Towards Continuous Speech Recognition Using Surface Electromyography

Szu-Chen Jou, Tanja Schultz, Matthias Walliczek, Florian Kraft, and Alex Waibel

International Center for Advanced Communication Technologies
Carnegie Mellon University, USA and Universität Karlsruhe, Germany
{scjou,tanja,ahw}@cs.cmu.edu, {walliczek,fkraft}@ira.uka.de

Abstract

We present our research on continuous speech recognition of the surface electromyographic signals that are generated by the human articulatory muscles. Previous research on electromyographic speech recognition was limited to isolated word recognition because it was very difficult to train phoneme-based acoustic models for the electromyographic speech recognizer. In this paper, we demonstrate how to train the phoneme-based acoustic models with carefully designed electromyographic feature extraction methods. By decomposing the signal into different feature space, we successfully keep the useful information while reducing the noise. Additionally, we also model the anticipatory effect of the electromyographic signals compared to the speech signal. With a 108-word decoding vocabulary, the experimental results show that the word error rate improves from 86.8% to 32.0% by using our novel feature extraction methods.

Index Terms: speech recognition, electromyography, articulatory muscles, 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 in small vocabulary due to the classification unit that is restrained to a whole utterance, instead of phonemes, which is a standard practice of LVCSR. In our previous work, we demonstrated a first phoneme-based system and analyzed it by studying the relationship of surface electromyography and articulatory features (AFs) on audible speech [7]. Additionally, we are exploring the effects of using different sub-word units for non-audible EMG speech recognition [8]. In this paper, we advance the research to a continuous EMG speech recognition system on normal audible speech, which makes use of phoneme-based acoustic models and feature extraction methods designed for continuous EMG speech.

In the next section, we describe our experimental setup, then in Section 3 the EMG feature extraction methods, followed by Section 4 for experiments and analyses. We present our conclusions in Section 5.

2. Experimental Setup

2.1. Data Acquisition

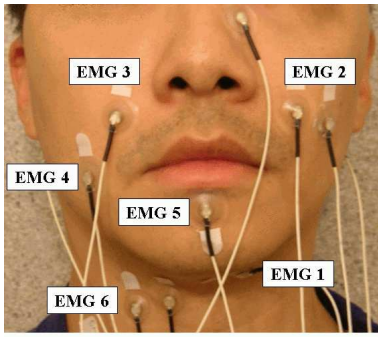
As shown in [6], EMG signals vary a lot across speakers, and even across recording sessions of the very same speaker. As a result, the performances across speakers and sessions may be unstable. To avoid this problem and to keep this research in a more controlled configuration, in this paper we report results of data collected from one male speaker in one recording session, 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 recorded with a parallel setup of an EMG recorder and a USB soundcard with a standard close-talking microphone attached to it, simultaneously. 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 speech was recorded with a Sennheiser HMD 410 close-talking headset.

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 *di-*

The authors wish to thank Peter Osztotics for his valuable contributions to this study.

[1] Strictly speaking, this method should be called *surface* EMG. However, we just use the term EMG for simplicity.

Figure 1: EMG electrode positioning



gastric (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 attaching to either the nose (EMG1) or to both ears (EMG 3,4,5). Note that the electrode positioning method follows [6], except that we do not use EMG5 in our final experiments because its signal is unstable, and one redundant electrode channel to EMG6 (EMG7 in [6]) has been removed because it did not provide additional gain on top of the other six [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 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 filters which can be used for the removal of line interference [6]. Also note that wearing the close-talking headset does not interfere with the EMG electrode attachment.

2.2. Audible Speech Recognizer

In order to forced-align the audible speech recordings, 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 continuous EMG 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. 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 do not exist in the language model training data.

3. EMG Feature Extraction

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 used in the experiments.

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.

