Interspeech 2018
2-6 September 2018, Hyderabad, India

Investigating Objective Intelligibility in Real-Time EMG-to-Speech Conversion

Lorenz Diener, Tanja Schultz

Cognitive Systems Lab, University of Bremen, Germany
lorenz.diener@uni-bremen.de

Abstract

This paper presents an analysis of the influence of various system parameters on the output quality of our neural network based real-time EMG-to-Speech conversion system. This EMG-to-Speech system allows for the direct conversion of facial surface electromyographic signals into audible speech in real time, allowing for a closed-loop setup where users get direct audio feedback. Such a setup opens new avenues for research and applications through co-adaptation approaches. In this paper, we evaluate the influence of several parameters on the output quality, such as time context, EMG-Audio delay, network-, training data- and Mel spectrogram size. The resulting output quality is evaluated based on the objective output quality measure STOI.

Index Terms: silent speech interfaces, surface electromyography, speech synthesis

1. Introduction

Speech is the most efficient and natural means for human communication. With advances in speech processing and machine learning, speech-based interfaces have grown in importance and entered everyday use. Still, there are situations where audible spoken communication and interfaces based on acoustic speech cannot be used:

- In the presence of loud noise, speech communication is severely hampered or entirely impossible.
- In public places or quiet environments, audible speech is undesirable.
- For speech impaired people (e.g. Laryngectomees), producing audible speech is not possible.

Speech interfaces that do not rely on the presence of an audible acoustic signal offer an alternative approach to processing speech that can mitigate some of the issues mentioned above. Such *Silent Speech Interfaces* (SSIs) use a range biosignals other than microphone-recorded audible speech to infer information from speech [1, 2]. Examples of such biosignals include Permanent Magnetic or Electromagnetic Articulography [3, 4, 5], lip reading from video [6], ultrasound imaging of the speech apparatus [7], electroencephalography- or functional near infrared spectroscopy based brain-computer-interfaces [8, 9] and, as in our work, surface electromyography [10] – the recording of electrical muscle activity using surface electrodes attached to the face.

Our *surface electromyography* (sEMG) based EMG-to-Speech SSI turns recorded sEMG signals directly into audible speech without an intermediate recognition step. Compared to a recognize-and-synthesize approach, this direct synthesis approach has several advantages:

- As there is no recognition vocabulary, the direct synthesis approach allows for operation independent of language constraints.

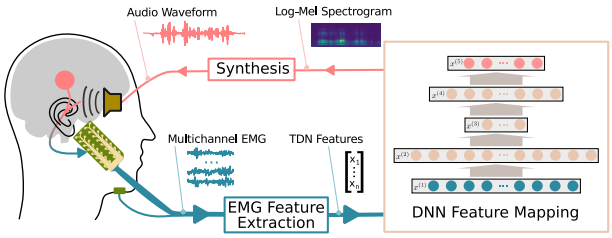

Figure 1: *Schematic of our EMG-to-Speech conversion system: A multichannel EMG signal is recorded from which TDN features are extracted. These features are then transformed to Mel spectra from which an audio waveform can be synthesized.*

- It allows for the retention of paralinguistic information, such as prosody and accentuation.
- The system output could be used to directly control conventional audible speech interface without the need for modification of those interfaces.
- There is no need to wait until an entire word or even sentence has been recognized – the system can output audio immediately.

For several reasons, we consider *real-time*, *low-latency* operation a crucial feature to further advance Silent Speech Interfaces:

- Practical uses such as silent telephony or speech restoration require that an audible speech signal is generated during speech production in real time. Additionally, delayed auditory feedback is known to inhibit speaking ability [11], necessitating low-latency operation.
- To train our EMG-to-Speech system, we currently rely on synchronously recorded sEMG and audible speech signals. However, it is known that muscle activity can differs significantly between audible and silent speech production due to the lack of auditory feedback [12]. This effect causes mismatches when the trained EMG-to-Speech system is applied to silently produced speech. Real-time low-latency auditory feedback is likely to weaken or eliminate the muscle activity differences.
- Due to the aforementioned silent-audible differences and to reduce setup and training time, it is necessary to explore means by which the system can adapt to the user or the user can adapt to the system (co-adaptation). This requires real-time feedback.

