EARS: ELECTROMYOGRAPHICAL AUTOMATIC RECOGNITION OF SPEECH

Szu-Chen Stan Jou and Tanja Schultz
International Center for Advanced Communication Technologies
Carnegie Mellon University, Pittsburgh, PA, USA
Karlsruhe University, Karlsruhe, Germany
scjou@cs.cmu.edu, tanja@cs.cmu.edu

Keywords: Electromyography, Speech Recognition, Articulatory Feature, Feature Extraction.

Abstract: In this paper, we present our research on automatic speech recognition of surface electromyographic signals that are generated by the human articulatory muscles. With parallel recorded audible speech and electromyographic signals, experiments are conducted to show the anticipatory behavior of electromyographic signals with respect to speech signals. Additionally, we demonstrate how to develop phone-based speech recognizers with carefully designed electromyographic feature extraction methods. We show that articulatory feature (AF) classifiers can also benefit from the novel feature, which improve the F-score of the AF classifiers from 0.467 to 0.686. With a stream architecture, the AF classifiers are then integrated into the decoding framework. Overall, the word error rate improves from 86.8% to 29.9% on a 100 word vocabulary recognition task.

1 INTRODUCTION

As computer technologies advance, computers have become an integral part of modern daily lives and our expectations for a user-friendly interface grow everyday. Automatic speech recognition (ASR) is one of the most efficient front-end for human-computer interface because it is natural for humans to communicate through speech. ASR is an automatic computerized speech-to-text process which converts human speech signals into written words. It has various applications, such as voice command and control, dictation, dialog systems, audio indexing, speech-to-speech translation, etc. However, these ASR applications usually do not work well in noisy environments. Besides, they usually require the user to speak aloud, which may be disturbing to bystanders and brings up concern of privacy loss. In this paper, we describe our research of integrating signals based on electromyography with traditional acoustic speech signals for the purpose of speech recognition.

The input speech signal of the traditional ASR process is usually recorded with a microphone, e.g., a close-talking headset or a telephone. However, from the ASR point of view, microphone recordings often suffer from ambient noise or in other words the noise robustness issue, because microphones pick up vibration from the air-transmitted channel; therefore, while picking up air vibration generated by human voices, microphones also pick up air-transmitted ambient noises. In most cases, ambient noise deteriorates the ASR performance and the decrease in performance depends on how badly the original voice signal has been corrupted by noise. Besides the noise robustness issue, microphone-based ASR often has applicability issues, by which we mean that it is often suboptimal to use microphones as the input device of speech applications in certain situation. For example, in an on-line shopping system, it is often required to input confidential information such as credit card numbers, which may be overheard if the user speak aloud via the air-transmitted channels. Usually this kind of overhearing results in confidentiality or privacy infringement. Besides, another issue of applicability is that speaking aloud usually annoys other people. Just imagine how annoying it would be if your officemate spends all day dictating to the computer to write a report, let alone many people dictate simultaneously.

In order to resolve the noise robustness and the applicability issues, we have applied electromyographic (EMG) method to our speech recognition research. The motivation is that the EMG method is inherently robust to ambient noise and it enables silent speech recognition to avoid disturbance and confidentiality issues. The EMG method measures muscular electric potential with a set of electrodes attached to the skin where the articulatory muscles underlie. In the physi-

ological speech production process, as we speak, neural control signals are transmitted to articulatory muscles, and the articulatory muscles contract and relax accordingly to produce voice. The muscle activity alters the electric potential along the muscle fibers, and the EMG method can measure this kind of potential change. In other words, the articulatory muscle activities result in electric potential change, which can be picked up by EMG electrodes for further signal processing, e.g., speech recognition. The EMG method is inherently robust to ambient noise because the EMG electrodes contact to the human tissue directly without the air-transmission channel. In addition, the EMG method has better applicability because the EMG method makes it possible to recognize silent speech, which means mouthing words without making any sound.

For silent speech recognition with EMG, Manabe et al. first showed that it is possible to recognize five Japanese vowels and ten Japanese isolated digits using surface EMG signals recorded with electrodes pressed on the facial skin (Manabe et al., 2003; Manabe and Zhang, 2004). EMG has been a useful analytic tool in speech research since the 1960's (Fromkin and Ladefoged, 1966), and the recent application of surface EMG signals to automatic speech recognition was proposed by Chan et al. They focused on recognizing voice command from jet pilots under noisy environment, so they showed digit recognition in normal audible speech (Chan et al., 2002). Jorgensen et al. proposed sub auditory speech recognition using two pairs of EMG electrodes attached to the throat. Sub vocal isolated word recognition was demonstrated with various feature extraction and classification methods (Jorgensen et al., 2003; Jorgensen and Binsted, 2005; Betts and Jorgensen, 2006). Maier-Hein et al. reported non-audible EMG speech recognition focusing on speaker and session independency issues. (Maier-Hein et al., 2005).

