Proceedings of Meetings on Acoustics

Volume 19, 2013

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 1aSCb: Digital Speech Processing (Poster Session)

1aSCb8. Detection for Lombard speech with second-order mel-frequency cepstral coefficient and spectral envelope in beginning of talking-speech
Takayuki Furoh*, Takahiro Fukumori, Masato Nakayama and Takanobu Nishiura

*Corresponding author's address: Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu, 525-8577, Shiga, Japan, is0038sv@ed.ritsumei.ac.jp

In noisy environments, the recorded speech is distorted by the additional noise and the Lombard effect. Thus, the automatic speech recognition (ASR) performance is degraded in noisy environments. To solve this problem, noise reduction methods have been proposed as the conventional study. However, in the conventional study, the improvement of ASR performance for the Lombard effect was not discussed well enough. In the present paper, we focus on the robustly detection for Lombard effect speech (Lombard speech). This is because the ASR system can employ a suitable acoustic model by detecting the Lombard speech. We previously proposed the detection for Lombard speech based on second-order mel-frequency cepstral coefficient (2nd-order MFCC) and fundamental frequency (f0). The previously proposed method however requires longer utterances to detect Lombard speech. We therefore newly propose the detection method for Lombard speech with 2nd-order MFCC and spectral envelope in beginning of talking-speech. To detect the Lombard speech at a short time, the proposed method employs variable weights corresponding to elapsed time for 2nd-order MFCC and spectral envelope. As a result of evaluation experiments, we confirmed that the detection time was reduced from the conventional method.

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

In recent years, hands-free devices have been developed [1] with rapid improvement of speech recognition techniques. The automatic speech recognition (ASR) performance in hands-free devices is, however, degraded due to noises in actual environment. Various techniques have been proposed [2, 3] for improving the ASR performance in actual environments. However, the ASR performance is also degraded with not only noises but also Lombard effects in actual environment. Lombard effects are one of speech distortion and strongly occur in case that a speaker gets no auditory feedback in noisy environments. Previous researches [4] have been reported that fundamental frequency (F0) and energy of Lombard speech are higher than ones of clean speech. Robust ASR system under noisy conditions is developed by reducing not only noises but also Lombard effects. In actual environments, detection of Lombard speech is necessary to improve the ASR performance for this speech. We previously proposed the detection method for Lombard speech with a second-order Mel-Frequency Cepstral Coefficient (2nd-order MFCC) and fundamental frequency (F0). The method, however, has a problem that it is unsuitable to process on real-time since it requires longer utterances. In this paper, we analyze the features to accurately detect Lombard speech with the beginning of talking-speech that Lombard effects dominantly appear. As a result of analysis, a 2nd-order MFCC and spectral envelope in beginning of talking-speech were effective features to detect Lombard speech. In addition, we accurately detected Lombard speech with a 2nd-order MFCC and spectral envelope which are weighted on the basis of utterance time.

FEATURES OF LOMBARD SPEECH

Lombard effects are one of speech distortion and strongly occur in case that a speaker gets no auditory feedback in noisy environments. In previous research [5], F0 and energy of Lombard speech are higher than ones of clean speech. In this section, we analyze a few features (F0, power spectrum and MFCC) of clean and Lombard speeches under the conditions listed in Table 1.

<table>
<thead>
<tr>
<th>TABLE 1. Analysis conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker</td>
</tr>
<tr>
<td>Utterance</td>
</tr>
<tr>
<td>Sampling rate</td>
</tr>
<tr>
<td>Quantization bits</td>
</tr>
<tr>
<td>Frame length</td>
</tr>
<tr>
<td>Shift length</td>
</tr>
</tbody>
</table>

FIGURE 1. F0 track of clean and Lombard speeches
Fundamental Frequency

F0 [6] is defined as the lowest frequency of periodic waveform which is produced with vibration of the vocal cords. F0 expresses the pitch of talking-speech for a speaker, and a F0 track designs an accent and intonation of speech. Figure 1 shows a F0 track of clean and Lombard speeches. From Fig. 1, F0 of Lombard speech was higher than one of clean speech. In addition, F0 widely varies depending on utterance. It is, thus, difficult to accurately detect Lombard speech with only F0.

Power Spectrum

Figure 2 shows power spectrum and spectral envelope in beginning of clean and Lombard speeches. From Fig. 2, the energy of Lombard speech was higher than one of clean speech. Furthermore, spectral envelope of Lombard speech was flatter than one of clean speech. The energy of speech is unsuitable for detection of Lombard speech since it is dependent on volume of utterance and recording conditions. There is, however, possibility that a spectral envelope is suitable to detect Lombard speech.

