SESSION INDEPENDENT NON-AUDIBLE SPEECH RECOGNITION USING SURFACE ELECTROMYOGRAPHY

Lena Maier-Hein, Florian Metze, Tanja Schultz, and Alex Waibel

Interactive Systems Labs
Universität Karlsruhe (TH), Carnegie Mellon University
{lena|tanja}@ira.uka.de

ABSTRACT

In this paper we introduce a speech recognition system based on myoelectric signals. The system handles audible and non-audible speech. Major challenges in surface electromyography based speech recognition ensue from repositioning electrodes between recording sessions, environmental temperature changes, and skin tissue properties of the speaker. In order to reduce the impact of these factors, we investigate a variety of signal normalization and model adaptation methods. An average word accuracy of 97.3% is achieved using seven EMG channels and the same electrode positions. The performance drops to 76.2% after repositioning the electrodes if no normalization or adaptation is performed. By applying our adaptation methods we manage to restore the recognition rates to 87.1%. Furthermore, we compare audibly to non-audibly spoken speech. The results suggest that large differences exist between the corresponding muscle movements. Still, our recognition system recognizes both speech manners accurately when trained on pooled data.

1. INTRODUCTION

Automatic Speech Recognition (ASR) has developed into a popular technology and is being deployed in a wide variety of every day life applications, including personal dictation systems, call centers or mobile phones. Despite the various benefits a conventional speech-driven interface provides to humans, there are three major drawbacks: Firstly, the audible (i.e. acoustic) speech signal prohibits a confidential conversation with or through a device. Besides that, talking can be extremely disturbing to others, especially in libraries or during meetings. Secondly, the speech recognition performance degrades drastically in adverse environmental conditions such as in restaurants, cars, or trains. Acoustic model adaptation can compensate for these effects to some degree, however the pervasive nature of mobile phones challenges this approach. Performance is also poor when sound production limitations occur, like under water. Last but not least, conventional speech-driven interfaces cannot be used by speech handicapped people, for example those without vocal cords.

To overcome these limitations, alternative methods are being investigated, which do not rely on an acoustic signal for ASR. Chan et al. [1] proved that the myoelectric signal (MES) from articulatory face muscles contains sufficient information to discriminate a given set of words accurately. This holds even when the words are spoken non-audibly, i.e. when no acoustic signal is produced [2].

To date, the practicability of MES based speech recognition is still limited. Firstly, the surface electrodes require a physical contact with the speaker's skin. Secondly, experiments are still restricted to isolated word recognition. Finally, today's systems are far from being robust, since they only work in matching training and test conditions. Just like conventional speech recognizers, the MES based systems are heavily influenced by speaker dependencies, such as speaking style, speaking rate, and pronunciation ideosyncrasies. Beyond that, the myoelectric signal is affected by even slight changes in electrode positions, temperature or tissue properties [3]. We will refer to this phenomenon as *session dependence* in analogy to the *channel dependence* of a conventional speech recognizer resulting from the microphone quality, the environmental noise, and the signal transmission of the acoustic signal.

According to our experience the loss in performance caused by session dependence in MES based speech recognition is significantly higher than that resulting from channel conditions in conventional systems. Despite this, only session dependent MES based speech recognition systems have been developed so far. In this paper we will address the session dependence by exploring methods for adjusting data from a new recording session to given training material from previous recording sessions.

The most important advantage of using the MES for speech recognition is the fact that it does not rely on the speaker to pronounce the words audibly. Coleman et al have established that the speech motor control plans for whispered speech and vocalized speech are similar [4]. Yet, no study has investigated the differences between audible and non-audible speech relevant for MES based speech recognition. This issue is therefore the second focus of our work.

