INTERSPEECH 2009 BRIGHTON

Impact of Different Speaking Modes on EMG-based Speech Recognition

Michael Wand[1], *Szu-Chen Stan Jou*[2], *Arthur R. Toth*[1], *Tanja Schultz*[1]

[1]Cognitive Systems Lab, University of Karlsruhe, Germany
[2]ATC, ICL, Industrial Technology Research Institute, Taiwan
mwand@ira.uka.de, stanjou@gmail.com, atoth@cs.cmu.edu, tanja@ira.uka.de

Abstract

We present our recent results on speech recognition by surface electromyography (EMG), which captures the electric potentials that are generated by the human articulatory muscles. This technique can be used to enable *Silent Speech Interfaces*, since EMG signals are generated even when people only articulate speech without producing any sound. Preliminary experiments have shown that the EMG signals created by audible and silent speech are quite distinct. In this paper we first compare various methods of initializing a silent speech EMG recognizer, showing that the performance of the recognizer substantially varies across different speakers. Based on this, we analyze EMG signals from audible and silent speech, present first results on how discrepancies between these speaking modes affect EMG recognizers, and suggest areas for future work.

Index Terms: speech recognition, surface electromyography, silent speech, articulation

1. Introduction

Automatic Speech Recognition (ASR) has now matured to a point where it is successfully deployed in a wide variety of everyday life applications, including telephone-based services and speech-driven applications on all sorts of mobile personal digital devices.

Despite this success, speech-driven technologies still face two major challenges: first, recognition performance degrades significantly in the presence of noise. Second, confidential and private communication in public places is difficult due to the clearly audible speech.

Both of these challenges may be approached by Silent Speech Interfaces (SSI). A Silent Speech Interface is an electronic system enabling speech communication to take place without the necessity of emitting an audible acoustic signal. In the past years, several techniques were proposed to recognize speech without producing clearly audible speech, among them the recognition of whispered speech with a throat microphone [1] or non-audible murmur with a special stethoscopic microphone [2]. Other approaches include using optical or ultrasound images of the articulatory apparatus, i.e. [3], or subvocal speech recognition [4].

In this paper, we present 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. This approach has two major advantages: firstly, it is able to recognize completely silent speech, where not even a whispering sound is uttered. Secondly, compared in particular to the optical or ultrasound approach, the required technology is relatively lightweight and comes at a manageable price.

Research in the area of EMG-based speech recognition has only a short history. In 2002, [5] showed that myoelectric signals can be used to discriminate a small number of words. The task of recognizing continuous speech via EMG was approached in 2006, when [6] showed that speaker dependent recognition of continuous speech via EMG is possible. Recent results include advances in acoustic modeling using a clustering scheme on *phonetic features*, which represent properties of a given phoneme, such as the place or the manner of articulation. In [7], we report that a recognizer based on such *bundled phonetic features* performs more than 30% better than a recognizer based on phoneme models only.

2. Purpose of This Study

With continuous-speech EMG recognition in place, tackling variations in the EMG signal is the next major goal. One kind of discrepancy arises with differences in the articulatory apparatus and varying tissue properties of multiple speakers and may be addressed by adaptation methods [8]. In this paper we address another source of variation in the EMG signal, namely the one caused by different speaking styles. We distinguish *audible EMG*, i.e. EMG signals recorded on normally pronounced speech, and *silent EMG*, i.e. signals from voicelessly mouthed speech. Since the capability of recognizing silent speech is a particular strength of EMG-based speech recognition, investigating how speech modes affect articulation and, eventually, the measured EMG signal, is of high interest to the silent speech research community.

In [9] we performed cross-modal experiments for the first time, i.e. we trained models on silent EMG and tested on audible EMG and vice versa. The results suggested that the EMG signals are impacted by the speaking modality. Furthermore, the cross-modal application gave better results for those speakers who had experience in speaking silently. We assume that the differences in audible versus silent EMG signals may stem from a larger variability in articulation which might be due to a lack of acoustic feedback when speaking silently. Therefore, we investigate in this paper on a larger number of subjects to what extent silent and audible EMG signals differ and how these differences impact the speech recognition performance. An example of audible and silent EMG signals is shown in figure 1.

