# Automatic Speech Recognition based on Electromyographic Biosignals

**Article** · January 2009

Source: OAI

CITATION
1

READS
37

**1 author:**

Tanja Schultz
Universität Bremen
**424** PUBLICATIONS **7,404** CITATIONS

SEE PROFILE

# Automatic Speech Recognition based on Electromyographic Biosignals

Szu-Chen Stan Jou[1] and Tanja Schultz[1,2]

[1] International Center for Advanced Communication Technologies,
Carnegie Mellon University, Pittsburgh, PA, USA
[2] Cognitive Systems Laboratory
Karlsruhe University, Karlsruhe, Germany
tanja@cs.cmu.edu

**Abstract.** This paper presents our studies of automatic speech recognition based on electromyographic biosignals captured from the articulatory muscles in the face using surface electrodes. We develop a phone-based speech recognizer and describe how the performance of this recognizer improves by carefully designing and tailoring the extraction of relevant speech feature toward electromyographic signals. Our experimental design includes the collection of audibly spoken speech simultaneously recorded as acoustic data using a close-speaking microphone and as electromyographic signals using electrodes. Our experiments indicate that electromyographic signals precede the acoustic signal by about 0.05-0.06 seconds. Furthermore, we introduce articulatory feature classifiers, which had recently shown to improved classical speech recognition significantly. We describe that the classification accuracy of articulatory features clearly benefits from the tailored feature extraction. Finally, these classifiers are integrated into the overall decoding framework applying a stream architecture. Our final system achieves a word error rate of 29.9% on a 100-word recognition task.

## 1 INTRODUCTION

Computers have become an integral part of our daily lives and consequently require user-friendly interfaces for efficient human-computer interaction. Automatic speech recognition (ASR) systems offer the most natural front-end for human-computer interface because humans naturally communicate through speech. ASR is the automatic process of transforming spoken speech into a textual representation of corresponding word sequences. It allows for applications, such as command and control, dictation, dialog systems, audio indexing, and speech translation.

However, traditional ASR is based on the acoustic representation of speech and thus comes with several challenges. First of all, it requires the user to speak audibly. This may disturb bystanders and may also jeopardize a confidential communication. For example, telephone-based service systems often require the user to provide confidential information such as passwords or credit card numbers. If the call is made in public places, this confidential information might be eavesdropped by others. At the same time, making a phone call might distract or annoy bystanders, for example if a phone call is made in a quiet environment such as in the theater or during a meeting. The second major challenge of traditional ASR is the lack of robustness in case the acoustic channel is disturbed by ambient noise. Since the input speech signal is transmitted over the air and usually is picked up by a standard microphone, all other air-transmitted acoustic signals are picked up by this microphone as well. In most cases it is impossible to accurately

extract the relevant speech signal from the overlapping noise. Usually, such a corrupted speech signal results in a dramatic decrease of ASR performance.

We address these challenges of traditional ASR by using a transmission channel that is robust against ambient noise. Instead of relying on the acoustic signal we switch to electromyographic biosignals emitted from our body when speaking. Electromyography (EMG) is a technique for measuring the electrical potential generated by muscle cells during muscle activity. Speech is produced by the activity of the articulatory apparatus, which is moved by a large variety of articulatory muscles. By placing surface electrodes on the relevant articulatory muscles, we measure the electrical potentials during the speech production process. Our recognition system learns the typical muscle activity patterns. After this training process it can then recognize a produced sound from the corresponding electromyographic signal.

Automatic speech recognition based on electromyographic biosignals is inherently robust to ambient noise because the EMG electrodes measure the muscle activity at the skin tissue and do not rely on any air-transmitted signals. In addition, there is another major advantage: since EMG-based ASR does not rely on any air-transmitted signal, it is no longer necessary to speak audibly. Rather it could be shown that it is possible to recognize spoken speech even if it is only mouthed without making any sound. As a result, EMG-based speech recognition provides answers to all three major challenges of traditional speech recognition. It is robust to ambient noise, it allows for confidential input in public places and the input process does not disturb any bystanders or quiet environments. In summary we believe that the proposed EMG-based interface driven by non-audible speech will be of significant benefit to the community. We see three major purposes:

1. Robustness and Environment: interfaces for non-audible speech will enable people to communicate silently without disturbing bystanders or contaminating the environment with noise.
2. Privacy and Security: silent speech interfaces keep confidential spoken input safe and secure.
3. Health and Aging: recognizing speech on the bases of electromyographic signals may offer an alternative for people with speech disabilities and also for elderly who need to preserve energy and want to speak with less effort.

## 2 RELATED WORK

EMG-based ASR is a very young discipline and several challenges have yet to be overcome. While capturing the eletromyographic biosignal has proved to be a very useful tool to analyze speech research since the 1960's [1], the application of surface EMG signals to automatic speech recognition happened very recently. It was first proposed by Chan et al. [2] in 2002. Their research focused on recognizing short commands and digits spoken by jet pilots during a flight mission, i.e. the speech was captured in an extremely noisy environment. Jorgensen et al. [3] proposed sub-auditory speech recognition using two pairs of EMG electrodes attached to the throat and was the first to demonstrate sub-vocal isolated word recognition using different feature extraction and classification methods [3–5]. Manabe et al. showed that silent speech recognition is feasible when the EMG is recorded at the skin surface using electrodes pressed to the skin [6, 7]. They applied this technique to discriminate five Japanese vowels and to recognize a small vocabulary of 10 Japanese digits, carefully spoken with pauses in between each event. Maier-Hein et al. investigated non-audible and audible EMG speech recognition and focused on important aspects, such as dependencies on speakers and effects of electrode re-positioning [8].