In our previous work, we have described a basic neural network based EMG-to-Speech conversion system, performing offline conversion [13, 10], as well as a first real time system [14]. While original analysis of the real-time system focused on timing performance, this paper will examine the impact of different parameters on output quality using small

10.21437/Interspeech.2018-2080

amounts of training data. This is important as the sEMG signal varies with speaker characteristics (e.g. tissue, muscles, vocal apparatus) and electrode positioning. For system performance reasons, recording training data, training or adapting the system with that data are currently performed within the same session, limiting the amount data that can be used.

2. Experiment Setup

2.1. EMG-to-Speech Conversion

The basic setup of our EMG-to-Speech conversion system can be seen in Figure 1. We record a multichannel surface EMG signal using an OT Bioelettronica Quattrocento EMG amplifier, with a 4 x 8 10 mm inter-electrode distance (IED) electrode array on the users cheek and an 8 electrode 5 mm IED strip below the chin. The signal is band-pass filtered to 10–900 Hz and sampled at 2048 Hz for digital processing. For training the system, an audible speech audio signal is recorded in parallel with the EMG signals. The two signals are time-synchronized using the EMG amplifiers trigger output.

2.1.1. EMG Features

From the raw multichannel EMG signal, we extract 32ms windows, with a frame shift shorter than that (examined in section 3.2), using a Blackman window. For each such window we extract a number of time domain (TD) features that have been shown to work well for EMG-based speech processing [15].

- Low frequency (up to 134 Hz) power
- Low frequency (up to 134 Hz) mean
- High frequency (above 134 Hz) power
- High frequency (above 134 Hz) zero-crossing rate
- High frequency (above 134 Hz) rectified mean

These features are then stacked with N past feature vectors for time context, resulting in a final set of modified TDN features. Note that, unlike standard TDN features, where the current frame is stacked with frames from the past and future, we can only stack with past frames, since future cannot be obtained for low latency run-on processing.

2.1.2. Audio Features

For the feature transformation, our system represents audio as a series of Mel spectrogram frames. For training the system, these features are calculated from the synchronously recorded audio by first windowing the signal with the same parameters as the EMG signal. We then compute the magnitude spectrogram for each window and finally apply a Mel filterbank to obtain the Mel spectrogram.

To turn a sequence of Mel spectrograms back into an audio waveform for playback, we first invert the Mel-filter. We then synthesize the waveform using the method proposed by Griffin and Lim [16] (iteratively generating phase information starting from a random signal, computing the short-term Fourier transform and replacing its magnitude spectrum with the input spectrum), implemented to operate on a continuous stream of data.

2.1.3. Feature transformation

To convert EMG TDN features to audio Mel features, our system uses a 5 layer (input $\rightarrow$3 hidden $\rightarrow$output) neural network architecture. The structure follows an hourglass shape with a bottleneck in the central layer. To prevent overfitting, dropout regularization is applied to every layer other than the output layer. The network is trained for up to 80 epochs, with early stopping being employed to abort training when the validation set error did not improve for 5 epochs, using stochastic gradient descent with a learning rate of 0.0001, a momentum of 0.9 and a mini-batch size of 1024 samples. Input and output features are normalized to zero mean and unit variance.

2.2. Data

Due to differences in electrode and skin condition, there is a high amount of variability between different EMG recording sessions even when signals are recorded from the same subject. To limit variability, we record the training utterance and perform the mapping within the same session, i.e. electrode positions remain fixed. In practice, this limits the amount of data that can be used to train the system, since every additional utterance adds both recording and training time.

