CONTINUOUS ELECTROMYOGRAPHIC SPEECH RECOGNITION WITH A MULTI-STREAM DECODING ARCHITECTURE

Szu-Chen Stan Jou, Tanja Schultz, and Alex Waibel

International Center for Advanced Communication Technologies
Carnegie Mellon University, Pittsburgh, PA
{scjou,tanja,ahw}@cs.cmu.edu

ABSTRACT

In our previous work, we reported a surface electromyographic (EMG) continuous speech recognition system with a novel EMG feature extraction method, E4, which is more robust to EMG noise than traditional spectral features. In this paper, we show that articulatory feature (AF) classifiers can also benefit from the E4 feature, which improve the F-score of the AF classifiers from 0.492 to 0.686. We also show that the E4 feature is less correlated across EMG channels and thus channel combination gains larger improvement in F-score. With a stream architecture, the AF classifiers are then integrated into the decoding framework and improve the word error rate by 11.8% relative from 33.9% to 29.9%.

Index Terms— speech recognition, electromyography, articulatory muscles, articulatory features, feature extraction

1. INTRODUCTION

As the research of automatic speech recognition (ASR) advances, computers are required to provide people a more convenient way to communicate. However, robustness and privacy have always been issues in speech based applications. To overcome this, efforts have been made to utilize whispered or non-audible silent speech for ASR with special recording devices. For example, "non-audible murmur" recognition using a stethoscopic microphone has been presented by Nakajima et al. [1], and we reported whispered speech recognition using a throat microphone [2]. Another approach is to make use of electromyographic (EMG) sensors to monitor the articulatory muscles in order to recognize non-audible silent speech. Chan et al. showed that such an approach can be used for small vocabulary isolated word recognition [3]. Other related work also showed different aspects of success on non-audible silent speech recognition [4, 5, 6].

However, these pioneering studies are limited to small vocabulary due to the classification unit that is restrained to a whole utterance, instead of phones, which is a standard practice of LVCSR. In our previous work, we demonstrated a first phone-based system and analyzed it by studying the relationship of surface electromyography and articulatory features (AFs) on audible speech [7]. Later, we have extended that work to an EMG phone-based continuous speech recognition system, which makes use of phone-based acoustic models and feature extraction methods designed for continuous EMG speech [8]. In this paper, we explore further by showing that our novel feature extraction method also performs well for EMG AF classification. Moreover, we show that the EMG AF classifiers can be integrated into the EMG acoustic model in a stream architecture for decoding to further improve the EMG ASR system.

The authors wish to thank Matthias Walliczek and Florian Kraft for their valuable contributions to this study.

2. EXPERIMENTAL SETUP

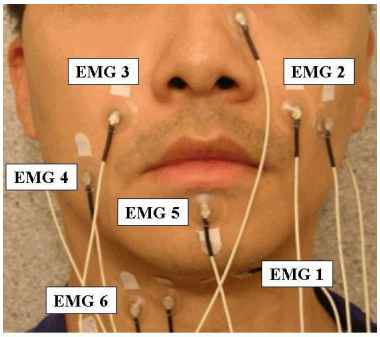
2.1. Data Acquisition

As shown in [6], EMG signals vary a lot across speakers, and even across recording sessions of the very same speaker. As a result, the performances across speakers and sessions may be unstable. To avoid this problem and to keep this research in a more controlled configuration, in this paper we report results of data collected from one male speaker in one recording session, i.e., the EMG electrode positions were stable and consistent during this whole session. In a quiet room, the subject read English sentences in normal audible speech, which was recorded simultaneously by an EMG recorder and a USB sound card with a standard close-talking microphone. 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 turns of a set of 38 phonetically-balanced sentences and 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 it is 600 Hz sampling rate, two bytes per sample, and linear PCM for the EMG signals.

The EMG signals were recorded with six pairs of Ag/Ag-Cl surface electrodes attached to the skin[1], 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) [3, 6]. 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). Note that the

[1] Strictly speaking, this method should be called *surface* EMG. However, we can simply use the term EMG without confusion in this paper.

Fig. 1. EMG electrode positioning



electrode positioning method follows [6], except that we do not use EMG5 in our final experiments because its signal is unstable in the recording session. In addition, we remove one electrode channel redundant to EMG6 (EMG7 in [6]).

