SESSION-INDEPENDENT EMG-BASED SPEECH RECOGNITION

Michael Wand and Tanja Schultz
Cognitive Systems Lab, Karlsruhe Institute of Technology, Adenauerring 4, 76131 Karlsruhe, Germany

Keywords: Electromyography, Silent Speech Interfaces, EMG-based Speech Recognition.

Abstract: This paper reports on our recent research in speech recognition by surface electromyography (EMG), which is the technology of recording the electric activation potentials of the human articulatory muscles by surface electrodes in order to recognize speech. This method can be used to create *Silent Speech Interfaces*, since the EMG signal is available even when no audible signal is transmitted or captured. Several past studies have shown that EMG signals may vary greatly between different recording sessions, even of one and the same speaker. This paper shows that *session-independent* training methods may be used to obtain robust EMG-based speech recognizers which cope well with unseen recording sessions as well as with speaking mode variations. Our best session-independent recognition system, trained on 280 utterances of 7 different sessions, achieves an average 21.93% Word Error Rate (WER) on a testing vocabulary of 108 words. The overall best session-adaptive recognition system, based on a session-independent system and adapted towards the test session with 40 adaptation sentences, achieves an average WER of 15.66%, which is a relative improvement of 21% compared to the baseline average WER of 19.96% of a session-dependent recognition system trained only on a single session of 40 sentences.

1 INTRODUCTION

Automatic Speech Recognition (ASR) has now reached a level of precision and robustness which allows its use in a variety of practical applications. Notwithstanding, speech recognition suffers of several drawbacks which arise from the fact that ordinary speech is required to be clearly audible and cannot easily be masked: on the one hand, recognition performance degrades significantly in the presence of noise. On the other hand, confidential and private communication in public places is difficult if not impossible. Even when privacy is not an issue, audible speech communication in public places frequently disturbs bystanders.

Both of these challenges may be alleviated by Silent Speech Interfaces (SSI). A Silent Speech Interface is a system enabling speech communication to take place without the necessity of emitting an audible acoustic signal, or when an acoustic signal is unavailable (Denby et al., 2010). Our approach to capture silent speech relies on surface ElectroMyoGraphy (EMG), which is the process of recording electrical muscle activity using surface electrodes. Since speech is produced by the activity of the human articulatory muscles, the EMG signal measured in a person's face may be used to retrace the corresponding speech, even when this speech is produced silently, i. e. articulated without any vocal effort. Application areas for EMG-based Silent Speech Interfaces include robust, confidential, non-disturbing speech recognition for human-machine interfaces and transmission of articulatory parameters for example by a mobile telephone for silent human-human communication.

Previous EMG-based speech recognition systems were usually limited to very small tasks and vocabularies. A main reason for this limitation was that those systems were usually *session-dependent*, i. e. they used training and test data from only one speaker and only one recording session. Here, the term *recording session* means that during the recording, the EMG electrodes were not removed or reattached. This paper presents our first *session-independent* and *session-adaptive* systems: We show that a system trained on multiple recording sessions of one and the same speaker yields a reasonable performance, and that a session-independent system recognizes test data from unseen sessions more robustly than a similarly large recognizer trained on data from just one session. We additionally prove that the increased robustness of a session-independent system also helps to cope with the difference between normal and silently articulated speech. Finally, we investigate how the system copes with increasing recognition vocabulary

Wand M. and Schultz T..
SESSION-INDEPENDENT EMG-BASED SPEECH RECOGNITION.
DOI: 10.5220/0003169702950300
In *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing* (BIOSIGNALS-2011), pages 295-300
ISBN: 978-989-8425-35-5

sizes and present results on an EMG-based speech recognition system with a vocabulary of more than 2000 words, which to the best of our knowledge is the largest vocabulary which has ever been used for recognizing speech based on EMG signals.

The remainder of this paper is organized as follows: In section 2, we give an overview of previous related works. Section 3 presents our data corpus, and section 4 describes the setup of our EMG-based speech recognizer. In section 5, we present our experiments and results, and section 6 concludes the paper and outlines possible future work.

2 RELATED WORK

The use of EMG for speech recognition dates back to the mid 1980s, however competitive performance was first reported by (Chan et al., 2001), who achieved an average word accuracy of 93% on a 10-word vocabulary of English digits. Good performance could be achieved even when words were spoken silently (Jorgensen et al., 2003), suggesting this technology could be used for Silent Speech Interfaces.

