Impact of Lack of Acoustic Feedback in EMG-based Silent Speech Recognition

Matthias Janke, Michael Wand, Tanja Schultz

Cognitive Systems Lab, Karlsruhe Institute of Technology,
Karlsruhe, Germany

matthias.janke@student.kit.edu, michael.wand@kit.edu, tanja.schultz@kit.edu

Abstract

This paper presents our recent advances in speech recognition based on surface electromyography (EMG). This technology allows for *Silent Speech Interfaces* since EMG captures the electrical potentials of the human articulatory muscles rather than the acoustic speech signal. Our earlier experiments have shown that the EMG signal is greatly impacted by the mode of speaking. In this study we extend this line of research by comparing EMG signals from audible, whispered, and silent speaking mode. We distinguish between phonetic features like consonants and vowels and show that the lack of acoustic feedback in silent speech implies an increased focus on somatosensoric feedback, which is visible in the EMG signal. Based on this analysis we develop a spectral mapping method to compensate for these differences. Finally, we apply the spectral mapping to the front-end of our speech recognition system and show that recognition rates on silent speech improve by up to 11.59% relative.

Index Terms: EMG, EMG-based speech recognition, Silent Speech Interfaces, somatosensoric feedback

1. Introduction

Automatic Speech Recognition (ASR) has matured to a point where it is successfully applied to ubiquitous applications and devices, such as telephone-based services and mobile personal digital assistants. Despite their success, speech-driven technologies still face two major challenges: recognition performance degrades significantly in the presence of noise, and confidential or private communications in public places are jeopardized by audible speech. Both of these challenges are addressed by Silent Speech Interfaces (SSI). A Silent Speech Interface is an electronic system enabling communication by speech without the necessity of emitting an audible acoustic signal.

In this paper, we report our most recent investigations in electromyographic (EMG) speech recognition, where the activation potentials of the articulatory muscles are directly recorded from the subject's face via surface electrodes[1].

While reliable automatic recognition of silent speech by means of electromygraphy is currently heavily investigated and recent performance results come within useful reach [1], little is known about the EMG signal variations resulting from differences in articulation between audible and silent speech production. The process of human speech production is very complex and subject to ongoing exploration, however, it is widely accepted that *acoustic feedback* plays a major role in uttering intelligible speech [2]. Therefore, this paper studies the variations in the EMG signal caused by *speaking modes*, i.e. by the difference between normal audible speech, whispered speech, and silently articulated speech, where no sound is heard. Since silent speaking results only in a movement of the articulatory muscles, we hypothesize that the lack of acoustic feedback is compensated by a stronger focus on somatosensoric feedback. The speaker therefore makes more prominent use of his proprioceptive and tactile impressions of skin or muscles in order to correctly reach the desired articulatory targets [2, 3].

Facial electromyography offers a way to quantify this effect. We extend the approach which we laid out in [4] by comparing the *signal energy* which is emitted while pronouncing different parts of speech, showing that in silent speaking mode, sounds which give a strong tactile sensation to the speaker are indeed articulated more powerfully than sounds for which the speaker gets less tactile feedback.

We then apply these experimental results to our silent speech recognizer, extending our *spectral mapping* algorithm [4] by a phoneme-based discrimination. This method allows to model audible and silent speech with joint acoustic models, which does not only improve the performance of silent speech recognition, but also makes it possible to seamlessly switch between audible and silent speaking mode.

This paper is organized as follows: In section 2 we describe the data corpus which we used for this study. Section 3 documents our EMG-based speech recognizer. Section 4 specifies our analytic experiments, and in section 5 we apply the results to the EMG-based speech recognizer. Section 6 concludes the paper.

2. The EMG-UKA Data Corpus

For our experiments, we recorded the *EMG-UKA* corpus [4] of EMG signals of audible, whispered, and silent speech of seven male speakers and one female speaker, aged between 24 and 28 years, which gives a total recording length of 243 minutes. Each speaker recorded between one and eleven sessions, resulting in a total amount of 22 sessions. The recording protocol was as follows: In a quiet room, the speaker read 50 English sentences for three times, first audibly, then in whispered speech, and at last silently mouthed. As an abbreviation, we call the EMG signals from these parts *audible EMG*, *whispered EMG* and *silent EMG*, respectively.