2. EMG BASED SPEECH PROCESSING

2.1. Surface EMG Measurement

Electromyography (EMG) is the process of recording the electrical activity of a muscle. When a muscle fiber is activated by the central nervous system, small electrical currents in form of ion flows are generated. Since electrical current moves through a resistance, the bodily tissue, it creates an electrical field. The resulting potential differences can be measured between certain regions on the body surface. A surface Electromyogram is the record obtained from

This work has been partly funded by the European Union under IST project No. FP6-506909 "CHIL – Computers in the Human Interaction Loop", http://chil.server.de. The authors wish to thank Christoph Mayer, Marcus Warga, Peter Osztotics and Artus Krohn-Grimberghe for their valuable contributions to this study.

measuring these voltages over time. The following equipment is needed to measure surface EMG (sEMG) [5]:

Surface electrodes convert the ionic currents generated by muscle contraction into electronic currents that can be fed into electronic devices. While two *detection electrodes* pick up the desired signal the *ground electrode* provides a common reference.

When detecting an EMG signal, amplification is necessary to optimize the resolution of the digitizing equipment. A *differential amplifier* subtracts the signals from two detection sites and amplifies the difference voltage between its two input terminals. As a consequence, signals common to both electrodes - such as noise originating far away from the detection sites - ideally produce a zero output, whereas local EMG signals are amplified. This way the signal-to-noise ratio is maximized.

A *high-pass filter* is applied to avoid aliasing artefacts whereas a *low-pass filter* is used to reduce movement artefacts in the signals.

2.2. Related Work

The body of published studies testing the potential of EMG for speech recognition is surprisingly small. Using an approach similar to that proposed here, Chan *et al.* [1, 6] proposed to perform ASR on the myoelectric signal for aircraft pilot communication. Five bipolar electrodes were embedded in pilots' oxygen masks and the myoelectric signals were recorded during audible pronunciation of the digits "zero" to "nine". An acoustic signal was also recorded and used to segment the utterances. The authors reported a maximum word accuracy of 93% for a linear discriminant analysis (LDA) classifier and of 86% for a hidden Markov model (HMM) classifier [6]. Moreover, they showed the potential of the MES to augment conventional speech recognition systems [1].

Jorgensen *et al.* [2, 7] investigated the recognition of non-audible speech. Their idea is to intercept nervous signal control signals sent to speech muscles using surface EMG electrodes placed on the larynx and sublingual areas below the jaw. Initially, they demonstrated the potential of non-audible speaker dependent isolated word recognition based on the MES with a Neural Network classifier [2]. They reported recognition rates of 92% for six control words [2] and of 73% on an extended vocabulary which additionally contains the ten English digits [7]. Recently, Jorgensen *et al.* expanded their earlier isolated word experiments to the recognition of vowels and consonants as a first step towards phoneme based speech recognition. Moreover, they developed a web browser interface that is controlled by myoelectric signals [7].

Manabe *et al.* [8] proposed the use of ring-shaped electrodes wrapped around the thumb and two fingers for non-audible speech recognition. In order for the electrodes to detect sEMG signals from facial muscles the fingers need to be pressed against the face in a specified manner. The authors investigated conventional ASR techniques for the recognition of the ten Japanese digits, achieving a maximum recognition rate of 64% [9]. They hope to perfect the system such that it develops to a mobile interface that can be used in both, silent and noisy environments.

3. METHODS

3.1. Data Acquisition

In this study, isolated word recognition was performed on a vocabulary consisting of the ten English digits "zero" to "nine". Three subjects, S1 (female, mother-tongue: German), S2 (male, mother-tongue: Hungarian), and S3 (male, mother-tongue: German), with no known speech disorders participated in the study. Each subject took part in five recording sessions on four different days, in morning and afternoon sessions. In four of their sessions the subjects pronounced the words non-audibly, i.e. without producing a sound. In the remaining sessions ordinary (i.e. audible) speech was recorded. Each audible session corresponds to one non-audible session in that the two were recorded in series without the electrodes being moved.

