Direct Conversion from Facial Myoelectric Signals to Speech using Deep Neural Networks

Lorenz Diener, Matthias Janke, Tanja Schultz
Cognitive Systems Lab, Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
Email: matthias.janke@kit.edu

***Abstract*—This paper presents our first results using Deep Neural Networks for surface electromyographic (EMG) speech synthesis. The proposed approach enables a direct mapping from EMG signals captured from the articulatory muscle movements to the acoustic speech signal. Features are processed from multiple EMG channels and are fed into a feed forward neural network to achieve a mapping to the target acoustic speech output. We show that this approach is feasible to generate speech output from the input EMG signal and compare the results to a prior mapping technique based on Gaussian mixture models. The comparison is conducted via objective Mel-Cepstral distortion scores and subjective listening test evaluations. It shows that the proposed Deep Neural Network approach gives substantial improvements for both evaluation criteria.**

I. Introduction

Over the last years *Silent Speech Interfaces* [1] - systems that enable speech communication when an acoustic signal is unavailable - have gathered intense public interest, since they offer solutions to problems faced by today's speech-driven technologies, in particular:

1) degradation of performance in the presence of noise,
2) disturbance of bystanders,
3) compromise of privacy and confidential information,
4) exclusion of speech-disabled persons from common speech processing systems.

Several kinds of techniques have been proposed to alleviate these problems. Our method of processing speech signals relies on surface electromyography (EMG) [2], where the activation potentials of the facial articulatory muscles are recorded with surface electrodes in order to retrace speech. Since this approach is solely based on the articulatory muscle activity, it also works when no audible speech is produced, i.e. the words are only mouthed. One target group for this type of interface are people, who have suffered from the loss of a phonation function, like patients from laryngectomy or tracheotomy. Achieving natural speech communication would be a great help to improve medical care and to retain social interaction.

We believe that paralinguistic information – like speaker identity, speaker's mood, etc. – is crucial for a natural communication and for an accepted usage of silent speech technologies. Previous work (e.g. [2], [3]) successfully implemented speech recognition systems that recognize the EMG-based input and give a text output, that can be synthesized using text-to-speech systems, but that ignores the natural paralinguistic information. We therefore propose a *direct* conversion from EMG to the acoustic domain [4].We expect this direct feature transformation technique to have the advantage of retaining the paralinguistic information, compared to an EMG-based speech recognition system. This approach, enabling a straight transformation of features, also benefits from the fact that there exist no vocabulary restrictions and no word recognition errors - a drawback which can be observed on speech recognition systems.

We proposed a first direct feature transformation from EMG to acoustic speech in [4]. A frame-based statistical mapping technique with Gaussian mixture models [5] was used, which was originally introduced in the Voice Conversion domain, where the voice parameters of one speaker are transformed to a different target voice. The input EMG features were transformed to acoustic features, while the fundamental frequency (F_0) was extracted from the simultaneously recorded acoustic speech signal. In previous work [6] we complemented this approach by generating F_0 from the EMG signal, but faced issues with the naturalness and prosody of the generated output. We also proposed a direct mapping technique [7] which is based on a Unit Selection approach and obtained promising results especially in terms of naturalness. Other research groups [8] used an EMG-based neural network approach, but for phone classification instead of regression. An articulatory-to-acoustic mapping approach based on Deep Neural Networks (DNN) was introduced by [9]. They trained on electromagnetic articulography (EMA) data which was recorded synchronously with the articulated speech sounds.

In this paper we investigate a direct EMG-to-speech mapping based on *Deep Neural Networks* and compare this approach to our previous Gaussian Mapping technique [10]. In the training stage we estimate the DNN parameters using information from corresponding EMG and speech data, collected during simultaneous data recordings. In the conversion stage a non-linear mapping is used to convert arbitrary facial EMG signals to acoustic speech features, namely the Mel Frequency Cepstral Coefficients (MFCCs) plus the fundamental frequency (F_0). These features are used to obtain the acoustic output using the Mel Log Spectrum Approximation (MLSA) filter method [11]. Figure 1 illustrates our mapping process.