In each part we recorded one *BASE* set of 10 sentences which were identical for all speakers and all sessions, and one *SPEC* set of 40 sentences, which varied across sessions. In each session, these sentence sets were the same for all three parts, so that the database covers all three speaking modes with parallel utterances. The total of 50 BASE and SPEC utterances in each part were recorded in random order. In all recognition experiments, the 40 SPEC utterances are used for training, and the 10 BASE utterances are used as test set.

[1] Strictly spoken, the technology is called *surface electromyography*, however we use the abbreviation *EMG* for simplicity.

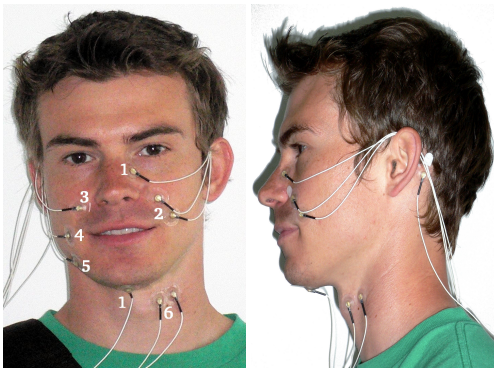

Figure 1: Electrode positioning

For EMG recording, we used a computer-controlled 6-channel EMG data acquisition system (Varioport, Becker-Meditec, Germany). All EMG signals were sampled at 600 Hz and filtered with an analog high-pass filter with a cut-off frequency at 60 Hz. We adopted the electrode positioning from [5] which yielded optimal results. Our electrode setting uses six channels and captures signals from the levator angulis oris (channels 2 and 3), the zygomaticus major (channels 2 and 3), the platysma (channel 4), the anterior belly of the digastric (channel 1) and the tongue (channels 1 and 6). Channels 2 and 6 use bipolar derivation, whereas channels 3, 4, and 5 were derived unipolarly, with two reference electrodes placed on the mastoid portion of the temporal bone (see Figure 1). Similarly, channel 1 uses unipolar derivation with the reference electrode attached to the nose. In the audible and whispered parts, we parallely recorded the audio signal with a standard close-talking microphone connected to a USB soundcard.

The *EMG-UKA* data corpus is intended to be comparable to the *EMG-PIT* corpus of EMG recordings of audible and silent speech [1]. However, the EMG-PIT corpus lacks whispered recordings, which we consider to be essential for our investigation in this paper.

3. The EMG-based Speech Recognizer

In this section we give a brief overview of the EMG-based speech recognizer which we use for our experiments. The modeling details are not required for the understanding of the remainder of this paper and are therefore omitted. The interested reader is refered to [1].

3.1. Feature Extraction

The feature extraction is based on *time-domain features* [6]. Here, for any given feature $\mathbf{f}$, $\bar{\mathbf{f}}$ is its frame-based time-domain mean, $\mathbf{P_f}$ is its frame-based power, and $\mathbf{z_f}$ is its frame-based zero-crossing rate. $S(\mathbf{f}, n)$ is the stacking of adjacent frames of feature $\mathbf{f}$ in the size of $2n + 1$ ($-n$ to n) frames.

For an EMG signal with normalized mean $x[n]$, the nine-point double-averaged signal $w[k]$ is defined as

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

The rectified high-frequency signal is $r[n] = |x[n] - w[n]|$. The final feature $\mathbf{TD15}$ is defined as follows [1]:

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

As in [7], frame size and frame shift were set to 27 ms respective 10 ms. In all cases, we apply LDA on the $\mathbf{TD15}$ feature to reduce it to 32 dimensions.

3.2. Training and Decoding

Our EMG speech recognition system is based on three-state left-to-right fully continuous Hidden-Markov-Models, which are used to model phonetic features (PFs) representing properties of phonemes. We use *bundled phonetic features* to optimally model phonetic units with little training data [1].

For decoding, we use the trained acoustic model together with a trigram Broadcast News language model giving a perplexity on the test set of 24.24. The decoding vocabulary is restricted to the words appearing in the test set, which results in a test vocabulary of 108 words.

3.3. Cross-Modal Initialization

Initializing an EMG-based Continuous Silent Speech recognizer is a challenging task since to initialize acoustic models representing sub-word units (phonemes or phonetic features), one needs a *time-alignment* of the training material, i.e. information about the phoneme boundaries in the training utterances. Our previous work on *audible* EMG data used a conventional speech recognizer on the parallely recorded audio stream to create such a time-alignment and then forced-aligned the training sentences. However, this method is infeasible for silent EMG.