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. In [10], the authors argue that the articulation process is defined by *auditory targets*, i.e. by phonemes which the speaker desires to utter, and present experimental results from American English which support this claim. The overview article [11] even goes one step further in saying that the process of speaking aims at achieving "sufficient perceptual contrast [...]

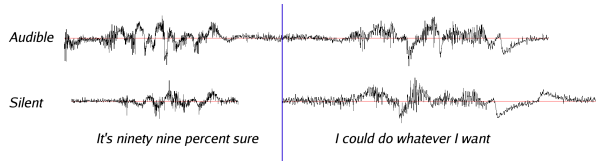

Figure 1: One-channel EMG signals for two sentences, audible: upper row, silent: lower row.

with minimal effort".

This means that when acoustic feedback fails, a major regulator for the articulation process is lacking. There exist studies which quantify the variability of articulation between different speaking modes. For example, in [12] the authors compare the duration and intensity of whispered and normally pronounced consonants, showing that whispered consonants have a prolonged duration compared to normally spoken consonants and asserting that these results are consistent across speakers. Further works deal with the effect of a disrupted acoustic feedback on the speech production process, see i.e. [13] and the references therein.

For the purpose of silent speech recognition, this means that one has to find suitable methods of dealing with these articulation differences. On the other hand, if a speaker is experienced in speaking silently, it may remedy part of the problem. In this paper we use a large corpus of audible and silent EMG recordings from many speakers (see section 3.1). We investigate several methods of training an initial EMG recognizer for silent speech and show that the recognition rates vary over a great range depending on the speaker. Then we compare the EMG signals of corresponding audible and silent utterances and show how variations in the signal properties relate to the recognizer accuracy when applied to silent speech.

The remainder of this paper is organized as follows: In section 3, we present our data corpus and method of data acquisition, and in section 4 we give an overview of the experiments we conducted on silent EMG recognition. The results section is split in two parts: In section 4.1 we present our results on different training approaches for silent EMG, while in section 4.2 we show the relation between EMG signal properties and recognition rate. Section 5 concludes the paper.

3. Data Corpus and Experimental Setup

3.1. The EMG-PIT Data Corpus

During the years 2007 - 2008 we collected a large database of EMG signals from 78 speakers. This collection was done in a joint effort with colleagues from the Department of Communication Science and Disorders at University of Pittsburgh [14]. The resulting data corpus bears the name *EMG-PIT*; to the best of our knowledge it is the largest corpus of EMG recordings of speech so far.

The collection was done in two phases, a pilot study with 14 speakers, each recording two sessions, and a main study with 64 speakers each recording one session. Each session consisted of two parts, one audible speech part and one silent speech part. In each part we recorded one BASE set of 10 sentences which were identical across all speakers, and one SPEC set of 40 sentences which were recorded by this speaker only. These sentence sets were the same for the audible and the silent speech part, so that the database covers both speaking modes with parallel utterances. The total of 50 BASE and SPEC utterances in each part were recorded in random order.

For EMG recording we used a computer-controlled 8-channel EMG data acquisition system (Varioport, Becker-Meditec, Germany). All EMG signals were sampled at 600 Hz. We adopted the electrode positioning from [9] which yielded optimal results. The audible utterances were simultaneously recorded with a conventional air-transmission microphone.

This article reports results on the *pilot study* data only. As in previous studies, we used the BASE sentences for testing and the SPEC sentences as training sets.

In order to study articulation differences between trained and untrained speakers, we augmented the EMG-PIT data by one further session of one speaker who had recorded several audible and silent speech EMG sessions before. In the following experiments, this speaker bears the speaker number 15. For this session we used the same recording setup as in the EMG-PIT corpus, however the session consisted of 500 sentences, half of which were pronounced audibly and the other half silently. Each part of 250 sentences was split into 60 test sentences and 190 training sentences.