The remainder of this paper is organized as follows: Sec. II presents the setup and describes the data corpus we used, followed by Sec. III which gives details about the compared feature mapping approaches. In Sec. IV, we present our experimental setup, followed by the results and evaluation in Sec. V. Sec. VI concludes the paper, outlines remaining problems and future work.

978-1-4799-1959-8/15/$31.00

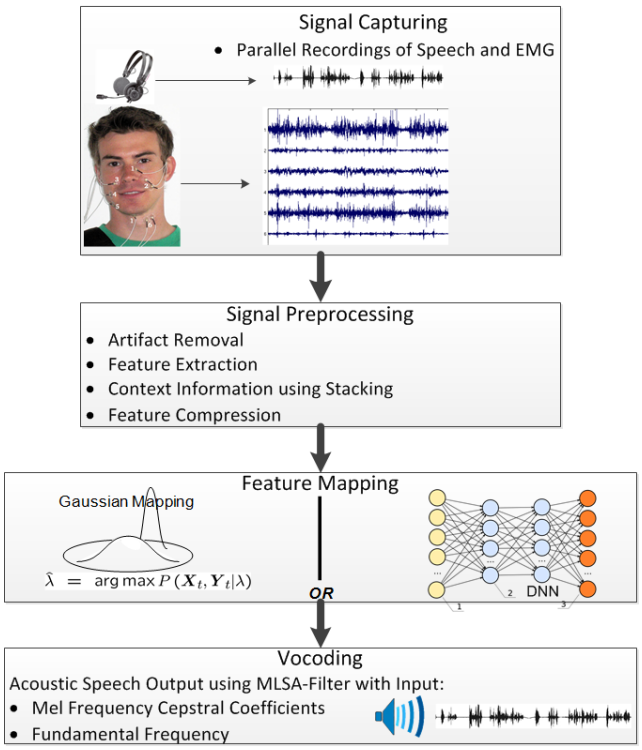

Fig. 1. *Process of mapping from electromyographic input to speech output.*

II. Recording Setup and Data

For our proposed Deep Neural Network based mapping approach we selected recording sessions from our previous work ([10], [12]) which contain more than 500 utterances of EMG signals recorded during audible speech. Since the training of DNNs requires a relatively large amount of data, we additionally use two recently recorded sessions with 1103 and 1978 utterances. In total the corpus contains six recording sessions, with data from two male speakers and one female speaker. Since the EMG signal shows high inter-individual differences, we only use it *session dependently*.

Fig. 2. ***left**: Single electrode positioning, black numbers indicate unipolar derivation with reference electrodes behind the mastoid bone (except channel 1), white numbers indicate bipolar derivation.* ***right**: Electrode array positioning, one large array is positioned on the cheek, one 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* setup and a novel *electrode array* setup. 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 anguli oris, 2) the zygomaticus major, 3) the platysma, 4) the anterior belly of the digastric and 5) the tongue, see Fig. 2 (left) for the electrode positioning. All EMG signals were sampled at 600 Hz and filtered with an analog 1 Hz high-pass filter. The electrode positioning which yielded optimal results was adopted from [13].

The electrode array acquisition device (EMG-USB2, OT Bioelettronica, Italy) recorded the EMG signals using a large electrode grid of four rows of eight electrodes each with 10 mm inter-electrode distance (IED) and a second smaller array with one row of eight electrodes with 5 mm IED. As illustrated in Fig. 2 (right) the large array was placed on the subject's cheek - similar to the positioning of a cell phone - while the smaller one was positioned under the chin to ensure the recording of the tongue. 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 [12].

In addition to the EMG signal, we simultaneously recorded the acoustic speech signal with a standard close-talking microphone at a sampling rate of 16 kHz. The audio signal is synchronized to the EMG signal using an additional analog marker channel. This marker system consists of a signal splitter with one analogue input, which is connected to a parallel port of the recording computer, and two outputs, one connected to the input of the EMG amplifier, and the other one connected to the second (stereo) channel input of the sound input device. This is a prerequisite for getting corresponding EMG and acoustic speech data.

