Session-Independent Array-Based EMG-to-Speech Conversion using Convolutional Neural Networks

Lorenz Diener, Gerrit Felsch, Miguel Angrick, Tanja Schultz

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

Abstract

This paper presents an evaluation of the performance of EMG-to-Speech conversion based on convolutional neural networks. We present an analysis of two different architectures and network design considerations and evaluate CNN-based systems for their within-session and cross-session performance. We find that they are able to perform on par with feedforward neural networks when trained and evaluated on a single session and outperform them in cross session evaluations.

1 Introduction

For most humans, speech is an effortless and natural way to communicate with other humans. Recent advances in speech processing and machine learning have extended this to human-machine interaction as speech-based interfaces in mobile phones and smart speakers have entered everyday use. However, despite its efficiency, there are situations in which audible speech communication is not an option:

- The presence of loud interfering noise, such as on a factory floor or at an airport, can make speech hard or impossible to understand.
- In some situations (e.g. a library or in public transport), audible speech is itself interfering noise and should be avoided.
- Some people (e.g. Laryngectomees) are simply not able to produce an audible speech signal unaided.

In such situations, it would be better to use systems that do not rely on the presence of an audible speech signal to function. Such *Silent Speech Interfaces* (SSIs) instead use a host of other – sometimes multiple – *speech-related biosignals* [1] to infer information about speech. Examples of such SSIs include interfaces based on brain activity recorded invasively using electrocorticography [2], ultrasound-based recording of tongue movements [3], lip reading with video cameras and, as in this work, muscle activity recorded using electromyography (EMG).

The SSI presented in this paper extends our previous work [4, 5] and is based on the recording of muscle activity in the face – facial surface Electromyography. It performs EMG-to-Speech conversion – direct conversion of facial sEMG data to audible speech. Compared to a recognition-based approach, this method has several advantages. With direct synthesis, it is possible to transport not only the textual content of speech, but also paralinguistic information such as stress and intonation. Additionally, while recognition systems are limited to a certain vocabulary and language, a direct conversion system does not suffer from such limitations. Finally, with such a system, it is possible to generate output with a low latency, since the system does not have to wait for some linguistic unit (e.g. a word or sentence) to be complete before starting output.

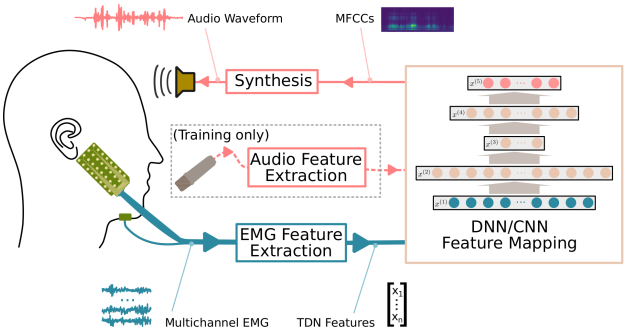

Figure 1: A system overview of our EMG-to-Speech conversion system.

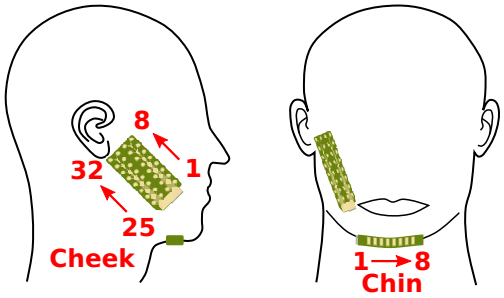

Figure 2: Electrode numbering for the multichannel array EMG recording used in this work.

2 System Description

In our previous work, we have shown and evaluated EMG-to-Speech conversion based on deep neural networks. In this work, we expand upon our previous work by demonstrating a system based on convolutional neural networks, exploiting the structure of EMG array electrodes we use in our recordings: Our EMG-to-Speech conversion system uses array electrodes – patches of electrodes arranged in a grid that can be easily attached to the face.