While the described pioneering studies show some of the potential of EMG-based speech recognition, they are limited to a very small vocabulary ranging from five to forty isolated spoken words. The main motivation for this limitation is to simplify the classification task by treating the complete utterance, spoken in isolation, as one class to be identified. In contrast, the standard practice in traditional large vocabulary continuous speech recognition (LVCSR) consists of modeling a word or utterance as a sequence of phones, i.e. in terms of the smallest possible sound unit. The rationale behind this modeling scheme is that more reliable models can be trained for

smaller units since they appear more often during training. Also, the complete set of phone units is usually small for most languages (around 50 units) and after training all phones of a language, each word of this language can be built by simply concatenating the corresponding phones.

Consequently, lifting the constraints of whole word models in EMG-based ASR by introducing phones as a basic recognition unit is one of the major stepping stones necessary to enable large vocabulary continuous speech recognition and thus to open up silent speech technologies to a large number of applications.

# 3 EXPERIMENTAL SETUP

This paper describes our efforts in developing an EMG-based speech recognition system for a 100-word vocabulary - a size which already allows for useful applications. We achieve this by creating a phone-based EMG ASR system that is based on a novel feature extraction method tailored toward EMG signals. This recognizer is benchmarked, analyzed, and enhanced by articulatory features (AF). Furthermore, we investigate the relationship between AFs and EMG signals, present issues and current limitations in the signal capturing process, discuss the extraction of relevant features and optimize the signal processing step for the purpose of speech recognition. Finally, we integrate the novel EMG features and the AF classifiers into the phone-based EMG speech recognition system using a stream architecture, and report speech recognition performance numbers in terms of word error rates.

## 3.1 Data Collection

For this study we collected data from one male speaker in a single recording session. The speaker was sitting in a quiet office room and read English sentences as prompted on a computer screen. To compare acoustic and electromyographic speech signals we recorded both signals simultaneously in a parallel setup. For this comparison to be valuable we recorded audibly spoken speech, so all results reported in this paper refers to audible speech, not silent speech. The acoustic signal was recorded with a Sennheiser HMD 410 close-speaking microphone in 16 kHz sampling rate, 16 bit resolution, and linear PCM encoding. The electromyographic signal was recorded with six pairs of Ag/Ag-Cl surface electrodes attached to the skin (see Fig. 1 with description below) using a Varioport EMG recorder [9], sampled at 600 Hz, with the same resolution and encoding as the acoustic signal. Additionally, a common ground reference for the EMG signals is connected via a self-adhesive button electrode placed on the left wrist. Synchronization of the signals was ensured by a push-to-talk scenario, where pushing the button generated a marker signal that was fed into both, the EMG recorder and the acoustic sound card. Furthermore, care was taken such that the close-speaking microphone did not interfere with the EMG electrode attachment.

The speaker read 10 times a set of 38 phonetically-balanced sentences and 10 times a set of 12 news article sentences. The resulting 380 phonetically-balanced utterances were used for training and the 120 news article utterances were used for testing. The total duration of the training and test set are 45.9 and 10.6 minutes, respectively. We also recorded ten utterances of 5-second silence for normalization purposes (see below).

## 3.2 Electrode Positioning

Fig. 1 shows the positions of the six pairs of Ag/Ag-Cl surface electrodes in the face of the speaker to pick up the electromyographic signals of the most relevant articulatory muscles: the levator angulis oris (EMG2,3), the zygomaticus major (EMG2,3), the platysma (EMG4), the orbicularis oris (EMG5), the anterior belly of the digastric (EMG1), and the tongue (EMG1,6). In earlier studies [2, 8] these and similar positions had shown to be the most effective ones. Two of these six channels (EMG2,6) are captured with the traditional bipolar configuration, with a 2cm center-to-center inter-electrode spacing. In the other four cases, one electrode is placed directly onto the articulatory muscle, while the other electrode is used as a reference attached to either the nose (EMG1) or to both ears (EMG 3,4,5). In order to reduce the impedance at the electrode-skin

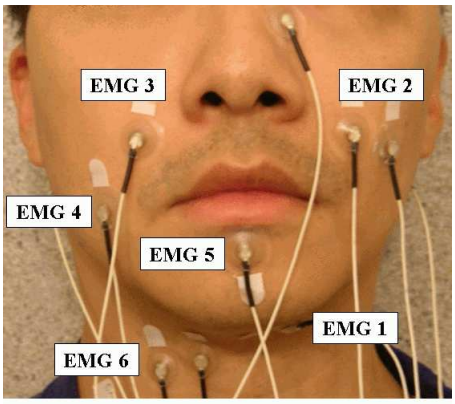


Fig. 1: EMG Electrode positioning

junctions, a small amount of electrode gel was applied. All the electrode pairs were connected to the EMG recorder [9], in which each of the detection electrode pairs pick up the EMG signal and the ground electrode provides a common reference. EMG responses were differentially amplified, filtered by a 300 Hz low-pass and a 1Hz high-pass filter. We chose to not apply a 50Hz notch filter to avoid the loss of information in this frequency band.

As described above the material for training and testing the recognizer was taken from the same recording session. This way we controlled the impact of electrode re-positioning, a challenge for state-of-the-art EMG speech recognition that we had previously studied in [8].

### 3.3 Modeling Units for EMG-based Speech Recognition

In our earlier work Walliczek et al.[10] compared different sound units for EMG-based speech recognition. Three model granularities were investigated, models based on full-words, on syllables, and based on phones. In addition, these models were refined to incorporate context information similar to traditional acoustic speech recognition. Three types of context models were explored, context independent, context dependent, and context clustered models. Experiment were conducted to recognize *seen* words, i.e. those words which had been seen in training and test. With a 32-word vocabulary, the word model performs the best with a word error rate of 17.1%, while context-dependent syllable model achieved 20.7% and context-clustered phone model gave 20.2% performance. However, when experiments were carried out on *unseen* words, i.e. on those test words which had not been seen during training but were covered by the test vocabulary, the picture changes drastically. In this case the full-word model does not have the flexibility to recognize unseen words, thus this test was performed on syllable and phone models only. Experimental result showed that the phone-based model outperforms the syllable-based model with a word error rate of 37.6% and 44.9% respectively. Consequently, we will focus the remainder of this work on phone-based speech recognition systems.

