EMG SIGNAL PROCESSING FOR AUDIO-EMG-BASED MULTI-MODAL SPEECH RECOGNITION

Zhipeng Zhang, Hiroyuki Manabe, Tsutomu Horikoshi, Tomoyuki Ohya
Multimedia Laboratories, NTT DoCoMo
3-5 Hikari-no-oka, Yokosuka, Kanagawa, 239-8536 Japan
Email: zzp@mml.yrp.nttdocomo.co.jp

ABSTRACT
This paper proposes robust methods for processing EMG (electromyography) signals in the framework of audio-EMG-based speech recognition. The EMG signals are captured when uttered and used as auxiliary information for recognizing speech. Two robust methods (Cepstral Mean Normalization and Spectral Subtraction) for EMG signal processing are investigated to improve the recognition performance. We also investigate the importance of stream weighting in audio-EMG-based multi-modal speech recognition. Experiments are carried out at various noise conditions and the results show the effectiveness of the proposed methods. A significant improvement in word accuracy over the audio-only recognition scheme is achieved by combining the methods.

KEY WORDS
Multimodal speech recognition, EMG, Spectral Subtraction, Cepstral Mean Subtraction

1. Introduction
Although high recognition accuracy can be obtained for clean speech using the state-of-the-art speech recognition technology, the accuracy achieved in practice is greatly degraded due to noise in the environment. Therefore, increasing the robustness to noise is one of the most important issues in automatic speech recognition (ASR). For dealing with this problem, a wide variety of noise robust methods have been proposed. Most approaches are based on the processing of just the audio component of speech. More robust speech recognition can be realized by combining speech signal with other information. Multi-modal speech recognition, in which acoustic features and other information are jointly used, has been investigated and found to increase robustness and thus the accuracy of ASR. Most multi-modal methods use visual features, typically lip information, in addition to acoustic features [1].

When speaking, the vocal and articulation organs are moved, and their actions are independent of background noise. This implies that speech can be recognized by observing the movement or activity of such organs. Several methods can be used to observe their movements. One is to use image capture and processing. Another is the use of EMG (electromyography). Manabe et al. [2] recently proposed using the power of EMG signals for speech recognition and successfully recognized Japanese five vowels. Furthermore, Chan et al. [3] have reported combining EMG and audio signal for speech recognition. Unfortunately, the improvement in ASR performance made possible by using the EMG signal is very limited due to the additive and convolutional noise. It is necessary to apply signal processing techniques to realize the full benefit of EMG. Furthermore, an important factor, the weights assigned to the audio and EMG signals were not investigated in [3].

This paper investigates two robust methods (Cepstral Mean Normalization and Spectral Subtraction) for EMG signal processing; we compare their performance for EMG-based speech recognition. They are applied to an audio-EMG multi-modal speech recognition system. This paper also investigates the importance of stream weighting in audio-EMG-based multi-modal speech recognition. We first explain EMG signal processing, and then report some experiments of audio-EMG-based multi-modal speech recognition at various noise and SNR condition. The paper concludes with a general discussion and issues related to future research.

2. EMG Signal Processing
2.1 EMG Signal
Because of the irregular nature of the myoelectric signal output by most muscles, it is appropriate to process an EMG signal as a stochastic process [4]. One can thus consider the temporal signature of the raw EMG signal as a time series generated by a stochastic process, implying that the techniques of time-series analysis may be used for EMG.

Though an EMG signal is neither linear nor stationary, second moment stationarity holds over short time intervals [5]. Due to the stochastic nature of raw EMG signals, various stochastic analysis methods have been applied for EMG-based pattern recognition. These methods assumed an autoregressive method for EMG signal processing and a fourth-order model has been widely adopted in various investigations.
2.2 Normalization for Cepstral Feature

Cepstral analysis techniques have been applied to EMG signals for classifying the patterns of movement [3,5]. Since the transfer function of the recording system introduces a frequency-dependent multiplicative factor in the overall transfer function, the net result, in terms of the cepstral, is an additive frequency-dependent factor which can be eliminated by averaging along the time axis. Let \( C_n \) be the cepstral coefficient for the \( n \)th time frame in an utterance having \( N \) time frames. Vector \( C_n \) can be represented as the sum of two factors, the first one representing the average values of total frames and second representing the variation of the vector from its average value. Thus, we calculate the average value of the cepstral coefficient and the normalized cepstral coefficient as follows:

\[
\overline{C}_n = \frac{1}{N} \sum_{k=1}^{N} C_k \\
\hat{C}_n = C_n - \overline{C}_n
\]

2.3 Spectral Subtraction for EMG Signal

Spectral subtraction is a noise suppression technique used to reduce the effect of added noise in speech [6,7]. We apply this technique to suppress the added noise in the EMG signal. The noise of an EMG signal is estimated during pauses and the short-time power spectrum \( |X(k)| \) of EMG signal can be obtained by subtracting its noise estimate \( |Y(k)| \) off from the corrupted speech \( |k| \):

\[
|X(k)|^2 = |Y(k)|^2 - |B(k)|^2
\]

3. Multi-modal Speech Recognition using Speech-EMG Signal

3.1 EMG During Utterance

Fig. 1 shows a speech signal and the EMG signal from under the jaw of a subject while a speaker uttered “nana” (“seven” in Japanese). It is clearly shown that the EMG signal can be captured when opening the mouth for uttering something. There is a delay between the EMG signal and the speech signal because EMG is preceded by muscle movement.