Convolutional neural networks are neural networks in which, instead of every neuron of a layer feeding into every neuron of the following layer, spatial information can be taken into consideration: Neurons are arranged in an n-dimensional grid and convolved with a number of hyperrectangular filters with learnable weights. This enables them to compensate for positional shifts and reduces the number of parameters that the network needs to learn. Both of these properties are desirable for EMG-to-Speech conversion:

- Shift invariance might reduce the influence of shifts in electrode position. This is especially useful in our case, as the electrode arrays we use consist of a number of electrodes that are arranged with fixed positions relative to each other (Compare Fig. 2).
- The lower number of parameters might help learning

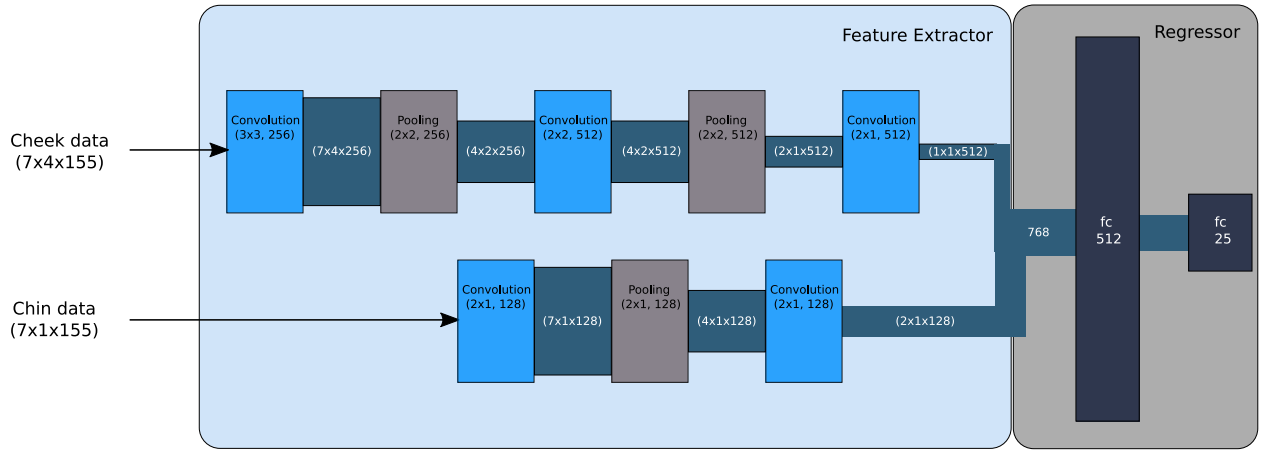

Figure 3: Architecture of our LeNet-inspired network used to convert sEMG features to MFCCs.

performance. This is important since EMG-to-Speech conversion, due to the nature of the recording equipment (electrodes can only be worn for some time before the signal starts to change due to changes in electrode and skin condition), is inherently a low-data problem.

The rest of this paper is organized as follows: Section 2 provides an overview of our EMG-to-Speech conversion system and introduces the network architectures we evaluate. Section 3 provides an overview of evaluation methodologies and presents results. Finally, the implications of these results are discussed in section 4.

Our EMG-to-Speech systems overall structure can be seen in Fig. 1. First, facial EMG data is recorded using a multichannel EMG amplifier (OT Bioelletronica EMG-USB2) at 2048 Hz using two *Array electrodes*: One 4 x 8 10 mm inter-electrode distance (IED) electrode array on the cheek and one 8 electrode 5 mm IED strip below the chin. Signals are measured using chained differential derivation (Compare electrode numbering in Fig. 2, with channels that go across a border being dropped), leading to a total of 35 input channels. From these, a set of EMG features is extracted and stacked (compare section 2.1) into feature vectors. These EMG feature vectors are converted to audio feature vectors (compare section 2.2) using a neural network. The resulting audio feature vectors can then converted to an audio waveform for playback using a *Mel-Log Spectrum Approximation* (MLSA) filter [6].

For training and evaluating our system, we use a corpus of parallel audible speech EMG and audio data. Further details about this data can be found in section 3.1.

