A comparison of EMG-to-Speech Conversion for Isolated and Continuous Speech

Lorenz Diener, Sebastian Bredehöft, Tanja Schultz

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

Abstract

This paper presents initial results of performing EMG-to-Speech conversion within our new EMG-to-Speech corpus. This new corpus consists of parallel facial array sEMG and read audible speech signals recorded from multiple speakers. It contains different styles of utterances — continuous sentences, isolated words, and isolated consonant-vowel combinations — which allows us to evaluate the performance of EMG-to-Speech conversion when trying to convert these different styles of utterance as well as the effect of training systems on one style to convert another. We find that our system deals with isolated-word/consonant-vowel utterances better than with continuous speech. We also find that it is possible to use a model trained on one style to convert utterances from another — however, performance suffers compared to training within that style, especially when going from isolated to continuous speech.

1 Introduction

Speech is the most natural form of human communication. Its use when communicating person-to-person is intuitive and speech-based human computer interfaces have become commonplace, being integrated into cellphones, smart speaker devices and general purpose computers. Thanks to advances in machine learning and processing power, such interfaces have not only been getting more common, but have also been getting ever closer to the ideal of just asking the computer a question as you would a fellow human.

Still, the ubiquity of these *speech interfaces* has also highlighted some problems with such interfaces. As they are intended for use with audible speech, they cannot be used (or can only be used with much reduced performance) in loud environments, such as on a factory floor or in a busy airport. Conversely, in environments where silence is expected (such as a library or in public transportation), their use is also hampered. They are unsuitable for exchanging confidential information like PIN codes or passwords, as there is the danger of a bystander overhearing that information. Finally, individuals who cannot produce audible speech (e.g. laryngectomees) cannot use such interfaces.

Silent Speech Interfaces (SSIs) — speech interfaces that do not rely on the presence of an audible acoustic speech signal — promise to address these issues. These interfaces instead use one or more alternative biosignals emitted by the body during speech production to infer information about the speech production process and, ultimately, about speech [1, 2].

Many different signal modalities have been proposed for use in SSIs. One approach is starting at the source, recording speech production and perception related brain activity via invasive electrocorticography [3, 4]. Another is to utilize bone-conduction and stethoscopic microphones to record non-audible murmur [5]. Then, there is the middle path of gathering information from those parts of the body

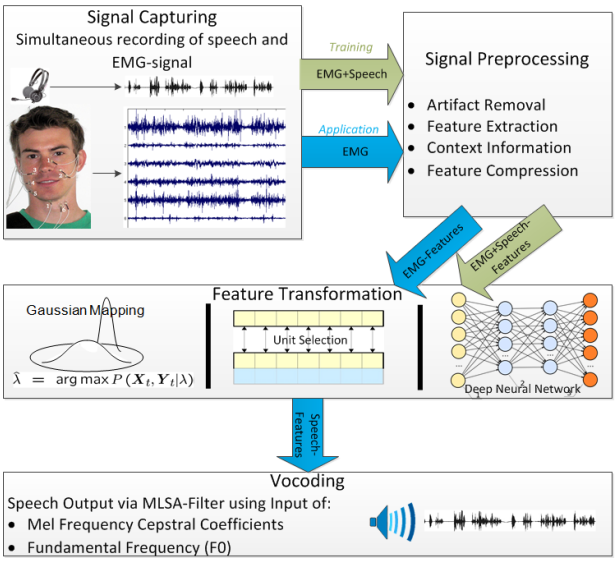

Figure 1: EMG-to-Speech conversion system overview.

that are directly involved with articulation. One approach in this space is to use ultrasound and video to capture the shape of the tongue [6] or capturing the shape of the lips [7] using video. The configuration of various articulators can also be captured using such techniques as permanent-magnetic articulography — tracking magnets attached to those articulators in 3D space [8].

The approach presented in this paper instead tracks the activity of the muscles that position the articulators, using surface electromyography (sEMG), using array electrodes [9, 10]. We then convert these sEMG signals directly to audible speech. The approach we have chosen has multiple advantages:

- Surface EMG sensing is fully noninvasive — there is no need to insert any markers or sensors into the body, attaching them to the surface of the skin is sufficient.
- Array electrodes allow for — compared to using several pairs of bi-polar electrodes — simple and fast electrode attachment.
- The *direct synthesis* approach, unlike any approaches to silent speech interfaces based on speech *recognition*, allows us to transport para-linguistic information such as intonation, stress or speaking rhythm.

