

Figure 4: *Mel cepstral distortions of EMG2Aud mapping approach, using different stacking widths.*

ing widths n for the $\mathbf{TDn}$ feature. Fig. 4 shows the results of the EMG-to-speech conversion for different contextual stacking widths with 27ms frame size, giving a mean MCD of 6.03 with $\mathbf{TD20}$. We also evaluated a frame size of 48ms, but obtain only slight differences to the results with 27ms.

Fig. 5 shows the results of the F_0 evaluation. Given are the amount of frames that are recognized as voiced (V) or unvoiced (U) frames. E.g. V/U denotes the amount of frames that are recognized as voiced, but are unvoiced in the reference. Using a higher stacking width improves the voicing accuracy, although no clear stacking optimum can be observed. Accuracies around 75% result in proper synthesized outputs, but still show room for improvement. An example for a synthesized output can be seen at the bottom of Fig. 6.

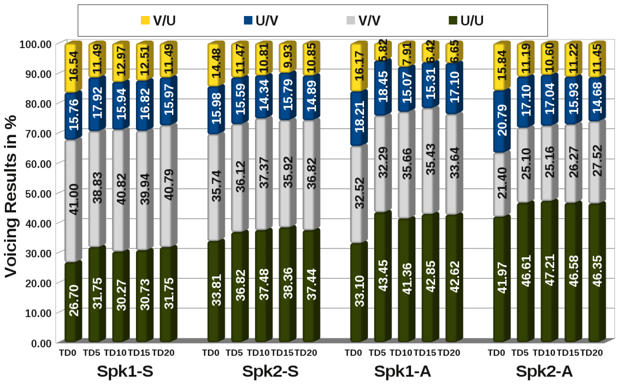

Figure 5: F_0 *results (top-to-bottom:V/U,U/V,V/V,U/U V=Voiced, U=Unvoiced Frames, order is hypothesis/reference) with EMG2Aud mapping approach, using different stacking widths.*

As a final experiment, we also evaluated the effects of different unit widths and shifts on the output performance. It can be observed that reducing the shift between the units gives a clear decrease of the MCD (about 10% relative improvement) for all recording sessions, whereas the variation of the unit width has only a slight effect on the output. On average, the reduction of the unit width from 11 to 9 frames yielded a relative improvement of 1.19%. The results for all investigated unit widths and shifts are shown in Fig. 7.

A comparison of these results to previous work [8], where a frame-based EMG-to-speech mapping is used and the same data (session *Spk1-S* and *Spk2-S*) is shared, shows that we obtain a MCD of 5.75 resp. 5.19 with our unit selection approach, compared to 5.92 resp. 5.6.

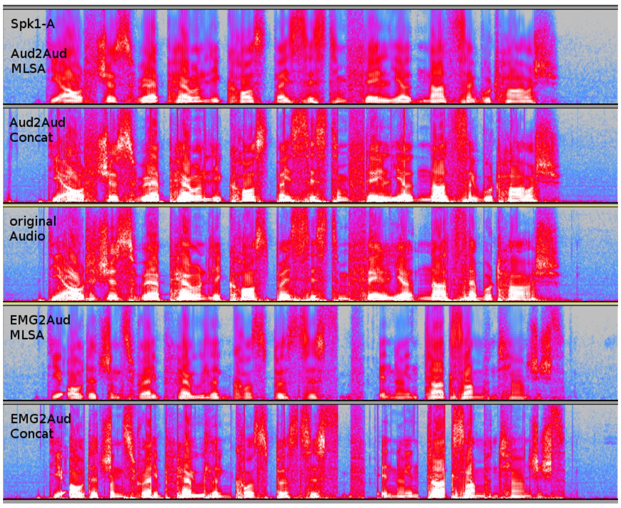

Figure 6: *Spectrograms of target audible speech (middle), Aud2Aud synthesis output (top) and EMG2Aud synthesis output (bottom), both using the MLSA-filter method and the direct wave-segment concatenation method of the utterance "The outages were apparently caused by system failure not sabotage."*

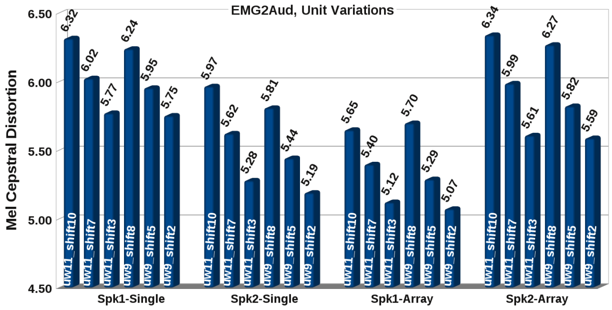

Figure 7: *Mel cepstral distortions of EMG2Aud unit selection approach with TD15 features, using different unit widths (uw) and unit shifts. See text for details.*

