A study of degraded-speech identification based on spectral centroid

Takayuki FUROH[1]; Takahiro FUKUMORI[1]; Masato NAKAYAMA[2]; Takanobu NISHIURA[2]

[1] Graduate School of Information Science and Engineering, Ritsumeikan University, Japan.

[2] College of Information Science and Engineering, Ritsumeikan University, Japan.

ABSTRACT

Hands-free speech interfaces are developed with the progress of speech recognition techniques. In the conventional automatic speech recognition (ASR) system, normal speech can be recognized with high accuracy. However, the ASR performance is degraded because the human speech is distorted by noise and speaking styles in noisy environments and crisis situations. This problem can be solved by applying suitable acoustic model corresponding to degraded speech. Therefore, we had previously proposed the identification for degraded-speech based on the fundamental frequency (F0), 2nd-order mel-frequency cepstral coefficient (MFCC) and rahmonic. The conventional method can identify normal speech, Lombard speech and shout speech, but it has an insufficient identification performance. This is because the conventional method utilizes acoustic features which are similar in Lombard speech and shout speech. In this paper, we therefore propose degraded-speech identification method based on the spectral centroid, F0, 2nd-order MFCC and rahmonic. The spectral centroid can represent the formant shift to the high-frequency spectrum. In the proposed method, the spectral centroid is utilized for identifying Lombard speech and shout speech. As a result of objective evaluation experiments, we confirmed the effectiveness of the proposed method towards the identification for degraded-speech.

Keywords: Spectral centroid, Degraded-speech, Lombard speech, Shout speech
I-INCE Classification of Subjects Number(s): 01.4

1. INTRODUCTION

Robust speech recognition has become very important, because it is essential for usable speech interfaces. In hands-free speech interfaces, automatic speech recognition (ASR) performance is, however, degraded because the human speech is distorted by noise and speaking styles in noisy environments and crisis situations (1). Various techniques for preventing this degradation have been proposed, such as the spectral subtraction method (2) for noisy environments and suitable acoustic models adaptation for speaking styles (3). Whereas, they have difficulty preventing degradation under unknown speaking styles. This problem can be solved by identifying speaking styles before ASR processing. In previous, the identification method for normal speech and Lombard speech has been proposed, which utilizes the fundamental frequency (F0) and 2nd-order mel-frequency cepstral coefficient (MFCC) (4). The F0 and 2nd-order MFCC are effective to identify normal speech and Lombard speech because the F0 of Lombard speech increases than normal speech due to increasing the pitch of Lombard speech and the 2nd-order MFCC of Lombard speech decreases than normal speech due to increasing the expiratory volume of Lombard speech. In addition, the identification method for normal speech and shout speech has been proposed, which utilizes MFCCs and rahmonic (5). MFCCs and the rahmonic are effective to identify normal speech and shout speech because MFCCs are generally utilized by ASR and the rahmonic of shout speech strongly arises than normal speech due to the vibration of the vocal cords periphery. We thus have previously proposed the identification method for normal speech, Lombard speech and shout speech based on the F0, 2nd-order MFCC and rahmonic for identifying multiple speaking styles. The characteristic of Lombard speech, nevertheless, is similar to that of shout speech because both Lombard speech and shout speech are the degraded-speech. For this reason, it is difficult to identify Lombard speech and shout speech. In this paper, therefore, we focus on the spectral centroid which can represent the formant shift to the high-frequency spectrum. The formant of shout speech is shifted to the high-frequency spectrum than that of normal speech and Lombard speech, since a human jaw is opened widely by shout speech. Therefore, it

[1] {is0038sv,cm013061}@ed.ritsumei.ac.jp
[2] {mnaka@fc,nishiura@is}.ritsumei.ac.jp

is effective to identify Lombard speech and shout speech by utilizing the spectral centroid. To identify normal speech, Lombard speech and shout speech with high accuracy, we first conduct to identify normal speech and degraded-speech (Lombard and shout speech), because the characteristic of Lombard speech is similar to that of shout speech. Second, we conduct to identify Lombard speech and shout speech. We therefore proposed the multi-stage identification method of degraded-speech based on the spectral centroid, F0, 2nd-order MFCC and rahmonic.