We evaluate our system on two sessions recorded with the live setup. A session consists of audible speech and synchronous sEMG signals of read German-accented English speech from the broadcast news domain. Each session contains 300 utterances with an average length of 4.09 seconds per utterances. From this, 50 utterances were selected for evaluation, split into 35 utterances for hyperparameter evaluation and 15 utterances for verifying results. From the remaining utterances, 15 were used as validation data and to perform early stopping during training. To evaluate the effect of adding more training data on the system, subsets of size 35 ($\sim$2.5 minutes), 135 ($\sim$9 minutes) or 235 ($\sim$16 minutes) of the utterances remaining after that were used for training the system.

3. Evaluation

3.1. Evaluation measure

To evaluate the performance of different sets of parameters, we employ the Short-Time Objective Intelligibility (STOI) index, a measure of time-frequency similarity that is known to correlate well with subjective intelligibility of noisy speech [17]. The STOI is calculated as a linear correlation of grouped discrete Fourier transform bins and ranges from 0 to 1, with higher values indicating better intelligibility.

3.2. Frame shift and Mel spectrogram dimensionality

Both frame shift and the number of coefficients in the Mel spectrogram can affect the output quality of a speech synthesis system. In this paper, we evaluate frame shifts of 2 ms, 5 ms and 10 ms. For each, we evaluate the dimensionality, with coefficient counts of 20, 50, 75 and 100. We evaluate performance both when directly re-synthesizing audio (i.e. quality just from extracting features and recreating audio waveforms, with no EMG-to-Speech mapping involved) as well as when performing EMG-to-Speech transformation (other parameters held constant at reasonable defaults – compare the other experiments in this paper as well as previous work [15] – with an EMG-Audio delay of 50 ms, a training set size of 135 utterances and a 2048 $\rightarrow$1024 $\rightarrow$2048 neuron network). The results of this evaluation are shown in Figure 2. Clearly, the results on audio resynthesized from reference Mel spectra are still vastly better than the EMG-to-Speech conversion results (note the different y-axis range). While frame shift and number dimensionality have an impact on the reference Mel spectrogram, there is lit-

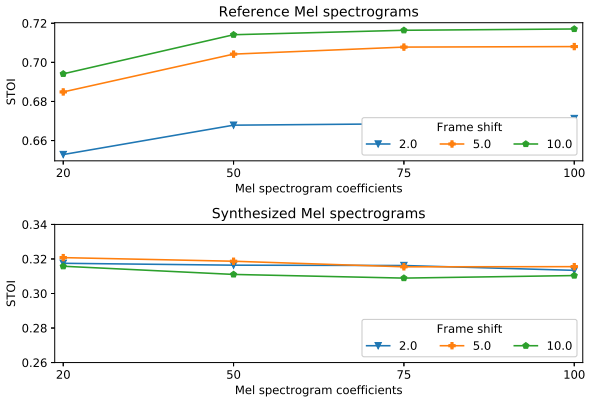

Figure 2: *STOI for different frame shifts and Mel coefficient counts, evaluated for audio synthesized from reference Mel spectra (on top) and Mel output of EMG-to-Speech conversion (on bottom).*

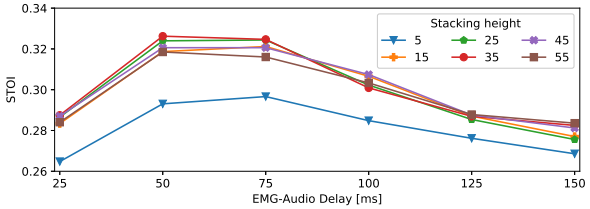

Figure 3: *STOI for different stacking heights and EMG-Audio delays.*

tle difference between different frame shifts for the synthesized Mel spectra, and the effect of the dimensionality of the Mel spectrogram seems negligible within the examined range.

