FURTHER INVESTIGATIONS ON EMG-TO-SPEECH CONVERSION

Matthias Janke, Michael Wand, Keigo Nakamura, Tanja Schultz

Cognitive Systems Lab, Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
matthias.janke@kit.edu

ABSTRACT

Our study deals with a *Silent Speech Interface* based on mapping surface electromyographic (EMG) signals to speech waveforms. Electromyographic signals recorded from the facial muscles capture the activity of the human articulatory apparatus and therefore allow to retrace speech, even when no audible signal is produced. The mapping of EMG signals to speech is done via a Gaussian mixture model (GMM)-based conversion technique.

In this paper, we follow the lead of EMG-based speech-to-text systems and apply two major recent technological advances to our system, namely, we consider *session-independent* systems, which are robust against electrode repositioning, and we show that mapping the EMG signal to whispered speech creates a better speech signal than a mapping to normally spoken speech. We objectively evaluate the performance of our systems using a spectral distortion measure.

Index Terms— Silent Speech, Electromyography, Speech Synthesis, Voice Conversion

1. INTRODUCTION

In the past decades, the robustness and performance of computer-based speech processing systems have improved substantially. However, speech-driven technologies today still face some major challenges, in particular: 1) the performance degrades in the presence of noise, 2) the clearly audible speech disturbs bystanders and compromises privacy, 3) speech-disabled persons may not be able to use these technologies. Several methods to alleviate these problems have been proposed, all with the purpose of creating a *Silent Speech Interface*[1], which is a system enabling speech communication when an acoustic signal is unavailable. Our method of processing silent speech relies on surface electromyography (EMG) [2], where the activation potentials of the human articulatory muscles are recorded with surface electrodes in order to retrace speech. Fig. 1 shows the typical setup of our EMG-based silent speech interface.

To implement a silent speech interface, input EMG signals have to be converted to text information [3] or to synthesized speech waveforms [4] so that a receiver, be it a computer or another human, can comprehend the intended message. In EMG-to-text conversion (also called EMG-based speech recognition), the main fields of investigation are currently the creation of *session-independent* systems, which are robust towards electrode repositionings, and the EMG-based investigation of *speaking mode* discrepancies between audible and silent speech [5]. In this paper, we follow a different approach and perform a direct conversion of EMG signals to speech waveforms [4]. The purpose of this paper is to apply the technological advances from EMG-to-text conversion to the direct EMG-to-speech mapping, to point out similarities, and to outline differences.

From an application standpoint, the EMG-to-speech approach is preferrable to the EMG-to-text method whenever human-to-human communication is intended, in particular, there are no vocabulary restrictions, and direct mapping of EMG signals to speech allows for the inclusion of prosody or emotional speech content, whereas in EMG-based speech

Fig. 1. Electrode positioning, black numbers indicate unipolar derivation with reference electrodes on mastoid portion of the temporal bone (except channel 1), white numbers indicate bipolar derivation.

recognition, only the pure textual content is preserved. Note that methods for the inclusion of prosodic information have been proposed in a previous paper [6] and are not dealt with here.

The paper is organized as follows. The GMM-based mapping method is briefly described in **Section 2**. **Section 3** describes our experiments and results, and section **Section 4** concludes the paper.

2. GMM-BASED MAPPING FROM EMG TO SPEECH

Voice Conversion (VC) is a feature modification technique that causes input speech (referred to as source speech) to sound as if it is uttered by another person (referred to as target speech). A classical VC method [7, 8] is based on Gaussian Mixture Models (GMMs). In this paper we modify this method to map *EMG signals* to speech signals, the algorithm is outlined in this section.

The GMM mapping method consists of the training and conversion parts. For training, one needs a corpus of utterances where the EMG signal and the acoustic signal have been synchronously recorded. The data for the GMM training consists of 32-dimensional EMG feature vectors as the source data and 25-dimensional vectors in form of Mel Cepstral Coefficients as the target data. Note that the 0-th Mel coefficient was not used in training and testing, since it represents the power of the acoustic signal, which is hard to estimate with EMG.