In this paper, we present a new corpus for evaluating sEMG based silent speech interfaces. This new corpus contains not only, as in our previous work, continuous speech, but also different types of isolated speech. Recording these different styles of utterances allows us to perform new evaluations that were not possible using continuous speech only.

The rest of this paper is organized as follows: Section 2 gives a brief overview of our EMG-to-Speech conversion system. Section 3 presents our new corpus on which the evaluations shown in this paper have been obtained. Section

4 presents these evaluations, and they are finally discussed in section 5. Section 6 summarizes the results and presents some potential avenues for future work.

2 EMG-to-Speech Conversion

Our system performs EMG-to-Speech conversion using a statistical mapping from a set of EMG features to audio features. A high-level overview of our system can be found in Fig. 1.

2.1 EMG Preprocessing

For sEMG, we calculate a set of 5 *time-domain* (TD) features [11] for each EMG channel. For this, we first split the signal into a low-frequency and a high-frequency part at a cut-off frequency of ~134 Hz. We then extract 32 ms Blackman-windowed frames from both parts with a frame shift of 10 ms (i.e. an overlap between frame of 22 ms). Based on these, we then calculate the low-frequency frame power and mean and the high-frequency frame power, rectified mean and zero-crossing rate. To add time context, we then stack each frame with 15 frames of past context and 15 frames of future context, resulting in the final TD15 feature vector, with $5 \times 31 = 155$ dimensions per channel.

To avoid large amounts of interference from broken channels, we visually inspect each sessions EMG signal and omit channels that are consistently broken from further processing. We also removed one utterance from a session where all channels of the EMG signal were faulty, possibly due to outside electromagnetic interference.

2.2 Audio Preprocessing

To represent audio, we use a variant of the *Mel-Frequency Cepstral Coefficients* (MFCCs) that are commonly used as audio features in speech recognition and the fundamental frequency (F_0) of the audio. First, the audio is split into frames in parallel with the EMG data (i.e. again with a frame shift of 10 ms and a frame length of 32 ms). We then extract MFCCs as the parameters for a Mel-Log Spectrum Approximation [12] filter, and the F_0 values using the YIN method [13]. Together, the excitation generated from the F_0 trajectory processed by the MLSA filter allow for the recreation of the audio waveform from the feature frames.

2.3 Feature Transformation

The evaluations presented in this paper were obtained using a Deep Neural Network (DNN) based transformation of EMG to Audio. We use a feed-forward network that contains 5 layers total (three hidden layers) with ReLU units. The overall architecture of the network follows a feature-extraction-then-conversion approach, resulting in an hourglass shape. The network is trained on parallel EMG and Audio feature vectors using stochastic gradient descent. These hyper-parameters were determined on a development set in our previous work [9].

Note that, as in our previous work, our statistical models require parallel EMG and audible speech data to train and are therefore trained on EMG data recorded during audible, not silent, speech.

3 Corpus Description

3.1 Session Contents

The new corpus consist of 4 parts: Continuous speech, consonant-vowel and vowel-consonant sequences, isolated words and digits.

For **continuous speech**, we used phonetically balanced sentences established in [9]. The sentences are from the broadcast news domain. The corpus contains 390 of these sentences; 300 for training, 50 for developing, and 40 for testing purposes. The entire test and development sets from our previous work are included, allowing for result comparability.

To get a high coverage of **consonant-vowel** and **vowel-consonant** sequences (CVs and VCs, respectively) with regards to our continuous speech block, we statistically examine the continuous speech training set utterances. We calculate a frequency distribution of CV and VC sequences and choose the most common sequences (within the 90th percentile, rounding up to the nearest 5). This results in 85 CVs and 75 VCs total.