### 3.4 Bootstrapping EMG-based Speech Recognition

In order to initialize the phone-based EMG speech recognizer we generated a forced alignment of the audible speech data with a pre-existing Broadcast News (BN) recognizer [11] that was trained using our Janus Speech Recognition Toolkit. Due to our simultaneous recording setup these forced-aligned labels can be used to bootstrap the EMG speech recognizer. Since the training set is small, we limited the recognizer to context-independent acoustic models, applying a 3-state Hidden Markov Model scheme using a total of 3.3k Gaussians divided among 136 states. The amount of Gaussians used to model one state is decided automatically in a data-driven fashion. For decoding, the resulting acoustic model was combined with a standard trigram BN language model and a vocabulary that was restricted to the words in the test set. In total the decoding

vocabulary contains 108 words of which 35 have been seen in the training. The test set was explicitly excluded from the language model corpus. We are aware that this setup ignores some aspects of large vocabulary speech recognition, such as the problems of out-of-vocabulary words. However, the aim of this study is to focus on the signal preprocessing and unit modeling aspects for electromyographic signals rather than on back-end language related challenges.

### 3.5 Articulatory Feature Classifier and Stream Architecture

Articulatory features are expected to be more robust than cepstral features since they represent articulatory movements, which are less affected by speech signal variation or noise. We derive the AFs from the IPA phonological features of phones as described in [12]. AFs have binary values, e.g. the values of the horizontal positions of the dorsum FRONT, CENTRAL, and BACK are either present or absent. To classify an AF as present or absent, we compare the likelihood score of the present model with that of the absent model. The models also consider priors based on the frequency of features seen in the training data. Similar to the phone units of the EMG recognizer, the AF classifiers are trained solely on the EMG signal, no acoustic signal is used. In total there are 29 AF classifiers, each is modeled by a Gaussian Mixture Model of 60 mixtures.

Finally the AF classifiers are combined with the phone-based HMMs using a stream architecture. This approach had proved to be successful for traditional ASR since AFs provide complementary information to the phone-based models. The stream architecture employs a list of parallel feature streams, each of which contains one of the phone-based or articulatory features. Information from all streams are combined with a weighting scheme to generate the final EMG acoustic score for decoding [12].

### 3.6 Feature Extraction for EMG

For baseline experiments we extracted traditional spectral features from the signals of each recorded EMG channel. First a 27ms hamming window with 10ms frame-shift is applied and a Short Time Fourier Transform is computed. From the resulting spectral features, 17 delta coefficients are calculated together with the mean of the time domain values in the 27ms observation window. This results in an 18-dimensional feature vector per channel. If the signals of more than one channel are used for classification, the corresponding feature vectors are concatenated to form the final vector.

Since the EMG signal differs substantially from the acoustic speech signal, we explored other features. First, we normalized the DC offset to zero by estimating the offset based on the 5-second silence utterances and subtract the value from all EMG signals. We denote the EMG signal with normalized DC as $x[n]$ and its short-time Fourier spectrum as $\mathbf{X}$. The nine-point double-averaged signal is given by $w[n]$, the high frequency signal $p[n]$, and the corresponding rectified signal $r[n]$. We define the time-domain mean features $\bar{\mathbf{x}}, \bar{\mathbf{w}}$, and $\bar{\mathbf{r}}$ of the signals $x[n], w[n]$, and $r[n]$, respectively. In addition, we use the power features $\mathbf{P_w}$ and $\mathbf{P_r}$ and we define $\mathbf{z}$ as the frame-based zero-crossing rate of $p[n]$. To improve the context modeling, we apply several contextual filters. The delta filter: $D(\mathbf{f}_j) = \mathbf{f}_j - \mathbf{f}_{j-1}$. The trend filter: $T(\mathbf{f}_j, k) = \mathbf{f}_{j+k} - \mathbf{f}_{j-k}$. The stacking filter: $S(\mathbf{f}_j, k) = [\mathbf{f}_{j-k}, \mathbf{f}_{j-k+1}, ..., \mathbf{f}_{j+k-1}, \mathbf{f}_{j+k}]$, where $j$ is the frame index and $k$ is the context width. After the feature extraction process we apply a linear discriminant analysis (LDA) on the final features to reduce the dimensionality to a constant value of 32 [13].

## 4 ARTICULATORY FEATURE CLASSIFIERS

In this chapter we describe the development and benchmarking of the Articulatory Feature Classifiers. The baseline AF classifiers are first trained using the forced-alignments derived from the acoustic speech recognizer using the BN system as described in section 3.4. The AF classifiers are trained on middle frames of the phones, since the alignment to the middle frames is assumed to be more stable than to the begin and end states. To evaluate the performance, the trained AF

classifiers are applied to the test data to generate frame-based hypothesis. The performance metrics are F-scores with $\alpha = 0.5$, i.e. F-score $= 2PR/(P+R)$, and precision $P = C_{\text{tp}}/(C_{\text{tp}}+C_{\text{fp}})$, recall $R = C_{\text{tp}}/(C_{\text{tp}}+C_{\text{fn}})$, $C_{\text{tp}}$ = true positive count, $C_{\text{fp}}$ = false positive count, and $C_{\text{fn}}$ = false negative count.