Thus the corpus of utterances which was used for this study has the following properties:

Speakers	14 speakers	1 speaker
Sessions	2 sessions per speaker	1 session
Average Length per session		
(total)	467 seconds	3762 seconds
(audible)	231 seconds	1770 seconds
(silent)	236 seconds	1992 seconds
Domain	Broadcast News	

3.2. The EMG Recognizer

The initial EMG recognizer was taken from [6]. It used an HMM-based acoustic modeling, which was based on fully continuous Gaussian Mixture Models. All experiments used *bundled phonetic features (BDPFs)* for the final training and decoding. Phonetic features (PFs) represent properties of a given phoneme, such as the place of articulation or the manner of articulation. The architecture we employ for the PF-based EMG decoding system is a *multi-stream* architecture [15], which means that the models draw their *acoustic probabilities* not from one single source, but from a weighted sum of various sources which correspond to acoustic models representing substates of PFs, like "middle of a vowel" or "end of a non-fricative". The conventional EMG phoneme-based recognizer contributes as well.

Phonetic feature bundling [7] is the process of pooling dependent features together, so that eventually we will end up with a set of PF acoustic models which represent *bundles* of PFs, like "voiced fricative" or "rounded front vowel". The algorithm which performs this pooling is a standard decision-tree based clustering approach [16], as it is successfully used in large vocabulary acoustic speech recognition to determine phoneme context clusters. On our corpus, the best average word error rate of this recognizer on *audible* utterances is 30.19%.

3.3. Feature Extraction

We use a feature extraction method based on *time-domain features* [6]. We use the following definitions [6]: For any feature f, $\bar{f}$ is its frame-based time-domain mean, P_f is its frame-based power, and z_f is its frame-based zero-crossing rate. $S(f, n)$ is the stacking of adjacent frames of feature 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 $r[n] = |x[n] - w[n]|$. In baseline experiments with audible EMG, the best word error rate is obtained with the following feature, which we use in this study as well:

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

3.4. Testing

For decoding, we use the trained acoustic model together with a trigram BN language model. We restricted the decoding vocabulary to the words appearing in the test set. This resulted in a test set of 10 sentences per speaker with a vocabulary of 108 words.

4. Experiments

4.1. Initializing a Silent Speech EMG Recognizer

The first batch of experiments deals with initializing a recognizer for silent speech EMG. This is a challenging task since in order 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. Previous works on *audible* EMG data used a conventional speech recognizer on the parallely recorded audio stream in order to create such a time-alignment and then forced-aligned the training sentences based on this information, see [6] for a detailed explanation. However for silent EMG, this method is unfeasible, and information on the phoneme boundaries is not readily available.

We investigate the following methods for creating a silent speech EMG recognizer. All of these methods rely on the fact that we have parallely recorded audible and silent utterances.

- *Cross-Modal Testing*: We train the recognizer on audible EMG and test it on silent EMG, so there is no training or adaptation to silent EMG. However, this method is clearly suboptimal if EMG signals for silent mode differ from those in audible mode.
- *Cross-Modal Labeling*: We use trained models from audible EMG to create a time-alignment for the silent EMG data. Then we forced-align the silent EMG data based on this information and do a full training run. This means that we create specific acoustic models for silent EMG.
- *Mapped Labels*: A direct way to obtain a time-alignment for silent EMG recordings is considering the corresponding *audible* utterance and mapping the phoneme boundaries on the corresponding silent utterance. This involves compensating for different utterance lengths as well as determining the exact speech onset and offset.

  In order to obtain exact boundaries of the utterances, we use an HMM-based silence detector which only trains two acoustic models, namely "silence" and "non-silence". Testing this recognizer on audible EMG data, using the audio time-alignment as ground truth, shows that the average absolute error of this silence detector is 14.69 milliseconds, which is less than 9 samples of the raw EMG signal.

  With this information, we paired up the audible and the silent utterance from the same speaker with the same content. Then we mapped the time-alignment from the audible utterance to the silent utterance, compensating for differing lengths by linearly growing or shrinking the phoneme lengths.
- *Speech synthesis:* A completely different way of handling EMG signals is presented in [17], where it is shown that a GMM-based transformation from the EMG signal to an audio signal can be trained. We therefore used the *transformed audio data* from these experiments to train and test a conventional speech recognizer based on MFCCs, using BDPF acoustic modeling as for the EMG recognizer.