However, these pioneering studies are limited to small vocabulary ranging from five to around forty isolated words. The main reason of this limitation is that the classification unit is restrained to a whole utterance, instead of a phone as a smaller and more flexible unit. As a standard practice of large vocabulary continuous speech recognition (LVCSR), the phone is a natural unit based on linguistic knowledge. From the pattern recognition's point of view, the phone as a smaller unit is preferred over a whole utterance because phones get more training data per classification unit for more reliable statistical inference. The phone unit is also more flexible in order to constitute any pronunciation combination of words as theoretically unlimited vocabulary for speech recognition. With the phone unit relaxation, EMG speech recognition can be treated as a standard LVCSR task and we can apply any advanced LVCSR algorithms to improve the EMG speech recognizer.

In this paper, we introduce such an EMG speech recognition system with the following research aspects. Firstly, we analyze the phone-based EMG speech recognition system with articulatory features and their relationship with signals of different EMG channels. Next, we demonstrate the challenges of EMG signal processing with the aspect of feature extraction for the speech recognition system. We then describe our novel EMG feature extraction methods which makes the phone-based system possible. Lastly, we integrate the novel EMG feature extraction methods and the articulatory feature classifiers into the phone-based EMG speech recognition system with a stream architecture. Notice that the experiments described in this paper are conducted on normal audible speech, not silent mouthing speech.

2 RESEARCH APPROACH

2.1 Data Acquisition

In this paper, we report results of data collected from one male speaker in one recording session, which means the EMG electrode positions were stable and consistent during this whole session. In a quiet room, the speaker read English sentences in normal audible speech, which was simultaneously recorded with a parallel setup of an EMG recorder and a USB soundcard with a standard close-talking microphone attached to it. When the speaker pressed the push-to-record button, the recording software started to record both EMG and speech channels and generated a marker signal fed into both the EMG recorder and the USB soundcard. The marker signal was then used for synchronizing the EMG and the speech signals. The speaker read 10 times of a set of 38 phonetically-balanced sentences and 10 times of 12 sentences from news articles. The 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 special silence utterances, each of which is about five seconds long on average. The format of the speech recordings is 16 kHz sampling rate, two bytes per sample, and linear PCM, while the EMG recording format is 600 Hz sampling rate, two bytes per sample, and linear PCM. The speech was recorded with a Sennheiser HMD 410 close-talking headset.

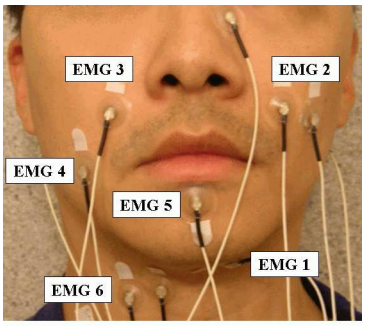

Figure 1: EMG positioning.

The EMG signals were recorded with six pairs of Ag/Ag-Cl surface electrodes attached to the skin, as shown in Fig. 1. Additionally, a common ground reference for the EMG signals is connected via a self-adhesive button electrode placed on the left wrist. The six electrode pairs are positioned in order to pick up the signals of corresponding 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) (Chan et al., 2002; Maier-Hein et al., 2005). Two of these six channels (EMG2,6) are positioned with a classical bipolar configuration, where a 2cm center-to-center inter-electrode spacing is applied. For the other four channels, one of the electrodes is placed directly on the articulatory muscles 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 junctions, a small amount of electrode gel was applied to each electrode. All the electrode pairs were connected to the EMG recorder (Becker, 2005), 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 and sampled at 600 Hz. In order to avoid loss of relevant information contained in the signals we did not apply a 50 Hz notch filter which can be used for the removal of line interference. Also note that all care was taken such that wearing the close-talking headset does not interfere with the EMG electrode attachment.