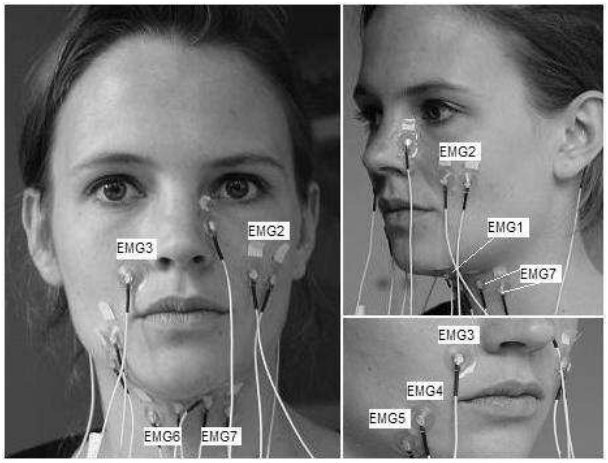

Fig. 1. Positioning for electrodes EMG1-EMG7. Section 3.1 enumerates the associated muscles.

In each recording session forty exemplars of each vocabulary word and forty exemplars of silence were recorded. The order of the words was randomly permuted and presented to the subject one at a time. A push-to-talk button controlled by the subject was used to mark the beginning and the end of each utterance. Subjects were asked to begin speaking approximately 1sec after pressing the button and to release the button about 1sec after finishing the utterance. When the pseudo-word silence appeared they were supposed keep all facial muscles relaxed for approximately 2sec.

EMG signal data was collected for each of the subjects using seven pairs of Ag/Ag-Cl electrodes. A self-adhesive button electrode placed on the left wrist served as a common reference. As shown in Figure 1 the electrodes were positioned such that they obtain the EMG signal of six articular muscles: the *levator anguli oris* (EMG2,3), the *zygomaticus major* (EMG2,3), the *platysma* (EMG4,5) the *depressor anguli oris* (EMG5), the *anterior belly* of the *digastric* (EMG1) and the *tongue* (EMG1,6,7) [10, 6]. For three of the seven EMG channels (EMG2,6,7) a classical bipolar electrode configuration with a 2cm center-to-center inter-electrode spacing was used. For the remaining four channels one of the detection electrodes was placed directly on the articulatory muscles and was referenced to either the nose (EMG1) or to both ears (EMG3,4,5) (Figure 1). The positioning of the electrodes was optimized in previous experiments, not reported here.

In order to ensure repeatability of electrode placements we have produced a gypsum mask for every speaker. Holes in the masks marked the electrode positions to be used. We found in previous experiments that using the mask for position identification gives slightly more reliable across-sessions results than using tape

measure.

For the purpose of impedance reduction at the electrode-skin junction a small amount of electrode gel was applied to each electrode. All electrode pairs were connected to a physiological data recording system [11]. EMG responses were differentially amplified, filtered by a 300Hz low-pass and a 1Hz high-pass filter and sampled at 600Hz. In order to avoid loss of relevant information contained in the signals we did not apply a 50Hz notch filter which can be used for the removal of line interference.

3.2. Feature extraction

The signal data for each utterance is transformed into feature vectors. For each channel, 18-dimensional channel feature vectors are extracted from 54ms observation windows with 4ms overlap. In terms of the number of coefficients per window this corresponds to a 32ms window at a sampling rate of 1000Hz which was used in previous experiments.

In order to obtain channel feature vector o_{ij} for channel j and observation window i the windowed Short Time Fourier Transform (STFT) is computed. Delta coefficients serve as the first 17 coefficients of o_{ij}. The 18th coefficient consists of the mean of the time domain values in the given observation window. The complete feature vector o_i for the observation window i is simply the concatenation of the channel feature vectors o_{ij}. The choice of these features is the result of intensive experiments in previous recording sessions. Ordinary STFT coefficients, cepstral coefficients, LPC coefficients, and the root-mean-squared value among others were also considered as features but did not add to the overall performance.

3.3. Feature Training

First order HMMs with Gaussian mixture models are used in most conventional ASR systems as classifiers because they are able to cope with both, variance in the time-scale and variance in the shape of the observed data. We trained a five-state left-to-right Hidden Markov Model λ_j with 12 Gaussians per state for every word W_j in the vocabulary using the Expectation Maximization (EM) algorithm. The number of iterations was chosen to be $N = 4$.