Mel-Frequency Cepstral Coefficients

Mel-Frequency Cepstral Coefficients (MFCCs) are cepstrums on the basis of Mel Filterbank [7]. MFCCs are used in speech analysis based on the auditory characteristics of human. Previous research [8] has shown that it is effective for accurate detection of Lombard speech with a 2nd-order MFCC. Figure 3 shows a 2nd-order MFCC track of clean and Lombard speeches. From Fig. 3, a 2nd-order MFCC of Lombard speech is globally lower than one of clean speech. This result shows that it is possible to detect Lombard speech by taking an average of the 2nd-order MFCC in shorter utterance.

FIGURE 2. Power spectrum and spectral envelope of clean and Lombard speeches
CONVENTIONAL METHOD FOR DETECTION OF LOMBARD SPEECH

We previously proposed the detection method for Lombard speech with a 2nd-order MFCC and F0 [5]. In previous method, Lombard speech was detected by using a clustering method on the basis of the difference of the 2nd-order MFCC and F0 in clean and Lombard speeches. A 2nd-order MFCC is possible to represent the amount of increase in higher frequency with Lombard effects since it expresses the power envelope in mel-frequency domain [9]. F0 is defined as the lowest frequency of periodic waveform which is produced with vibration of the vocal cords. In previous section, F0 of Lombard speech was higher than one of clean speech. The previous method, however, has a problem that it is unsuitable to process on real-time since it requires longer utterances.

DETECTION FOR LOMBARD SPEECH WITH 2ND-ORDER MFCC AND SPECTRAL ENVELOPE IN BEGINNING OF TALKING-SPEECH

The previous method for the detection of Lombard speech has a problem that it is unsuitable to detect shorter Lombard speech. To overcome this problem, we focus on the spectral envelope which is possible to become an effective candidate for detection of Lombard speech with shorter utterance. This is because spectral envelope of Lombard speech was flatter than one of clean speech. In the proposed method, Lombard speech was detected by following steps, as illustrated in Fig. 4.

Step 1. Prepare clean and Lombard speeches
First, clean and Lombard speeches are prepared as the training data to design the detection criteria for Lombard speech. Lombard speech is recorded under conditions that a speaker mounts headphone which emits emitted louder noise because Lombard effects generally occur in case that people speak in the noisy environments.

Step 2. Measure 2nd-order MFCC and spectral envelope
Each average of the 2nd-order MFCC and spectral envelope is calculated from the beginning of talking-speech prepared in Step 1.

Step 3. Calculate feature vectors for Lombard speech
To detect Lombard speech, we calculate feature vectors on the basis of the features measured in Step 2. The feature vectors $u(t; k)$ with speaking style $k$ and utterance time $t$ are derived from Eq. (1).
FIGURE 4. Overview of the proposed method

\[
\mathbf{u}(t;k) = \begin{bmatrix}
w_{\text{MFCC}}(t) \cdot \frac{1}{N} \sum_{i=0}^{N} u_{\text{MFCC}}(i;k) \\
w_{\text{SE}}(t) \cdot \frac{1}{N} \sum_{i=0}^{N} u_{\text{SE}}(i;k)
\end{bmatrix},
\]

where \(u_{\text{MFCC}}(i;k)\) and \(u_{\text{SE}}(i;k)\) represent a 2nd-order MFCC and spectral envelope of training sample \(i\) with speaking style \(k\). \(t\) represents a utterance time, and \(N\) represents number of training samples. In proposed method, the feature vectors are calculated with a 2nd-order MFCC and spectral envelope which are weighted on the basis of utterance time because Lombard effects to a 2nd-order MFCC and spectral envelope greatly depends on utterance time. The weighted coefficients to these features are determined from Eqs. (2) and (3).

\[
w_{\text{MFCC}}(t) = \alpha + \beta,
\]

\[
w_{\text{SE}}(t) = 1 - w_{\text{MFCC}}(t),
\]

where \(w_{\text{MFCC}}(t)\) and \(w_{\text{SE}}(t)\) represent the weighted coefficients to a 2nd-order MFCC and spectral envelope of Lombard speech with utterance time \(t\). Coefficients \(\alpha\) and \(\beta\) are calculated by performing regression analysis.

**Step 4. Detect Lombard speech with the feature vectors**

Finally, the feature vectors designed in Step 3 are used to detect Lombard speech. In this step, clean and Lombard speeches are measured as testing samples. After that, a 2nd-order MFCC and spectral envelope are calculated in the Step 2. The feature vectors \(\mathbf{x}\) of testing samples are derived from Eq. (4).

\[
\mathbf{x} = \begin{bmatrix} w_{\text{MFCC}}(t) \cdot x_{\text{MFCC}} \\
w_{\text{SE}}(t) \cdot x_{\text{SE}} \end{bmatrix},
\]

where \(x_{\text{MFCC}}\) and \(x_{\text{SE}}\) represent a 2nd-order MFCC and spectral envelope for testing samples. Lombard speech is, then, detected on the basis of trained and test features from Eq. (5).

\[
D = \arg \min_{k} d(\mathbf{x}, \mathbf{u}(k)),
\]

where \(d(\mathbf{x}, \mathbf{u}(k))\) represents Euclidean distance between \(\mathbf{x}\) and \(\mathbf{u}(k)\). A detection result \(D\) is the cluster \(k\) which takes the minimum \(d(\mathbf{x}, \mathbf{u}(k))\).
EVALUATION EXPERIMENT

In this research, objective experiments were performed to evaluate whether the proposed method can accurately detect Lombard speech. In preliminary experiment, we determined weighted coefficients for a 2nd-order MFCC and spectral envelope in Eqs. (2) and (3).

