Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

The Lombard Speech Recognition Based on the Voice Conversion Towards Neutral Speech

Yuji Uemura(1), Masanori Morise(2), Takanobu Nishiura(2)

(1) Graduate School of Science and Engineering, Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu, Shiga, 525-8577 Japan
(2) College of Information Science and Engineering, Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu, Shiga, 525-8577 Japan

PACS: 43.72.NE

ABSTRACT

The automatic speech recognition (ASR) under noisy environments is focused as one of the challenging topics. Especially, the observed speech under noisy environments much distorts compared with neutral observed speech under quiet one. This distortion is called Lombard effects, and ASR performance degrades by them. They should strongly occur subject to no auditory feedback for speaker. In conventional research, their features tend to be ascent of power, ascent of fundamental frequency (F0), flat of spectral envelope and higher-frequency shift of the first formant frequency (F1) and the second formant frequency (F2). Therefore, the ASR performance without any especially operations degrades by affecting such features. To overcome this problem, they had proposed to reconstruct the acoustic model with many Lombard speech signals, to adopt acoustic model with some Lombard ones, and so on. However, it is indispensable for above both approaches to supply the sufficient amount of Lombard speech signals in advance. Thus in this paper, we propose the new approach based on the voice conversion from Lombard speech to neutral speech. This approach has the advantages which are without not only the operation for acoustic model but also the sufficient amount of Lombard speech signals in advance. We try to improve the Lombard speech recognition performance with the voice conversion technique towards neutral speech. We firstly analyse F0, F1 and F2 features with Lombard speech and neutral one in detail. As a result, we confirmed that F0, F1 and F2 features with Lombard speech are higher frequency than neutral ones. In addition, we also confirmed the standard deviations of the ascending rate for F1 and F2 features are smaller than the one for F0 feature. Therefore, we decided to employ F1 and F2 as the feature for voice conversion, and we finally converted Lombard speech to neutral speech by equalizing the ascending rates for F1 and F2 features. We carried out evaluation experiments. As a result of experiments, we confirmed the ASR performance increases to 10 % for female speakers and 4 % for male one with proposed method. We therefore confirmed that the Lombard speech can be robustly recognized without reconstruct the acoustic model.

INTRODUCTION

In recent years, the automatic speech recognition (ASR) technology has already developed and been used as speech guide system. However, ASR performance under noisy environments degrades compared with ASR performance under quiet one. The voice user interface in speech guide system will be realized, provided that ASR performance under noisy environments was improved. Also, that will improve the barrier-free interface.

ASR performance degrades, if the observed speech includes the background noise. It is difficult to keep highly signal-to-noise ratio (SNR) because microphone position is far from speaker. To overcome this problem, spectral subtraction (SS) was proposed to reduce the background noise from observed speech [1]. For example, ASR performance toward car navigation system [2] is improved by using SS. However, ASR performance is also degraded by not only background noise but also distortion called Lombard effects [3-6]. Lombard effects should occur strongly subject to no auditory feedback for speaker. In conventional research, Lombard effects tend to be ascent of power, ascent of fundamental frequency (F0), varies of spectral tilt and higher-frequency shift of the first formant frequency (F1) and the second formant frequency (F2). The ASR performance is degraded by changing these features. Therefore, ASR performance is also improved by compensating Lombard effects. The methods for improvement of ASR performance had been proposed that reconstructs the acoustic model with many Lombard speech signals and adopts acoustic model with some Lombard ones [5]. However, it is indispensable for two approaches to supply the sufficient amount of Lombard speech signals in advance.

In this paper, we propose the new approach based on the voice conversion from Lombard speech to neutral speech. However, detailed Lombard features haven't been analysed. First of all, we recorded Lombard speech and designed Lombard speech corpus. Then, we analysed differences between Lombard speech features and neutral ones designed corpus in detail. And we try to improve ASR performance by converting Lombard speech to neutral speech based on analysed features.