The resulting word error rates are charted in figure 2 and show that Cross-Modal Labeling generally gives the best results so far. While for the majority of speakers the word error rates (WER) exceed 80%, we find that in particular for speakers 6 and 11, the word error rates are in a reasonable range, and that for speaker 15, who is the one speaker who had recorded several silent speech sessions before, the best WER achieved on silent EMG is 30.10%, which is about the same as for the audible utterances of this speaker.

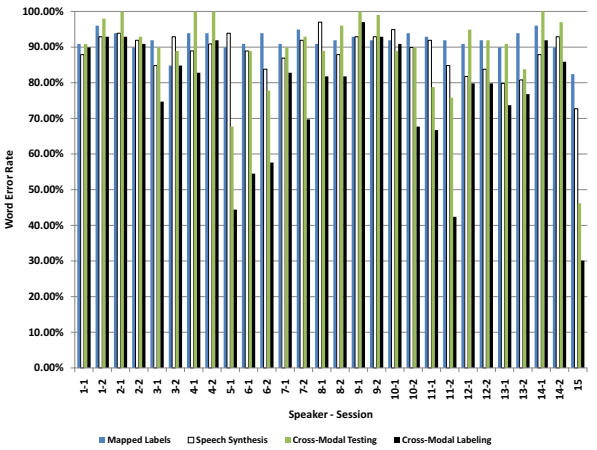

Figure 2: Word Error Rates per Speaker and Session for Different Recognizer Initialization Methods

From having such a large variance in error rates, one can conclude that the properties of audible and silent articulation greatly vary across speakers. In particular, the good result for speaker 15 supports the assertion in [9] that training to articulate silent speech leads to better results. In the following section, we investigate the differences between audible and silent EMG recordings and give some suggestions as to what makes this difference.

4.2. Analysis of Silent vs Audible EMG Signals

As a first step, we paired corresponding audible and silent utterances and compared their respective durations. The results are charted on the left-hand side of figure 3 and show a high correlation between the durations of corresponding audible and silent utterances. The correlation coefficient is 0.77.

Next, we analyzed the average time-domain magnitude of EMG channel 1, again for corresponding utterances. This channel is particularly interesting since it mainly reflects the opening

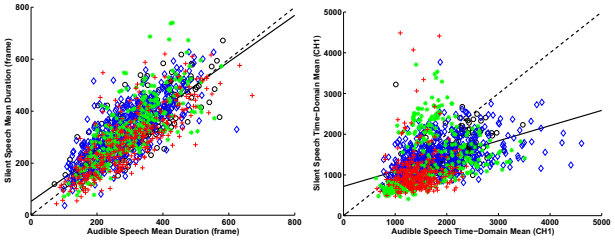

Figure 3: Comparison between durations (left) and time-domain means (right) of corresponding audible and silent utterances, with regression lines. The dots are labeled with respect to speaker-wise WERs. *circle*: WER $< 60\%$, *diamond*: $60\% <$ WER $< 80\%$, *star*: $80\% <$ WER $< 90\%$, *cross*: WER $> 90\%$.

and closing of the mouth, giving an estimate of the power the speaker puts into articulation. The result is charted on the right-hand side of figure 3 and shows a more interesting picture: One sees that on average, the magnitude of silent utterances is significantly *lower* than that of corresponding audible utterances. The correlation coefficient is 0.36. Interestingly, one gets a very similar graph when one charts the average magnitude of corresponding utterances from session 2 vs session 1 of one speaker, which may be explained be the fact that after about two hours of recording, the speakers got tired or lost concentration.

Based on this observation, in figure 4 we chart the *absolute difference* of the average magnitudes of audible and silent utterances versus the utterance-based Word Error Rate. One sees that while for utterances with high Word Error Rate, the differences are almost uniformly distributed in a rather wide range, for most "good" utterances with low WER the distance of the average magnitudes is quite low.