3.3. EMG context stacking and EMG-Audio delay

The sEMG signal precedes muscle movement with a time delay depending on muscle and type of movement [18] – this effect is known as *electromechanical delay* (EMD). For EMG-to-Speech conversion, this means that we need to account for a delay between the EMG and the audio signal. There are two ways in which this can be done: The EMG signal can be shifted against the audio signal by some amount (50 ms has been found to be a good shift to use for EMG-based speech processing [15]), or stacking can be employed to increase the time context available within each EMG feature vector so that all important information for synthesizing audio features for the current frame is available. In the context of a run-on, real-time system, EMD can be an advantage for sEMG based SSIs: If the sEMG signal is processed within the EMD, the system's outputs audible speech at the time the listener expects the original voice.

To evaluate the effect of EMG-Audio delay and stacking height on output quality, we performed EMG-to-Speech conversion with delays from 25 ms to 150 ms and stacking heights between 5 and 55 frames, with the other parameters held constant (5 ms frame shift, 135 training utterances, 2048 →1024 →2048 neuron network). The results of this evaluation can be found in Figure 3.

We find a delay of around 50 ms to be ideal, with no improvement beyond 75 ms (on the contrary, beyond a delay of 100 ms, the STOI decreases, likely due to important information being shifted back outside the stacked feature vector). Higher

Table 1: *Neural network layer sizes (compare Figure 1).*

	Architecture		
Layer	1	2	3
$x^{(1)}$	(EMG dimensionality)		
$x^{(2)}$	1024	2048	4096
$x^{(3)}$	512	1024	2048
$x^{(4)}$	1024	2048	4096
$x^{(5)}$	(Audio dimensionality)		

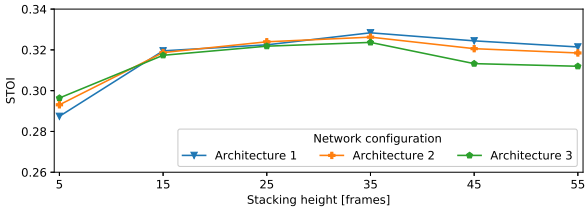

Figure 4: *STOI for different stacking heights and network configurations (compare section 3.3).*

stacking heights dampen the influence of the EMG-Audio delay somewhat. 25 to 35 frames of past context appear to be sufficient to contain all information useful for EMG-to-Speech transformation.

3.4. EMG context stacking and network size

In our neural network based offline EMG-to-Speech conversion system [13], we have found an "hourglass" shaped network architecture (modeling a feature extractor – regressor architecture) to work well. Here, we evaluate how different variations of this architecture perform for given input dimensionalities (e.g. EMG context stacking heights).

We evaluate three architectures with different amounts of neurons for the inner layers – compare Table 1 for the specific sizes used. For each, we evaluate stacking heights of 5 to 55 frames (Other parameters are held constant at 50 ms EMG-Audio delay, 5 ms frame shift and a train set size of 150 ms). The results of this evaluation can be seen in Figure 4.

As in section 3.3, increased context seems to slightly improve conversion quality. Increased network size beyond the 2048 →1024 →2048 (Architecture 1) neuron network seems to have a detrimental effect.

3.5. Network size and amount of training data

Larger networks means a larger amount of parameters, which are best trained with more data samples. We evaluate the effect of adding more training data to our system by performing EMG-to-Speech conversion with training set sizes of 35 utterances, 135 utterances and 235 utterances, for stacking heights of both 15 and 45 frames and for the same network configurations and with the same constant parameters as in section 3.4. Figure 5 shows the results of this evaluation. Adding more data increases quality, however, even the largest evaluated data set and high input dimensionality did not allow for the larger network configurations to outperform the smaller ones.