To allow for a more consistent pronunciation of these sequences, we added a context around the combinations (e.g. *T_**AK**_E* for AK or *_**FE**_DERAL* for FE) for prompting during recording — note, however, that participants were instructed to read only the CV or VC, not the surrounding context. In a few exceptions the resulting words were infeasible for use in our corpus because of their structure, e.g. words with a dental fricative (”th” sound) or diphthongs like ”EO” or ”OU”, where the CV or VC goes across the boundary of the sound, or where one of the letters in the CV/VC was silent. Examples are *T_**HI**_S*, *FO_**UN**_D* or *PE_**OP**_LE*. In these specific cases, we used the next common word without these drawbacks.

As a step between continuous sentences and isolated CV/VCs, we use **Isolated words**. The words we include in our corpus were selected from a set of words used for intelligibility evaluations in telephony [14]. The original corpus contains 300 words in total — 150 by variation of initial (phonetic) elements and 150 by variation of final elements, in groups of six words. For this corpus, we selected 180 words (30 groups of six), 90 words with initial variation and 90 words with final variation. An example of a group with variations of initial elements is: *LED - SHED - RED - BED - FED - WED* and a group with final variation: *BAT - BAD - BACK - BASS - BAN - BATH*. This specific setup allows for multiple-choice intelligibility testing, with similar or dissimilar words.

Finally, the new corpus contains **digits** from 0 to 9, which can act as a simple reference set that can be recorded in a short amount of time.

3.2 Recording Setup

To record sessions suitable for training our EMG-to-Speech conversion system, we use an OT Bioelletronica Quattrocento multi-channel EMG amplifier. We use two EMG array electrodes: One 8×4 matrix on the cheek, and one 1×8 strip below the chin. The signal is acquired using chained differential derivation along the longer axis of each electrode (resulting in a total of $7 \times 5 = 35$ EMG channels), processed with a DC offset removal filter at 10 Hz and anti-aliasing filter at 900 Hz and then sampled at 2048 Hz.

			Words				Sentences			
Speaker	Gender	Digits	Initial Var.	Final Var.	CV	VC	Train	Dev	Test	Total
Spk1	f	00:20	02:42	02:39	02:15	01:58	25:00	04:06	03:44	42:44
Spk2	m	00:20	02:42	02:46	02:14	01:57	21:19	03:26	03:09	37:53
Spk3	f	00:26	03:32	03:32	03:02	02:44	23:04	03:45	03:23	43:28
Spk4	m	00:22	03:00	03:04	02:30	02:14	19:43	03:12	02:50	36:55
Spk5	m	00:21	02:45	02:48	02:30	02:14	23:29	03:48	03:27	41:22
Spk6	m	00:19	02:40	02:43	02:42	02:12	20:16	03:15	03:00	37:07
Total (hh:mm:ss)										03:59:29

Table 1: Data corpus breakdown for recorded utterances (mm:ss)

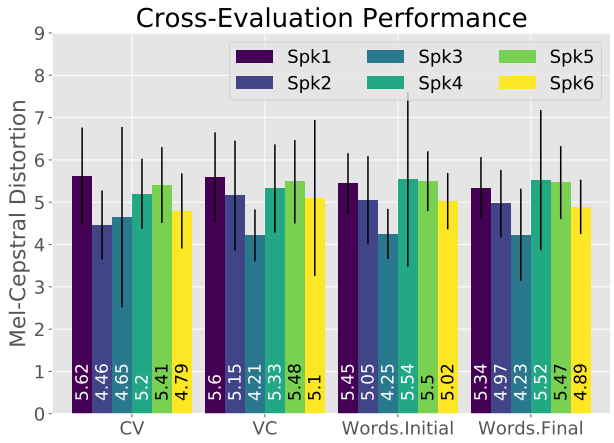

Figure 2: Results of EMG-to-Speech conversion on isolated speech, obtained using 8-fold cross evaluation. Bars indicate utterance standard deviation, lower is better.