Jou et al. (Jou et al., 2007) successfully demonstrated that phonemes can be used as modeling units for EMG-based speech recognition, thus allowing recognition of continuous speech. Phoneme models can be improved by using a clustering scheme on *phonetic features*, which represent properties of a given phoneme, such as the place or the manner of articulation (Schultz and Wand, 2010); this modeling scheme has also been also employed for this study.

There exist some studies on speaker adaptation for EMG-based speech recognition tasks (Maier-Hein et al., 2005; Wand and Schultz, 2009). Generally speaking, these experiments show that when data of different speakers is combined, the recognition performance degrades severely. In this paper we instead propose session-independent EMG-based speech recognition systems as a goal which is both tractable and practically relevant.

3 DATA CORPUS

Our corpus is based on a subset of the *EMG-UKA* corpus (Janke et al., 2010a) of EMG signals of speech. This subset consists of 32 recording sessions of those two speakers who had recorded a large number of sessions. Each of these 32 sessions consists of 40 training utterances and 10 test utterances, as described below. We call these sessions *small sessions*. Additionally, each speaker recorded a *large session* with 500 training utterances.

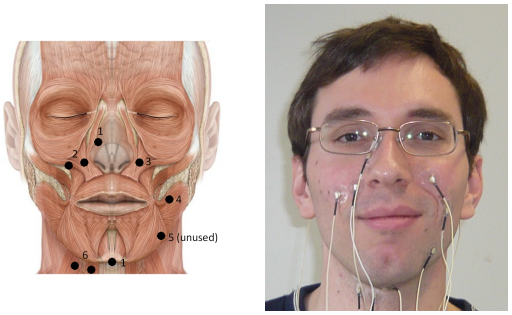

Figure 1: Overview of electrode positioning and captured facial muscles (muscle chart adapted from (Schünke et al., 2006)).

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 (Maier-Hein et al., 2005) which yielded optimal results, using five channels and capturing 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). In the audible and whispered parts, we parallely recorded the audio signal with a standard close-talking microphone connected to a USB soundcard.

The recording protocol for each "small" session of the EMG-UKA corpus was as follows: In a quiet room, the speaker read a series of 50 English sentences. These sentences were recorded either only once, in normal (audible) speaking style, or 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 batch of 10 *BASE* sentences which were identical for all speakers and all sessions, and one batch of 40 *SPEC* sentences, which varied across sessions. The sentence sets were identical for all parts of a session, 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.

The two additional "large" sessions followed the same protocol, with the only difference that the set of SPEC sentences was enlarged to 500 sentences. The two large sessions only contain audible EMG recordings. In all cases, the SPEC sentences (or a subset of them) were used as training respectively adaptation data, and the BASE sentences were used as test data.

Table 1: Statistics of the data corpus.

Speaker	**1**		**2**	
Total # of Sessions	24	1	8	1
Sessions with a Silent EMG part	11	0	2	0
Audible EMG Training Sentences per Session	40	500	40	500
Audible EMG Test Sentences per Session	10	10	10	10
Silent EMG Test Sentences per Session (where present)	10	-	10	-
Average Duration of Audible Training Data per Session	149s	1641s	146s	1625s
Average Duration of Audible Test Data per Session	42s	40s	40s	38s
Average Duration of Silent Test Data per Session (where present)	45s	-	45s	-

Note that we did not use the whispered EMG recordings for this study, and that from the silent EMG parts, we only used the test set. The final corpus which we used for this study is summarized in table 1.

4 THE EMG-BASED SPEECH RECOGNIZER

In this section we give a brief overview of our EMG-based speech recognizer.

The feature extraction is based on *time-domain features* (Jou et al., 2006). 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 $w[n]$ is defined as

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

The rectified high-frequency signal is $r[n] = |x[n] - w[n]|$. The final feature **TD15** is defined as follows (Schultz and Wand, 2010):

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

Frame size and frame shift are set to 27 ms respective 10 ms. In all cases, we apply LDA on the **TD15** feature to reduce it to 32 dimensions.

The recognizer is based on three-state left-to-right fully continuous Hidden-Markov-Models. All experiments used *bundled phonetic features (BDPFs)* for training and decoding, see (Schultz and Wand, 2010) for a detailed description.

For decoding, we used 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 was restricted to the words appearing in the test set, which resulted in a test vocabulary of 108 words, with the exception of the final experiment, where we extended the decoding vocabulary to 2102 words (the entire vocabulary of the full EMG-UKA corpus). We applied lattice rescoring to obtain the best weighting of language model and acoustic model parameters.