The recorded text corpus is based on [14] and consists of phonetically balanced English sentences which originated from the broadcast news domain. For the larger sessions, further text data from the Arctic corpus [15] was added. The longest session with more than 2 hours of recorded EMG and speech data additionally used utterances from the TIMIT corpus [16]. 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. We additionally defined a development (*dev*) set from session *Spk1-Array*, which is used for the parameter optimization. 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 I lists the durations of the six recorded sessions and the number of utterances per session.

III. Feature Mapping Approaches

We implement a direct EMG-to-speech feature transformation, we simply refer to as *mapping*. In the training process we estimate the model parameters using information from the simultaneously recorded EMG and speech data. Thus, in the final conversion stage the acoustic speech output is created from the unseen input EMG data using the estimated model. This section introduces the proposed mapping techniques.

TABLE I. *Data corpus information for the recorded utterances, including speaker/session breakdown.*

Session	Accumulated data length, in (mm:ss)			# of train/eval utterances		
	Train	Eval	Dev	Train	Eval	Dev
Spk1-Single	27:10	01:19		500	20	
Spk2-Single	26:54	00:49		496	13	
Spk1-Array	31:01	00:47	01:59	500	10	30
Spk2-Array	25:44	01:10		500	20	
Spk1-Array-Large	76:44	00:48		1093	10	
Spk3-Array-Large	123:04	00:45		1968	10	
Total	310:37	05:38	01:59	5057	83	30

A. Gaussian Mapping

Our previous feature mapping approach [10] is based on a transformation via Gaussian mixture models (GMM), a technique that is successfully used in the Voice Conversion domain and also in similar speech feature transformations [5]. Like all feature mapping approaches described in this paper, this *Gaussian Mapping* approach consists of two parts:

1) a training stage,
2) a conversion stage on unseen data.

For training we use EMG and acoustic data that was simultaneously recorded (see section II). The training data consists of 32-dimensional EMG feature vectors as source data and 25-dimensional Mel-Cepstral Coefficients as target data. See Sec. IV-B and IV-A for details on the used EMG and acoustic features.

For the conversion stage we define a static source and target feature vector at frame t as $\boldsymbol{x}_t = [x_t(1), \cdots, x_t(d_x)]^\top$ and $\boldsymbol{y}_t = [y_t(1), \cdots, y_t(d_y)]^\top$, respectively. d_x and d_y denote the dimension of $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$, respectively. After preparing the training data, a GMM is trained to describe the joint probability density of the source and the target feature vectors. The conversion stage is based on a minimum mean-square error criterion: $\hat{\boldsymbol{y}}_t = \sum_{m=1}^{M} P\left(m|\hat{\boldsymbol{x}}_t, \lambda\right) \boldsymbol{E}_{m,t}^{(Y)}$, where $\hat{\boldsymbol{y}}_t$ is the estimated target feature vector at frame t from input feature vector $\hat{\boldsymbol{x}}_t$. m denotes the mixture component index, M denotes the total number of the mixture components, λ represents the parameter set of the GMM, which consists of weights, mean vectors, and full covariance matrices for individual mixture components.

B. Neural Network Mapping

In this work, we propose to use a deep neural network (DNN) which is implemented using the Computational Network Toolkit [17] to perform the mapping from EMG features to audible speech. This application reflects a regression problem instead of the classification usually done via DNNs. We use a five layer feed forward neural network with bottleneck layer topology to obtain the mapping function between the source and the target vectors. Standard backpropagation learning is used to adjust the weighting parameters of the DNN so as to minimize ϵ, i.e., the mean squared error between the desired and the actual output values. We performed experiments with different combinations of epochs, mini-batch sizes and learning rates. Parameter results are reported in Sec. V.

Figure 3 shows the architecture of the employed five-layer neural network which we use for mapping the electromyographic features to the acoustic space of audible speech.

