The Impact of Audible Feedback on EMG-to-Speech Conversion

Lorenz Diener

Cognitive Systems Lab, University of Bremen

lorenz.diener@uni-bremen.de

Abstract

In this extended abstract, I describe the topic of my PhD thesis: Evaluating what the impact of audible feedback is on silent speech production and how audible feedback affects the conversion of surface electromyographic data to audible speech. I willl perform recordings using a novel data acquisition protocol and develop a real-time low-latency EMG-to-Speech conversion system and then use and evaluate this system in feedback experiments. A proposed timeline including information on what work has already been completed is presented in the conclusion of this extended abstract.

1. Introduction

Audible speech – be it face to face or via telephony – is the primary way in which humans communicate with each other. With advancements in computing power and speech technology research, speech communication has become even more important as speech-based man-machine-interfaces have become ubiquitous. Where available, speech interfaces offer a natural, mobile, hands- and eyes-free alternative to touch-based means of interaction.

Most people, in most situations, do not have any trouble using these speech interfaces – however, there are exceptions to this rule. In places where silence is expected, such as in a library or on public transport, speech interfaces cannot be used. They also fail to work in environments where background noise drowns out the speech signal – i.e. on a factory floor. Finally, people who cannot produce speech, e.g. laryngectomees, cannot use speech interfaces at all.

In those situations, where regular speech interfaces fail to deliver, *Silent Speech Interfaces* (SSIs) – speech interfaces that do not rely on the presence of an audible acoustic signal to function – can continue to function, expanding the scope of situations in which speech can be used to communicate.

SSIs have been built using many different modalities – examples include ultrasound [1, 2], permanent-magnetic articulography [3], microwave radar [4], *surface electromyography* (sEMG, with muscle movement [5] or sub-vocal [6]), non-audible murmur recorded with a throat microphone [7] or even electrocorticography [8].

In my thesis, I focus on one specific type of SSI – the direct conversion of surface electromyographic signals to audible speech, or EMG-to-Speech for short.

In previous work [5], such interfaces have been built, with some success, to operate and be tested entirely off-line, with no limits on available training data and time. It is, however, known that there are differences between speaking modes – articulatory muscle movements differ between *modal* (i.e. audible) speech and *silent* speech. This difference is caused by a lack of audible feedback [9] when speaking silently – humans would usually use their own perception of their spoken voice to adjust their articulation, this of course isn't possible when no audible speech is produced.

The goal of my thesis is to produce a workable silent-operating and low latency EMG-to-Speech system that is (unlike an off-line silent speech system) capable of producing feedback and to then use that system to evaluate the effects of feedback on users. The system will record multi-channel EMG data and convert it directly to an audible waveform that can then be played back to the user with low latency (see Figure 1 for a broad overview of this system structure). With this user-in-the-loop system design, I will investigate the effect of feedback on EMG-to-Speech conversion. The following sections describe the baseline EMG-to-Speech conversion approach as well as the requirements for the system described in my thesis.

2. System Overview

2.1. Baseline EMG-to-Speech Conversion

I first describe my baseline off-line EMG-to-Speech conversion system [10]. The system is based on a neural network conversion approach: It first splits the input data into frames and then uses a three-hidden-layer deep neural network to convert a set of stacked EMG time-domain features (TD15 features [11]) to convert these into audio features (Mel-Frequency Cepstral Coefficients [12]) that, using a vocoder, can then be converted back to an audio waveform.

While the system described is the state of the art in EMG-to-Speech conversion, it does have shortcomings: While it does not require any linguistic information, it still does require a large amount of parallel *audible* training data. Due to differences in the EMG signal caused by different speaking modes, silent operation of a system built this way is not possible. Because of between-session and between-speaker variance in the sEMG signal, this also means that the system is restricted to offline operation and evaluation on pre-recorded data.

2.2. Real-Time Operation

One of the requirements for a system that can be used to investigate the effects of feedback on EMG-to-Speech conversion is that the system can provide such feedback. To do this, it has to operate with low latency (ideally close to the electromechanical delay [13]) and in real time. It must possible to use the system after providing only a small amount of data and without a lengthy training phase.

2.3. Session Adaptation

My approach to solving this problem is to build a session-adaptive system with a background model trained from a large amount of data. This background model can be trained offline as before, and is then adapted to new sessions (and possibly even new users) with a small amount of registration data. I evaluate

Acknowledgements: I would like to thank my thesis advisor, Prof. Dr-Ing. Tanja Schultz, for her valuable feedback on my thesis proposal.

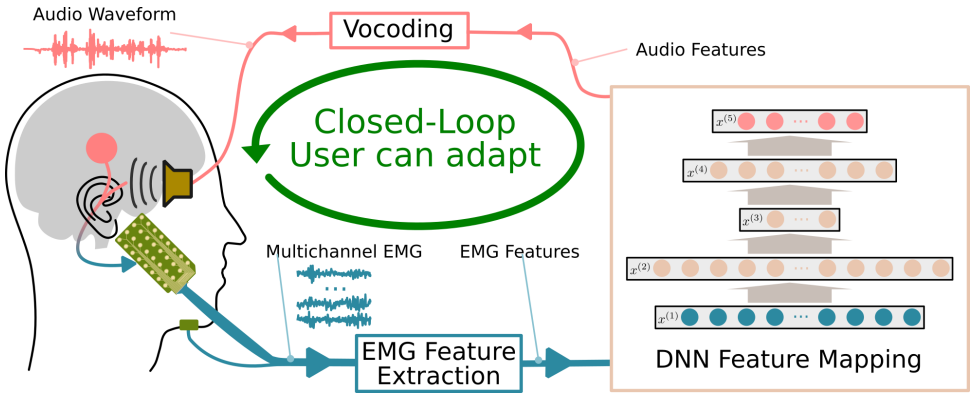

Figure 1: *High-level overview of a user-in-the-loop EMG-to-Speech conversion system.*

different models with regards to how well and how quickly they can be re-trained and with regards to how much data is necessary for this. I also investigate methods for making this system robust against broken EMG channels, a common problem when recording with array EMG electrodes.