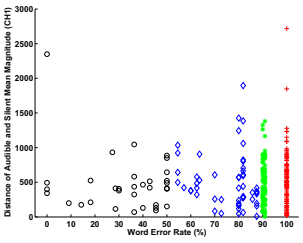

Figure 4: Utterance-based Absolute Deviation of EMG Channel 1 versus Word Error Rate

5. Conclusions

In this paper we presented first results on training and testing continuous-speech EMG-based recognizers for *silent* speech, comparing four different methods for initializing such a recognizer. Our findings show that initializing such a recognizer is possible, however variance in articulation between audible and silent EMG is still a major problem to be overcome.

Observing that the WER of a silent speech EMG recognizer varies substantially across different speakers, with the best results for the most experienced speaker, we analyzed the utterances in our corpus, pointing out that comparing corresponding utterances, audible EMG generally has a higher signal magnitude than silent EMG. This information can on the one hand be used to further investigate how to map from audible EMG signals to silent EMG signals; on the other hand it could also be used to provide feedback on a subject's articulation for diagnostic purposes.

6. Acknowledgements

Szu-Chen Stan Jou is supported by Project 8353C41220 conducted by ITRI under sponsorship of the Ministry of Economic Affairs, Taiwan, R.O.C.

7. References

[1] S.-C. Jou, T. Schultz, and A. Waibel, "Whispery Speech Recognition Using Adapted Articulatory Features," in *Proc. ICASSP*, 2005.

[2] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Non-Audible Murmur Recognition," in *Proc. Eurospeech*, 2003.

[3] T. Hueber, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Continuous-Speech Phone Recognition from Ultrasound and Optical Images of the Tongue and Lips," in *Proc. Interspeech*, 2007, pp. 658–661.

[4] C. Jorgensen and K. Binsted, "Web Browser Control Using EMG Based Sub Vocal Speech Recognition," in *Proceedings of the 38th Hawaii International Conference on System Sciences*, 2005.

[5] A. Chan, K. Englehart, B. Hudgins, and D. Lovely, "Hidden Markov Model Classification of Myolectric Signals in Speech," *Engineering in Medicine and Biology Magazine, IEEE*, vol. 21, no. 9, pp. 143–146, 2002.

[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, Sep 2006.

[7] T. Schultz and M. Wand, "Modeling Coarticulation in Large Vocabulary EMG-based Speech Recognition," Speech Communication Journal, 2009, to Appear.

[8] M. Wand and T. Schultz, "Towards Speaker-Adaptive Speech Recognition Based on Surface Electromyography," in *Proc. Biosignals*, 2009.

[9] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, "Session Independent Non-Audible Speech Recognition Using Surface Electromyography," in *Proc. ASRU*, 2005.

[10] 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.

[11] J. Perkell, M. Matthies, H. Lane, F. Guenther, R. Wilhelms-Tricaricoa, J. Wozniaka, and P. Guioda, "Speech Motor Control: Acoustic Goals, Saturation Effects, Auditory Feedback and Internal Models," *Speech Communication Journal*, vol. 22, pp. 227 – 250, 1997.

[12] S. T. Jovicic and Z. Saric, "Acoustic Analysis of Consonants in Whispered Speech," *Journal of Voice*, vol. 22, no. 3, pp. 263 – 274, 2006.

[13] J. A. Jones and D. Striemer, "Speech Disruption During Delayed Auditory Feedback with Simultaneous Visual Feedback," *J Acoust Soc Am*, vol. 122, pp. 135 – 141, 2007.

[14] M. Dietrich, "The Effects of Stress Reactivity on Extralaryngeal Muscle Tension in Vocally Normal Participants as a Function of Personality," Ph.D. dissertation, University of Pittsburgh, 2008.

[15] S.-C. S. Jou, T. Schultz, and A. Waibel, "Continuous Electromyographic Speech Recognition with a Multi-Stream Decoding Architecture," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2007), Honolulu, Hawaii, US, April 15-20, 2007*, 2007.

[16] L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahmoo, and M. A. Picheny, "Decision Trees for Phonological Rules in Continuous Speech," in *Proc. ICASSP*, 1991.

[17] A. Toth, M. Wand, and T. Schultz, "Speech Synthesis from EMG Signals," in *Proc. Interspeech*, Brighton, UK, 2009.