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

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 $\mathbf{X}$. A nine-point double-averaged signal is defined as

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

A high frequency signal is defined as

$$p[n] = x[n] - w[n]$$

and the corresponding rectified signal is

$$r[n] = \begin{cases} p[n] & \text{if } p[n] \geq 0, \\ -p[n] & \text{if } p[n] < 0. \end{cases}$$

Since all the features are frame-based, the time indices 0 and N represent the beginning and the length of the frame, respec-

Figure 2: Word Error Rate on Spectral Features

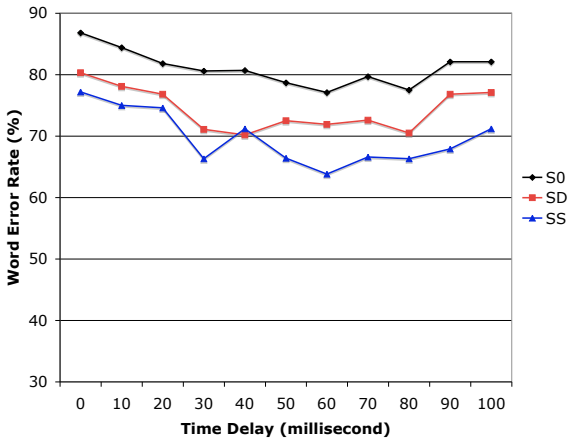
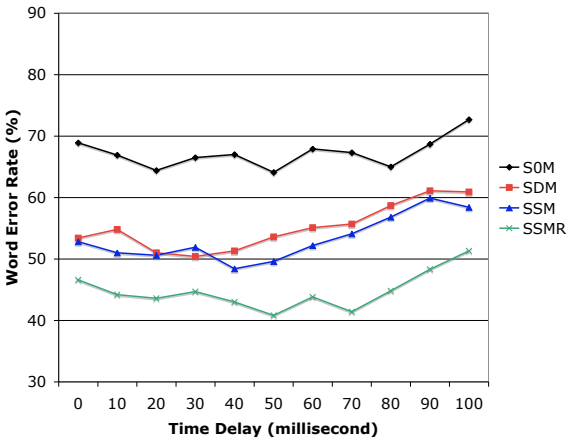


Figure 3: Word Error Rate on Spectral+Temporal Features



tively. The time-domain mean feature is defined as

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$$

Similarly we define

$$\bar{\mathbf{w}} = \frac{1}{N}\sum_{n=0}^{N-1} w[n] \quad \text{and} \quad \bar{\mathbf{r}} = \frac{1}{N}\sum_{n=0}^{N-1} r[n]$$

Besides, we use the power features

$$\mathbf{P_w} = \sum_{n=0}^{N-1} |w[n]|^2 \quad \text{and} \quad \mathbf{P_r} = \sum_{n=0}^{N-1} |r[n]|^2$$

and the frame-based zero-crossing rate of $p[n]$

$$\mathbf{z} = \text{zero-crossing count of } (p[0], p[1], ..., p[N-1])$$

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 LDA on the final feature.

4. Experiments and Analyses

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

4.1. EMG ASR Systems Using Spectral Features

In our previous work, we had reported that the spectral coefficients are better than cepstral and LPC coefficients on EMG speech recognition [6]. Therefore, we used the spectral features as baseline in this paper. As their WER is shown in Fig. 2, the spectral features are

$$\mathbf{S0} = \mathbf{X}$$
$$\mathbf{SD} = [\mathbf{X}, D(\mathbf{X})]$$
$$\mathbf{SS} = S(\mathbf{X}, 1)$$

We can see that the contextual features improve WER. Additionally, adding time delays for modeling the anticipatory effects also helps. This is consistent to our previous work [7].

4.2. EMG ASR Systems Using Spectral+Temporal Features

We had reported that the time-domain mean feature provided additional gain to spectral features [6]. Here we also added the time-domain mean feature, as their WER is shown in Fig. 3:

$$\mathbf{S0M} = \mathbf{X_m}$$
$$\mathbf{SDM} = [\mathbf{X_m}, D(\mathbf{X_m})]$$
$$\mathbf{SSM} = S(\mathbf{X_m}, 1)$$
$$\mathbf{SSMR} = S(\mathbf{X_{mr}}, 1)$$

where $\mathbf{X_m} = [\mathbf{X}, \bar{\mathbf{x}}]$ and $\mathbf{X_{mr}} = [\mathbf{X}, \bar{\mathbf{x}}, \bar{\mathbf{r}}, \mathbf{z}]$.