2.2 EMG-based Speech Recognition

We used the following approach to bootstrap the phone-based EMG speech recognizer. First of all, the forced alignment of the audible speech data is generated with a Broadcast News (BN) speech recognizer (Yu and Waibel, 2000), which is trained with the Janus Recognition Toolkit (JRTk). Since we have parallel recorded audible and EMG speech data, the forced-aligned labels of the audible speech were used to bootstrap the EMG speech recognizer. Since the training set is very small, we only trained context-independent acoustic models. The trained acoustic model was used together with a trigram BN language model for decoding. Because the problem of large vocabulary continuous speech recognition is still very difficult for state-of-the-art EMG speech processing, we restricted the decoding vocabulary to the words appearing in the test set in this study. This approach allows us to better demonstrate the performance differences introduced by different feature extraction methods. To cover all the test sentences, the decoding vocabulary contains 108 words in total. Note that the training vocabulary contains 415 words, 35 of which also exist in the decoding vocabulary. Also note that the test sentences were not applied for language model training.

2.3 Articulatory Feature Classifier and Stream Architecture

Compared to widely-used cepstral features for automatic speech recognition, articulatory features are expected to be more robust because they represent articulatory movements, which are less affected by speech signal variation or noise. Instead of measuring the AFs directly, we derive them from phones as described in (Metze and Waibel, 2002). More precisely, we use the IPA phonological features for AF derivation. In this work, we use AFs that have binary values. For example, each of the positions of the dorsum, namely FRONT, CENTRAL and BACK is an AF that has a value either present or absent. To classify the AF as present or absent, the likelihood scores of the corresponding present model and absent model are compared. Also, the models take into account a prior value based on the frequency of features in the training data.

The training of AF classifiers is done on middle frames of the phones only, because they are more stable than the beginning or ending frames. Identical to the training of EMG speech recognizer, the AF classifiers are also trained solely on the EMG signals without speech acoustics. There are 29 AF classifiers, each of which is a Gaussian Mixture Model (GMM) containing 60 Gaussians. To test the performance, the AF classifiers are applied and generate frame-based hypotheses.

The idea behind the stream architecture with AF classifiers is that the AF streams are expected to provide additional robust phonological information to the

phone-based hidden Markov model (HMM) speech recognizer. The stream architecture employs a list of parallel feature streams, each of which contains one of the acoustic or articulatory features. Information from all streams are combined with a weighting scheme to generate the EMG acoustic model scores for decoding (Metze and Waibel, 2002).

2.4 Feature Extraction

2.4.1 Traditional Spectral Feature

The recorded EMG signal is transformed into 18-dimensional feature vectors, with 54-ms observation window and 10-ms frame-shift for each channel.

For each channel, hamming-windowed Short Time Fourier Transform is computed, and then its delta coefficients serve as the first 17 coefficients of the final feature. The 18th coefficient consists of the mean of the time domain values in the given observation window (Maier-Hein et al., 2005). In the following experiments, features of one or more channels can be applied. If more than one channel are used for classification, the features of the corresponding channels are concatenated to form the final feature vector.

2.4.2 Special EMG Feature

Since the EMG signal is very different from the speech signal, it is necessary to explore feature extraction methods that are suitable for EMG speech recognition. Here we describe the signal preprocessing steps and feature extraction methods we designed for EMG signals.

As noted above, the EMG signals vary across different sessions. Nonetheless, the DC offsets of the EMG signals vary, too. In the attempt to make the DC offset zero, we estimate the DC offset from the special silence utterances on a per session basis, then all the EMG signals are preprocessed to subtract this session-based DC offset. Although we only discuss a single session of a single speaker in this paper, we expect this DC offset preprocessing step makes the EMG signals more stable.

To describe the features designed for EMG signals, we denote the EMG signal with normalized DC as $x[n]$ and its short-time Fourier spectrum as $\mathbf{X}$. We also denote the nine-point double-averaged signal $w[n]$, high frequency signal $p[n]$, and the corresponding rectified signal $r[n]$.

We then 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. Besides, 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 better model the context, we use the following contextual filters, which can be applied on any feature to generate a new one. 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. Note that we always apply linear discriminant analysis (LDA) on the final feature in order to reduce the dimensionality to 32.

3 EXPERIMENTS AND ANALYSES

The performance metrics used in this paper are F-score and word error rate (WER). F-score ($\alpha = 0.5$) is reported for the AF performances and WER is reported for the speech recognition performances[1].

3.1 Articulatory Feature Analysis

3.1.1 Baseline System

First of all, we forced-aligned the speech data using the aforementioned BN system. In the baseline system, this time-alignment was used for both the speech and the EMG signals. Because we have a marker channel in each signal, the marker signal is used to offset the two signals to get accurate time-synchronization. Then the aforementioned AF training and testing procedures were applied both on the speech and the six-channel concatenated EMG signals. The averaged F-scores of all 29 AFs are 0.814 for the speech signal and 0.467 for the EMG signal. Fig. 2 shows individual AF performances for the speech and EMG signals along with the amount of training data. We can see that the amount of training data (given in frames of 10 ms) has an impact on the EMG AF performance.