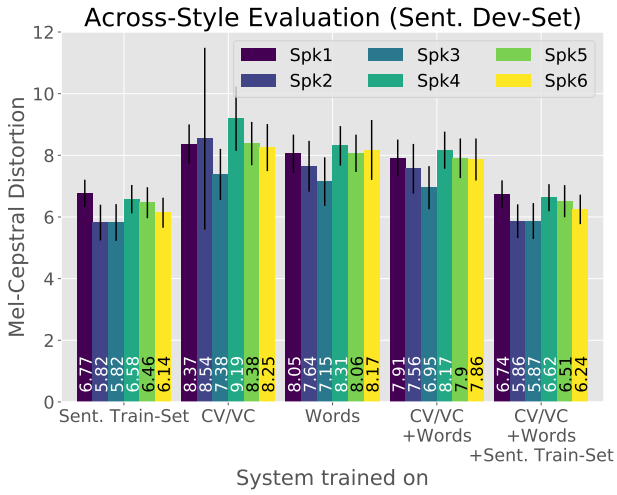

Figure 3: Results of EMG-to-Speech conversion on continuous speech (on the sentences development set), with the system being trained on different combinations of training data. Bars indicate utterance standard deviation, lower is better.

3.3 Recorded Corpus

Using the setup and corpus described in this section, we have recorded six sessions of parallel EMG and Audio data to evaluate our system on. Our subjects (Four male, two female) were between ~20 and ~30 years old and are all non-native English speakers. All of the recorded subjects were healthy and reported never having had any speech disorders. Subjects were thoroughly informed about the recording procedure and experimental evaluations to be done with recorded data and informed consent of all subjects was obtained before recording. In total, we recorded ~4 hours of data. A detailed breakdown into the different parts of the corpus for all recorded speakers can be found in Tab. 1.

4 Initial Evaluations

We present some initial results of performing EMG-to-Speech conversion on our new corpus, including a comparison between isolated and continuous speech. To evaluate our systems, we compare the *Mel-Cepstral Distortion* (MCD) scores of the systems output [15]. The MCD score is a distance measure in MFCC space; a lower MCD means that the system output is more similar to the reference audio. MCD scores obtained in EMG-to-Speech conversion typically fall into the range between 4.5 and 7.0.

Since the EMG signal varies greatly between speakers and depends strongly on skin condition, all evaluations presented in this paper are session-dependent, i.e. no data is shared between speakers.

4.1 Cross-Evaluation on Isolated Speech

To evaluate the performance of our system when both training on and converting isolated words or CV/VCs, we perform 8-fold cross evaluation training on these subsets (splitting utterances into folds). The MCD scores of the resulting audio can be found in Fig. 2.

4.2 Across-Style mapping

To evaluate the performance of our system when training on one utterance style and evaluating on another, we use two different test sets.

Fig. 3 shows MCD scores obtained on continuous speech (the sentences development set) using systems trained on different combinations of training data. We show scores when training on the sentences training set, the isolated CV/VCs, the isolated words, words + CV/VCs and finally, the sentences training set + words + CV/VCs all together.

Fig 4 shows a similar evaluation for isolated speech (the Words set). Here, we use systems trained on CV/VCs and on CV/VCs + the sentences training set. The word cross-evaluation results are provided as a reference.

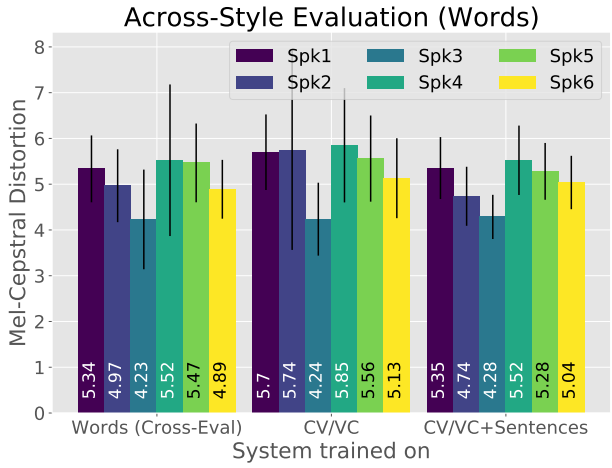

Figure 4: Results of EMG-to-Speech conversion on isolated speech (on the Words set), with the system being trained on different combinations of training data. Words 8-fold cross-evaluation provided for reference. Bars indicate utterance standard deviation, lower is better.

5 Discussion

For speech recognition, it is known that the task of recognizing isolated words is very different from the task of large-vocabulary continuous speech recognition [11].

Comparing Fig. 2 to Fig. 3, we can see that this also holds true for EMG-based speech synthesis. The MCD scores for all categories of isolated speech are lower than the results for continuous speech — even though the training set available for continuous speech is much larger (Compare Tab. 1).