4.3. 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

Figure 4: Word Error Rate on EMG Features

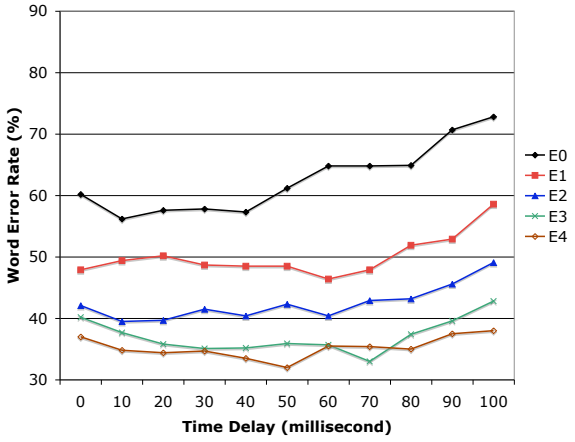
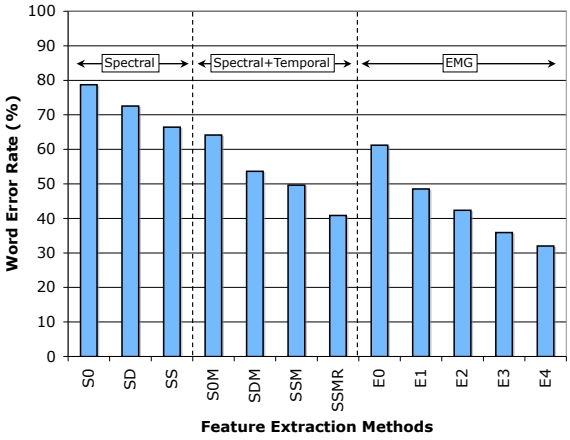


Figure 5: WER of Feature Extraction Methods with 50-ms Delay



in Fig. 4, where the EMG features are

$$\mathbf{E0} = [\mathbf{f0}, D(\mathbf{f0}), D(D(\mathbf{f0})), T(\mathbf{f0}, 3)], \quad \text{where } \mathbf{f0} = [\bar{\mathbf{w}}, \mathbf{P_w}]$$

$$\mathbf{E1} = [\mathbf{f1}, D(\mathbf{f1}), T(\mathbf{f1}, 3)], \quad \text{where } \mathbf{f1} = [\bar{\mathbf{w}}, \mathbf{P_w}, \mathbf{P_r}, \mathbf{z}]$$

$$\mathbf{E2} = [\mathbf{f2}, D(\mathbf{f2}), T(\mathbf{f2}, 3)], \quad \text{where } \mathbf{f2} = [\bar{\mathbf{w}}, \mathbf{P_w}, \mathbf{P_r}, \mathbf{z}, \bar{\mathbf{r}}]$$

$$\mathbf{E3} = S(\mathbf{E2}, 1)$$

$$\mathbf{E4} = S(\mathbf{f2}, 5)$$

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 our previous work. 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. 5, in which all the feature extraction methods apply a 50-ms delay.

5. Conclusions

We have presented a continuous EMG speech recognition system which makes use of feature extraction methods designed for EMG speech signals. With the restricted 108-word decoding vocabulary, we explored various feature extraction methods that are better representing the EMG signals for continuous speech recognition. Modeling the EMG anticipatory effect also improves the performance. The WER of the plain spectral-feature system is 86.8% and the WER of the best EMG-feature system drops to 32.0%. In the future, we expect to further model the channel-specific anticipatory effect to improve EMG feature extraction.

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] M. Walliczek, F. Kraft, S.-C. Jou, T. Schultz, and A. Waibel, "Sub-word unit based non-audible speech recognition using surface electromyography," in *Proc. Interspeech*, Pittsburgh, PA, Sep 2006.

[9] "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.