We therefore have to create time-alignment for silent EMG using the EMG signal only. Our method is as follows:

- A "base speaker dependent recognizer" is trained in advance on audible or whispered EMG, using data from the same speaker and the same session.
- We use the trained models from this recognizer to create a time-alignment for the silent EMG data of that speaker.
- Then we forced-align the silent EMG data and do a full training run for a particular speaker. This means that we create specific acoustic models for silent EMG.

Clearly, the quality of the produced time-alignments is affected by the discrepancy between audible and silent EMG. However, in [7] we found this initialization method to be the most suitable one which can be applied before the EMG signal mapping which we develop in this paper.

Figure 2 gives an example for a recorded EMG signal (channel 1) of the utterance "We can do it". At the bottom of the signal the aligned phones are listed.

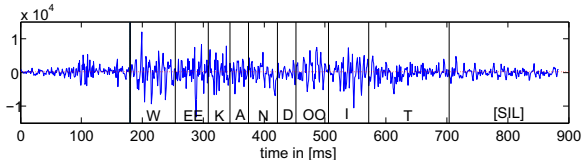

Figure 2: Example for the EMG signal (channel 1) of the utterance "We can do it"

4. Spectral Density Comparison of Audible, Whispered and Silent EMG

In this section, we consider the power spectral density (PSD) of EMG signals to obtain a measure of the articulatory power. In

order to get sufficiently fine-grained results, we break down the raw signal into phoneme units before computing the PSD.

The PSD computation is based on Welch's method [8] and works as follows:

- The input signal is divided into frames with a length of 16 samples.
- Each frame is windowed with a Hamming window to reduce spectral distortion.
- On each frame, the Fast Fourier Transform is computed, yielding 9 Fourier coefficients, which represent the EMG signal spectrum from 0 Hz to 300 Hz (the Nyquist Frequency of the signal).

According to [4], we investigated the PSD of EMG signals of audible, whispered and silent speech. We split up the signals into frames with a length of 27ms and a frame shift of 10ms, which is the same setup as used by the EMG recognizer.

The resulting frames can then be grouped according to the phonemes or phonetic features they represent. Clearly, the quality of this grouping depends on the quality of the underlying time alignment, which must be computed in advance (see section 3.3). Note that from the set of six available EMG channels, we picked channels 2 and 4 (see Figure 1) for our experiments since these electrodes are closest to the muscles which control the lips, and since we assumed that the lip movement is one major source of somatosensoric feedback while speaking.

The difference between vowels and consonants is among the most fundamental distinctions of English phonology. Therefore, our first experiment compares the PSDs of consonants and vowels. We picked the designated phones from the according sessions and computed the mean audible, whispered and silent PSD over all sessions. Figure 3 shows the PSDs for consonants and vowels in channel 2 and channel 4. It can be seen that audible EMG in general has the highest spectral density. Whispered EMG seems to take the assumed role of an in-between of audible and silent speech, since the whispered PSDs are in the range between the other speech modes. While the PSD shapes of silent and audible consonants differ only little, there is a noticeable difference in the vowel PSD chart. A reason for the higher vowel PSD in audible EMG could be the fact that (English) vowels are syllable peaks and thus major articulatory targets. When the speaker lacks acoustic feedback while articulating, this might have the consequence that acoustic targets are not fully reached any more, thus causing a less intense vowel articulation.

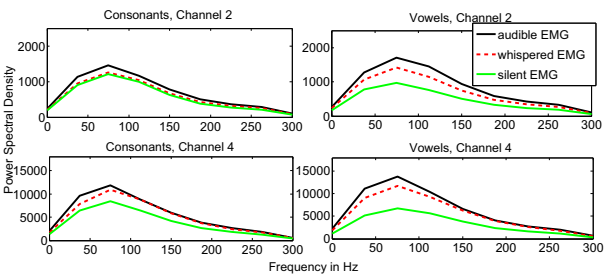

Figure 3: Comparison of PSDs of vowels and consonants in audible, whispered and silent EMG

For further experiments we subdivided the classes of consonants and vowels. We performed these subdivisions according to the degree of somatosensoric feedback which we considered to occur during the pronunciation of the respective phonemes.