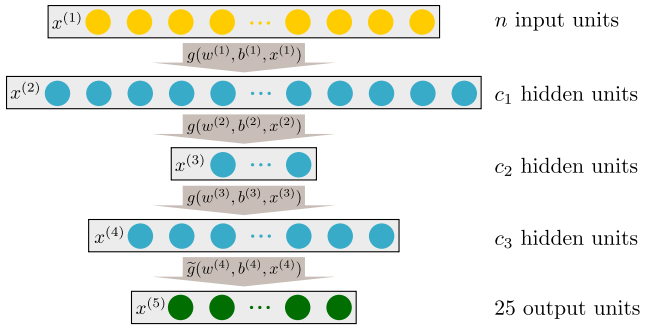

Fig. 3. *Structure of the neural network used to convert electromyographic features to target Mel Frequency Cepstral Coefficients.*

The neural network is trained to map the EMG features to the target features of audible speech, i.e., if $G(\boldsymbol{x}_t)$ denotes the mapping of $\boldsymbol{x}_t$, then the error of the mapping is given by $\epsilon = \sum_t \|\boldsymbol{y}_t - G(\boldsymbol{x}_t)\|^2$. $G(\boldsymbol{x}_t)$ is defined as

$$\begin{aligned} G(\boldsymbol{x}_t) = &\widetilde{g}(\boldsymbol{w}^{(4)}, \boldsymbol{b}^{(4)}, \\ &g(\boldsymbol{w}^{(3)}, \boldsymbol{b}^{(3)}, \\ &g(\boldsymbol{w}^{(2)}, \boldsymbol{b}^{(2)}, \\ &g(\boldsymbol{w}^{(1)}, \boldsymbol{b}^{(1)}, \boldsymbol{x}_t)))) \end{aligned}$$

where

$$\widetilde{g}(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{x}) = \boldsymbol{w} * \boldsymbol{x} + \boldsymbol{b}$$

and

$$g(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{x}) = \mathrm{ReL}(\widetilde{g}(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{x}))$$

Here, $\boldsymbol{w}^{(n)}$ and $\boldsymbol{b}^{(n)}$ represent the weight and bias matrices of the hidden and output layers and ReL denotes the rectified linear activation function $\mathrm{ReL}(x) = max(0, x)$. We use the EMG feature vector input described in detail in section IV-B, followed by three hidden layers g with different sizes concluded by a final regression layer $\widetilde{g}$ having as many nodes as the number of acoustic output parameters, resulting in an "hourglass" configuration with a $c_2 = 32$ node bottleneck in the center surrounded by a $c_1 = 2500$ node computation layer between input and bottleneck and a $c_3 = 1024$ node computation layer between bottleneck and output. We used this three-hidden-layer DNN structure based on our previous work on the conversion of whispered to audible speech [18].

Once the training process has converged, we get a set of weight and bias matrices which represent the mapping function from source EMG features to target acoustic speech features. These matrices can be used in the conversion stage to transform an EMG feature vector to a feature vector of the audible speech. To avoid bias towards numerically larger EMG- or audio features, the signal is normalized to zero mean and unit variance for training.

IV. Experiment Setup

A. Acoustic features

In the acoustic signal domain, an excitation-filter model of speech is considered. 25 Mel Frequency Cepstral Coefficients (MFCCs) [19] are extracted as filter parameters and fundamental frequency (F_0) estimates are derived as excitation features

for every 10 ms in 32 ms frame. These features represent the acoustic speech information and will be used to obtain the acoustic output: The Mel Log Spectrum Approximation (MLSA) filter method [11] takes the generated F_0 and MFCCs as input and generates the final output speech waveform.