2.1 EMG features

To represent a channel of the EMG signal as a series of feature vectors, we first window it using a 32 ms Blackman filter, with a window shift of 10 ms (i.e. an overlap of 22 ms). We then extract a set of time-domain (TD) features [7]:

- 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, extracted for all EMG channels, make up a single *TD0* feature vector. We then stack these feature vectors with a stacking height of 15 frames into both the past and the future to provide a total of 31 frames of time context, resulting in the final *TD15* feature vector. While using recurrent neural networks might seem like a more obvious approach in this case, it is important to remember that EMG-to-Speech conversion is generally a low-data problem. This lack of data means that training more complex networks can be intractable and that simpler models should be preferred.

2.2 Audio features

Audio is represented in our system as a series of Mel-frequency Cepstral coefficients (MFCCs) and fundamental frequency (F0) values. To extract these, we first window the audio signal (sampled at 16 kHz) with the same parameters as the EMG signal (window size 32 ms, window shift 10ms, Blackman window). We then extract MFCCs following Imai and Abe [6] and F0s using the YIN algorithm [8]. Together, these two feature sequences allow for the resynthesis of a waveform from features using the MLSA filter.

Note that in this paper, we are only concerned with optimizing the performance of our system with regards to MFCC output quality – F0 is not considered, as the focus of this paper is on intelligibility, which depends primarily on correctly predicted MFCC values. F0 values have no influence on the metric used for evaluation.

2.3 Network Architectures

We evaluate two different convolutional neural network architectures: One based on LeNet [9] and another based on an encoder-decoder structure [10].

Both architectures extract features from both input arrays separately, which are then processed by two fully connected layers to compute the MFCCs. The dimensions of the input layers are determined by the array sizes and the input feature dimensions. In our case they are $7 \times 4 \times 155$ for the cheek array and $7 \times 1 \times 155$ for the chin array. The size of the output layer is determined by the number of predicted MFCCs. Note that, since our problem is not a recognition task, but instead a regression task – the last layer therefore uses a linear activation function.

After we compared several network structures differentiating in layer number, sizes and types on on the corpus' development set, the LeNet-5 and the encoder-decoder architectures were chosen for performing best. Layer sizes

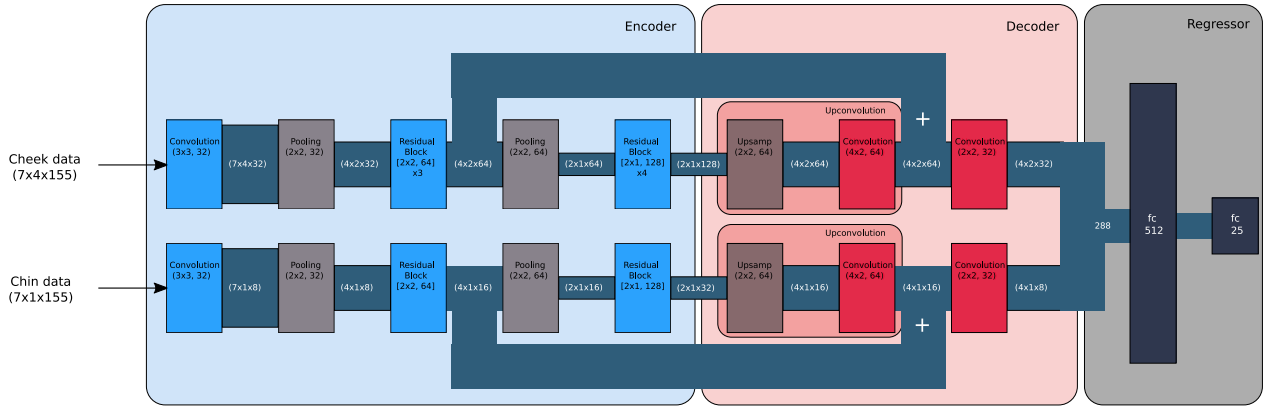

Figure 4: Architecture of our encoder-decoder neural network used to convert sEMG features to MFCCs.

and activation functions were further empirically tuned to optimize performance. The networks were trained until the error on a hold out validation set stopped decreasing.

To train the LeNet-inspired network, we used Adam [11] with a learning rate of 0.003, β_1 of 0.9 and β_2 of 0.999. For training the encoder-decoder architecture a learning rate of 0.002, β_1 of 0.9 and β_2 of 0.999 was used. Both architectures were initialized using a He normal initializer [12]. The mean squared error was used as the loss function for parameter updates. Training was stopped after the loss on a hold-out validation set saturated.