We first considered two groups of consonants, namely *velar* (i.e. 'g', 'k' and 'ng') and *bilabial* consonants ('p','b' and 'm'). Since the articulation of bilabial sounds involves the complete closure at the lips, we hypothesize that a missing acoustic feedback may be compensated by the tactile feedback of the lip closure. This would imply that we can measure a relatively stronger articulation of bilabial consonants than of velar consonants, which are articulated with the back part of the tongue, with significantly smaller somatosensoric feedback.

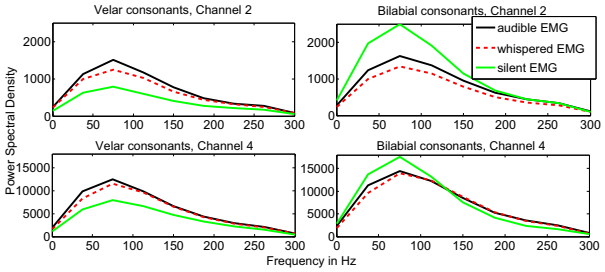

Figure 4: Comparison of PSDs of bilabial and velar consonants in audible, whispered and silent EMG

Figure 4 charts the PSDs of EMG channels 2 and 4. The result markedly shows that there is a significant difference between bilabial and velar consonants: While for velar consonants, we have the usual pattern of decreasing PSD from audible to silent speaking mode, this order is reversed for bilabial consonants, where the articulation power is much higher in silent speaking mode than in any other mode. Also note that the PSDs of bilabial and velar consonants in audible and whispered speaking mode are almost identical. This shows that bilabial consonants are indeed hyperarticulated in silent speech, confirming our hypothesis that a speaker makes use of an increased tactile feedback when speaking silently.

As a next step we split up the class of vowels in an analog way. Rounded vowels are produced when the lips form a circular opening, while unrounded vowels are pronounced with relaxed lips. Again, the assumption is that the lip rounding is more pronounced in silent speech than in other speaking modes.

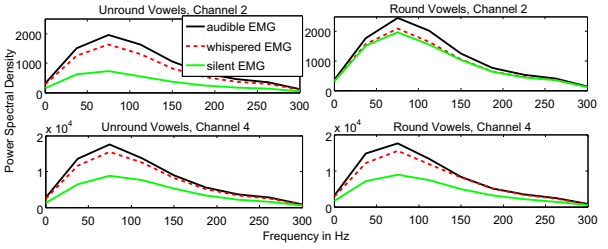

Figure 5: Comparison of PSDs of round and unround vowels in audible, whispered and silent EMG

Figure 5 charts the PSDs of channel 2 and 4 for rounded and unrounded vowels. Although the effect is less obvious than for bilabial and velar consonants, it can be seen that the silent articulation of rounded vowels is relatively stronger than the one of unrounded vowels. One can also see that for audi-

ble and whispered speaking mode, there is again no difference between rounded and unrounded vowels. This again supports the hypothesis that rounded vowels tend to provide a higher somatosensoric feedback than unrounded vowels and are therefore hyperarticulated when speaking silently.

The analysis quantifies the relative articulation power in audible and silent speech for individual consonants. Since the PSD charts always show a peak in the frequency range of approximately 75 Hz, we computed the PSD ratio of audible to silent EMG at this frequency bin and plot the mean ratio in channel 2 and 4 for ten different consonants in Figure 6. Additionally the error ellipses are given, which represent a confidence region based on the covariance matrices.

The PSD ratio between audible and silent EMG is a measure of the discrepancy in articulation between audible and silent speech. Assuming that audible speech gives the "normal" power of articulation for a specific phoneme, a PSD ratio smaller than 1 means that a phoneme is hyperarticulated, while a high ratio means that this phoneme is relatively weakly pronounced when speaking silently. Our hypothesis, as described in the previous paragraphs, is that a small PSD ratio correlates with a high somatosensoric feedback. This is confirmed by the scatter plot: The three consonants of the bilabial group show the smallest ratio, whilst consonants with small somatosensoric feedback – like 'g' ,'ng' or 'hh'– give a high ratio.