B. Electromyographic features

We evaluate a feature which is based on a composition of *time-domain features* [3]. 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

$$w[n] = \frac{1}{9}\sum_{k=-4}^{4} v[n+k], \quad \text{where} \quad v[n] = \frac{1}{9}\sum_{k=-4}^{4} x[n+k].$$

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 $\mathbf{TD15}$ is defined as follows:

$$\mathbf{TD15} = S(\mathbf{f2}, 15), \text{where } \mathbf{f2} = [\bar{\mathbf{w}}, \mathbf{P_w}, \mathbf{P_r}, \mathbf{z_p}, \bar{\mathbf{r}}].$$

Frame size and frame shift were set to 32 ms and 10 ms respectively. This $\mathbf{TD15}$ feature is computed from each of the EMG channels, then a fused vector is formed by stacking all channel vectors. For the DNN mapping, we use this TD15 feature as the input to our mapping system. To compare our implicit dimensionality reduction approach with a more explicit feature reduction, we also apply Linear Discriminant Analysis (LDA) to reduce the dimensionality of the TD15 data to 32, as in our prior work [10], and perform training and mapping on the low-dimensional post-LDA EMG data.

Prior work ([20], [21]) indicates 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 when the EMG signal is aligned to the audible speech data. Every EMG channel is delayed by the same amount of time.

Note that the electrode array recordings provide 35 channels instead of the six EMG channels given by the single electrodes. For two of the array sessions, we visually inspected the EMG signals and discarded channels that we deemed too noisy, resulting in reduced channel sets. Further details about the positioning and processing of the electrode array signals can be found in [12].

C. Experimental results

For the *objective evaluation* of the proposed EMG-to-speech conversion we use the Mel-Cepstral Distortion (MCD) [22]. The MCD is a scaled Euclidean distance between the spectral features of the target audible speech and the spectral features (i.e. MFCCs) of the converted EMG speech in decibel.

$$\mathbf{MCD} = 10/\ln 10 \sqrt{2 \cdot \sum_{k=2}^{25} (\mathbf{mc}_{est}[k] - \mathbf{mc}_{tar}[k])^2}$$

$\mathbf{mc}_{est}[k]$ and $\mathbf{mc}_{tar}[k]$ denote the k-th Mel-Cepstral coefficient of target and estimated data. Smaller numbers imply better results. Also note that the first coefficient was not included, since it represents the power of the acoustic signal.

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 simultaneously, 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 *subjective estimation* is evaluated using AB preference listening tests comparing the Gaussian Mapping output (see Sec. III-A) to the proposed DNN-converted speech (see Sec. III-B). Each participating subject listens to the original target audio file and compares the mapping outputs A and B to decide which one is preferred. Each utterance is presented in randomized order. If no preference can be perceived, a third neutral option is available, so the listening subject is not forced to make a decision.

V. DNN Mapping

A. Parameter optimization

We initially chose the three-hidden-layer structure of the DNN based on our experience with converting whispered to audible speech, as reported in [18]. We increase the size of the hidden layers to accommodate the higher input dimensionality of our EMG TD15 feature data. The bottleneck structure is chosen to mirror the feature-extraction-followed-by-mapping structure of the previously used LDA-GMM approach - without, however, requiring LDA training based on label information. Layer sizes and activation functions are then empirically tuned on a development set from session *Spk1-Array*, which was held out from training, to optimize performance. The target function used for training is the square error between normalized network-estimated Mel-Cepstral coefficients and reference audio data.

To train the network, we use stochastic gradient descent training, with a momentum of 0.9 and a learning rate of 0.01, initial network parameters being chosen uniformly at random from range $[-0.5, 0.5]$. Holding these things constant, we search for an optimal training epoch count on a development set held out from training for this purpose, arriving at 20 epochs as a setting producing good results for all sessions.

Figure 4 gives an example of square error values obtained during training with a different number of epochs. It can be seen that while the training error continues to decrease, the error on the development set starts rising past 20 epochs of training as overtraining sets in. Consequently we stopped the training of our DNN-based mapping after 20 epochs.