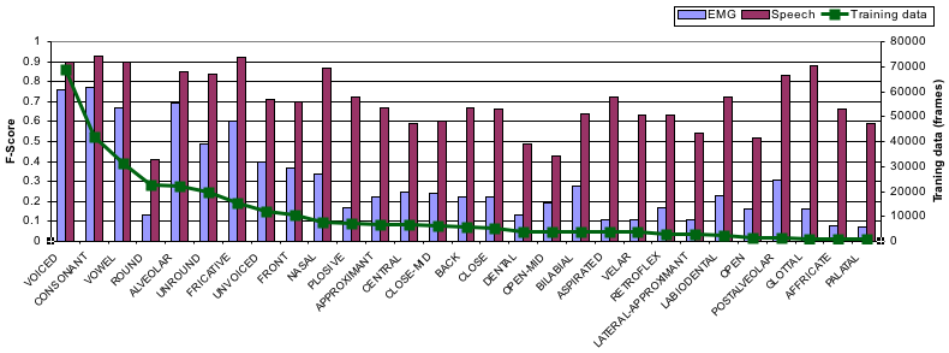


Fig. 2: AF classification performance (F-score) for acoustic and EMG signals

The AF training and test procedure was applied to both, the acoustic speech and the EMG signals. For the EMG signals we concatenated all six channels (EMG1 - EMG6) into a single feature vector as described in section 3.6. The average F-score for all 29 AFs is 0.814 for the acoustic signal and 0.467 for the EMG signal. Fig. 2 gives a breakdown of F-scores for each single AF for both acoustic and EMG signal. It also shows the amount of data available to train each feature (frame counts, 10ms per frame).

First, we observe that the F-score of the AFs trained on acoustic signals significantly outperforms the F-score trained on the EMG signal. This is true for all articulatory features, however for some AFs the decrease is less severe than for others. The AFs VOICED, CONSONANT, VOWEL, and ALVEOLAR seem to be less effected than the rest. Second, we see that the amount of training data is naturally biased toward more general categories, such as VOICED-UNVOICED, VOWEL-CONSONANT.

### 4.1 Time Delay Between Acoustic and EMG signals

We investigated why some of the EMG-based AF classifiers show a larger decrease in performance compared to the acoustic ones than others. For the evaluation above we used the offset between the two signals as given by the marker channel of our simultaneous recording setup, i.e. all EMG channels are synchronized to the acoustic channel. However, the human articulator movements are anticipatory to the acoustic signal since the speech signal is a product of articulator movements and source excitation [2]. Consequently, the time alignment we used for bootstrapping our EMG-based system might lead to mis-alignments for the EMG signals. Therefore, we investigated the delay time of the acoustic signal to the EMG signals by applying different offset times for the forced-alignment labels of the acoustic signal. As shown on the left-hand side of Fig. 3, the initial time-alignment (delay = 0) does not give the best F-score. The best F-scores are achieved with time delays between 0.02 to 0.12 seconds. After 0.12 seconds the performance decreases drastically. These results suggest that EMG signals precede the acoustic speech signal.

In the next set of experiments we explore if the anticipatory effect between the articulator movements and the acoustic speech differs for particular articulatory muscles. We assumed that this would be the case since some muscle activity result in longer range movements than others and thus may be more sensitive to a matching time delay. To explore the impact of the time delay on single EMG channels we conducted the same experiments as above, this time separate for each single EMG channel. The right-hand side of Fig. 3 depicts the impact of the time delay for each of the six EMG channels. As expected, some EMG signals are more sensitive to time delay

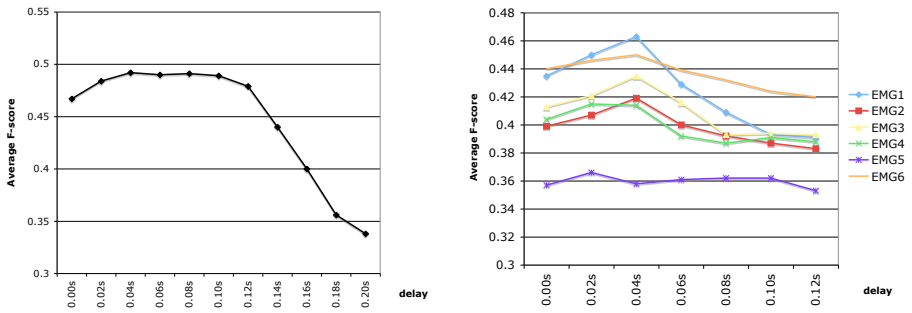


Fig. 3: Classification performance (F-score) over different time delays between EMG and acoustic signals. Six-channel concatenated EMG (left) and individual EMG channels (right)

than others, e.g. EMG1 (digastric, the muscle which moves the jaw, among other functions) has a clear peak at 0.04 seconds delay, while EMG6 (tongue muscle) is more consistent over the various time delays. The time delays vary over the channels but the performance peaks range between 0.2 and 0.10 seconds. When the optimal time delay is applied to each EMG channel, the F-score increases to 0.502 compared to the baseline of 0.467. It also outperforms a uniform delay of 0.04 seconds which gave 0.492.