The used conversion is based on our previous work [6]: 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 as follows:

$$P(\boldsymbol{x}_t, \boldsymbol{y}_t|\lambda) = \sum_{m=1}^{M} w_m \mathcal{N}\left([\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top]^\top; \boldsymbol{\mu}_m^{(X,Y)}, \boldsymbol{\Sigma}_m^{(X,Y)}\right),$$

$$\boldsymbol{\mu}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \qquad \boldsymbol{\Sigma}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix},$$

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. m denotes the mixture component index, and M denotes the total number of the mixture components. The parameter set of the GMM is denoted by λ, which consists of weights w_m, mean vectors $\boldsymbol{\mu}_m^{(X,Y)}$ and full covariance matrices $\boldsymbol{\Sigma}_m^{(X,Y)}$ for individual mixture components. $\boldsymbol{\mu}_m^{(X)}$ and $\boldsymbol{\mu}_m^{(Y)}$ represent the mean vectors of the mth mixture component for the source and the target features, respectively. $\boldsymbol{\Sigma}_m^{(XX)}$ and $\boldsymbol{\Sigma}_m^{(YY)}$ represent the covariance matrices and $\boldsymbol{\Sigma}_m^{(XY)}$ and $\boldsymbol{\Sigma}_m^{(YX)}$ represent the cross-covariance matrices of the mth mixture component for the source and the target features, respectively.

The conversion method we are using is adapted from [7, 8] and is based on a minimum mean-square error criterion:

$$\begin{aligned} \hat{\boldsymbol{y}}_t &= \sum_{m=1}^{M} P\left(m|\boldsymbol{x}_t, \lambda\right) \boldsymbol{E}_{m,t}^{(Y)}, \\ P\left(m|\boldsymbol{x}_t, \lambda\right) &= \frac{w_m \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(X)}, \boldsymbol{\Sigma}_m^{(XX)}\right)}{\sum_{n=1}^{M} w_n \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}_n^{(X)}, \boldsymbol{\Sigma}_n^{(XX)}\right)}, \\ \boldsymbol{E}_{m,t}^{(Y)} &= \boldsymbol{\mu}_m^{(Y)} + \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)^{-1}} \left(\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(X)}\right), \end{aligned}$$

where $\hat{\boldsymbol{y}}_t$ is the estimated target feature vector at frame t. Thus $\hat{\boldsymbol{y}}_t^\top$ is the finally estimated Mel Cepstral feature vector.

3. EXPERIMENTAL EVALUATION

3.1. Experimental Conditions

For EMG recording, we used a computer-controlled 6-channel EMG data acquisition system (Varioport, Becker-Meditec, Germany). We adopted the electrode positioning which yielded optimal results from [3].

We used six channels and captured signals from 1) the levator angulis oris, 2) the zygomaticus major, 3) the platysma, 4) the anterior belly of the digastric and 5) the tongue, see Figure 1 for the electrode positioning. All EMG signals were sampled at 600 Hz and filtered with an analog high-pass filter.

For this study we use a subset of the EMG-UKA corpus [9], namely a subset of five different male speakers. In a quiet room each of the speakers read the same 50 sentences from the Broadcast News domain in two different speaking modes: audible (normally spoken) and in whispered speech. 45 sentences were used for GMM training, and the remaining 5 sentences were used for testing. Note that the test sentences were not included in the training.

In the audible and whispered part, we parallely recorded the acoustic signal with a standard close-talking microphone connected to a USB soundcard. The audio signal is synchronized to the EMG signal with an analog marker. We also applied a 50 ms delay between audio and EMG signal, since the muscle activity precedes the acoustic sound and this number gave best results in EMG-based speech recognition experiments [3].