3.1.2 Channel Synchronization

It is observed that human articulator movements are anticipatory to the speech signal as speech signal is a product of articulator movements and source excitation (Chan et al., 2002). This means the time alignment we used for bootstrapping our EMG-based

[1] With $\alpha = 0.5$, F-score $= 2PR/(P+R)$, where precision $P = C_{tp}/(C_{tp}+C_{fp})$, recall $R = C_{tp}/(C_{tp}+C_{fn})$, C_{tp} = true positive count, C_{fp} = false positive count, C_{fn} = false negative count.

WER $= \frac{S+D+I}{N}$, where S = word substitution count, D = word deletion count, I = word insertion count, N = number of reference words.

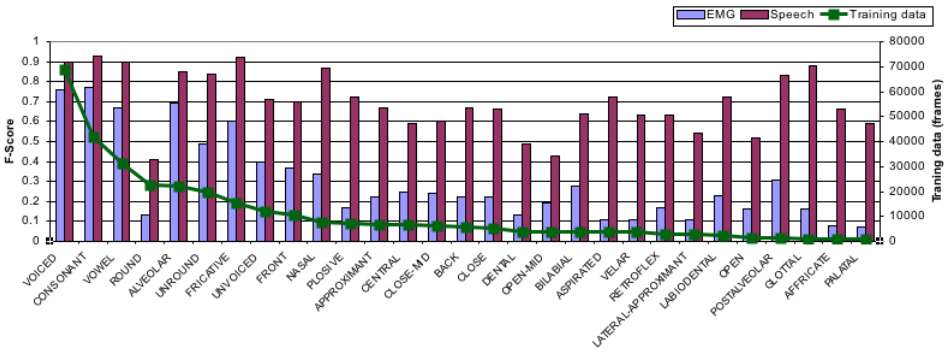

Figure 2: Baseline F-scores of the EMG and speech signals vs. the amount of training data.

system is actually mis-aligned for the EMG signals, because the speech and the EMG signals are inherently asynchronous in time. Based on this, we delayed the EMG signal with various duration to the forced-alignment labels of speech signal, and conducted the training and testing experiments respectively. As shown in Fig. 3, the initial time-alignment does not have the best F-score, while the best F-scores come with time delays around 0.02 second to 0.12 second. This result suggests that a time-delayed effect exists between the speech and the EMG signals.

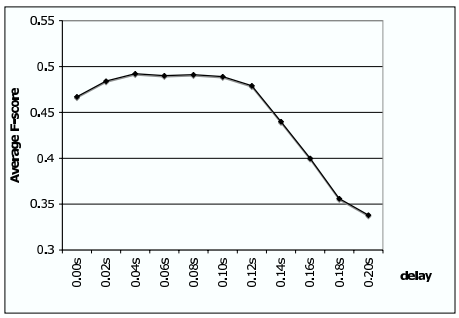

Figure 3: F-scores of concatenated six-channel EMG signals with various time delays (a delay of 0.1 means that the EMG signal is delayed to the acoustic signal by 0.1 seconds).

3.1.3 Articulator-Dependent Synchronization

To explore the time-delayed effect of EMG signals, we conducted the same experiments on the level of single EMG channels, instead of previously concatenated six-channels. The rationale is that articulators' behaviors are different from each other, so the resulted time delays are different on the corresponding EMG signals. The effect of different time delays can be seen in Fig. 4. We observed that some EMG signals are more sensitive to time delay than others, e.g. EMG1 vs. EMG6, where EMG6 is more consistent with different time delays. The delays to achieve peak performance vary for each channel and the variation is within the range of 0.02 to 0.10 seconds. To further show the time-delay effect, we also conducted an experiment which is identical to the baseline, except each channel is offset with its known best time delay. This approach gave a better F-score of 0.502 than the baseline's 0.467. It also outperforms the uniform delay of 0.04 second which gave 0.492.

Figure 4: F-scores of single-channel EMG signals with various time delays with respect to the speech signals.

3.1.4 Complementary EMG Pairs

As suggested in (Maier-Hein et al., 2005), concatenated multi-channel EMG features usually work better than single-channel EMG features. Therefore, based on aforementioned time-delayed results, we conducted experiments on EMG-pairs in which each EMG signal is adjusted with its best single-channel time offset. The first row of values in Table 1 shows the F-scores of single-channel baseline (i.e. without any time delay) and the second row shows those with the best single-channel time delay, while the rest of the values are F-scores of EMG pairs. The F-scores suggest that some EMG signals are complementary to each other, e.g. EMG1-3 and EMG2-6, which pairs perform better than both their single channels do.