The LeNet-inspired architecture consists of two parts, one feature extractor composed of three convolutional and two pooling layers and a regression part consisting of two fully connected layers. The structure of the LeNet-inspired architecture can be seen in Fig. 3.

The encoder-decoder architecture uses the same regression part. Its feature extraction part is divided into an encoder part that is based on ResNet-32 and a decoder part consisting of a combination of unpooling an convolutional layers. Fig. 4 shows the structure of our encoder-decoder network.

3 Evaluation

For the objective evaluation of the results we use the Mel-Cepstral Distortion (MCD) score [13], defined as the scaled euclidean distance between the genuine and predicted MFCCs, with the first coefficient excluded. Since the MCD is a distance measure, a lower MCD implies a better feature mapping. MCD scores have been found to correlate with subjective estimates of intelligibility.

It is important to note that MCD scores upwards of 5.0 and 6.0 are common when driving speech synthesis as part of a silent speech interface, as the task of converting non-audio biosignals to speech differs considerably from i.e. text-to-speech synthesis or voice conversion.

3.1 Data corpus

The data corpus used holds five sessions from one speaker with a total of 1950 utterances that consist of 649,983 samples. Each session is split into a training, development and test set. A detailed breakdown can be found in Tab. 1. The development set was used during parameter optimization, whereas the test holdout was used only for final performance evaluations. In sum, our corpus contains ~ 88 minutes of training data.

	Nb. Utterances			Nb. Samples		
Session	Train	Dev	Test	Train	Dev	Test
Small1	140	30	30	32,322	7,589	7,307
Small2	140	30	30	39,876	8,272	7,999
Large1	450	50	40	151,870	16,231	16,296
Large2	413	38	19	148,029	13,423	6,668
Large3	450	50	40	158,324	16,691	14,816

Table 1: Data corpus breakdown

As our baseline, we compare the two convolutional architectures with a feedforward deep neural network (DNN) conversion approach as described in our previous work [4].

3.2 Average Pooling vs. Max-Pooling

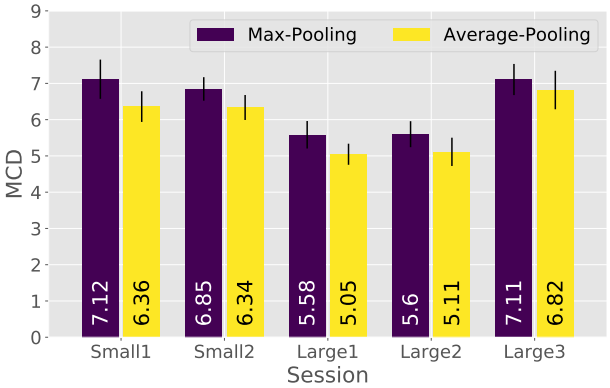

Figure 5: Mel-Cepstral distortions of the LeNet network with max-pooling and with average pooling. Lower is better.

We evaluate two different kinds of pooling layers for the architectures, max-pooling and average pooling with regards to session dependent performance using the LeNet-inspired network. Fig. 5 shows the MCD score of the four resulting architectures for each session.

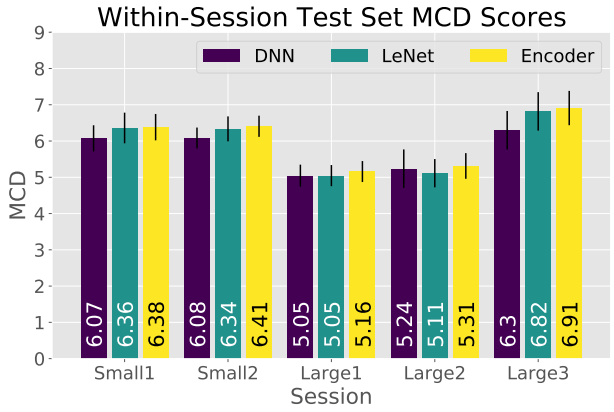

Figure 6: Within-Session Mel-Cepstral distortions for different architectures. Lower is better.