5 EXPERIMENTS

This section is structured as follows: First, we show how a session-independent system performs in comparison to large session-dependent systems. Second, we prove that these encouraging results can be improved even further by means of MLLR adaptation. Third, we present recognition results on silent EMG, showing that a session-independent system may cope with the difference between audible and silent EMG. Fourth, we outfit our best systems with an extended vocabulary and show that even then, EMG-based speech recognition is still possible.

5.1 Session-Independent EMG-based Speech Recognition

We compared the following three kinds of systems:

- A *session-dependent (SD)* system is trained and tested on data from one single session, during which the recording electrodes were not removed.
- A *multi-session (MS)* system uses training data from multiple sessions. The session on which such a system is tested is always part of the training sessions. Note that as described in section 3, the training data set and the test data set are disjoint.
- A *session-independent (SI)* system uses training data from one or more sessions. No data from the session on which the system is tested may form part of the training corpus.

The multi-session system may be considered an intermediate step towards a session-independent system. We expect that when the amount of training data is

the same, a multi-session system should perform better than a session-independent system, and this is indeed the case.

For our first experiment, we subdivided the 32 "small" sessions of the data corpus into blocks of 2, 4, or 8 sessions. Each session forms part of exactly one block of each size. On each block of n sessions, we trained n session-independent systems, each of which was characterized by testing on the test data of one particular session, and training on the training data of the remaining $n-1$ sessions. Thus e. g. from the 24 sessions of speaker 1, we obtain 24 SI systems which were trained on 1 session (different from the test session), 24 SI systems which were trained on 3 sessions, and 24 SI systems which were trained on 7 sessions. Consequently, we have got SI systems trained on 40, 120, and 280 training sentences.

In order to obtain comparable results, we trained multi-session systems in the same way, using the same amount of training sentences. This means that from a block of e. g. 8 sessions, we left out one session (either the first one or the last one) and trained a multi-session system on the remaining 7 sessions. This system is then tested on the test data of one of the sessions included in the training set. Finally, we obtain as many MS systems as SI systems.

For comparison, we used the two "large" sessions of our corpus to train and test session-dependent systems with 40, 120, or 280 training sentences.

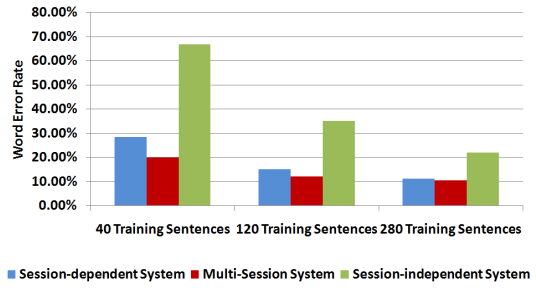

Figure 2: Average system performances (Word Error Rates) for session-dependent, session-independent, and multi-session systems.

Figure 2 shows the Word Error Rates (WERs) of the systems described above, averaged over all sessions of both speakers. It can be seen that a small session-independent system with 40 training sentences performs quite badly, with 66.93% WER versus 28.43% WER for the session-dependent system of the same size. However when the number of training sessions increases, the performance of the session-independent system also increases and approaches the performance of the session-dependent system. The average word error rate of our largest SI systems with 280 training sentences is 21.94%, compared to a WER of 11.28% for the speaker-dependent system of the same size. The multi-session systems unsurprisingly perform best, with a WER of 10.45% for 280 training sentences.

A session-independent (SI) system is a system where the sets of training sessions and test sessions are disjoint. However, by definition such an SI system can be trained on data from many sessions, or just on data from one large session. From similar observations in acoustic speech recognition, it can be hypothesized that when multiple sessions are used, the system "sees" more different data "shapes" and therefore becomes more robust than a system trained on only one session.

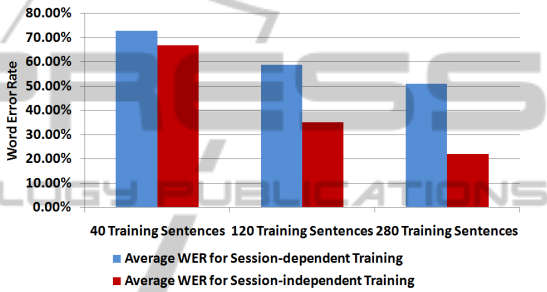

Figure 3: Average performance of single-session and multi-session systems when tested on data from unseen sessions.