To recognize an unknown signal the corresponding sequence of feature vectors (o_k) was computed. Next, the Viterbi alignment for each vocabulary word W_j was determined and the word corresponding to the best Viterbi score was output as the hypothesis. Feature extraction, HMM training, and signal recognition were performed using the *Janus Recognition Toolkit* (JRTk) [12].

4. EXPERIMENTS AND RESULTS

To ensure comparability of results from different experiments the same number of samples was used for each classifier training, namely thirty exemplars of each word. Whenever training and testing were performed on the same session(s), a round robin procedure was applied to get reliable results. When the testing session was different from the training session(s), the training data was split into a disjoint set of training sets each satisfying the condition from above (i.e. each containing thirty exemplars of each vocabulary word) and the results for the training sets were averaged.

4.1. Baseline system

The system described in section 3 serves as our baseline system. Table 1 shows the word accuracies for within-session testing for each speaker using different numbers of channels for recognition. The term within-session refers to a matching training/test condition, i.e. training and testing are performed on the same session. The results for each speaker are averaged over the corresponding four non-audible sessions. Due to the fact that we applied the round robin algorithm for within-session testing (number of sessions sN, number of round robin sets rN) and used thirty samples per word for training the classifier (number of words per set uN), we had a total of $sN \cdot rN \cdot uN = 4 \cdot 4 \cdot 100 = 1600$ test samples per speaker. The table presents the results for (a) each individual channel, (b) the combination of all channels, and (c) the best combination of $k = 2, 3, 4, 5, 6$ channels. We used a greedy procedure to identify the best combination of k channels: Initially, we simply chose the channel yielding the best individual within-session results. We then added the remaining channels one by one, in the order that gave the best (within-session) performance when combined with the already selected channels.

Channels	S1	S2	S3	Avg
Individual Channels				
EMG1	74.2	92.1	**77.4**	**81.2**
EMG2	64.1	90.7	69.4	74.7
EMG3	**76.1**	**93.8**	72.9	**81.0**
EMG4	61.2	83.1	71.6	71.9
EMG5	62.4	73.4	63.6	66.5
EMG6	63.6	64.4	52.3	60.1
EMG7	59.8	66.3	60.0	62.0
Avg EMG1-EMG7	65.9	80.5	66.7	71.1
Channel Combination				
Best 1 (EMG1)	74.2	92.1	77.4	81.2
Best 2 (EMG1,3)	93.5	97.6	90.1	93.7
Best 3 (EMG1,3,6)	97.1	98.1	91.3	95.5
Best 4 (EMG1,3,4,6)	**97.5**	98.3	93.4	96.4
Best 5 (EMG1,2,3,4,6)	97.3	98.6	95.5	97.1
Best 6 (EMG1,2,3,4,5,6)	97.4	**98.8**	**96.2**	**97.4**
All 7 channels	97.2	98.8	96.0	97.3

Table 1. Within-session word accuracies (in %) averaged over four sessions for each speaker.

Speaker S2 achieved the best recognition results. This speaker had already recorded several non-audible sessions before participating in this study. He stated that he had developed a particular speaking style for non-audible speech over time. In fact, we noticed for all speakers that an increasing level of experience improved the performance. The results in Table 1 indicate a significant variation in performance for the individual channels. Channels EMG1 and EMG3 yield the best recognition results for all speakers. These two channels correspond to different muscle groups, and therefore provide orthogonal information. The results from the best channel combination in table 1 reveal that it is crucial to apply more than one electrode (highly significant difference between Best 1 and Best 2). Even between 2 and 3 electrodes we see a highly significant performance increment on the 9.56E-05 $\cdot$ 100 % level, while the performance differences for 5, 6 or 7 electrodes are insignificant.