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 [9], in which each of the detection electrode pairs picks 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 is usually used for the removal of line interference [6].

2.2. Audible Speech Recognizer

In order to forced-align the audible speech acoustic waveforms, we used a Broadcast News (BN) speech recognizer trained with the Janus Recognition Toolkit (JRTk) [10]. In this system, Mel-frequency cepstral coefficients (MFCC) with vocal tract length normalization (VTLN) and cepstral mean normalization (CMN) is used to get the frame-based feature. On top of that, linear discriminant analysis (LDA) is applied to a 15-frame (-7 to +7 frames) segment to generate the final feature for recognition. The recognizer is HMM-based, and makes use of quintphones with 6000 distributions sharing 2000 codebooks. The baseline performance of this system is 10.2% WER on the official BN test set (Hub4e98 set 1), F0 condition.

2.3. EMG Speech Recognizer

We used the following approach to bootstrap the EMG continuous speech recognizer. First of all, the forced-aligned labels of the audible speech data is generated with the aforementioned BN speech recognizer. 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. Context dependency is beyond the scope of this paper. 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 the state-of-the-art EMG speech processing, in this study, we restricted the decoding vocabulary to the words appearing in the test set (108 words). This approach allows us to better demonstrate the performance differences introduced by different feature extraction methods. Note that the training vocabulary contains 415 words, 35 of which also exist in the decoding vocabulary. The test sentences do not exist in the language model training data. Also note that the EMG speech recognizer is trained solely on the signals captured by the EMG electrodes, i.e., trained without speech acoustics.

2.4. Articulatory Feature Classifier

Compared to widely-used cepstral features, 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 [11]. 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 dorsum position FRONT, CENTRAL and BACK is an AF that has a value either present or absent. Moreover, these AFs do not form an orthogonal set because we want the AFs to benefit from redundant information. 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 [11].

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 GMM containing 60 Gaussians. To test the performance, the AF classifiers are applied and generate frame-based hypotheses. F-score ($\alpha = 0.5$) is reported in our experiments as the performance metric[2].

2.5. The Stream Architecture

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 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 [11].

3. EMG FEATURE EXTRACTION

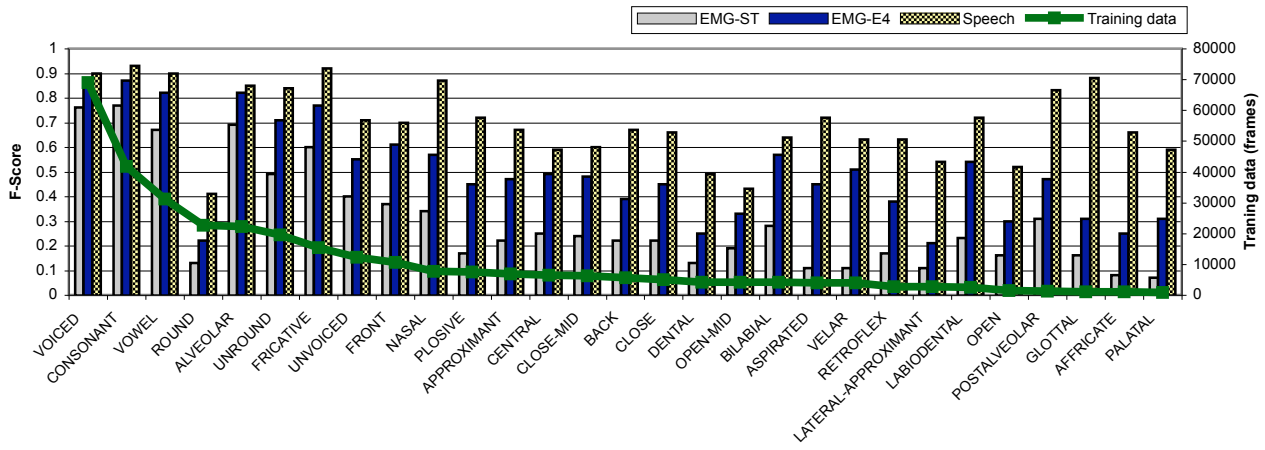
Since the EMG signal is very different from the speech acoustic signal, it is necessary to explore feature extraction methods that are suitable for EMG speech recognition. Here we describe the signal preprocessing steps and feature extraction methods we used in the experiments.

Besides the EMG signals vary across different sessions, 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 remove 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.

