ARTICULATORY FEATURE CLASSIFICATION USING SURFACE ELECTROMYOGRAPHY

Szu-Chen Jou, Lena Maier-Hein, Tanja Schultz, and Alex Waibel

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

ABSTRACT

In this paper, we present an approach for articulatory feature classification based on surface electromyographic signals generated by the facial 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. On average, we found that the signals to be time delayed by 0.02 to 0.12 second. Furthermore, it is shown that different articulators have different anticipatory behavior. With offset-aligned signals, we improved the average F-score of the articulatory feature classifiers in our baseline system from 0.467 to 0.502.

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 studied by Nakajima et al. [1]. 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 [2]. Other related work also showed different aspects of success on non-audible silent speech recognition [3, 4, 5]. 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 order to overcome this problem, we built a first phoneme-based system and analyzed it by studying the relationship of surface electromyography and articulator features (AFs) on audible speech.

In the next section, we describe our experimental setup. followed by Section 3 for experiments and analyses. We present our conclusion in Section 4.

2. EXPERIMENTAL SETUP

2.1. Data Acquisition

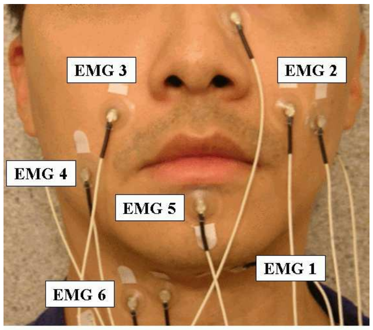
As shown in [5], EMG signals vary a lot across speakers, and even across recording sessions of the very same speaker. To reduce this effect, in this paper we report results of data collected from one male speaker in one recording session, which means the EMG electrode positions were stable and consistent 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 presses the push-to-record button, the software starts to record both EMG and speech channels and generates a marker signal fed into both the EMG recorder and the USB soundcard. The marker signal is 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 are used for training and the 120 news article utterances are used for testing. The total duration of the training and test set are 45.9 and 10.6 minutes, respectively. 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, 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 articular 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) [2, 5]. 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 articular 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

The authors wish to thank Christoph Mayer, Marcus Warga, Peter Osztotics and Artus Krohn-Grimberghe for their valuable contributions to this study.

1-4244-0469-X/06/$20.00 ©2006 IEEE

Fig. 1. EMG positioning



follows [5], except the EMG5 position is different and one redundant electrode channel to EMG6 (EMG7 in [5]) has been removed because it did not provide additional gain on top of the other six [5]. The idea of changing the EMG5 position is to more closely monitor the *orbicularis oris*, which controls the lips movement.

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 [6], 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 [5]. Also note that wearing the close-talking headset does not interfere with the EMG electrode attachment.

2.2. Feature Extraction

The recorded EMG signal is tranformed into 18-dimensional feature vectors, with 54-ms observation window and 10-ms frame-shift for each channel. We have changed the frame-shift from 4 ms to 10 ms from the original setting in order to align the speech and EMG signals.

For each channel, hamming-windowed Short Time Fourier Transform (STFT) 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 [5]. 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.

On the speech counterpart, Mel-frequency cepstral coefficients (MFCC) with vocal tract length normalization (VTLN) and cepstral mean normalization (CMN) were used to get the frame-based feature, where each frame is 16-ms long, hamming-windowed, with 10-ms frame-shift. On top of that, a linear discriminant analysis (LDA) is applied to a 15-frame (-7 to +7 frames) segment to generate the final feature vector for classification.

2.3. 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 differences or noise [7]. Instead of measuring the AFs directly, we derive them from phonemes as described in [8]. More precisely, we use the IPA phonological features for AF derivation. In this work, we use AFs that have binary values [8]. 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 [8].

The training of AF classifiers is done on middle frames of the phones only, because they are acoustically more stable than the beginning or ending frames. 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.

3. EXPERIMENTS AND ANALYSES

3.1. Baseline system

First of all, we forced-aligned the speech data using a Broadcast News English speech recognizer trained with the Janus Recognition Toolkit [9]. 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.

3.2. 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 [2]. This means the time alignment we used for bootstrapping our EMG-based system is actually mis-aligned for the EMG signals, because the speech and the EMG signals are inherently off-synchronized in time. Based on this, we delayed the EMG signal with various duration to the forced-alignment labels of speech signal,

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

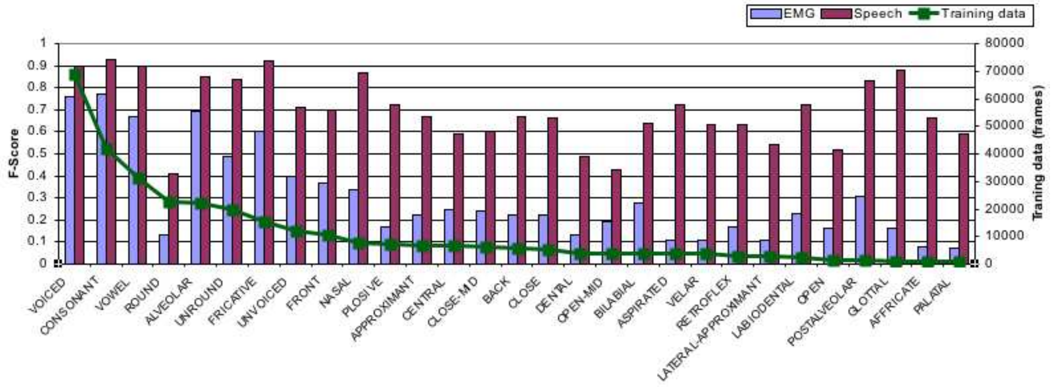


Fig. 3. F-scores of concatenated six-channel EMG signals with various time delays with respect to the speech signals