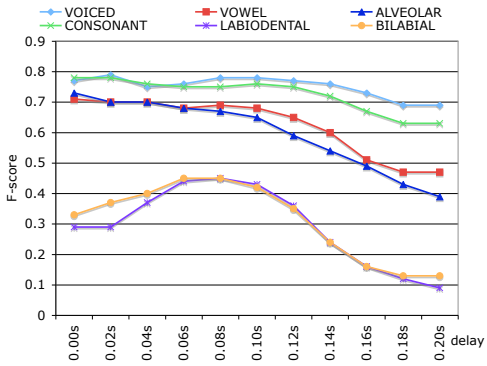


Fig. 4: Performances of six representative AFs

### 4.2 Impact of Time Delay on Articulatory Features

In the last set of experiments on time delay we investigated the impact of the delay from a different angle. Rather than looking at single EMG channels and averaging the F-score over all AF classifiers, we explored the impact on the single Articulatory Features and averaged over all channels. Fig. 4 shows the analysis for six AFs that represent different characteristics of performance changes with different delays. For example the F-scores for VOICED are rather stable across various delays, while BILABIAL and ALVEOLAR are rather sensitive. We were not able to find a conclusive explanation yet, but will investigate this further on a multiple speaker data set.

### 4.3 EMG Channel Combination

Results in [8] for EMG speech recognition based on full-word models had indicated that the concatenation of multiple EMG channels into one feature vector usually outperforms single EMG channel features. In the following experiments we wanted to explore if this finding holds for the combination of time-delayed signals and for AF classifiers. For this purpose we conducted experiments on EMG-pairs in which each EMG signal is adjusted with its optimal time offset.

The first row in Table 1 shows the F-scores averaged over all AFs for the single channel baseline when no time delay is applied. The second row gives the F-scores when the optimal time delay for the individual channel is applied. The third to last row give the F-scores for combinations of two EMG channels, i.e. the F-score 0.464 in row EMG2 + ... column EMG6 shows that the combination of channel EMG2 + EMG6 outperforms the single channel EMG2 (F-score = 0.419) and the single channel EMG6 (F-score = 0.450). This indicates that the two channels carry complementary information. Similar effects could be observed with the combination of EMG1 and EMG3.

Table 1: F-scores for single EMG Channels and Pairwise Combination

| F-Scores | EMG1 | EMG2 | EMG3 | EMG4 | EMG5 | EMG6 |
|---|---|---|---|---|---|---|
| delay = 0 | 0.435 | 0.399 | 0.413 | 0.404 | 0.357 | 0.440 |
| opt. delay | **0.463** | 0.419 | 0.435 | 0.415 | 0.366 | 0.450 |
| EMG1 + ... | | 0.439 | **0.465** | 0.443 | 0.417 | 0.458 |
| EMG2 + ... | | | 0.440 | 0.443 | 0.414 | **0.464** |
| EMG3 + ... | | | | 0.421 | 0.414 | 0.449 |
| EMG4 + ... | | | | | 0.400 | 0.433 |
| EMG5 + ... | | | | | | 0.399 |

Table 2 lists the top-4 articulators with the best F-scores. For single channels, EMG1 performs the best across these top-performance articulators, while EMG1-3, EMG1-6, and EMG2-6 are the best pairwise channel combinations. Interestingly, even though EMG5 performs the worst as a single channel classifier, EMG5 can be complemented with EMG2 to improve the classification performance for VOWEL.

Table 2: AF Classification Performance on single EMG Channels and Combination

| AFs | VOICED | | CONSONANT | | ALVEOLAR | | VOWEL | |
|---|---|---|---|---|---|---|---|---|
| Sorted F-Score (single EMG) | 1 | 0.80 | 2 | 0.73 | 1 | 0.65 | 1 | 0.59 |
| | 6 | 0.79 | 3 | 0.72 | 3 | 0.61 | 2 | 0.59 |
| | 3 | 0.76 | 1 | 0.71 | 2 | 0.59 | 6 | 0.56 |
| | 4 | 0.75 | 6 | 0.71 | 6 | 0.56 | 3 | 0.52 |
| | 2 | 0.74 | 4 | 0.69 | 4 | 0.55 | 4 | 0.51 |
| | 5 | 0.74 | 5 | 0.63 | 5 | 0.45 | 5 | 0.51 |
| Sorted F-Score (Paired EMGs) | 1-6 | 0.77 | 1-6 | 0.76 | 1-3 | 0.69 | 2-6 | 0.64 |
| | 1-3 | 0.76 | 2-3 | 0.75 | 1-6 | 0.67 | 2-4 | 0.62 |
| | 1-2 | 0.76 | 3-6 | 0.74 | 1-2 | 0.66 | 2-5 | 0.62 |
| | 2-6 | 0.75 | 2-4 | 0.74 | 2-6 | 0.66 | 1-6 | 0.62 |
| | 3-6 | 0.75 | 2-6 | 0.74 | 2-3 | 0.65 | 1-3 | 0.61 |

# 5 Feature Extraction for EMG Speech Recognition

In this chapter we investigate the extraction of relevant features for EMG-based speech recognition. We report performance numbers based on the Word Error Rate (WER) of the recognizer. Word error rate is given as WER $= \frac{S+D+I}{N}$, with $S =$ word substitution count, $D =$ word deletion count, $I =$ word insertion count, $N =$ number of reference words. In the following experiments the final EMG features are generated by stacking single-channel EMG features of the channels EMG1, EMG2, EMG3, EMG4, and EMG6. The channel EMG5 was not considered because it was found to be rather noisy. After stacking, an LDA was applied to reduce the dimensions to 32 throughout the experiments.

**Spectral Features:** In earlier work we found that spectral coefficients outperform cepstral and LPC coefficients on EMG-based speech recognition [8]. Therefore, we use the spectral features as baseline in this paper. The spectral features are denoted by $\mathbf{S0} = \mathbf{X}$, $\mathbf{SD} = [\mathbf{X}, D(\mathbf{X})]$, and $\mathbf{SS} = S(\mathbf{X}, 1)$. The left-hand side of Fig. 5 depicts the Word Error Rates for the spectral features. It shows that the contextual features improve the performance. Additionally, adding time delays for modeling the anticipatory effects also helps, which is consistent with the results of the articulatory feature analysis.