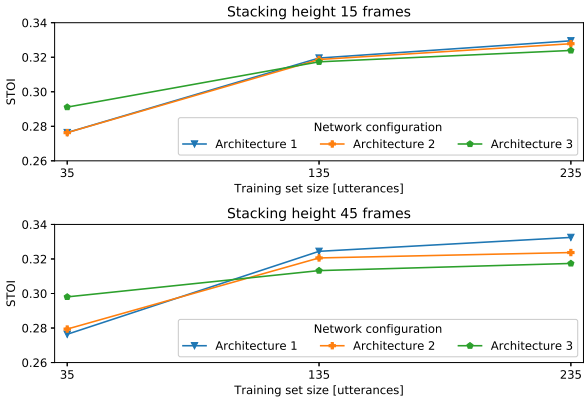

Figure 5: *STOI for different training set sizes and network configurations (compare section 3.3), for a context stacking height of 15 frames (on top) and 45 frames (on bottom).*

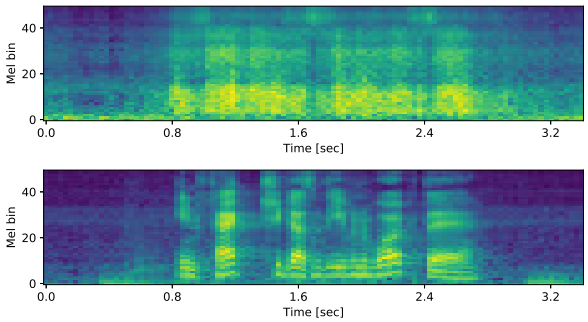

Figure 6: *Spectrogram view of the real-time system output (on top) and reference audio (on bottom).*

3.6. Verification of parameters, latency and spectrogram comparison

We verify our results on 15 utterances held out for this purpose, selecting parameters found to be good in the rest of this paper: 5 ms frame shift, 50 ms EMG-Audio delay, a stacking height of 35 frames, 50 Mel coefficients, 250 training utterances and architecture 1. We find a STOI of ~0.313, compared to a STOI of ~0.338 for the 35 utterance set used in parameter optimization, indicating some overfitting of parameters. To increase the robustness of our results, we plan to collect more data in the future.

Fig. 6 shows a set of aligned Mel spectrograms of reference audio and EMG-to-Speech conversion output from the verification run. While a large amount of noise remains and details are lost, it can be seen that the basic low-frequency structure of the spectrogram can be recreated in real-time.

We additionally measure the latency of our system in this configuration with a test signal. We find that the delay between an EMG signal going into the amplifier and the system producing an audio response is ~79 ms (measurement repeated five times). Accounting for EMD, this is well within the range of allowable delays for live speech feedback.

4. Discussion

Creating an EMG-based speech interface that can be trained and used within one session and that runs in real time requires trade-offs.

In section 3.2 we examined the influence of frame shift on output quality. Our analysis of synthesis from the reference Mel spectra shows that, for high-quality input data, a frame shift of 10 ms is a good choice. For the synthesized Mel spectrograms, there is little difference. Overall, a shift of 5 ms seems like a reasonable choice for further experiments. The dimensionality of the Mel spectrogram seems to have little influence on output quality beyond 50 coefficients and even less influence for Mel spectra generated from EMG.

Section 3.3 explored the relationship between EMG-Audio delay and stacking height. We find that more context is beneficial, but this is, of course, at the price of a higher input feature dimensionality. Compared to Jou et al. [15], we too find a delay of around 50 ms but possibly up to 75 ms to give best results. the implications of an increased delay (and thus, increased future context) on latency remain to be investigated.

We further investigated the effects of context stacking in sections 3.4 and 3.5, again finding that adding more context has a positive effect on output quality. We also find that the learning behaviour of our different network architectures does not follow a clear pattern with regards to their size, leading us to believe that, while the current architectures do perform well for Mel-Frequency Cepstral Coefficients and larger sessions, a close look at their performance with regards to Mel spectrum output on small data sets might be necessary.

Adding more training data to a machine learning system is often the easiest and most fruitful way to improve performance. In the case of our EMG-to-Speech transformation system, it is not that simple: As the sEMG signal is highly session-dependent, all data that is used to train a system for a session must presently be recorded in that session, meaning that any additional data increases pre-use recording and training time. In section 3.5, we found that adding more data improved performance. While initially, a small amount of training data may be preferable, collecting a larger amount of training data may be a good way to increase system robustness. Other possible solutions that we will explore in the future are signal level adaptation [19] and initializing the real-time system with session-independent models [20] to use the incoming training data for unsupervised adaptation during run-time.