3.2 Feature Extraction

In the audio-EMG-based speech recognition system considered here, audio signals are sampled at 16kHz with 16bit resolution. The frame length of audio signal is 25ms and the frame interval is 10ms. Each frame is converted into 25 acoustic parameters: 12 MFCCs (Mel Frequency Cepstral Coefficient, their derivatives, and normalized log-energy).

EMG signals are sampled at 2kHz with 16bit resolution. The frame length of EMG signal is 25ms and the frame interval is 10ms. Normalized log-energy and cepstral features (or spectral feature) are extracted.

3.3 Recognition System

Figure 2 shows the system. The audio and EMG parameters are merged into a single vector; the EMG features are combined with acoustic features in each frame. A set of Hidden Markov Models (HMMs) is used as a model in our system. The HMM has 4 mixtures. The observation probability \( b_j(O_{AE}) \) of generating audio-EMG feature \( O_{AE} \) is given by the following equation:

\[
b_j(O_{AE}) = b_{Aj}(O_A)^{\lambda_A} \times b_{Ej}(O_E)^{\lambda_E}
\]

where \( b_{Aj}(O_A) \) and \( b_{Ej}(O_E) \) are the probabilities of generating acoustic vector \( O_A \) and EMG vector \( O_E \) in state \( j \), respectively, and \( \lambda_A \) and \( \lambda_E \) are weighting factors for the audio and EMG streams, respectively. By properly controlling these weighting factors according to the noise condition, improvements in the recognition accuracy compared to audio-only ASR are expected.

4. Experiments

4.1 Task

The task of the system is the recognition of 13 isolated Japanese digits. We record the speech and a single EMG signal as simultaneously uttered by three speakers. Each subject performed 10 trials, in which each digit was uttered 10 times. We captured EMG from under the jaw as this area
overlays a key muscle. EMG was captured by pressing electrodes mounted on the subject's fingers to his face.

4.2 Spectral Subtraction for EMG only ASR

The first experiment used 12 Fourier transform-based filter-bank [8] spectral features for EMG-based speech recognition and employed the leave-one-out technique: data from one set was used for testing while data from the other sets were used for training. This process was rotated to cover all possible combinations.

Table 1 shows the accuracy for the “Baseline” that is the existing method using spectral feature and proposed “SS” method that performs spectral subtraction for the spectral domain. The noise of an EMG signal is estimated using the first 5 frames of EMG signal which are assumed to be silence. “SS” had better performance than “Baseline” for all test speakers. A 4.2% improvement in word accuracy was achieved by the spectral subtraction method. It confirms the effectiveness of subtracting noise of EMG which can be seen in EMG signal of Fig1.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Baseline using spectral feature</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>63.7</td>
<td>68.6</td>
</tr>
<tr>
<td>2</td>
<td>65.4</td>
<td>69.2</td>
</tr>
<tr>
<td>3</td>
<td>68.2</td>
<td>72.1</td>
</tr>
</tbody>
</table>

Table 1: Comparison of “Baseline” using spectral feature and proposed “SS” spectral subtraction method for EMG

4.3 Cepstral Normalization for EMG only ASR

The second experiment used 12 MFCCs for EMG-based speech recognition. Table 2 shows the result for three speakers. “Baseline using cepstral feature” is the existing method using MFCC while “CMS” indicates the result gained with proposed method. The table shows that the “CMS” method improved the recognition accuracy for all speakers. A 5.6% improvement in word accuracy was achieved by using the normalization technique.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Baseline using cepstral feature</th>
<th>CMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68.5</td>
<td>73.3</td>
</tr>
<tr>
<td>2</td>
<td>62.3</td>
<td>67.3</td>
</tr>
<tr>
<td>3</td>
<td>67.2</td>
<td>74.3</td>
</tr>
</tbody>
</table>

Table 2: Comparison of “Baseline” using cepstral feature and proposed cepstral mean subtraction method for EMG

4.4 Audio-EMG-based Speech Recognition

Audio-EMG-based multimodal experiments were conducted. Training data were clean audio data; audio data for testing were contaminated to yield various SNR levels. In these experiments, the normalized cepstral feature (CMS) of EMG was used. Figures 3 and 4 show the digit accuracy (ACC) for one speaker (speaker 1) in the station and hall noise condition (3 SNR levels). The horizontal axis shows the audio stream weight. Squares and triangles indicate audio-only and EMG-only speech recognition results, respectively.

Recognition results of the multi-modal method are shown by the star. It shows that its performance varies with audio stream weight and the multi-modal method improves the recognition accuracy at the best audio stream weight in all conditions. In particular, the addition of the EMG signal is effective in the low SNR condition. The best audio stream weight falls with the SNR value. The improved recognition result confirms the effectiveness of audio-EMG-based multi-modal speech recognition and also indicates the importance of audio stream weighting.

Fig 3. Result of multi-modal recognition (station noise-added speech)
The same experiments were performed with two other speakers. Tables 3 and 4 show the best multi-modal recognition results averaged over the three speakers at different SNR levels. These results indicate the effectiveness of audio-EMG-based speech recognition.

5. Conclusions

This paper has introduced robust methods (SS and CMS) for EMG signal processing in the framework of audio-EMG-based speech recognition. We also investigated the importance of stream weighting in audio-EMG-based multi-modal speech recognition. Experimental results show the effectiveness of proposed methods at various SNR and noise conditions.

Future research includes increasing the volume of test data to cover more conditions and creating the best stream weight decision method.

6. Acknowledgements

The authors would like to thank Prof. S. Furui of the Tokyo Institute of Technology for his kind advice.

References