Table 2 shows the within-session and naive across-sessions results for speaker S3. Naive across-sessions testing refers to testing without any normalizations and adaptations. The large performance differences between within-session results (values on the diagonal in bold face) and across-sessions results (values in the remaining cells) illustrate the problem of session dependence.

	session I	session II	session III	session IV
session I	**94.5**	74.3	83.0	58.8
session II	67.5	**93.5**	80.5	73.8
session III	48.8	59.5	**97.5**	77.8
session IV	60.5	67.0	91.8	**98.5**

Table 2. Word accuracies (in %) for within-session testing and naive (no normalization) across-sessions testing for speaker S3 using all seven channels. Training session (row), Test session (column).

The results for naive across-sessions testing for all speakers are summarized in Tables 3 and 4 for all channels and for individual channels respectively (*method=BASE*). The numbers represent the average word accuracy when one session is used for training and another session is used for testing. Thus, in Table 3 each cell corresponding to method *BASE* represents the results for $sN \cdot (sN-1) = 4 \cdot 3 = 12$ experiments. In Table 4 on the other hand, the entries represent the results for $cN \cdot sN \cdot (sN-1) = 7 \cdot 4 \cdot 3 = 84$ experiments, where cN represents the number of channels.

Again, the results for across-sessions testing are significantly worse than those for within-session testing. We address this crucial problem of session dependence in the next section and will show that we achieve significant improvement across sessions by normalizing data and adapting our models.

4.2. Session Independence

As already mentioned above the signal obtained from surface EMG measurements depends on a number of different factors which cannot be held constant over several recording sessions. Exact electrode positioning plays a particularly crucial role [3]. Although gypsum masks were used to improve placement repeatability, the poor across-sessions results indicate existing variation in the positioning. In fact, experiments showed an across-sessions deviation of up to 5mm. Furthermore, other factors like the amount of applied electrode gel may vary from session to session. Moreover, the speakers' speech patterns produced on different days may differ from each other. Subject S3, for example, stated that he had the impression that he pronounced the non-audibly spoken words differently in different recording sessions.

We investigated the following normalization and adaptation procedures to compensate for the described session dependent variations:

1. *Session Combination (SC)*: The data to train the classifiers is shared across three sessions, each contributing the same number of samples (ten samples per vocabulary word).
2. *Session Selection (SS)*: A conventional HMM classifier C_i is trained for every training session i. The incoming unknown signal is then decoded by each classifier C_i, giving a hypothesis W_i and a corresponding Viterbi score v_i. The word with the overall best viterbi score is output as the hypothesis. $W_{hyp} = W_l; l = \arg\max_n v_n$.
3. *VN in combination with SC (SC&VN)*: For each training session two normalization vectors are computed; one containing the mean of each feature vector coefficient for the session's training samples and one containing the variance of each feature vector coefficient. Similarly, two normalization vectors are computed for all test session data. Prior to Viterbi path computation during training or testing, the obtained vectors are applied to normalize the extracted feature vectors o_i.
4. *VN with enrollment data and SC (SC&VN_enr)*: Similar to *SC&VN* but the normalization vectors for the test session are computed on enrollment data rather than on the test data itself. The enrollment data set consisted of two examples for each vocabulary word including silence.
5. *Supervised Feature Space Adaptation and SC (SC&FSA_sup)*: Feature Space Adaptation (FSA) is a constrained Maximum Likelihood (ML) transformation of input features. In analogy to Speaker Adaptive Training (SAT) [13] we perform *session adaptive training*. First, an initial classifier is computed on three training sessions. Then, we iteratively (a) adapt each training session to the current classifier (beginning with the initial classifier) and (b) recompute the classifier models using the adapted training data. After four iterations, the final classifier is used for a supervised computation of an adaptation matrix for the test data. During testing, only adapted test data is used.
6. *Unsupervised FSA and SC (SC&FSA_unsup)*: Like *SC& FSA_sup* but unsupervised adaptation is performed on the test data using hypothesis from the computed classifier.
7. *FSA with enrollment data and SC (SC&FSA_enr)*: Like *SC&FSA_sup* but the adaptation matrix is computed on an enrollment data set consisting of two samples per vocabulary word including silence.
8. *FSA with enrollment data, iterative learning, and SC (SC& FSA_enr_it)*: Like *SC&FSA_enr* but the adaptation matrix for the test data is recomputed after each hypothesis computation for a test signal.
9. *Combinations of the above methods*: When both, *VN* and *FSA* are applied, the features are first normalized and then adapted to the model.