Figure 3 shows that this is indeed the case: We trained session-dependent recognizers on the two "large" sessions, using 40, 120, or 280 training sentences, and tested these recognizers on the test sets of the "small" sessions of the respective speakers. We compared these results to the performance of the SI systems trained on data from several "small" sessions. On 40 training sentences, the recognizers have quite similar performance, with a word error rate of 66.93% respectively 72.91%. On 280 training sentences, however, the average WER of the single-session SI recognizer is 51.04%, whereas the average WER of the multi-session SI recognizer drops to 21.93%! Even though there is some variation between different sessions, the result shows that a recognizer trained on multiple sessions is much better than a similarly-sized recognizer trained on one session. Moreover, the performance difference increases with the number of training sessions, supporting our claim that increasing the number of training sessions indeed increases the robustness of the recognizer.

5.2 Session-adaptive Recognition

In this section we investigate whether the session-independent systems may be improved by using lim-

ited amounts of training data from the sessions on which the respective systems are to be tested. The classical method is to *adapt* the trained models of the SI system towards the training data from the test session. Due to its ability to deal with varying amounts of adaptation data, we used *Maximum Likelihood Linear Regression* (Leggetter and Woodland, 1995), a standard method in acoustic speech recognition.

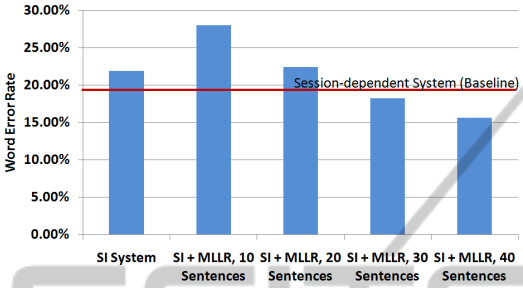

Figure 4: Average performance of session-adaptive systems with various numbers of adaptation sentences. The performance of a speaker-dependent system with 40 training sentences is given as a baseline.

Figure 4 shows the results of our experiments in session adaptation. We started with our largest SI system, trained on 280 utterances, and used 10, 20, 30, or 40 training utterances to adapt the system towards the data of the respective test session. The testing was performed on the full test set of the respective session.

For 10 adaptation sentences, the performance of the system actually degrades, probably due to undertraining artifacts. However for 30 or 40 sentences, MLLR adaptation has a beneficial effect, yielding better systems than the original SD system trained on 40 sentences of training data. The best session-adaptive system gives a WER of 15.66%, which is significantly better than the 19.96% WER of the SD system.

5.3 Robust Recognition of Silent Speech

The next experiment answers the question whether the increased robustness of a session-independent recognizer may also help in recognizing speech with different speaking modes. With *speaking mode*, we refer to *audible* and *silent* speech, as described in section 3. The variation between these two speaking modes has been shown to have a very large impact in EMG-based speech recognition (Janke et al., 2010a; Janke et al., 2010b).

For this experiment, we used the multi-session and session-independent systems with 40, 120, and 280 training sentences which are described above. We took the acoustic models from these systems, which had been trained on audible EMG, and applied them to the available silent EMG test sets. This implies that these experiments were limited to the 13 sessions for which silent EMG data is available (see section 3). According to (Janke et al., 2010a), we call this setup *Cross-Modal Testing*.

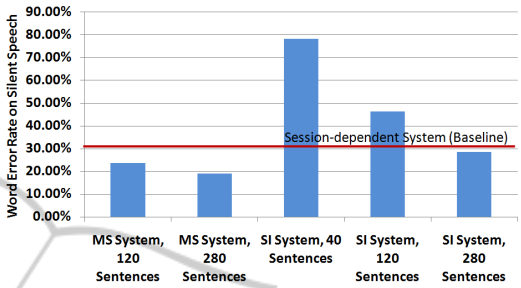

Figure 5: Average performance of different systems on *silent* EMG. The performance of a speaker-dependent system with 40 training sentences is given as a baseline.

The recognition results are presented in figure 5 and show a very encouraging picture: While our baseline cross-modal recognition performance on SD systems is at 36.21% WER, the multi-session systems perform significantly better, with the best system with 280 training sentences yielding 19.04% WER. Furthermore, even the session-independent system with 280 training sentences performs better than the SD system, yielding a WER of 28.45% on silent EMG.

5.4 Performance on Large Vocabularies

As a final experiment, we extended our recognition vocabulary to 2102 words, which is the whole set of words which occurs in the complete EMG-UKA corpus. We used the best speaker-dependent and speaker-independent systems with 280 training sentences and tested these models on the same test set as before, but with a testing vocabulary of 2102 words.