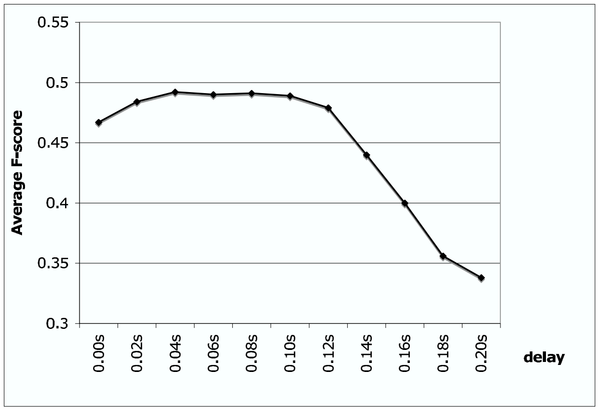
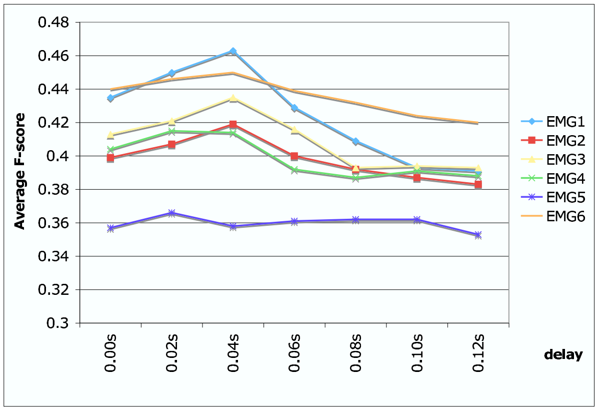


Fig. 4. F-scores of single-channel EMG signals with various time delays with respect to the speech signals



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.

3.3. 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 to 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 peak performance varies for each channel while happens around 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.

3.4. Complementary EMG Pairs

As suggested in [5], 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.

3.5. Performance with Respect to Individual Articulators

In Table 2 and 3, we list the top-5 articulators that have the best F-scores. For single channels, EMG1 performs the best

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

across these top-perfomance articulators, while EMG1-3, EMG1-6, and EMG2-6 perform as well as the 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 explaination on the relation between the AFs and the delays. Further exploration shall be conducted.

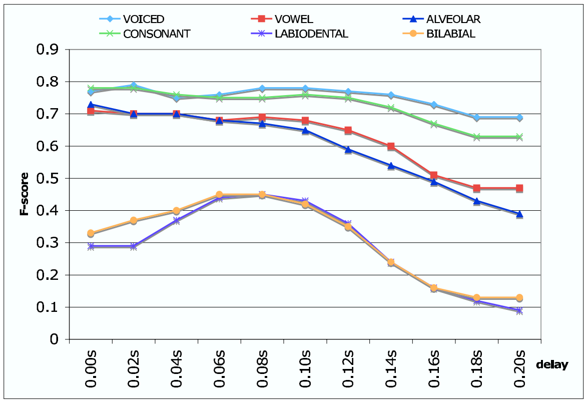
Table 2. Best F-Scores of Single EMG channels w.r.t. AF

AFs	VOICED	CONSONANT	ALVEOLAR	VOWEL	FRICATIVE
Sorted F-score	1 0.80	2 0.73	1 0.65	1 0.59	1 0.52
	6 0.79	3 0.72	3 0.61	2 0.59	2 0.50
	3 0.76	1 0.71	2 0.59	6 0.56	3 0.50
	4 0.75	6 0.71	6 0.56	3 0.52	6 0.50
	2 0.74	4 0.69	4 0.55	4 0.51	4 0.45
	5 0.74	5 0.63	5 0.45	5 0.51	5 0.39

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

AFs	VOICED	CONSONANT	ALVEOLAR	VOWEL	FRICATIVE
Sorted F-Score	1-6 0.77	1-6 0.76	1-3 0.69	2-6 0.64	1-3 0.57
	1-3 0.76	2-3 0.75	1-6 0.67	2-4 0.62	1-6 0.57
	1-2 0.76	3-6 0.74	1-2 0.66	2-5 0.62	3-6 0.56
	2-6 0.75	2-4 0.74	2-6 0.66	1-6 0.62	2-3 0.56
	3-6 0.75	2-6 0.74	2-3 0.65	1-3 0.61	2-6 0.56

Fig. 5. Performances of six representative AFs with delays



4. CONCLUSIONS

We have presented a study on the articulatory feature classification on surface electromyographic signals. The study showed that time offsets among articulators and speech waves need to be carefully considered. With carefully chosen articulator specific delays, we improved the average F-score of the articulatory feature classifiers from 0.467 to 0.502. Additionally, complementary EMG pairs can improve AF classification. We observed that the anticipatory effect and the AF performance are related and they are AF-specific. For example, as we expected, EMG6 on the throat works well on VOICED and VOWEL, which usually have longer duration so EMG6 is not affected much in terms of the anticipatory effect. This can be seen in Fig. 4 as the EMG6 performance varies slowly with different time offsets. Additionally, designed to monitor *orbicularis oris*, EMG5 does not work well as expected. Since lip movement is an important articulator, one of the significant problems is how to improve the classification on *orbicularis oris*.

5. 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] 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.

[3] C. Jorgensen and K. Binsted, "Web browser control using EMG based sub vocal speech recognition," in *Proc. HICSS*, Hawaii, Jan 2005.

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

[5] 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.

[6] "Varioport," http://www.becker-meditec.de.

[7] K. Kirchhoff, *Robust Speech Recognition Using Articulatory Information*, Ph.D. thesis, University of Bielefeld, Germany, July 1999.

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

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