Finally, visual and aural inspections of system output reveal a large amount of background noise in the system output. One factor contributing to this may be due to within-session changes in signal quality, either gradual (e.g. drift in electrode-skin impedance) or sudden (e.g. partial electrode detachment). While we already try to lessen the impact of such problems (e.g. by using a high-pass filter for DC offset removal), it may be prudent to further investigate methods by which such changes can be detected and compensated for.

5. Conclusion

In this paper, we have given a brief overview of our real-time EMG-to-Speech conversion system. We have examined how different parameters influence the output quality of this system and discussed the trade-offs involved in converting facial sEMG data to audible speech in real time and with a latency short enough that, accounting for EMD, audio output is near instant. In the future, we hope to explore the new applications and research into co-adaptation that were not possible with an offline batch conversion system.

6. 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, K. D. J., 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, Nov 2017.

[3] J. A. Gonzalez, L. A. Cheah, J. M. Gilbert, J. Bai, S. R. Ell, P. D. Green, and R. K. Moore, "Direct speech generation for a silent speech interface based on permanent magnet articulography," in *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies*, vol. 4, 2016, pp. 96–105.

[4] R. Hofe, S. R. Ell, M. J. Fagan, J. M. Gilbert, P. D. Green, R. K. Moore, and S. I. Rybchenko, "Evaluation of a Silent Speech Interface Based on Magnetic Sensing," in *Proc. Interspeech*, 2010.

[5] M. Kim, B. Cao, T. Mau, and J. Wang, "Speaker-independent silent speech recognition from flesh-point articulatory movements using an lstm neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2323–2336, 2017.

[6] R. Bowden, S. Cox, R. Harvey, Y. Lan, and E.-J. Ong, "Recent Developments in Automated Lip-Reading," *SPIE Security+ Defence. International Society for Optics and Photonics*, 2013.

[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 Communication*, vol. 52, pp. 288–300, 2010.

[8] J. S. Brumberg, A. Nieto-Castanon, P. R. Kennedy, and F. H. Guenther, "Brain-Computer Interfaces for Speech Communication," *Speech Communication*, vol. 52, pp. 367–379, 2010.

[9] 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, pp. 1–11, 2015.

[10] 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, Nov 2017.

[11] G. Fairbanks and N. Guttman, "Effects of delayed auditory feedback upon articulation," *Journal of Speech, Language, and Hearing Research*, vol. 1, no. 1, pp. 12–22, 1958.

[12] M. Janke, M. Wand, and T. Schultz, "Impact of lack of acoustic feedback in EMG-based silent speech recognition," in *11th Annual Conference of the International Speech Communication Association*, 2010.

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

[14] L. Diener, C. Herff, M. Janke, and T. Schultz, "An initial investigation into the real-time conversion of facial surface EMG signals to audible speech," in *Engineering in Medicine and Biology Society (EMBC), 2016 38th Annual International Conference of the IEEE*, Aug 2016.

[15] S.-C. S. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards continuous speech recognition using surface electromyography," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2006, pp. 573–576.

[16] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 2, pp. 236–243, 1984.

[17] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4214–4217.

[18] P. Cavanagh and P. Komi, "Electromechanical delay in human skeletal muscle under concentric and eccentric contractions," *European Journal of Applied Physiology and Occupational Physiology*, vol. 42, no. 3, pp. 159–163, 1979.

[19] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, "Session independent non-audible speech recognition using surface electromyography," in *2005 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005, pp. 331–336.

[20] M. Wand, C. Schulte, M. Janke, and T. Schultz, "Compensation of recording position shifts for a myoelectric silent speech recognizer," in *The 39th International Conference on Acoustics, Speech, and Signal Processing*, 2014, iCASSP 2014.