Table 1: F-Score of EMG and EMG Pairs.

F-Scores	EMG1	EMG2	EMG3	EMG4	EMG5	EMG6
single	0.435	0.399	0.413	0.404	0.357	0.440
+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

3.1.5 Performance of Individual Articulators

In Table 2 and 3, we list the top-4 articulators that have the best F-scores. For single channels, EMG1 performs the best across these top-perfomance articulators, while EMG1-3, EMG1-6, and EMG2-6 are the better paired channels. Interestingly, even though EMG5 performs the worst as a single channel classifier, EMG5 can be complemented with EMG2 to form a better pair for VOWEL. In Fig. 5, we show six AFs that represent different characteristics of performance changes with different delays. For example, VOICED's F-scores are rather stable with various delay values while BILABIAL is rather sensitive. However, we do not have conclusive explanation on the relation between the AFs and the delays. Further exploration shall be conducted.

Table 2: Best F-Scores of Single EMG channels w.r.t. AF.

AFs	VOICED	CONSONANT	ALVEOLAR	VOWEL
	1 0.80	2 0.73	1 0.65	1 0.59
Sorted	6 0.79	3 0.72	3 0.61	2 0.59
F-score	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

Table 3: Best F-Scores of Paired EMG Channels w.r.t. AF

AFs	VOICED	CONSONANT	ALVEOLAR	VOWEL
	1-6 0.77	1-6 0.76	1-3 0.69	2-6 0.64
Sorted	1-3 0.76	2-3 0.75	1-6 0.67	2-4 0.62
F-Score	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

3.2 Feature Extraction Experiments

In the following experiments, the final EMG features are generated by stacking single-channel EMG features of channels 1, 2, 3, 4, 6. We do not use channel 5 because it is relatively noisy for this experiment. The final LDA dimensions are reduced to 32 for all the experiments.

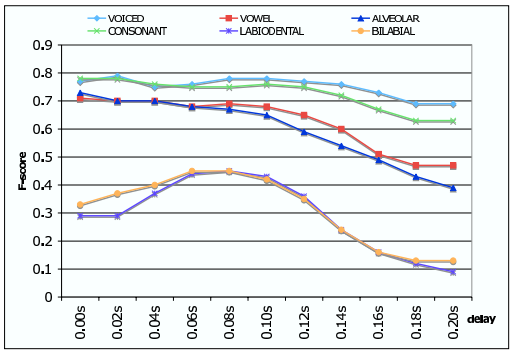

Figure 5: Performances of six representative AFs.

3.2.1 EMG ASR using Spectral Features

It was reported that the spectral coefficients are better than cepstral and LPC coefficients on EMG speech recognition (Maier-Hein et al., 2005). Therefore, we use the spectral features as baseline in this paper. As their WER is shown in Fig. 6, the spectral features are $\mathbf{S0} = \mathbf{X}$, $\mathbf{SD} = [\mathbf{X}, D(\mathbf{X})]$, and $\mathbf{SS} = S(\mathbf{X}, 1)$. We can see that the contextual features improve WER. Additionally, adding time delays for modeling the anticipatory effects also helps. This is consistent to the articulatory feature analysis above.

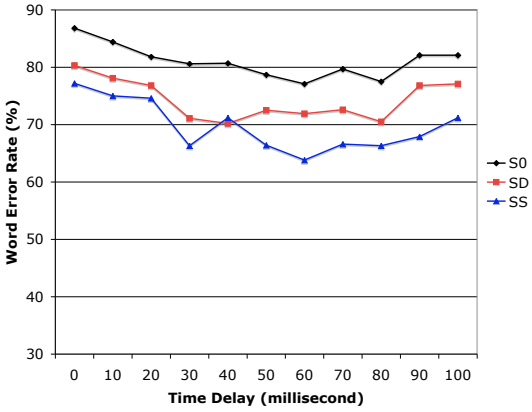

Figure 6: Word Error Rate on Spectral Features.

3.2.2 EMG ASR Systems using Spectral+temporal (ST) Features

Is was also reported that the time-domain mean feature provided additional gain to spectral features (Maier-Hein et al., 2005). Here we also added the time-domain mean feature, as their WER is shown in Fig. 7: $\mathbf{S0M} = \mathbf{X_m}$, $\mathbf{SDM} = [\mathbf{X_m}, D(\mathbf{X_m})]$, $\mathbf{SSM} = S(\mathbf{X_m}, 1)$, and $\mathbf{SSMR} = S(\mathbf{X_{mr}}, 1)$. where $\mathbf{X_m} = [\mathbf{X}, \bar{\mathbf{x}}]$ and $\mathbf{X_{mr}} = [\mathbf{X}, \bar{\mathbf{x}}, \bar{\mathbf{r}}, \mathbf{z}]$.