3.2. Preprocessing

The feature extraction for source EMG signals is based on *time-domain features* [3]. For any given feature $\mathbf{f}$, $\overline{\mathbf{f}}$ 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. Note that classical Voice Conversion typically uses first oder second order delta coefficients. We expect that the construction of adjacent feature vectors captures more complex 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 **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 27 ms respective 10 ms. In all cases, we apply Linear Discriminant Analysis (LDA) on the TD15 feature to reduce it to a final feature vector with 32 coefficients.

3.3. Experimental Results

For evaluation of our different setups we use Mel Cepstral Distortions (MCD). The MCD is a scaled Euclidian distance between the spectral features of the target audible/whispered speech and the spectral features of the synthesized EMG speech in decibel, given by the following equation:

$$\mathbf{MCD} = 10/\ln 10 \sqrt{2 \cdot \sum_{k=1}^{24} (\mathbf{mc}_t[k] - \mathbf{mc}_s[k])^2}$$

$mc_t[k]$ and $mc_s[k]$ denote the k-th mel cepstral coefficient of target and synthesized data. Smaller numbers implicate better results. Since we use only the MCD metric of Mel coefficients 1 - 24 for evaluation, we do not evaluate prosody.

3.3.1. Gaussian mixtures

In a first experiment we obtain the optimal number of Gaussian mixture components. Since we want to optimize this number towards a large session-independent system, we use the two speakers from the EMG-UKA corpus with a preferably high amount of recorded data. Both speakers additionally recorded 520 unique audible sentences. Note that we also did a prosodic analysis of the fundamental frequency with the same data in our previous work[6]. We vary the number of mixture components between 4 and 512 and train the GMM-based conversion with 500 sentences using the remaining 20 sentences for testing. Figure 2 shows the MCD of those mixture components.

Speaker 2 gets best results with 256 Gaussian mixture components, whilst Speaker 1 shows a minimal MCD with 128 mixtures – with only little difference to 256 mixture components.

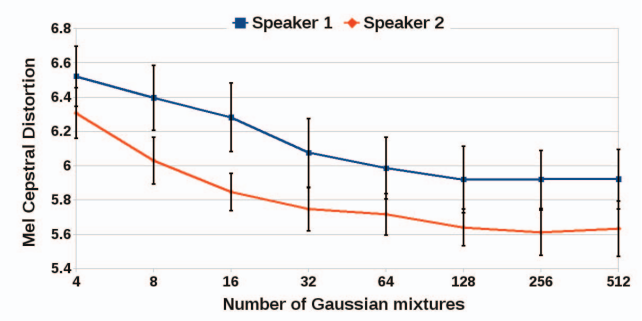

Fig. 2. Mel Cepstral Coefficients with different numbers of Gaussian mixtures

3.3.2. Synthesis of audible vs whispered speech from EMG

Based on our analysis of different speaking modes in [5] we transformed EMG-derived features from several speakers to audible and whispered speech. Since whispered speech nearly has no fundamental frequency and since we focus in this study on spectral features only, this gives us an interesting comparison on EMG to whispered vs audible speech conversion. Another motivation for having a closer look at the whispered speech synthesis is that whispered speech can be regarded as an in-between of audible and silent speech.

For this evaluation we want to compare several speakers and therefore use data from all five subjects of our subset EMG-UKA corpus. For each speaking mode we use 45 sentences for training and 5 sentences for testing our GMM-based voice conversion. Additionally we train the GMMs with combined data of 90 whispered and audible data, using the remaining 10 sentences for testing. Due to the lower amount of training data we set the number of Gaussian mixtures to 32. The MCDs with standard deviations are listed in Table 1. The best result on this data set shows speaker 2

Table 1. Mel Cepstral Coefficients (with standard deviations) of EMG-to-Speech conversion from five different speakers

Spk	EMG-to-AUD	to-WHIS	to-AUD/WHIS
1	6.164 (0.29)	5.920 (0.25)	6.204 (0.20)
2	4.999 (0.29)	4.528 (0.20)	5.220 (0.48)
3	5.927 (0.34)	5.519 (0.10)	6.094 (0.43)
4	7.357 (0.53)	6.021 (0.28)	7.265 (0.99)
5	6.537 (0.49)	5.630 (0.17)	6.542 (0.80)

with an MCD of 4.53 dB on whispered speech. In general EMG-to-Whisper gives better results compared to the EMG-to-Audio conversion. It's also noticable that there are large MCD variations between the speakers. The difference between the best and worst speaker in EMG-to-AUD is 2.36 and it is unclear what are the reasons for this discrepancy.