2. CONVENTIONAL METHOD

We previously proposed the identification method for normal speech, Lombard speech and shout speech based on the F0, 2nd-order MFCC and rahmonic. First, in the conventional method, normal speech, Lombard speech and shout speech models are trained by Gaussian mixture model (GMM) based on feature vectors from Eqs. (1, 2, 3).

$$g_{NS}(\boldsymbol{y}) = \sum_{m=1}^{M} w_m \Phi(\boldsymbol{y} \mid \boldsymbol{\mu}_{NS,m}, \boldsymbol{\sigma}^2_{NS,m}) \,, \tag{1}$$

$$g_{LS}(\boldsymbol{y}) = \sum_{m=1}^{M} w_m \Phi(\boldsymbol{y} \mid \boldsymbol{\mu}_{LS,m}, \boldsymbol{\sigma}^2_{LS,m}) \,, \tag{2}$$

$$g_{SS}(\boldsymbol{y}) = \sum_{m=1}^{M} w_m \Phi(\boldsymbol{y} \mid \boldsymbol{\mu}_{SS,m}, \boldsymbol{\sigma}^2_{SS,m}) \,, \tag{3}$$

where *NS*, *LS* and *SS* represent speaking styles as normal speech, Lombard speech and shout speech, respectively, $g(\boldsymbol{y})$ represents a probability density distribution with feature vector $\boldsymbol{y}$, $\Phi(\boldsymbol{y} \mid \boldsymbol{\mu}_m, \boldsymbol{\sigma}^2_m)$ represents a normal distribution function, M represents a mixture number, w_m, $\boldsymbol{\mu}_m$ and $\boldsymbol{\sigma}^2_m$ represent weight, mean and variance of regular distribution m, respectively. Feature vectors are normalized and utilized with equal weights. Second, the likelihood P_d of input speech is calculated by normal speech and degraded-speech models from Eq. (4).

$$P_d = \frac{\Phi(\boldsymbol{y} \mid \boldsymbol{\mu}_{d,m}, \boldsymbol{\sigma}^2_{d,m})}{\sum_{m=1}^{M} w_m \Phi(\boldsymbol{y} \mid \boldsymbol{\mu}_{d,m}, \boldsymbol{\sigma}^2_{d,m})}, \tag{4}$$

where d represents *NS*, *LS* and *SS*. Finally, the speaking style of input speech is identified according to maximum likelihood from Eq. (5).

$$\hat{d} = \underset{d}{\operatorname{argmax}}(P_d), \tag{5}$$

where $\hat{d}$ represents an identification result.

F0 is defined as the lowest frequency of the periodic waveform which is produced by the vibration of the vocal cords. The 2nd-order MFCC represents the increase power in the high-frequency spectrum with increasing the expiratory volume. The rahmonic represents the vibration of the vocal cords periphery. However, these acoustic feature distributions of Lombard speech are similar to that of shout speech. Therefore, it is difficult to identify Lombard speech and shout speech with these conventional acoustic features.

3. PROPOSED METHOD

Figure 1 shows the process of the proposed method. First, in the proposed method, we utilize GMM to train normal speech and degraded-speech models for the first stage. Second, we utilize GMM to train Lombard speech and shout speech models for the second stage. Third, input speech is identified by normal speech and degraded-speech models in the first stage. Finally, degraded-speech which is identified correctly in the first stage is identified by Lombard speech and shout speech models in the second stage.