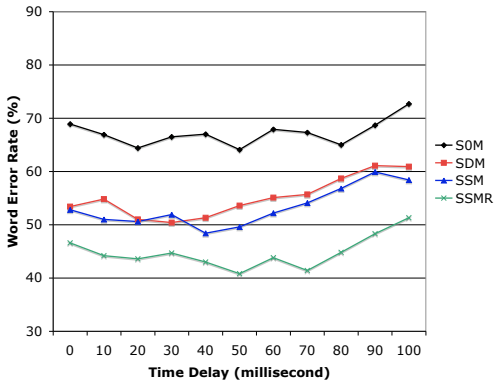

Figure 7: Word Error Rate on Spectral+Temporal Features.

3.2.3 EMG ASR Systems using EMG Features

We have observed that even though the spectral features are among the better ones, they are still very noisy for acoustic model training. Therefore we designed the EMG features that are normalized and smoothed in order to extract features from EMG signals in a more robust fashion. The performance of the EMG features are shown in Fig. 8, where the EMG features are

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

The essence of the design of feature extraction methods is to reduce noise while keeping the useful information for classification. Since the EMG spectral feature is noisy, we decide to first extract the time-domain mean feature, which is empirically known to be useful in literature. By adding power and contextual information to the time-domain mean, **E0** is generated and it already outperforms all the spectral-only features. Since the mean and power only represent the low-frequency components, we add the high-frequency power and the high-frequency zero-crossing rate to form **E1**, which gives us another 10% improvement. With one more 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. We summarize by showing the performance of all the presented feature extraction methods in Fig. 9, in which all the feature extraction methods apply a 50-ms delay.

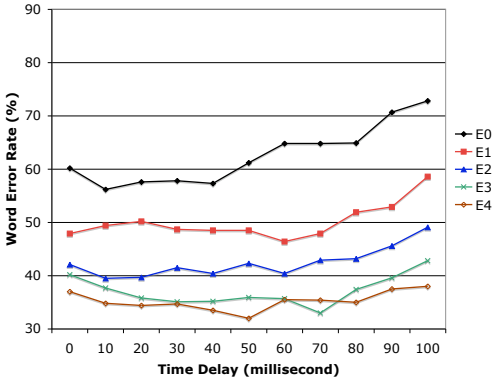

Figure 8: Word Error Rate on EMG Features.

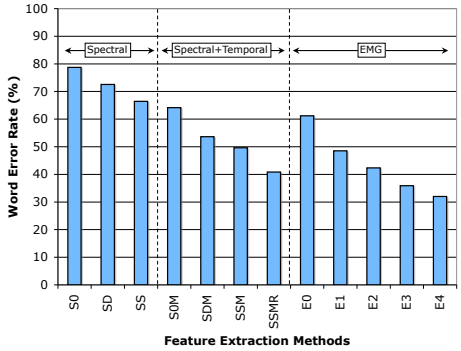

Figure 9: WER of Feature Extraction Methods.

3.3 Integration of Special EMG Feature and AF Classifiers

3.3.1 AF Classification with the E4 Feature

Identical to the aforementioned experiments, we forced-aligned the speech data using the BN speech recognizer. In the baseline system, this time-alignment was used for both the speech and the EMG signals. Because we have a marker channel in each signal, the marker signal is used to offset the two signals to get accurate time-synchronization. Then the AF training and testing procedures were applied both on the speech and the five-channel concatenated EMG signals, with the ST and E4 features. The averaged F-scores of all 29 AFs are 0.492 for EMG-ST, 0.686 for EMG-E4, and 0.814 for the speech signal. Fig. 10 shows individual AF performances for the speech and EMG signals along with the amount of training data in frames. The E4