THE LOMBARD SPEECH

Lombard ASR has been studied about improvement of ASR performance in noisy factory [7]. The study utilized the pattern matching technique as means to improve Lombard ASR performance. As a result, it succeeded to improve Lombard ASR performance and operation performance of system in factory. However, these targets are limited such as speaker dependent and contents dependant. Therefore, in this paper, to improve ASR performance under noisy environments, we designed Lombard speech corpus that phoneme balanced sentences and connected digit speech as multipurpose.

Design of Lombard speech corpus

In this paper, we recorded Lombard speech for Lombard speech corpus in the soundproof room. Figure 1 shows recording environments and Tab. 1 shows recording conditions. The corpus includes 20 female speakers and 20 male speakers. We recorded Lombard speech with close-talking microphone and distant microphone at the same time. We set close-talking microphone nearby mouth and distant microphone at 100 cm far from speaker. Also, we prepared three kinds of background noises. These are car interior noise (CAR.), department store's basement food floor noise (DEP.), and pink noise (PIN.). In addition, we fix presentation levels are 60 dBA, 70 dBA and 75 dBA each noises. Utterance contents are phoneme balanced sentences and connected digit speech.

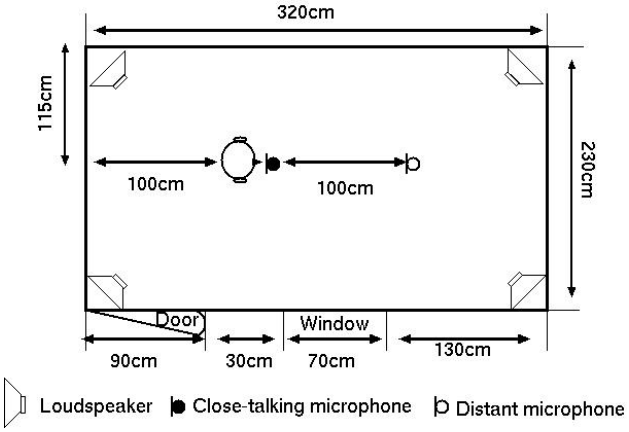

Figure 1. Recording environment

Table 1. Recording conditions

Close-taliking mic.	SONY ECM-360
Distant mic.	SONY ECM-77B
Dummy head	NEUMANN KU100
Headphone	SONY MDR-IF 8000
Sampling rate	48 kHz
Quantization	16 bits

Recorded Lombard speech are as follows.

- Neutral-Clean speech (NC)

NC means neutral-clean speech recorded with close-talking microphone in quiet environment.

- Lombard-Clean speech (LC)

LC means Lombard-clean speech recorded with close-talking microphone in quiet environment. However, speaker received noise with headphone.

- Lombard-Noisy speech (LN)

LN means Lombard-noisy speech recorded with distant microphone in noisy environments.

Analysis of Lombard features

To confirm differences among NC features and LC ones, we analysed NC and LC in the corpus. In conventional research, their features tend to be ascent of power, ascent of F0, varies of spectral tilt and higher-frequency shift of F1 and F2 [3]. In this paper, we specifically geared toward F0, F1 and F2 based on conventional researches.

The fundamental frequency with Lombard speech

We analysed F0 differences between LC and NC. As F0 features depend on gender, we separately calculated F0 for female speakers and male speakers. Figure 2 shows F0 averages and standard deviations. For female speakers with NC, F0 denotes 266 Hz and F0 with LC denotes 302 Hz on average. On the other hand, for male speakers with NC, F0 denotes 153 Hz and F0 with LC denotes 203 Hz on average. We confirmed that F0 with LC increases 15 % for female speakers and 30 % for male speakers compared with F0 with NC.