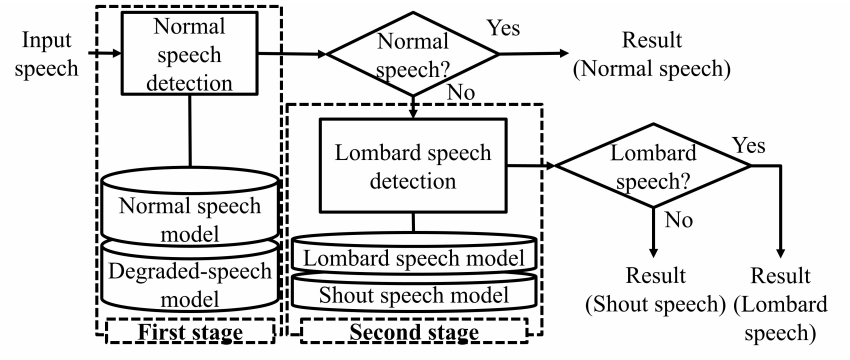

Figure 1 – Multi-stage degraded-speech identification.

3.1 First stage

The proposed method identifies normal speech and degraded-speech (Lombard speech and shout speech) in the first stage. First, normal speech and degraded-speech models are trained by GMM for the first stage based on feature vectors from Eqs. (1, 6).

$$g_{DS}(\boldsymbol{y}) = \sum_{m=1}^{M} w_m \Phi(\boldsymbol{y} \mid \boldsymbol{\mu}_{DS,m}, \boldsymbol{\sigma}^2_{DS,m}) \,, \tag{6}$$

where DS represents a speaking style as degraded-speech. Second, the likelihood of input speech is calculated by normal speech and degraded-speech models from Eq. (4). Finally, the speaking style of input speech is identified according to maximum likelihood from Eq. (5).

It is possible to identify normal speech and degraded-speech with the F0, 2nd-order MFCC and rahmonic, because these acoustic features are utilized in the conventional method. These acoustic features are calculated as same as the conventional method (4, 5). In addition, the formant of shout speech is shifted to the high-frequency spectrum than that of normal speech and Lombard speech. We therefore consider that the spectral centroid is effective to identify normal speech and degraded-speech.

3.1.1 Spectral centroid

The spectral centroid is calculated from Eq. (7).

$$C_t = \frac{\sum_{k=K_s}^{K_e} M_t[k] \times k}{\sum_{k=K_s}^{K_e} M_t[k]}, \tag{7}$$

where C_t represents the spectral centroid of frame number t, $M_t[k]$ represents the power spectrum with frequency bin number k, K_s represents the lower limitation of frequency bandwidth bin number and K_e represents higher one. In this paper, we calculate the spectral centroid by utilizing 0-8 [kHz], because its frequency bandwidth represents characteristics of speech (6).

3.2 Second stage

The proposed method, in the second stage, identifies Lombard speech and shout speech from degraded-speech which is identified correctly in the first stage. First, Lombard speech and shout speech models are trained by GMM for the second stage based on feature vectors in the same way as the first stage from Eqs. (2, 3). Second, the likelihood P_d of input speech is calculated by normal speech and degraded-speech models from Eq. (4). Finally, degraded-speech which is identified correctly in the first stage is identified according to maximum likelihood from Eq. (5).

It is difficult to identify Lombard speech and shout speech by utilizing the 2nd-order MFCC and rahmonic, because 2nd-order MFCC and rahmonic distributions of Lombard speech are similar to that of shout speech. On the other hand, the spectral centroid can represent that the formant is shifted to high-frequency spectrum by shout speech. Moreover, F0 of shout speech increases than that of Lombard speech. We thus consider that the spectral centroid and F0 are effective to identify Lombard speech and shout speech.

Table 1 – Experimental conditions.

Speaking style		Normal, Lombard and shout
Speaker		3 female and 6 male speakers
Training samples		900 [sample]
Testing samples(open)		450 [sample]
Feature vectors		Spectral centroid, F0, 2nd-order MFCC and rahmonic
Model numbers	Conventional method	3 (Normal , Lombard and shout)
	Proposed method	First stage : 2 (Normal and degraded) Second stage : 2 (Lombard and shout)
Mixture number		32

Table 2 – Combinations of feature vectors for constructing acoustic models.