Determination of the Weighted Coefficients

In this section, we determine weighted coefficients for a 2nd-order MFCC and spectral envelope by utterance time. In this research, weighted coefficients are calculated by taking the highest F-measure in each utterance time. F-measure represents the harmonic average of precision and recall as detection of Lombard speech. F-measure is calculated with Eq. (6).

\[
F\text{-}measure = \frac{2 \cdot p_{\text{precision}} \cdot p_{\text{recall}}}{p_{\text{precision}} + p_{\text{recall}}},
\]

where \(p_{\text{precision}}\) and \(p_{\text{recall}}\) represent the precision and recall as detection of Lombard speech. Figure 5 shows an optimum weighted coefficient for a 2nd-order MFCC by utterance time. From Fig. 5, a weighted coefficient for a 2nd-order MFCC tends to decrease in shorter utterance. On the basis of this result, a regression analysis was performed to formulate the weighted coefficient using a 2nd-order MFCC and utterance time. Coefficients \(\alpha = 5.0 \times 10^{-2}\) and \(\beta = 5.3 \times 10\) in Eq. (2) were obtained in regression analysis. In the next experiment, we conducted detection of Lombard speech with these weighted coefficients.

Detection of Lombard Speech with 2nd-order MFCC and Spectral Envelope

In this section, we conducted the detection experiment of Lombard speech with a 2nd-order MFCC and spectral envelope. Experimental conditions are shown in Table 2. In this experiment, we evaluated the detection performance for Lombard speeches with seven kinds of utterance time. We recorded 400 speech samples that 20 females uttered 10 phonetically-balanced sentences with two kinds of speaking-styles. The speech recordings were conducted with 16 kHz sampling and 16 bits quantization. F-measure was used to evaluate the detection performance.

Experimental Results and Discussions

Figure 6 shows the result with F-measure of clean and Lombard speeches with seven kinds of the utterance time. From Fig. 6, the proposed method had higher F-measure than conventional method under all conditions. As a result, we could confirm that proposed method provides much better detection performance than conventional method. In addition, F-measure of more than 90 % was achieved by using proposed method under the condition with utterance time of more than 500 ms. It means that only 500 ms from the beginning of taking-speech is possible to accurately detect Lombard speech.

![Optimum weighted coefficient of 2nd-order MFCC](image-url)
TABLE 2. Experimental conditions

<table>
<thead>
<tr>
<th>Speaking style</th>
<th>Clean and Lombard</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Speaker</strong></td>
<td>20 females</td>
</tr>
<tr>
<td><strong>Utterance</strong></td>
<td>Phonetically-balanced</td>
</tr>
<tr>
<td></td>
<td>sentence 10 [sentence]</td>
</tr>
<tr>
<td><strong>Utterance time</strong></td>
<td>100, 300, 500, 700, 900, 1,000 and 5,000 [ms]</td>
</tr>
<tr>
<td><strong>Features</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Conventional method</strong></td>
<td>2nd-order MFCC and F0</td>
</tr>
<tr>
<td><strong>Proposed method</strong></td>
<td>2nd-order MFCC and Spectral envelope</td>
</tr>
<tr>
<td><strong>Evaluation index</strong></td>
<td>F-measure</td>
</tr>
<tr>
<td><strong>Sampling rate</strong></td>
<td>16 [kHz]</td>
</tr>
<tr>
<td><strong>Quantization bits</strong></td>
<td>16 [bit]</td>
</tr>
<tr>
<td><strong>Frame length</strong></td>
<td>25 [ms]</td>
</tr>
<tr>
<td><strong>Shift length</strong></td>
<td>10 [ms]</td>
</tr>
</tbody>
</table>

FIGURE 6. Experimental results for F-measure

CONCLUSIONS

In this research, we proposed the detection method for Lombard speech with a 2nd-order MFCC and spectral envelope which are weighted on the basis of utterance time. As a result of experiment, proposed method provides much better detection performance than conventional method. Moreover, we could confirm that it is possible to accurately detect Lombard speech with only 500 ms in beginning of talking-speech. In the future work, we intend to analyze other features such as dynamic fundamental frequency for improvement of detection performance for Lombard speech.

ACKNOWLEDGMENTS

This work was partly supported by Grants-in-Aid for Scientific Research funded by MEXT, Japan.
REFERENCES