In our previous work, we showed the anticipatory effects of the EMG signals when compared to speech signals [7]. We also

[2]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.

Fig. 2. F-scores of the EMG-ST, EMG-E4 and speech articulatory features vs. the amount of training data



demonstrated modeling this anticipatory effect improves the F-score of articulatory feature classification. In this paper, we model the anticipatory effect by adding frame-based delays to the EMG signals when the EMG signals is forced-aligned to the audible speech labels. Only channel-independent delay is introduced in this paper, i.e., every EMG channel is delayed by the same amount of time.

3.1. The E4 Feature

We have shown in our previous work that the traditional spectral plus time-domain mean feature (ST) is very noisy. Therefore we designed the E4 feature that are normalized and smoothed in order to extract features from EMG signals in a more robust manner [8]. The E4 feature is defined as:

$$\mathbf{E4} = S(\mathbf{f2}, 5), \text{where } \mathbf{f2} = [\bar{\mathbf{w}}, \mathbf{P_w}, \mathbf{P_r}, \mathbf{z}, \bar{\mathbf{r}}]$$

where $S(\mathbf{f}, n)$ is the stacking of adjacent frames of feature $\mathbf{f}$ in the size of $2n+1$ ($-n$ to n) frames, $\mathbf{w}$ is the nine-point double-averaged signal, $\bar{\mathbf{w}}$ is the frame-based time-domain mean of $\mathbf{w}$, $\mathbf{P_w}$ is the power of $\mathbf{w}$, $\mathbf{P_r}$ is the power of the rectified signal, $\mathbf{z}$ is the zero-crossing rate of the high frequency signal, and $\bar{\mathbf{r}}$ is the time-domain mean of the rectified signal.

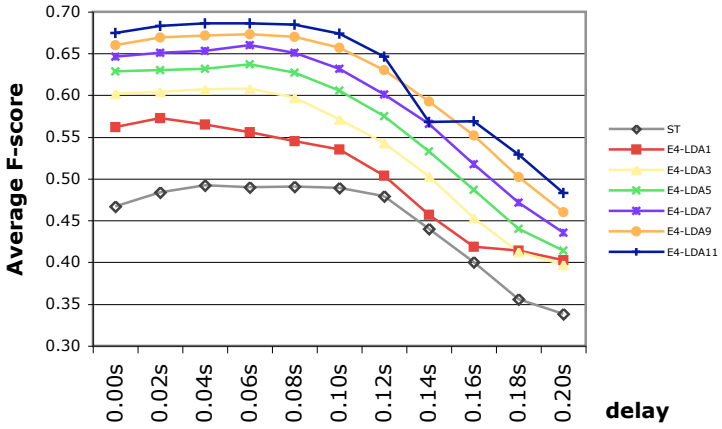
4. EXPERIMENTS AND ANALYSES

In the following experiments, the final EMG features are generated by stacking single-channel E4 features of channels 1, 2, 3, 4, and 6. Then LDA is applied and the final LDA dimensions are reduced to 32 for all experiments, in which the frame size is 27 ms and frame shift is 10 ms.

4.1. AF Classification with the E4 Feature

First of all, we forced-aligned the speech data using the aforementioned Broadcast News English 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 aforementioned 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. 2 shows individual AF performances for the speech and EMG signals along with the amount of training data in frames. We can see that E4 significantly outperforms ST in that the EMG-E4 feature performance is much closer to the speech feature performance.

Fig. 3. 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 our previous work to investigate the EMG vs. speech anticipatory effect. Fig. 3 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.

4.2. 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. 4. It shows E4 outperforms ST 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 paired

Fig. 4. F-scores of the EMG-ST and EMG-E4 articulatory features on single EMG channel and paired EMG channels