Method	S1	S2	S3	Avg
BASE	74.5	83.7	70.3	76.2
SC	84.6	90.1	77.6	84.1
SS	85.2	88.3	77.3	83.7
SC&VN	83.4	94.3	83.7	87.1
SC&VN_enr	84.3	90.3	79.6	84.7

Table 3. Word accuracies (in %) for across-sessions testing using all channels for recognition. Four non-audible sessions are used for each speaker and the across-sessions results are averaged.

The data set for the experiments on session independence consists of the four non-audible sessions from each speaker. We examined both, across-sessions recognition using all seven channels

(Table 3) and across-sessions recognition using only one channel (Table 4). In the latter case, the word accuracies for the individual channels were averaged. Due to the fact that FSA computations led to numerical instabilities when high-dimensional data was used (seven channels correspond to 126 dimensions), we did not apply feature space adaptation based methods when using all seven channels for recognition. Initial experiments using an LDA for dimensionality reduction decreased word accuracies.

As shown in Tables 3 and 4, normalization and adaptation improve performance for all speakers. In fact, the χ^2-test confirms that the results for *BASE* and *SC* are different at a significance level of 2.93E-20% (table 3). The additional application of *VN* leads to another increment on a significance level of 2.84E-03%.

Method	S1	S2	S3	Avg
BASE	37.0	53.5	41.3	43.9
SC	40.3	59.3	44.2	47.9
SS	43.4	61.4	48.6	51.1
SC&FSA_sup	42.5	62.7	47.7	51.0
SC&FSA_unsup	42.0	62.3	47.0	50.5
SC&FSA_enr	42.3	62.5	47.1	50.6
SC&FSA_enr_it	42.1	62.5	47.2	50.6
SC&VN	40.2	61.6	47.1	49.6
SC&VN_enr	38.8	60.5	45.5	48.3
SC&VN&FSA_sup	42.6	65.0	49.9	52.5
SC&VN&FSA_unsup	42.0	64.6	49.5	52.0
SC&VN_enr&FSA_enr	41.2	63.7	48.2	51.0
SC&VN_enr&FSA_enr_it	41.3	64.1	48.5	51.3

Table 4. Word accuracies (in %) for across-sessions testing using one channel for recognition and four sessions from each speaker. Each cell represents the average over all seven channels.

As in ASR, combining data from several sessions improves performance considerably (Session Combination *SC*). Session Selection (*SS*) leads to significant improvements in performance as well. However, this method requires three times as much training material and the training of three times as many parameters. Consequently, *SS* is not directly comparable to the other methods. In fact, we obtained an improvement of 1.9% (1.5% absolute) for all channels and and 4.6% (2.2% absolute) for individual channels when we used the same amount of training material for combination (*SC*) as for selection *SS* (thirty samples per word from each session). We therefore did not combine *SS* with *VN* and *FSA*. Experiments suggest, however, that a similar increase in word accuracy as with *SC* can be achieved.

Both tables show a significant improvement in word accuracy when Variance Normalization (*VN*) is applied. However, the method fails to increase word accuracies for speaker S1. We attribute this to large deviations in recording lengths for speaker S1 which leads to significant deviations in the amount of silence relative to the amount of speech in different recording sessions. This in turn leads to an unreliable estimation of the *VN* normalization vector.

Feature Space Adaptation based methods increase the performance for all speakers. Interestingly, supervised adaptation performs equally well as unsupervised adaptation. Combining *FSA* and *VN* leads to further improvements, yet the improvements are not additive, i.e. both methods address similar artifacts. In order to apply FSA based methods when several channels are used for recognition, we will explore feature dimensionality reduction techniques for EMG speech data in the future.