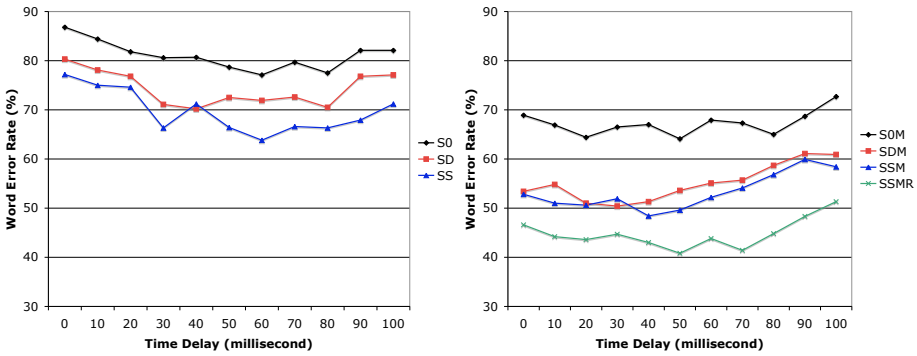


Fig. 5: Word Error Rate on Spectral (left) and Spectral+Temporal (right) Features

**Spectral + Temporal (ST) Features:** Following the results of [8] we added the following time-domain features: $\mathbf{S0M} = \mathbf{X_m}$, $\mathbf{SDM} = [\mathbf{X_m}, D(\mathbf{X_m})]$, $\mathbf{SSM} = S(\mathbf{X_m}, 1)$, and $\mathbf{SSMR} = S(\mathbf{X_{mr}}, 1)$. where $\mathbf{X_m} = [\mathbf{X}, \bar{\mathbf{x}}]$ and $\mathbf{X_{mr}} = [\mathbf{X}, \bar{\mathbf{x}}, \bar{\mathbf{r}}, \mathbf{z}]$. The performance of the resulting speech recognition system is shown on the right-hand side of Fig. 5. Enhancing the spectral features by time-domain features improves the performance quite substantially.

**Specialized EMG Features:** We observed that the spectral features are still noisy for the model training of EMG-based speech recognition. Therefore we designed specialized EMG features that are normalized and smoothed to extract relevant information from EMG signals more robustly. The performance of these EMG features are given on the left-hand side of Fig. 6, where the EMG

features are

$$\mathbf{E0} = [\mathbf{f0}, D(\mathbf{f0}), D(D(\mathbf{f0})), T(\mathbf{f0}, 3)], \quad \text{where } \mathbf{f0} = [\bar{\mathbf{w}}, \mathbf{P_w}]$$
$$\mathbf{E1} = [\mathbf{f1}, D(\mathbf{f1}), T(\mathbf{f1}, 3)], \quad \text{where } \mathbf{f1} = [\bar{\mathbf{w}}, \mathbf{P_w}, \mathbf{P_r}, \mathbf{z}]$$
$$\mathbf{E2} = [\mathbf{f2}, D(\mathbf{f2}), T(\mathbf{f2}, 3)], \quad \text{where } \mathbf{f2} = [\bar{\mathbf{w}}, \mathbf{P_w}, \mathbf{P_r}, \mathbf{z}, \bar{\mathbf{r}}]$$
$$\mathbf{E3} = S(\mathbf{E2}, 1)$$
$$\mathbf{E4} = S(\mathbf{f2}, 5)$$

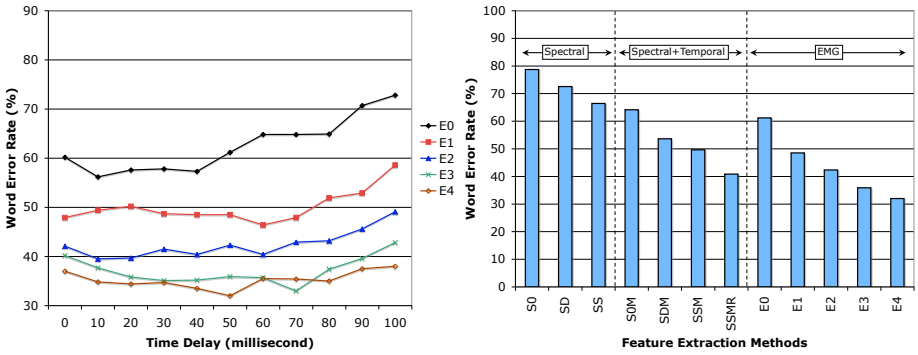


Fig. 6: Word Error Rate on EMG Features over Delay (left) and Summary of Improvements (right)

The essence of the design of these specialized EMG feature extraction methods is to reduce noise while preserving the most relevant information for classification. Since the EMG spectral feature is noisy, we first extract the time-domain mean feature, which is empirically known to be useful. By adding power and contextual information to the time-domain mean, **E0** is generated and it by itself outperforms all the spectral-only features. Since the mean and power represent low-frequency components only, we added the high-frequency power and the high-frequency zero-crossing rate to form **E1**, which gives another 10% improvement. With one additional feature of the high-frequency mean, **E2** is generated. **E2** again improves the WER. **E1** and **E2** show that the specific high-frequency information can be helpful. **E3** and **E4** use different approaches to model the contextual information, and they show that large context provides useful information for the LDA feature optimization step. They also show that the features with large context are more robust against the EMG anticipatory effect. The performance of the specialized EMG features are summarized on the right-hand side of Fig. 6. The delay was set to a constant value of 50 ms for this summary.