2.4. Training for Silent Operation

Silent operation is one of the aspects of an EMG-to-Speech conversion system which feedback may improve by replacing the audible feedback present in the case of modal speech with generated speech output, however, having a base system that works even without feedback is desirable. For this reason, another topic I explore is how to train a system when parallel audio data is not easily available.

One approach I take is to record data in a speak-along setup: Users are first recorded while speaking audibly. Then, the users are prompted to silently speak along with their previously recorded voice. Another approach I evaluate is to obtain alignment data based on lip reading. In one of these ways, I obtain an alignment between audible speech data and silent EMG data. I then use this parallel data to train a system that works on silent EMG data. I compare this system with a cross-modally trained system (i.e. a system that is simply trained on parallel audible EMG and audio and then evaluated on silent EMG).

3. Experiments

3.1. Data corpus

To build and evaluate my systems, I use different corpora:

- The EMG-UKA [14] corpus, a large set of audible EMG speech recordings performed with a single electrode setup, some of which include video recordings.
- The CSL-EMG-Array corpus, a audible EMG speech corpus that contains several recording sessions performed with an array EMG setup, some of which also contain silently recorded speech EMG data.
- A new corpus that includes audible EMG (parallel audible speech and EMG) as well as silent EMG (EMG recorded during silent speech production) recorded using the speak-along recording protocol described above.

I also perform feedback experiments during which new data for session adaptation is collected.

3.2. Approaches to Evaluation

Evaluating an EMG-to-Speech conversion system is not an easy task. In the first place, for paralinguistic parameters such as intonation, simply comparing the generated signal to a reference signal is not sufficient – not every deviation from the reference is necessarily bad. Additionally, in the case of silent operation, no reference is available.

The gold standard in audio evaluation is, of course, human listening tests. I perform these for the final, tuned methods to be evaluated and of course collect human feedback during feedback experiments. This kind of subjective evaluation is, unfortunately, not practical during development. Here, I rely on objective measures such as the Mel-Cepstral Distortion Score and a new measure that I introduce and evaluate, the trajectory-label accuracy, which compares two intonation trajectories with regards to speech naturalness.

3.3. Feedback Experiments

I finally plan to use my low-latency real-time user-in-the-loop EMG-to-Speech conversion system to evaluate the effects of feedback on users, trying to prove the hypothesis that feedback improves system output quality as users are able to learn to use the system.

Specifically, I plan to evaluate two types of feedback: ”Simple” feedback, where only a speech-related buzzing noise is produced according to mouth movements, and ”Complex” feedback where a fully trained EMG-to-Speech system is used to provide audio feedback that is as good as possible. I plan to perform these feedback experiments with a minimum of three users per variant, with two recording sessions per user performed on different days to investigate long-term learning effects.

4. Conclusion

I have described the topic of my thesis: Building a low-latency real-time EMG-to-Speech conversion system and performing experiments to evaluate the effect that audible feedback has on silent speech production and EMG-to-Speech conversion.

Some of the work described above has already been done: The baseline neural network based system has been built and brought to state of the art performance, and I have developed a base framework for the real-time system. Additionally, I have developed and evaluated a new measure for evaluating the naturalness of generated intonation contours. Currently in progress is the recording of speak-along audio, which I will then start to analyze and integrate in all other experiments. The rest of the experiments and analysis described in this abstract will be performed over the next year, with the goal of completing results by July 2020.

5. References

[1] T. Grósz, G. Gosztolya, L. Tóth, T. G. Csapó, and A. Markó, "F0 estimation for dnn-based ultrasound silent speech interfaces," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 291–295.

[2] 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.

[3] 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.

[4] P. Birkholz, S. Stone, K. Wolf, D. Plettemeier, K. Wolf, D. Plettemeier, S. Stone, and P. Birkholz, "Non-invasive silent phoneme recognition using microwave signals," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 12, pp. 2404–2411, 2018.

[5] 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.

[6] A. Kapur, S. Kapur, and P. Maes, "Alterego: A personalized wearable silent speech interface," in *23rd International Conference on Intelligent User Interfaces*. ACM, 2018, pp. 43–53.

[7] 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*, 2005, pp. 1957–1960.

[8] 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.

[9] C. Herff, M. Janke, M. Wand, and T. Schultz, "Impact of different feedback mechanisms in emg-based speech recognition," in *12th Annual Conference of the International Speech Communication Association*, 2011, interspeech 2011.

[10] 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.

[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, 1983, pp. 93–96.

[13] P. R. Cavanagh and P. V. 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.

[14] M. Wand, M. Janke, and T. Schultz, "the EMG-UKa Corpus for Electromyographic Speech Processing," in *Proc. Interspeech*, 2014, pp. 1593–1597.