Both, *FSA_unsup* and *VN* require the whole set of test data for initial computations. Obviously, this is impractical. We therefore examined the use of enrollment data for the computation of normalization vectors and adaptation matrices. According to Table 4 only a small decrease in word accuracy results when enrollment data is used. However, *VN_enr* performs significantly worse than *VN* when all channels are used for recognition. Unfortunately, this cannot be explained satisfyingly by the current experiments, we therefore plan to investigate this more in the future.

In conclusion, we were able to improve word accuracies for across-sessions testing by 18.5% (8.1% absolute) for individual channels and by 14.3% (10.9% absolute) for all seven channels by sharing training data across sessions and by applying methods based on Variance Normalization and Feature Space Adaptation. This indicates, that conventional speech recognition methods can be transferred to EMG based recognition systems and achieve comparable word error rate reductions.

4.3. Audible vs Non-Audible Speech

To investigate the influence of speech manner (audible vs non-audible) on the performance of EMG based speech recognition, we recorded one audible and one non-audible session for each speaker. These two "sessions" were in fact recorded as one session with the exact same electrode placement, i.e. the electrodes were not removed between the two parts. The only difference was the speech manner. We now investigate the following aspects: (1) do the EMG signals produced by audible speech differ from those produced by non-audible speech and (2) is the recognition performance of audible speech different from that of non-audible speech. To investigate the first aspect we determined the recognition results across speech manners, i.e. models trained on audible speech were applied to non-audible speech and vice versa. To examine the second issue we compared the recognition results between the two speech manners in a matching condition, i.e. the models were trained and tested on the same speech manner. In a third experiment, we shared the training data across speech manners from each speaker to determine the performance of a recognizer that works on both, non-audible and audible speech. In the latter case we trained two systems; one with the same number of parameters as our baseline system and one with twice as many parameters.

The results of our experiments are shown in Table 5 for all channels and in Table 6 for individual channels respectively. It is noticeable that speakers S1 and S3 have much better recognition rates for audible speech than for non-audible speech. By contrast, there is no significant difference in performance for speaker S2. We believe that this relies to the fact that speaker S2 had the most experience in speaking non-audibly. As alluded to above, we noticed an improvement in performance with increasing experience for all speakers. We deduce from this, that MES based recognition of non-audible speech can work just as well as MES based recognition of audible speech (on our vocabulary) provided that the speaker is accustomed to the speaking manner.

The relatively low results in the mismatched condition suggest that muscle movements corresponding to audible speech differ from muscle movements corresponding to non-audible speech. However, the results for the mixed systems indicate that a recognizer can be trained for both, audible and non-audible speech, with reasonable results. The comparison of the 12-Gaussian vs the 24-

Gaussian systems suggests to increase the numbers of parameters for the mixed system.

Speech manner	S1	S2	S3	Avg
non-audible	97.0	99.8	93.5	96.8
audible	99.5	98.8	96.0	98.1
audible on non-audible	72.8	84.5	64.3	73.8
non-audible on audible	67.2	92.5	69.3	76.3
mixed; 12 Gaussians	96.1	98.1	91.8	95.3
mixed; 24 Gaussians	96.1	98.4	93.5	96.0

Table 5. Word Accuracies (in %) of non-audible and audible speech using all seven channels.

Speech Manner	S1	S2	S3	Avg
non-audible	63.0	83.4	60.0	68.8
audible	73.9	84.7	70.3	77.5
audible on non-audible	43.3	59.4	39.2	47.3
non-audible on audible	39.0	60.9	32.7	44.2
mixed; 12 Gaussians	62.6	79.3	57.3	66.4
mixed; 24 Gaussians	64.7	81.1	59.7	68.5

Table 6. Word Accuracies (in %) for non-audible and audible speech using one channel for recognition. Each entry represents the average over all seven channels.