5. Conclusions and future work

We successfully showed the feasibility of a unit selection approach to convert surface EMG signals of the articulatory muscles to audible speech. Using only a small amount of training data (approx. 30 min) yields promising conversion results with an average MCD of 5.4. There will be more work on evaluating the framework by conducting listening tests including a more detailed comparison to the frame-based approach [8]. To further improve the conversion framework we plan to extend the amount of data, to use EMG and acoustic signals from multiple recording sessions and to evaluate different cost functions for target and concatenation cost.

The unit selection approach gives us a number of other opportunities for improvement, like the usage of additional information in the unit database (e.g., part-of-speech tagging, additional prosodic information) or even the possibility to integrate other signal sources to enable a multi-modal silent speech interface.

6. References

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, and J. Gilbert, "Silent Speech Interfaces," Speech Communication, 52(4):270 – 287, 2010.

[2] A. D. C. Chan, K. Englehart, B. Hudgins, D. F. Lovely, D.F., "Hidden Markov Model Classification of Myolectric Signals in Speech", IEEE Engineering in Medicine and Biology Society, vol. 21, no. 5, pp. 143–146, 2002.

[3] A. R. Toth, M. Wand, and T. Schultz, "Synthesizing Speech from Electromyography using Voice Transformation Techniques," Proceedings of Interspeech 2009, pp. 652–655, 2009.

[4] T. Toda, H. Saruwatari, and K. Shikano, "Voice Conversion Algorithm based on Gaussian Mixture Model with Dynamic Frequency Warping of STRAIGHT Spectrum", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 841–844, 2001.

[5] K. Nakamura, M. Janke, M. Wand, and T. Schultz "Estimation of Fundamental Frequency from Surface Electromyographic Data: EMG-to-F0," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),pp. 573–576, 2011.

[6] A. J. Hunt and A. W. Black, "Unit Selection in a Concatenative Speech Synthesis System using a Large Speech Database", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 373–376, 1996.

[7] T. Toda, K. Shikano, "NAM-to-Speech Conversion with Gaussian Mixture Models", European Conference on Speech Communication and Technology (Interspeech), pp. 1957–1960, 2005.

[8] M. Janke, M. Wand, K. Nakamura, and T. Schultz, "Further Investigations on EMG-to-Speech Conversion", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 365–368, 2012.

[9] M. Wand, C. Schulte, M. Janke, and T. Schultz "Array-based Electromyographic Silent Speech Interface", International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS), pp.89–96, 2013.

[10] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, "Session Independent Non-Audible Speech Recognition Using Surface Electromyography", IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 331–336, 2005.

[11] T. Schultz, and M. Wand, "Modeling Coarticulation in Large Vocabulary EMG-based Speech Recognition", Speech Communication, 52(4):341 – 353, 2010.

[12] A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm, IEEE Transactions on Information Theory, 13(2), pp.260–269, 1967.

[13] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li, "Exemplar-based Unit Selection for Voice Conversion utilizing Temporal Information", Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech), pp. 3057–3061, 2013.

[14] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An Adaptive Algorithm for Mel-Cepstral Analysis of Speech," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 137–140, 1992.

[15] S. Imai, "Cepstral Analysis Synthesis on the Mel Frequency Scale", IEEE International Conference on Acoustics, Speech, and Signal Processing Vol. 8, pp. 93–96, 1983.

[16] S. -C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards Continuous Speech Recognition using Surface Electromyography,", Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech), pp. 573–576, 2006.

[17] S. -C. Jou, L. Maier-Hein, T. Schultz, and A. Waibel, "Articulatory Feature Classification using Surface Electromyography", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 605–608, 2006.

[18] E. J. Scheme, B. Hudgins, and P.A. Parker, "Myoelectric Signal Classification for Phoneme-based Speech Recognition", IEEE Transactions on Biomedical Engineering, 54(4), pp. 694–699, 2007.

[19] R. F. Kubichek, "Mel-Cepstral Distance Measure for Objective Speech Quality Assessment", IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, pp. 125–128, 1993.