## 6 Integration of AF Classifiers and Special EMG Features

In the following experiments we describe the integration of the newly developed specialized EMG features into the Articulatory Feature classifiers. Similar to the experiments described above, we bootstrapped the EMG recognizer from the forced alignments based on the acoustic data. The baseline system is created with setting the delay to zero. Different from the experiments described in section 4 we only applied five EMG channels (all but EMG5) and compared the spectral plus

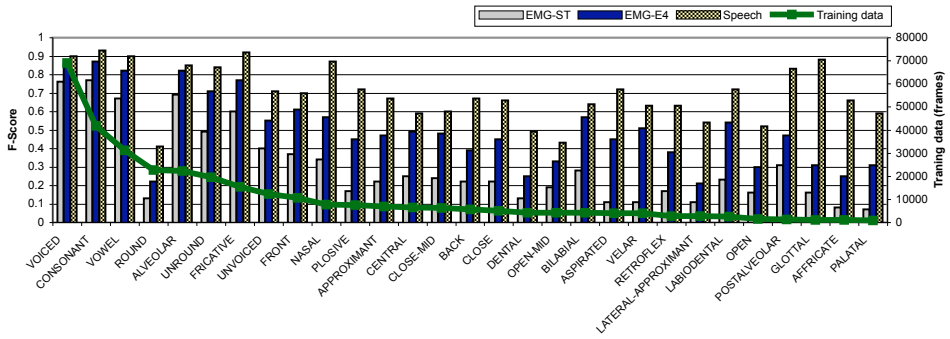


Fig. 7: AF classification performance (F-score) for acoustic and EMG signals with E4 features

time domain features (EMG-ST) with the specialized EMG features (EMG-E4). The average F-scores over the five channels of all 29 AFs are 0.492 for EMG-ST, 0.686 for EMG-E4, and 0.814 for the acoustic signal (the acoustic signal remains the same, so the F-score has the same value as reported in section 4). Fig. 7 shows the classification performance for the individual AFs for both, the EMG and the acoustic signal, along with the amount of training data in frames. The results indicate that EMG-E4 significantly outperforms EMG-ST. Also, performance of the EMG-based recognizer closes the large gap to the acoustic based recognizer that was initially observed in our baseline experiments (see Fig.2).

We also conducted time delay experiments similar to the ones described above to investigate the anticipatory effect of EMG signals [14]. Fig. 8 shows the F-scores of E4 over various LDA frame sizes and delays. We observe similar anticipatory effects of E4-LDA and ST with time delays around 0.02 to 0.10 seconds. Compared to the 90-dimensional ST feature, E4-LDA1 has a dimensionality of 25 but demonstrates a much higher F-score. Fig. 8 also indicates that a wider LDA context width provides a higher F-score and seems to be more robust for modeling the anticipatory effect. We believe this results from that fact that LDA is able to pick up useful information from the wider context.

### 6.1 EMG Channel Combination

In order to analyze E4 for individual EMG channels, we trained the AF classifiers on single channels and pairwise channel combinations. The F-scores are shown in Fig. 9 and prove that the E4 features outperform ST features in all configurations. Moreover, E4 on single-channel EMG 1, 2, 3, 6 are already better than the all-channel ST's best F-score 0.492. For ST, the pairwise channel combination provides only marginal improvements; in contrast, for E4, the figure shows significant improvements of pairwise channels combinations compared to single channels setups. We believe this significant improvements comes from a better decorrelated feature space provided by the E4 features.

### 6.2 Decoding in the Stream Architecture

Finally we conducted full decoding experiments by integrating the phone-based and the AF-based information into our stream architecture. The test set was divided into two equally-sized subsets, on which the following procedure was done in two-fold cross-validation. On the development

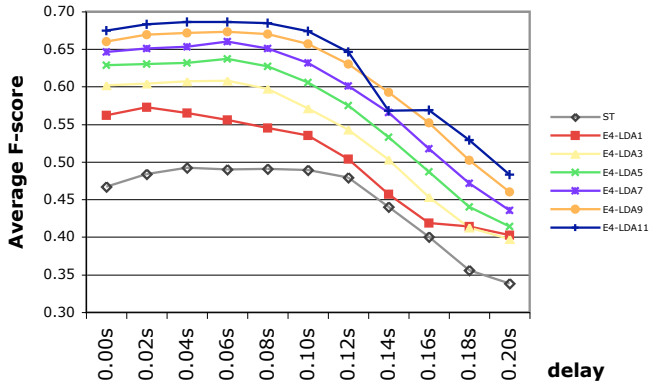


Fig. 8: AF Classification (F-scores) for five-channel EMG-ST and EMG-E4 over different LDA frame sizes and time delays

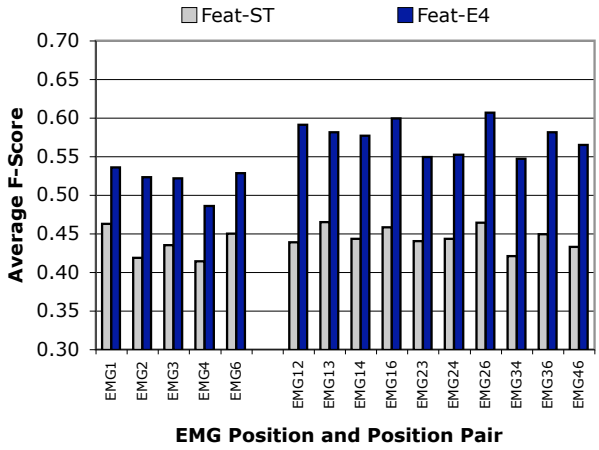


Fig. 9: AF Classification (F-scores) on single and combined EMG channels for EMG-ST and EMG-E4 features

subset, we incrementally added the AF classifiers one by one into the decoder in a greedy fashion, i.e., the AF that achieves the best WER was kept in the stream. After the WER improvement was saturated, we fixed the AF sequences and applied them on the test subset. Fig. 10 shows the word error rate and its relative improvements averaged over the two cross-validation turns. With five AFs, the WER tops 11.8% relative improvement, but there is no additional gain by

adding more AFs. Among the selected AFs, only four are selected in both cross-validation turns. This inconsistency suggests a further investigation of the AF selection procedure to ensure better generalization.