B. Input EMG-Feature Comparison

Our previous mapping approaches use Linear Discriminant Analysis (LDA) to reduce the input TD15 feature set to a 32-dimensional vector. The LDA matrix is computed on the training data of each session divided into classes based on the 45 English phones, plus one silence phone. For the computation, each EMG feature vector needs to be assigned to one phone, implying that phone-based label information

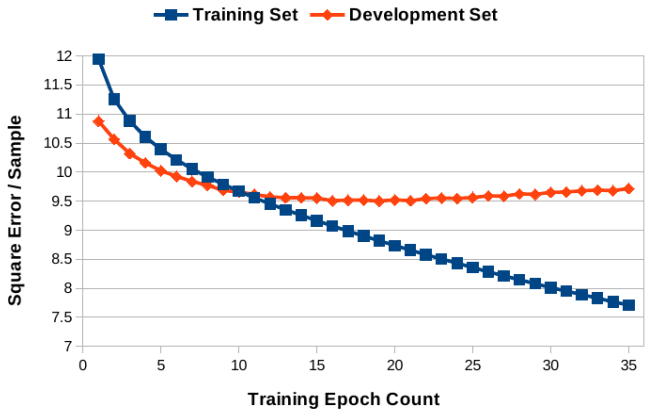

Fig. 4. Average square error per sample between DNN output and normalized reference audio for Spk1-Array when evaluating the network after training with varying epoch count, calculated on the training set and the development set.

is needed. This label information is obtained by performing forced alignment of phone labels obtained from the known utterance text to the audio recordings. Thus a pre-trained speech recognition system is used. This means that performing training on plain parallel recordings of EMG and audio data with no known utterance text, or training on data where the text is known but no speech recognition system is available, is impossible. Additionally, since the automatic alignment is not necessarily correct for every frame, it introduces an additional error source. Thus, it would be preferable to eliminate the dependence on the LDA.

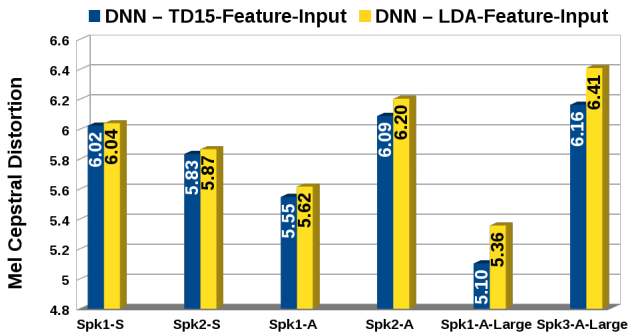

Fig. 5. Mel-Cepstral distortions of DNN-based EMG-to-speech mapping output: 32-dimensional EMG input versus high-dimensional TD15 EMG feature.

We compare two different DNN-based mappings, whose parameters where optimized individually on the development set. The first approach uses the post-LDA 32-dimensional input EMG features, the second one the high-dimensional TD15 features. Although the MCD scores differ only slightly (average MCD of 5.79 versus 5.91), the TD15 feature based mapping manages to consistently outperform the LDA approach in our experiments with a relative improvement of 2.05%, implying that no LDA computation is necessary for proper mapping results. Figure 5 gives the MCDs for each speaker/session of the mapping output.

C. Comparison to Gaussian Mapping

We compare our DNN-based mapping results to our previous work [10], where a frame-based EMG-to-speech mapping is used and some of the data used in this paper (sessions *Spk1-Single* and *Spk2-Single*) is shared (see Section III-A for details). For our Gaussian Mapping, we apply the post-LDA 32-dimensional EMG feature and use 128 Gaussian mixtures for the mapping to the final 25-dimensional MFCCs.