3.3 Within-Session Performance

We compare the session dependent (i.e. training on the training set of one session and then evaluating the MCD score when converting the test set of that same session) performance of the convolutional architectures and the baseline DNN. The resulting MCD scores can be seen in Fig. 6.

3.4 Cross-Session Performance

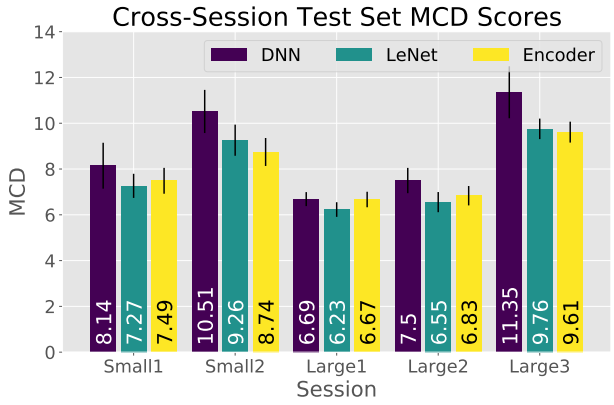

Figure 7: Cross-Session Mel-Cepstral distortions for different architectures. Lower is better.

To evaluate the performance of our CNN approach on data from entirely unseen sessions, we train a sEMG-Feature-to-MFCC mapping on the training data of all but one held out session and evaluate the MCD score on the test set of the held-out session. The evaluated systems are, in effect, session independent: The sEMG data being converted to audio features consists of unseen sentences from an unseen session. Such systems would allow a user to use EMG-to-Speech conversion without first having to record data and train a new system – an important step towards real-time EMG-to-Speech conversion [14]. The results of this evaluation are shown in Fig. 7.

4 Discussion

In section 3.2, we compared the performance of convolutional neural networks when using different pooling methods. Where most of the common network architectures perform best using max-pooling, our networks (which perform a regression task instead of a recognition task) perform significantly better using average pooling (verified at a significance level of p=0.05). We suspect that this is due to max-pooling layers reducing variance, whereas average pooling layers increase variance instead – the latter is preferable for varied regression output.

In section 3.3, we compare the performance of our two CNN architectures to a baseline DNN architecture when training session-dependent systems. Both the encoder-decoder architecture as well as the LeNet architecture are able to perform on par with, though are unable to outperform, the DNN. We suspect that this is because the convolutional network is unable to play its strengths here: Within a session, there is no positional shift of the array, and the already tuned and properly regularized DNN system generalizes to the relatively similar test set well.

Finally, section 3.4 compares the performance of the CNNs and the DNN when training in a session-independent manner, where generalization is harder than in the within-session case. It can be seen that here, the LeNet approach manages to significantly (one sided dependent sample t-test produces p-values smaller than 0.05) outperform the deep neural network approach for all sessions. The Encoder approach is unable to consistently significantly outperform the DNN (though performance is significantly better for all sessions but session Large2). Note that in every case, the cross-session performance is still significantly worse than the performance of a session-dependent system.

4.1 Conclusion

We have presented EMG-to-Speech conversion based on convolutional neural networks that can perform EMG to Speech conversion for unseen sentences on unseen sessions. One of the two systems presented, based on an LeNet architecture, is able to outperform a plain deep neural network based conversion system on this task. In the future, we hope to further improve the performance of our cross-session system by using greater amounts of data and by investigation session normalization e.g. using autoencoding or domain-adversarial adaptation. We also plan to investigate session adaptation approaches to quickly create systems that work very well for a specific speaker based on a background baseline system – this would be an important step towards real world EMG-to-Speech conversion systems. For evaluation, we would like to perform subjective intelligibility evaluations using listening tests.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[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, pp. 2375–2385, nov 2017.

[6] 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.

[7] 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.

[8] 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.

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[10] X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Advances in neural information processing systems*, pp. 2802–2810, 2016.

[11] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *ArXiv e-prints*, Dec. 2014.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," *ArXiv e-prints*, Feb. 2015.

[13] 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.

[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.