1. Spectral centroid	8. F0 + 2nd-order MFCC
2. F0	9. F0 + rahmonic
3. 2nd-order MFCC	10. 2nd-order MFCC + rahmonic
4. Rahmonic	11. Spectral centroid + F0 + 2nd-order MFCC
5. Spectral centroid + F0	12. Spectral centroid + F0 + rahmonic
6. Spectral centroid + 2nd-order MFCC	13. F0 + 2nd-order MFCC + rahmonic
7. Spectral centroid + rahmonic	14. Spectral centroid + F0 + 2nd-order MFCC + rahmonic

4. EVALUATION EXPERIMENT

In this paper, objective experiments were performed to evaluate whether the proposed method could accurately identify normal speech, Lombard speech, and shout speech.

4.1 Experimental condition

Table 1 shows experimental conditions. We conducted identification experiments of degraded-speech and evaluated identification rates of combinations of feature vectors in the first and second stages. Table 2 shows combinations of feature vectors (spectral centroid, F0, 2nd-order MFCC and rahmonic) for constructing acoustic models. Acoustic models are trained with normal speech, Lombard speech, and shout speech of 9 speakers (each 300 samples). We evaluated the identification performance with normal speech, Lombard speech and shout speech (each 150 samples) in the first stage. In the second stage, we evaluated the identification performance with degraded-speech which is identified correctly in the first stage. Identification methods for normal speech, Lombard speech and shout speech are defined as follows:

- Conventional method without spectral centroid : Acoustic features are the F0, 2nd-order MFCC and rahmonic.
- Conventional method with spectral centroid : Acoustic features are the spectral centroid, F0, 2nd-order MFCC and rahmonic.
- Proposed method without spectral centroid : Acoustic features are the F0, 2nd-order MFCC and rahmonic in the first stage, and acoustic features are the F0 and 2nd-order MFCC in the second stage.
- Proposed method with spectral centroid : Acoustic features are the spectral centroid, F0, 2nd-order MFCC and rahmonic in the first stage, and acoustic features are the spectral centroid and F0 in the second stage.

4.2 Experimental result

Figure 2 (a) shows identification results of normal speech and degraded-speech with combinations of 4 acoustic features in the first stage. The vertical axis shows the identification rate, and the horizontal axis shows the number of combinations of acoustic features in Tbl. 2. From Fig. 2 (a), utilizing the spectral centroid, F0, 2nd-order MFCC, and rahmonic obtained the identification rate (98 %) higher than others. As a result of the

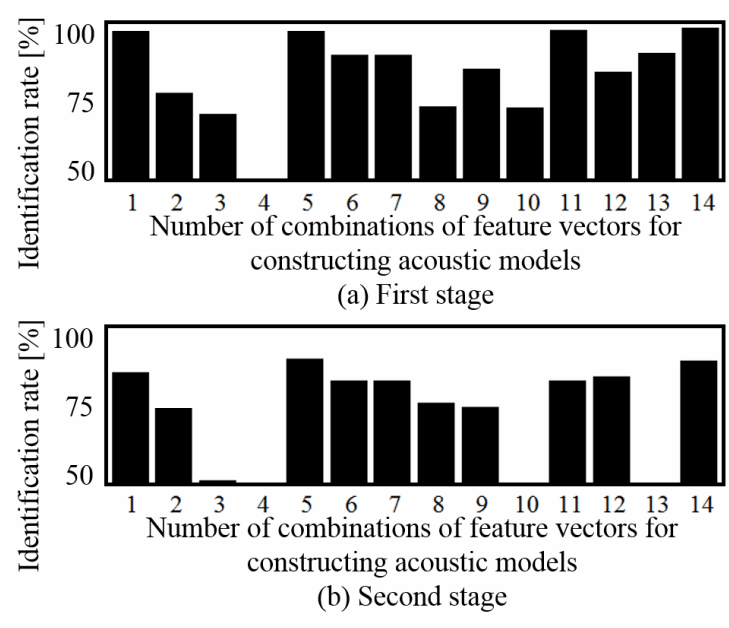

Figure 2 – Experimental results for combinations of four feature vectors in the proposed method.