Therefore a closer look at e.g. articulatory features would be possible, instead of evaluating the whole sentence.
Surprisingly the results on audible speech in general are better compared to the prior experiment with 500 sentences. Apart from speaker 4 the MCDs of the combined speaking modes give worse results compared to the separate speaking modes.

3.3.3. Session independent evaluation

As a last experiment we investigate how a *session-independent* system will perform on our data. The term *session* indicates that during the recording of this session, the EMG electrodes were not reattached or removed. Since the training and test data for one speaker was only used from one single session, we only obtained session-dependent systems in the previous experiments. We now use the speaker from the EMG-UKA corpus with the highest number of recordings and train our GMM-based Voice Conversion with all that data. In total we use 16 sessions of speaker 1 resulting in a training set of 1480 sentences and a testing set of 370 sentences. The results with different numbers of Gaussian mixtures can be seen in Table 2.

Table 2. Mel Cepstral Coefficients (and standard deviations) of Session independent EMG-to-Speech conversion with different numbers of Gaussian mixtures.

# of mixtures	MCD (SD)
512	6.036 (0.99)
256	6.046 (1.01)
128	6.082 (1.02)
64	6.188 (1.01)
32	6.207 (1.03)

It is noticable that the MCD of around 6.1 is only slighty worse compared to the 500 sentence training session-dependent result from the same speaker, which gave an MCD of around 6. Thus we can state that the GMM-based voice conversion approach performs robust even with minor changes in the electrode placement or other influences.

[10] introduced a similar synthesis approach using NAM-to-speech conversion. They achieved an MCD of 5.99, improving this number to 5.77 using visual information. Compared to those numbers, we can see that with using only electromyographic data we can achieve reasonable performance in speech synthesis.

4. CONCLUSION

In order to convert surface electromyographic signals captured from a subject's face to audible speech, we successfully used a Gaussian Mixtured Model based Voice Conversion approach. The results of our experimental evaluation indicated that the optimal number of Gaussian mixtures depends on the amount of training data. We also achieved better results, in terms of a Mel Cepstral Distance (MCD) metric, with whispered speech compared to normal audible speech. A session-independent evaluation was introduced, which used recording sessions, between which the EMG electrodes were removed and reattached. The results on this data showed quite reasonable performance compared to similar session dependent conversions.

5. 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 Interspeech 2006, 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] M. Wand, A. Toth,S.C. Jou, and T. Schultz, "Impact of Different Speaking Modes on EMG-based Speech Recognition," Proceedings of Interspeech 2009, pp. 648–651, 2009.

[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] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous Probabilistic Transform for Voice Conversion," IEEE Transaction on Speech and Audio Processing (SAP), vol. 6, no. 2, pp. 131–142, 1998.

[8] A. Kain and M. W. Macon, "Spectral Voice Conversion for Text-to-speech Synthesis, " Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 285–288, 1998.

[9] M. Janke, M. Wand, and T. Schultz, "Spectral Energy Mapping for EMG-based Recognition of Silent Speech," Proceedings of First International Workshop on Bio-inspired Human-Machine Interfaces and Healthcare Applications, 2010.

[10] V. A. Tran, A. G. Bailly, H. Loevenbruck, and T. Toda, "Multimodal HMM-based NAM-to-speech conversion", Proceedings of Interspeech 2009, pp. 656–659, 2009.

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