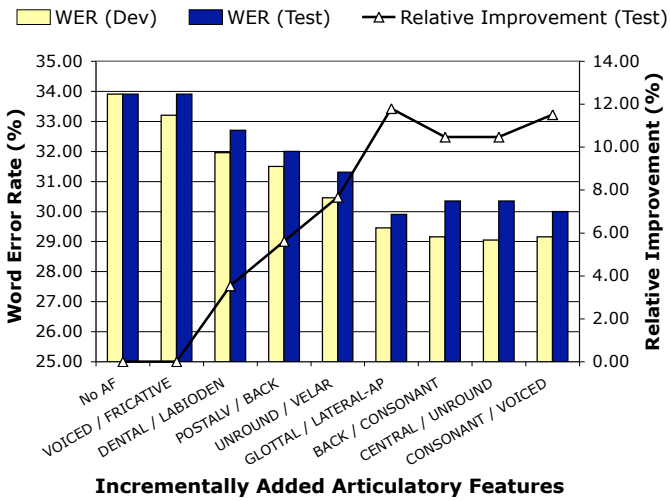


Fig. 10: EMG recognition performance (WER) and relative improvements for incrementally added AF classifiers. The two AF sequences correspond to the best AF-insertion on a two-fold cross-validation

# 7 CONCLUSIONS AND FUTURE WORK

We presented our studies of automatic speech recognition based on electromyographic biosignals captured from the articulatory muscles in the face using surface electrodes. We designed our data collection and experiments such that audibly spoken speech was simultaneously recorded as acoustic and as electromyographic data. This way we could study the time delay between the acoustic signals captured by a microphone and the EMG signals resulting from the corresponding articulatory muscle activity. Our experiments indicate that the EMG signal precedes the acoustic signal by about 50 to 60ms. Furthermore, we found that the time delay depends on the captured articulatory muscle.

Based on the collected dataset we develop a phone-based EMG speech recognizer and described how the performance of this recognizer improves by carefully designing and tailoring the extraction of relevant speech features toward electromyographic signals. The specialized EMG features gave significant improvements over spectral and time-domain features. Furthermore, we introduce articulatory feature classifiers, which had recently shown to improve classical speech recognition significantly. We describe that the classification accuracy of articulatory features clearly benefits from the tailored feature extraction.

Finally, we integrated the AF classifiers into the overall decoding framework by applying a stream architecture. The AF classifiers gave a 11.8% relative improvement over the phone-based

recognizer. Our final EMG based speech recognition system achieves a word error rate of 29.9% on a 100-word recognition task and thus comes within performance ranges useful for practical applications.

EMG-based speech recognition is a very young research area and many aspects are waiting to get explored. In the near future we expect to study effects such as the impact of speaker dependencies, electrode re-positioning, and kinematic differences between audible and non-audible speech. In order to investigate some of these aspects we are currently collecting a database of EMG speech that targets a large number of speakers uttering audible and non-audible speech.

## References


1. Fromkin, V., Ladefoged, P.: Electromyography in speech research. Phonetica **15** (1966)
2. Chan, A., Englehart, K., Hudgins, B., Lovely, D.: Hidden Markov model classification of myoelectric signals in speech. IEEE Engineering in Medicine and Biology Magazine **21**(4) (2002) 143–146
3. Jorgensen, C., Lee, D., Agabon, S.: Sub auditory speech recognition based on EMG signals. In: Proc. IJCNN, Portland, Oregon (July 2003)
4. Jorgensen, C., Binsted, K.: Web browser control using EMG based sub vocal speech recognition. In: Proc. HICSS, Hawaii (January 2005)
5. Betts, B., Jorgensen, C.: Small vocabulary communication and control using surface electromyography in an acoustically noisy environment. In: Proc. HICSS, Hawaii (January 2006)
6. Manabe, H., Hiraiwa, A., Sugimura, T.: Unvoiced speech recognition using EMG-Mime speech recognition. In: Proc. CHI, Ft. Lauderdale, Florida (April 2003)
7. Manabe, H., Zhang, Z.: Multi-stream HMM for EMG-based speech recognition. In: Proc. IEEE EMBS, San Francisco, California (September 2004)
8. Maier-Hein, L., Metze, F., Schultz, T., Waibel, A.: Session independent non-audible speech recognition using surface electromyography. In: Proc. ASRU, San Juan, Puerto Rico (November 2005)
9. Becker, K.: Varioport. http://www.becker-meditec.de (2005)
10. Walliczek, M., Kraft, F., Jou, S.C., Schultz, T., Waibel, A.: Sub-word unit based non-audible speech recognition using surface electromyography. In: Proc. Interspeech, Pittsburgh, PA (September 2006)
11. Yu, H., Waibel, A.: Streaming the front-end of a speech recognizer. In: Proc. ICSLP, Beijing, China (2000)
12. Metze, F., Waibel, A.: A flexible stream architecture for ASR using articulatory features. In: Proc. ICSLP, Denver, CO (September 2002)
13. Jou, S.C., Schultz, T., Walliczek, M., Kraft, F., Waibel, A.: Towards continuous speech recognition using surface electromyography. In: Proc. Interspeech, Pittsburgh, PA (September 2006)
14. Jou, S.C., Schultz, T., Waibel, A.: Continuous electromyographic speech recognition with a multi-stream decoding architecture. In: Proc. ICASSP, Honolulu, Hawai'i (April 2007)