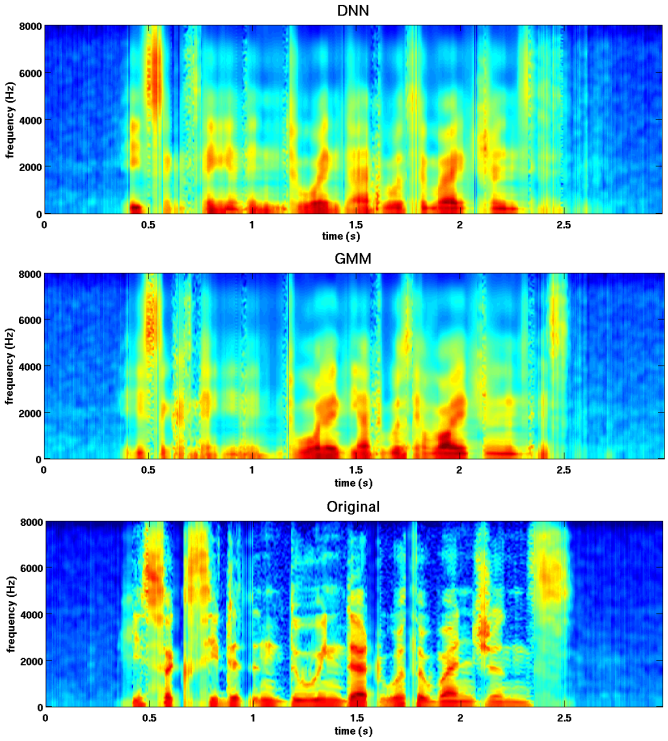

Fig. 6. Exemplary spectrograms of DNN-based and Gaussian Mapping based EMG-to-speech mapping output plus the original target audio file (top to bottom) of the utterance: “He succeeded in doing that with a vengeance.”

Figure 6 shows spectrograms of the converted output from the Gaussian mapping, as well as from the DNN-based feature transformation and the original target audio file from the exemplary test utterance “He succeeded in doing that with a vengeance.”, taken from session *Spk1-Array-Large*. The final output was synthesized using the MLSA filter, based on the converted MFCC output and the target F_0 information that was extracted from the parallel audio file. It can be seen that both conversion approaches show similar results and lack detailed spectral information reconstruction.

With the six sessions we obtain an average MCD of 5.79 with our DNN approach, compared to an average MCD of 5.94 with the Gaussian Mapping output. This corresponds to a relative improvement of 2.47%. Figure 7 gives the MCDs for each speaker/session.

Since objective MCD scores do not perfectly correspond to human acoustic perception, we perform an subjective AB preference listening test (adapted from [23]) between converted speech from DNN-based output and from Gaussian Mapping output. We also included a third neutral option, when no preference was perceived. Each participating subject listens to

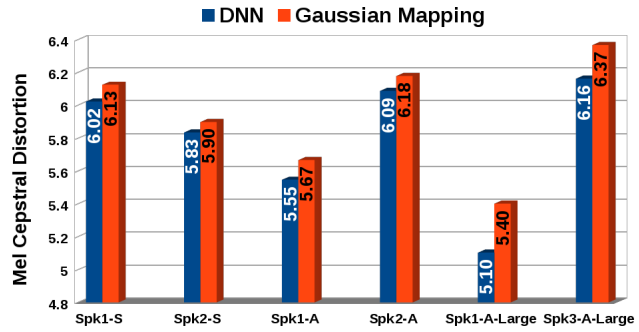

Fig. 7. Mel-Cepstral distortions of EMG-to-speech mapping output: DNN approach versus Gaussian Mapping approach.

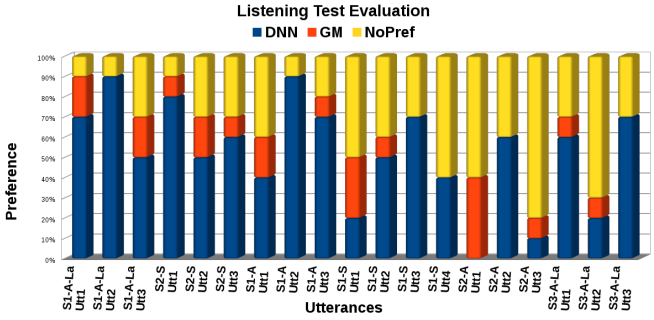

Fig. 8. Listening test preferences of the converted output per test utterance: DNN output versus Gaussian Mapping (GM) output versus no preference (NoPref). S1/S2/S3 = Speaker1/2/3, S/A/A-La = Session with Single-Electrodes/Array/Array-Large.

the target audio file and compares the two mapping outputs to decide which one is preferred. We randomly selected three to four utterances from the test set of each session, resulting in 19 utterance for the listening test and performed the listening test on 10 subjects. This results in 190 utterances, from which 52.6% of DNN output were preferred and only 11.6% of the Gaussian Mapping system were preferred. For 35.8% of the utterances no distinct preference was made.