That there is a qualitative difference between isolated and continuous speech can also be seen in the results obtained when using different utterance styles in training and testing, as in Fig. 3. Using the isolated Words set in training significantly improves performance compared to using just CV/VCs, and combining both is better still — however, none of the systems trained on isolated speech perform the task as well as the system trained on continuous speech. In fact, adding the isolated speech sets to the sentences training set does not significantly improve performance versus using just the sentences training set alone.

That the difference between continuous and isolated speech is not merely an effect of adding more training data can be seen by examining the results presented Fig. 4: Here, the isolated CV/VCs are used to train a system that is then used to convert the isolated words. While the CV/VC based system does perform significantly worse than a system trained on the words themselves (again, in 8-fold cross evaluation), the drop in performance is not as severe as the drop between isolated and continuous speech.

Note that in some of our evaluations, the standard deviation appears very high. This is due to rare utterances that have extreme MCD scores, caused by artifacts in that utterances EMG signals. Due to low amounts of training data, the influence of these artifacts on MFCC output can be severe. For this reason, it will be important to develop automated algorithms in place to detect and alleviate this effect.

The results discussed in this section were tested for statistical significance using a one-tailed dependent sample t-test, at a significance level of $p < 0.05$.

6 Conclusion

In this paper, we have introduced a new corpus and initial recordings that we have performed using this corpus. We have presented results that have shown while it is possible to perform EMG-to-speech conversion of continuous speech when training only on isolated segments of speech, performance suffers in this case. We have also shown results that suggest that EMG-based speech processing of isolated speech is, in general, an easier task than EMG-based processing of continuous speech.

In the future, we hope to investigate different architectures that can be better trained using smaller data sets. We also hope to investigate methods to automatically detect and suppress or replace artifacts and faulty channels, which will be especially important for further work towards real-time EMG-to-Speech conversion [16]. Finally, we will investigate intelligibility using forced-choice listening tests and perform further comparisons between continuous and isolated speech.

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, pp. 2257–2271, nov 2017.

[3] C. Herff, G. Johnson, L. Diener, J. Shih, D. Krusienski, and T. Schultz, "Towards direct speech synthesis from ecog: A pilot study," in *Engineering in Medicine and Biology Society (EMBC), 2016 38th Annual International Conference of the IEEE*, Aug 2016.

[4] M. Angrick, C. Herff, G. Johnson, J. Shih, D. Krusienski, and T. Schultz, "Interpretation of Convolutional Neural Networks for Speech Regression from Electrocorticography," in *ESANN 2018 – 26th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, (Brugge, Belgium), pp. 7–12, 2018.

[5] T. Toda and K. Shikano, "Nam-to-speech conversion with gaussian mixture models," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 1957–1960, 2005.

[6] D. Fabre, T. Hueber, L. Girin, X. Alameda-Pineda, and P. Badin, "Automatic animation of an articulatory tongue model from ultrasound images of the vocal tract," *Speech Communication*, vol. 93, pp. 63–75, 2017.

[7] M. Wand, J. Koutník, and J. Schmidhuber, "Lipreading with long short-term memory," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 6115–6119, IEEE, 2016.

[8] 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," *Computer Speech & Language*, vol. 39, pp. 67–87, 2016.

[9] 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*, pp. 1–7, 2015. IJCNN 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, pp. 2375–2385, nov 2017.

[11] S.-C. S. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards continuous speech recognition using surface electromyography," in *In proceedings of the 9th ISCA International Conference on Spoken Language Processing*, 2006.

[12] S. Imai, “Cepstral analysis synthesis on the mel frequency scale,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 8, pp. 93–96, 1983.

[13] A. De Cheveigné and H. Kawahara, “Yin, a fundamental frequency estimator for speech and music,” *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[14] A. S. House, C. Williams, M. H. L. Hecker, and K. D. Kryter, “Psychoacoustic speech tests: A modified rhyme test,” *The Journal of the Acoustical Society of America*, vol. 35, no. 11, pp. 1899–1899, 1963.

[15] R. Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1, pp. 125–128 vol.1, May 1993.

[16] 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.