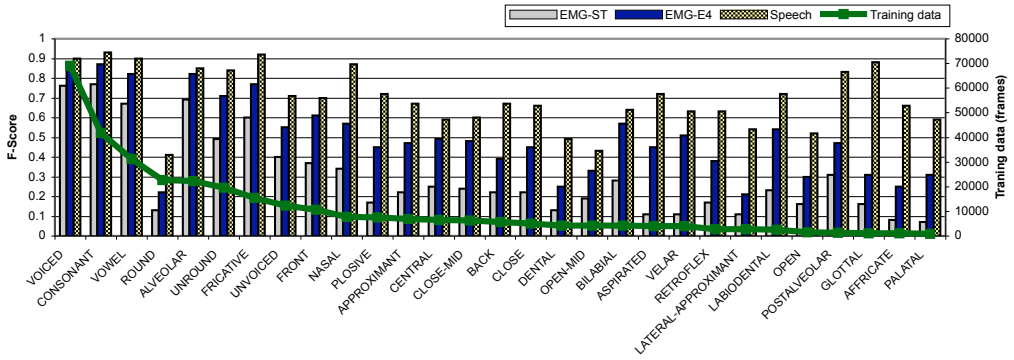

Figure 10: F-scores of the EMG-ST, EMG-E4 and speech articulatory features vs. the amount of training data.

significantly outperforms ST in that the EMG-E4 feature performance is much closer to the speech feature performance.

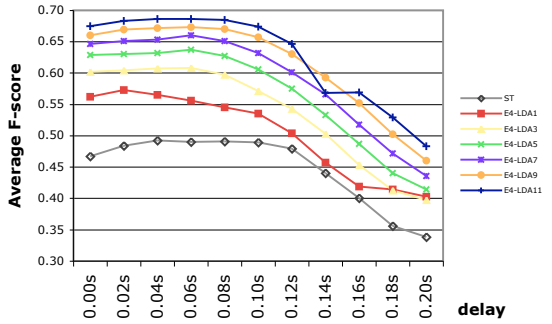

Figure 11: F-scores of concatenated five-channel EMG-ST and EMG-E4 articulatory features with various LDA frame sizes on time delays for modeling anticipatory effect.

We also conducted the time-delay experiments as done in previous ones to investigate the EMG vs. speech anticipatory effect. Fig. 11 shows the F-scores of E4 with various LDA frame sizes and delays. We observe similar anticipatory effect of E4-LDA and ST with time delay around 0.02 to 0.10 second. Compared to the 90-dimension ST feature, E4-LDA1 has a dimensionality of 25 while having a much higher F-score. The figure also shows that a wider LDA context width provides a higher F-score and is more robust for modeling the anticipatory effect, because LDA is able to pick up useful information from the wider context.

3.3.2 EMG Channel Pairs

In order to analyze E4 for individual EMG channels, we trained the AF classifiers on single channels and channel pairs. The F-scores are shown in Fig. 12. It shows E4 outperforms ST in all configurations. Moreover, E4 on single-channel EMG 1, 2, 3, 6 are already

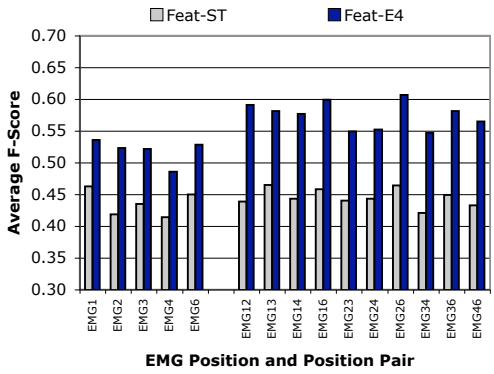

Figure 12: F-scores of the EMG-ST and EMG-E4 AFs on single EMG channel and paired EMG channels.

better than the all-channel ST's best F-score 0.492. For ST, the paired channel combination only provides marginal improvements; in contrast, for E4, the figure shows significant improvements of paired channels compared to single channels. We believe this significant improvements come from a better decorrelated feature space provided by E4.

3.3.3 Decoding in the Stream Architecture

We then conducted a full decoding experiment with the 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 subset, we incrementally added the AF classifiers one by one into the decoder in a greedy approach, i.e., the AF that helps to achieve the best WER was kept in the streams for later experiments. After the WER improvement was saturated, we fixed the AF sequence and applied them on the test subset. Fig. 13 shows the WER and its relative improvements averaged on the two cross-validation turns. With five AFs, the WER tops 11.8% relative improvement, but

there is no additional gain with more AFs.

Among the selected AFs, only four of them are selected in both cross-validation turns. This inconsistency suggests a further investigation of AF selection is necessary for generalization.

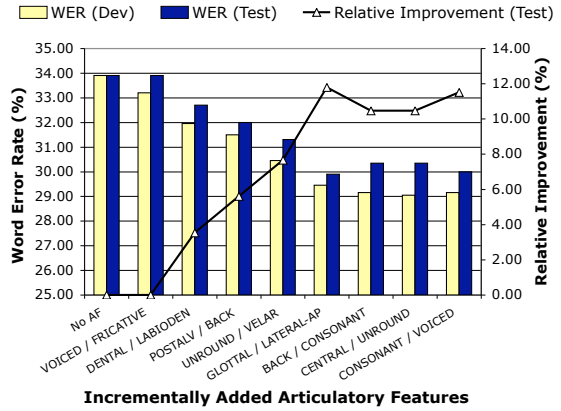

Figure 13: Word error rates and relative improvements of incrementally added EMG articulatory feature classifiers in the stream architecture. The two AF sequences correspond to the best AF-insertion on the development subsets in two-fold cross-validation.