Table 3 – Experimental results for degraded-speech identification.

Conventional method without spectral centroid		**Identification rate**
Total		72 % (324/450)
Conventional method with spectral centroid		**Identification rate**
Total		85 % (382/450)
Proposed method without spectral centroid		**Identification rate**
First stage	Normal	90 % (135/150)
	Degraded	84 % (252/300)
Second stage	Lombard	78 % (86/111)
	Shout	82 % (115/141)
Total		74 % (336/450)
Proposed method with spectral centroid		**Identification rate**
First stage	Normal	100 % (150/150)
	Degraded	96 % (288/300)
Second stage	Lombard	97 % (135/140)
	Shout	82 % (121/148)
Total		90 % (406/450)

experiment, not only conventional acoustic features which are the F0, 2nd-order MFCC and rahmonic but also the spectral centroid are effective to identify normal speech and degraded-speech.

Figure 2 (b) shows identification results of Lombard speech and shout speech with combinations of 4 acoustic features in the second stage. The vertical axis shows the identification rate, and the horizontal axis shows the number of combinations of acoustic features in Tbl. 2. From Fig. 2 (b), utilizing the spectral centroid, and F0 obtained the identification rate (90 %) higher than others. As a result of the experiment, the spectral centroid is effective to identify Lombard speech and shout speech because the spectral centroid of shout speech increases compared with Lombard speech due to the formant shift to high-frequency spectrum. Furthermore, the F0 is effective to identify Lombard speech and shout speech because the F0 of shout speech increases compared with Lombard speech.

Table 3 shows identification results of the conventional method and proposed method. The number in parentheses in Tbl. 3 shows "Number of correctly identified samples / Number of evaluation samples". As a result of the experiment, we could confirm that the proposed method which utilized the spectral centroid

improved the identification rate by 18 % compared with the conventional method.

5. CONCLUSIONS

In this paper, we proposed the identification method of degraded-speech based on the spectral centroid, F0, 2nd-order MFCC, and rahmonic. As a result of the experiment, proposed method increased 18 % identification performance than the conventional method. In the future work, we intend to analyze other acoustic features such as the dynamic frequency spectrum for improving the identification performance of degraded-speech.

ACKNOWLEDGEMENTS

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

REFERENCES

1. Jean-Claude Junqua, "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex," Speech Communication, vol. 20, pp.13-22, 1996.
2. Steven Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. ASSP, vol. 27, pp. 113-120, 1979.
3. Yasuhiro Shimizu, Shuji Kajita, Kazuya Takeda, and Fumitada Itakura, "Robust speech recognition based on space diversity taking room acoustics into account," Institute of Electronics, Information, and Communication Engineers, vol. J83-DII, pp. 2448-2456, 2000.
4. Takayuki Furoh, Takahiro Fukumori, Masato Nakayama, and Takanobu Nishiura, "Detection for Lombard speech with second-order mel-frequency cepstral coefficient and spectral envelope in beginning of talking-speech," ICA 2013, PaperID:1aSCb8, 2013.
5. Naoto Kakino, Takahiro Fukumori, Masato Nakayama, and Takanobu Nishiura, "Experimental study of shout detection with the Rahmonic structure," ICA 2013, PaperID:1aSCb6, 2013.
6. Hisao Kuwabara, and Tohru Takagi, "Quality Control of Speech by Modifying Formant Frequencies and Bandwidth," 11th Inter. Congress of Phonetic Sciences, pp.281 -284, 1987.