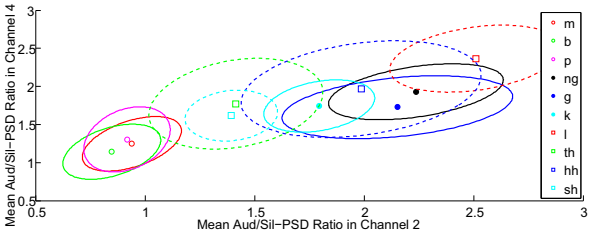

Figure 6: PSD ratios of single consonants in channel 2 (X-axis) and channel 4 (Y-axis) — a small ratio indicates that this phoneme is articulated strongly in silent speech

5. Speech Recognition Experiments

Based on the analysis in section 4, we extended our *spectral mapping* algorithm from [4] by a phoneme class distinction. The algorithm is applied to each channel of a *silent* EMG signal and works as follows:

1. We define a "base speaking mode" (audible or whispered) and a "mapping phonetic feature (PF)" (e.g. vowels) for this experiment.
2. According to the time alignment, every frame of an utterance is binarily classified as "mapping PF" or not. On the class of "mapping PF" frames, we compute the average PSD ratio between "base speaking mode" and target silent EMG (as a function of the frequency). We call this ratio *mapping factor*.
3. Each frame of an utterance is transformed into the frequency domain by the Fast Fourier Transform (FFT). Each frequency component is multiplied by the corresponding mapping factor, if the frame belongs to the "mapping PF" group. Otherwise the frame remains unchanged. The resulting frequency representation of the signal is transformed back into the time domain by applying the inverse FFT.
4. After this procedure, the transformed signal is used for the training and testing process as usual [1]. The details of this process have been described in section 3.

Note that the computation of the mapping factors is done independently for each session of each speaker and for each channel of the EMG signals.

We applied this mapping with consonants and vowels as "mapping PF". Table 1 gives the resulting Word Error Rates averaged over all speakers on silent EMG utterances.

Mapping Method	*WER*	Δ *Baseline*
Baseline	61.67%	
Consonant-Mapping	57.16%	**7.31%**
Vowel-Mapping	54.52%	**11.59%**

Table 1: Mean Word Error Rates for the used mapping methods

6. Conclusions

We compared the discrepancies between different speaking modes, analyzing the spectral densities of audible, whispered and silent EMG signals. While there is only little difference between whispered and audible speaking mode, we proved that in silent speech, sounds whose articulation gives a strong tactile feedback are articulated relatively stronger than others. This result confirms our assumption that the speaker compensates the lack of acoustic feedback by focusing on the somatosensoric feedback—indicated by emphasizing those sounds which provide more somatosensoric feedback than others. We achieved a spectral mapping of selected phonetic features, which gives an improvement of 11.59% WER relative to our baseline system.

7. References

[1] T. Schultz and M. Wand, "Modeling Coarticulation in Large Vocabulary EMG-based Speech Recognition," *Speech Communication*, vol. 52, no. 4, pp. 341 – 353, 2010.

[2] F. H. Guenther, S. S. Ghosh, and J. A. Tourville, "Neural Modeling and Imaging of the Cortical Interactions underlying Syllable Production," *Brain and Language*, vol. 96, pp. 280 – 301, 2006.

[3] F. H. Guenther, M. Hampson, and D. Johnson, "A Theoretical Investigation of Reference Frames for the Planning of Speech Movements," *Psych.Rev.*, vol. 105, pp. 611 – 633, 1998.

[4] M. Janke, M. Wand, and T. Schultz, "A Spectral Mapping Method for EMG-based Recognition of Silent Speech," in *First International Workshop on Bio-inspired Human-Machine Interfaces and Healthcare Applications, B-INTERFACE*, 2010.

[5] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, "Session Independent Non-Audible Speech Recognition Using Surface Electromyography," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, San Juan, Puerto Rico, 2005, pp. 331 – 336.

[6] S.-C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards Continuous Speech Recognition using Surface Electromyography," in *Proc. Interspeech*, Pittsburgh, PA, 2006, pp. 573 – 576.

[7] M. Wand, S.-C. S. Jou, A. R. Toth, and T. Schultz, "Impact of Different Speaking Modes on EMG-based Speech Recognition," in *Proc. Interspeech*, 2009.

[8] P. Welch, "The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method based on Time Averaging over Short, Modified Periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.