Figure 8 depicts the single preferences on all 19 listening test utterances. Obviously most utterances of the DNN output (blue) are preferred, although for some utterances no clear preference is made. Even though there is no clear preference for some utterances (such as S2-A-Utt1), where both systems outputs are not satisfactory, none obtain absolute majority with the Gaussian Mapping output.

VI. Conclusion

In this paper we investigated Deep Neural Networks (DNN) to convert surface EMG signals of the articulatory muscles to audible speech. An objective and subjective comparison to a Gaussian Mixture model based mapping technique shows a relative improvement of 2.47% yielding a Mel-Cepstral distortion (MCD) of 5.79. A listening test also shows significant preference for the proposed DNN-based mapping system. While previous approaches used a Linear Discriminant Analysis (LDA) for input feature dimensionality reduction, our DNN-based method achieves similar results with high-dimensional input features. Hence we can omit LDA computation and thus do not require any further information beyond synchronously recorded audio- and surface EMG data.

In the future we plan to compare DNNs to Unit Selection approaches with pre-recorded target speech segments. To further improve the conversion framework we plan to extend the amount of data and evaluate different kinds of EMG input features, since the proposed TD15 feature is highly optimized for EMG-based speech recognition, rather than for synthesis. We also plan to use other types of neural networks that model time-dependent contextual information, e.g. Recurrent Neural Networks and Long Short Term Memories.

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] 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.

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

[5] T. Toda and K. Shikano, “NAM-to-Speech Conversion with Gaussian Mixture Models,” Proceedings of Interspeech 2005, pp. 1957–1960, 2005.

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

[7] M. Zahner, M. Janke, M. Wand, and T. Schultz, “Conversion from Facial Myoelectric Signals to Speech: A Unit Selection Approach”, Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech), pp. 1184–1188, 2014.

[8] T. Tsuji, N. Bu, J. Arita, and M. Ohga. “A speech synthesizer using facial EMG signals”, International Journal of Computational Intelligence and Applications, pp. 1–15, 2008.

[9] F. Bocquelet, T. Hueber, L. Girin, P. Badin, and B. Yvert, “Robust Articulatory Speech Synthesis using Deep Neural Networks for BCI Applications”, Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech), pp. 2288–2292, 2014.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] J. Kominek, and A. W. Black, “The CMU Arctic speech databases” In Fifth ISCA Workshop on Speech Synthesis, pp. 223–224, 2004.

[16] J. Garofolo, L. Lamel, W. Fisher, L. Fiscus, D. Pallett, and N. Dahlgren “DARPA TIMIT: Acoustic-Phonetic Continuous Speech Corpus”, National Institute of Standards and Technology, 1993.

[17] D. Yu, A. Eversole, M. Seltzer, K. Yao, Z. Huang, B. Guenter, O. Kuchaiev, Y. Zhang, F. Seide, H. Wang, J. Droppo, G. Zweig, C. Rossbach, J. Currey, J. Gao, A. May, B. Peng, A. Stolcke, M. Slaney, "An Introduction to Computational Networks and the Computational Network Toolkit", Microsoft Technical Report MSR-TR-2014-112, 2014.

[18] M. Janke, M. Wand, T. Heistermann, T. Schultz, K. Prahallad, "Fundamental Frequency Generation for Whisper-to-Audible Speech Conversion" IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2598–2602, 2014

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] S. Kraft, U. Zlzer: "BeaqleJS: HTML5 and JavaScript based Framework for the Subjective Evaluation of Audio Quality", Linux Audio Conference, 2014.