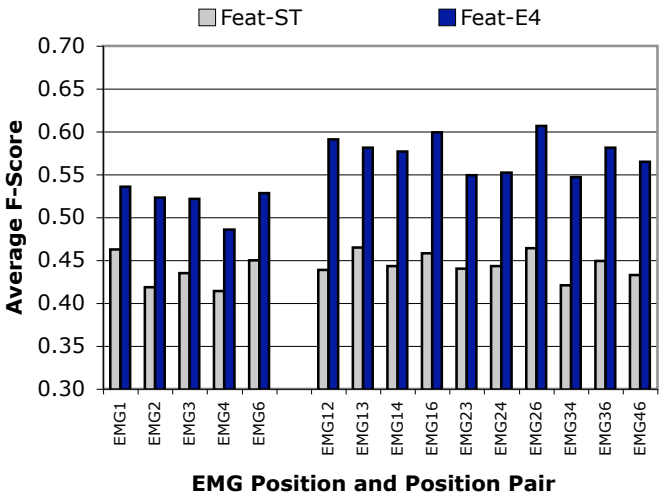
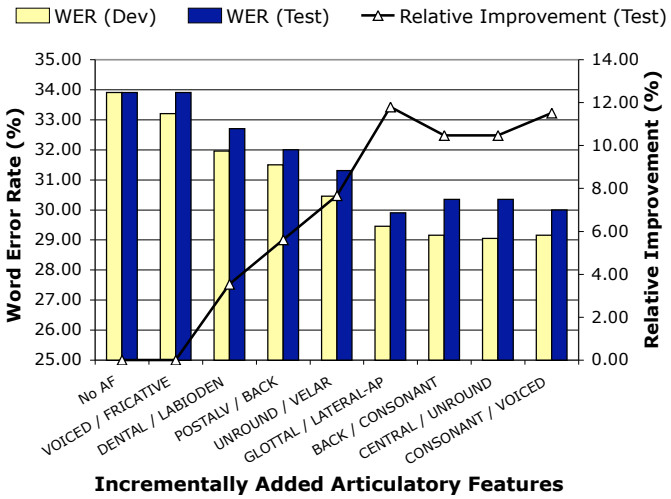


Fig. 5. Word error rates and relative improvements of incrementally added EMG articulatory feature classifiers in the stream architecture. The two AF sequences correspond to the best AF-insertion on the development subsets in two-fold cross-validation.



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.

4.3. 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. 5 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.

5. CONCLUSIONS

We have presented an EMG continuous speech recognition system with a multi-stream architecture, which makes use of the information from EMG articulatory feature classifiers besides HMM. The proposed E4 EMG feature extraction method has been shown to outperform the traditional ST method for the EMG HMM decoder, the EMG AF classifiers, and the combination of both. We have shown that E4 improves the F-score of the EMG AF classifiers from 0.492 to 0.686, and E4 is better for EMG channel combination. With the stream architecture consisting of HMM and AF classifiers, the WER improves 11.8% relative from 33.9% to 29.9%.

In the future, we plan to investigate feature selection and weighting schemes in order to effectively make use of the stream architecture for the EMG phone-based continuous speech recognizer.

6. REFERENCES

[1] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin," in *Proc. ICASSP*, Hong Kong, 2003.

[2] S.-C. Jou, T. Schultz, and A. Waibel, "Whispery speech recognition using adapted articulatory features," in *Proc. ICASSP*, Philadelphia, PA, March 2005.

[3] A.D.C. Chan, K. Englehart, B. Hudgins, and D.F. Lovely, "Hidden Markov model classification of myoelectric signals in speech," *IEEE Engineering in Medicine and Biology Magazine*, vol. 21, no. 4, pp. 143–146, 2002.

[4] B. Betts and C. Jorgensen, "Small vocabulary communication and control using surface electromyography in an acoustically noisy environment," in *Proc. HICSS*, Hawaii, Jan 2006.

[5] H. Manabe, A. Hiraiwa, and T. Sugimura, "Unvoiced speech recognition using EMG-mime speech recognition," in *Proc. HFCS*, Ft. Lauderdale, Florida, 2003.

[6] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, "Session independent non-audible speech recognition using surface electromyography," in *Proc. ASRU*, San Juan, Puerto Rico, Nov 2005.

[7] S.-C. Jou, L. Maier-Hein, T. Schultz, and A. Waibel, "Articulatory feature classification using surface electromyography," in *Proc. ICASSP*, Toulouse, France, May 2006.

[8] S.-C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards continuous speech recognition using surface electromyography," in *Proc. Interspeech*, Pittsburgh, PA, Sep 2006.

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

[10] H. Yu and A. Waibel, "Streaming the front-end of a speech recognizer," in *Proc. ICSLP*, Beijing, China, 2000.

[11] F. Metze and A. Waibel, "A flexible stream architecture for ASR using articulatory features," in *Proc. ICSLP*, Denver, CO, Sep 2002.