4 COLLECTING MORE DATA

We are making efforts on larger-scale data collection of EMG speech. The targeted total number of speakers is in dozens and the recording modalities include acoustic speech, EMG, and video. Each speaker participates in two recording sessions, each of which includes a part of normal audible speech recording and a part of silent mouthing speech recording. In each part, two sets of phonetically balanced sentences are collected. One set is referred to as the general set and it exists in every part of every speaker. The other set is a speaker specific set, which is different for different speakers. Per part, the general set contains 10 sentences and the speaker specific set contains 40 sentences.

The data collection process is designed to be as unbiased as possible, e.g., to eliminate the fatigue factor. The two sessions are recorded one week apart. Besides, the order of the silent part and the audible part is reversed in the two sessions. In each recording part, the two sentence sets are mixed together into a set of 50 sentences and the sentences appear in random order. Table 4 shows the data details per speaker.

With this larger EMG corpus, we expect to be able to study the effects of speaker dependency, session dependency, audible versus mouthing speech kinematics, just to name a few.

Table 4: Data per speaker.

Speaker	
Session 1	Session 2
Part 1 audible speech rand(10+40 sentences)	Part 1 silent speech rand(10+40 sentences)
Part 2 silent speech rand(10+40 sentences)	Part 2 audible speech rand(10+40 sentences)

5 CONCLUSIONS

We have presented our recent advances on EMG speech recognition research, which has the advantages of better noise robustness and better applicability compared to traditional acoustic speech recognition. With the special EMG feature extraction methods and articulatory feature analyses, we have advanced the EMG speech recognition research from isolated word recognition to phone-based continuous speech recognition. Besides, the introduction of anticipatory effect modeling also plays an important role in this study. In summary, the EMG articulatory feature performance improves from 0.467 to 0.686 and the overall speech recognition word error rate improves from 86.8% to 29.9%.

This research topic is relatively new and unexplored with many questions waiting for answers. Although the proposed special EMG feature extraction methods do improve the performance, we believe they are still sub-optimal. Designing a better EMG feature extraction method for speech recognition is still an open problem and we are continuously working on it. Another issue is that the multi-channel EMG signals are inherently asynchronous with respect to articulatory apparatus movements. How to model this asynchronicity remains an open problem. We believe this modeling would benefit the study of speech recognition as well as articulatory kinematics.

REFERENCES

Becker, K. (2005). Varioport. *http://www.becker-meditec.de*.

Betts, B. and Jorgensen, C. (2006). Small vocabulary communication and control using surface electromyography in an acoustically noisy environment. In *Proc. HICSS*, Hawaii.

Chan, A., Englehart, K., Hudgins, B., and Lovely, D. (2002). Hidden Markov model classification of myoelectric signals in speech. *IEEE Engineering in Medicine and Biology Magazine*, 21(4):143–146.

Fromkin, V. and Ladefoged, P. (1966). Electromyography in speech research. *Phonetica*, 15.

Jorgensen, C. and Binsted, K. (2005). Web browser control using EMG based sub vocal speech recognition. In *Proc. HICSS*, Hawaii.

Jorgensen, C., Lee, D., and Agabon, S. (2003). Sub auditory speech recognition based on EMG signals. In *Proc. IJCNN*, Portland, Oregon.

Maier-Hein, L., Metze, F., Schultz, T., and Waibel, A. (2005). Session independent non-audible speech recognition using surface electromyography. In *Proc. ASRU*, San Juan, Puerto Rico.

Manabe, H., Hiraiwa, A., and Sugimura, T. (2003). Unvoiced speech recognition using EMG-Mime speech recognition. In *Proc. CHI*, Ft. Lauderdale, Florida.

Manabe, H. and Zhang, Z. (2004). Multi-stream HMM for EMG-based speech recognition. In *Proc. IEEE EMBS*, San Francisco, California.

Metze, F. and Waibel, A. (2002). A flexible stream architecture for ASR using articulatory features. In *Proc. ICSLP*, Denver, CO.

Yu, H. and Waibel, A. (2000). Streaming the front-end of a speech recognizer. In *Proc. ICSLP*, Beijing, China.