Figure 6 shows the results of this experiment. Clearly, there is a decay in system performance when the vocabulary is increased. This decay appears to be similar for both the SD and the SI systems. The session-dependent system still performs best, with a WER of 33% on the large vocabulary. The session-independent system yields a WER of 50.48%.

6 CONCLUSIONS

In this paper we presented an EMG-based speech recognition system which works *session-independently*: It uses training data from multiple

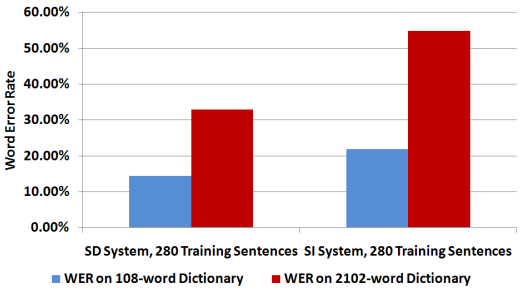

Figure 6: Average performance of systems on extended testing vocabulary.

EMG recording sessions, between which the EMG electrodes have been removed and reattached. We demonstrated that session-independent EMG-based speech recognition yields a suitable performance, and that in particular, when testing is performed on *unseen* sessions, the session-independent system performs significantly better than a similarly large session-dependent system, which shows that the session-independent training approach indeed increases the robustness of the system. We also showed that adapting a session-independent system towards a specific test session further improves the system performance.

This technology allows us to create larger EMG-based speech recognition systems than the ones previously investigated. We have shown that our current best system can deal with vocabulary sizes ranging up to 2.100 words, which brings EMG-based speech recognition within a performance range which makes spontaneous conversation possible.

Further steps in the field of EMG-based speech processing may include a systematic study of the discrepancies between different recording sessions, which could not only improve the systems presented in this paper, but also give further insight in what causes these discrepancies. Second, transiting to true speaker-independent systems is another major goal for the future. In order to achieve it, however, further studies on the behavior of the EMG signals of the articulatory muscles are needed.

REFERENCES

Chan, A., Englehart, K., Hudgins, B., and Lovely, D. (2001). Myoelectric Signals to Augment Speech Recognition. *Medical and Biological Engineering and Computing*, 39:500 – 506.

Denby, B., Schultz, T., Honda, K., Hueber, T., and Gilbert, J. (2010). Silent Speech Interfaces. *Speech Communication*, 52.

Janke, M., Wand, M., and Schultz, T. (2010a). A Spectral Mapping Method for EMG-based Recognition of Silent Speech. In *Proc. B-INTERFACE*.

Janke, M., Wand, M., and Schultz, T. (2010b). Impact of Lack of Acoustic Feedback in EMG-based Silent Speech Recognition. In *Proc. Interspeech*.

Jorgensen, C., Lee, D., and Agabon, S. (2003). Sub Auditory Speech Recognition Based on EMG/EPG Signals. In *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, pages 3128 – 3133, Portland, Oregon.

Jou, S.-C., Schultz, T., Walliczek, M., Kraft, F., and Waibel, A. (2006). Towards Continuous Speech Recognition using Surface Electromyography. In *Proc. Interspeech*, pages 573 – 576, Pittsburgh, PA.

Jou, S.-C. S., Schultz, T., and Waibel, A. (2007). Continuous Electromyographic Speech Recognition with a Multi-Stream Decoding Architecture. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 401 – 404, Honolulu, Hawaii.

Leggetter, C. J. and Woodland, P. C. (1995). Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer Speech and Language*, 9:171–185.

Maier-Hein, L., Metze, F., Schultz, T., and Waibel, A. (2005). Session Independent Non-Audible Speech Recognition Using Surface Electromyography. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 331 – 336, San Juan, Puerto Rico.

Schultz, T. and Wand, M. (2010). Modeling Coarticulation in Large Vocabulary EMG-based Speech Recognition. *Speech Communication*, 52:341 – 353.

Schünke, M., Schulte, E., and Schumacher, U. (2006). *Prometheus - Lernatlas der Anatomie*, volume [3]: Kopf und Neuroanatomie. Thieme Verlag, Stuttgart, New York.

Wand, M. and Schultz, T. (2009). Towards Speaker-Adaptive Speech Recognition Based on Surface Electromyography. In *Proc. Biosignals*, pages 155 – 162, Porto, Portugal.