5. CONCLUSIONS AND FUTURE WORK

In this paper we introduced a speech recognition system based on myoelectric signals. To cope with one of the main challenges of surface electromyography based speech recognition, namely session dependence, we investigated a variety of signal normalization and model adaptation methods. Our results suggest that methods used in conventional speech recognition systems for channel and speaker adaptation can be used for session adaptation in EMG based speech recognizers. Sharing training data across sessions and applying methods based on Variance Normalization and Maximum Likelihood adaptation improve across-sessions performance. We achieved an average word accuracy of 97.3% for within-session testing using seven EMG channels. Across-sessions testing without any adaptation yielded an average of 76.2%. By applying our normalization and adaptation methods we were able to bring recognition rates back up to 87.1%. Comparative experiments indicate that applying more than two electrodes is crucial, while using more than five electrodes does not lead to significant performance improvements.

Furthermore, our experiments indicate significant differences between the muscle movement corresponding to non-audible and the muscle movement corresponding to audible speech. While our recognizer performs slightly better on audible speech than on non-audible data, it is possible to merge training data and improve the robustness of the resulting recognizer. We also see large performance differences across speakers, however, as EMG-based speech recognition targets applications based on personal devices, speaker independence is not crucial.

To demonstrate the potential of this technology we are currently implementing a prototype "silent" mobile phone. An EMG speech recognizer is trained on a set of sentences typically used for answering a phone call during a meeting, for instance "I'm in a meeting", "is it urgent?" and "I'll call back later". This "silent" mobile phone application enables the user to conduct confidential phone calls without disturbing others nearby. The presented results are very promising but several limitations still need to be overcome. Among the biggest challenges are the usage of robust non-contact sensors to avoid clinging electrodes to the user's face. Another challenge is to move beyond discrete speech recognition and approach continuously spoken large vocabulary tasks.

6. REFERENCES

[1] A.D.C. Chan, K.Englehart, B. Hudgins, and D.F. Lovely, "Myoelectric Signals to Augment Speech Recognition," *Medical and Biological Engineering and Computing*, vol. 39, pp. 500–506, 2001.

[2] C. Jorgensen, D. Lee, and S. Agabon, "Sub Auditory Speech Recognition Based on EMG/EPG Signals," in *Proc. of the International Joint Conference on Neural Networks*, 2003.

[3] B. Leveau and G.B.J. Andersson, "Output Forms: Data Analysis and Applications," in *Selected Topics in Surface Electromyography for Use in the Occupational Setting: Expert Perspective*. U.S. Department of Health and Human Services, 3 1992, DHHS(NIOSH) Publication No 91-100.

[4] J. Coleman, E. Grabe, and B. Braun, "Larynx movements and intonation in whispered speech," 2002, Summary of research supported by British Academy grant SG-36269.

[5] C. De Luca, "Surface Electromyography: Detection and Recording," Tech. Rep., DelSys Inc., 2002.

[6] A.D.C. Chan, K.Englehart, B. Hudgins, and D.F. Lovely, "Hidden Markov Model Classification of Myolectric Signals in Speech," *Engineering in Medicine and Biology Magazine, IEEE*, vol. 21, pp. 143–146, 9 2002.

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

[8] H. Manabe, A. Hiraiwa, and T. Sugimura, "Unvoiced Speech Recognition using EMG - Mime Speech Recognition -," in *Proc. of the 2003 Conference on Human Factors in Computing Systems, Ft. Lauderdale, Florida, USA*, 2003.

[9] H. Manabe and Z.Zhang, "Multi-stream HMM for EMG-Based Speech Recognition," in *Proc. of the 26th IEEE EMBS Conference, San Francisco, CA, USA*, 2004.

[10] UCLA Phonetics Laboratory, "Dissection of the Speech Production Mechanism," Tech. Rep., Department of Linguistics, University of California, Los Angeles, 2002.

[11] K.Becker, "Varioport ™," http://www.becker-meditec.de.

[12] M. Finke, P. Geutner, H. Hild, T. K emp, K. Ries, and M. Westphal, "The Karlsruhe Verbmobil Speech Recognition Engine," in *Proc. of the ICASSP*, München; Germany, 4 1997, IEEE.

[13] H. Jin, S. Matsoukas, R. Schwartz, and F. Kubala, "Fast Robust Inverse Transform SAT and Multi-stage Adaptation," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA; USA, 1998.