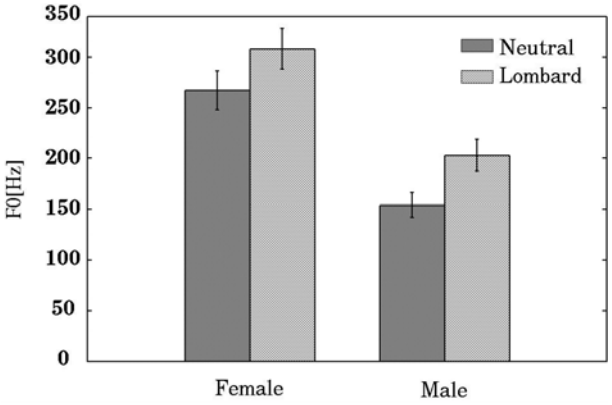

Figure 2. F0 averages and standard deviations

The formant frequency with Lombard speech

We analysed F1 and F2 for Japanese vowels (/a/, /i/, /u/, /e/, /o/) to confirm differences among formant frequency for NC and one for LC. Figure 3 shows F1 and F2 averages and standard deviations for male speakers. We confirmed both F1 and F2 ascent compared with NC. As F1 and F2 depend on gender as well as F0, the ascending rate for female speakers and male speakers are different. Figure 4 shows averages and standard deviations for the F0 and formant frequency ascending rate. From Fig. 4, the F0 ascending rate individually varies. On the other hand, standard deviations of the formant frequency ascending rate are smaller than F0 those.

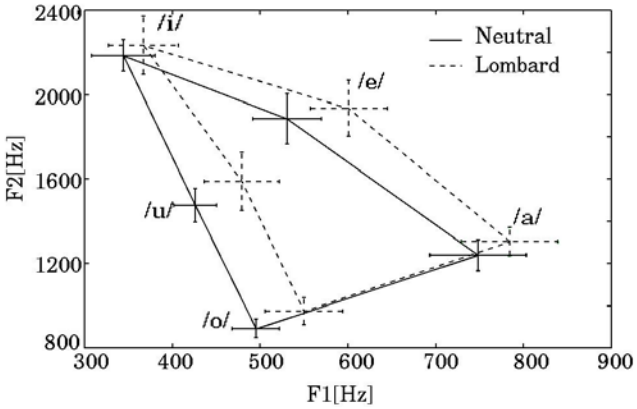

Figure 3. F1 and F2 averages and standard deviations

VOICE CONVERSION BY PICOLA AND RESAMPLING

In general, the methods for improvement of ASR performance have been proposed that to reconstruct the acoustic model with many Lombard speech signals or to adopt acoustic model with some Lombard ones [5]. However, it is indispensable for above approaches to supply the sufficient amount of LC in advance. In this paper, we propose the new approach based on the voice conversion from LC to NC without any acoustic model reconstructions or adaptations.

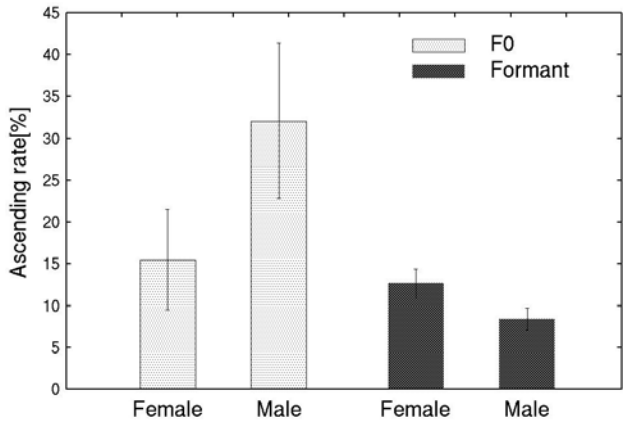

Figure 4. F0 and formant frequency ascending rate

This approach has two advantages which are a few computational costs for acoustic model and preparation without the sufficient amount of Lombard speech signals in advance. We try to improve ASR performance with LC by using the voice converting to NC. Thus, we proposed a method to convert F0 and formant frequency features based on PICOLA (Pointer Interval Controlled OverLap and Add) [8] and resampling. We defined resampling by two processes.

- Input signal $x(n)$ sampled by f_{s_0} of certain sampling rate is converted to analog signal $s_a(t)$ interpolated by sinc function based on Eq. (1) and Eq. (2).
- Analog signal $s_a(t)$ is converted to $y(n)$ sampled f_{s_1} of new sampling rate based on Eq. (3) and Eq. (4).

$$s_a(t) = \sum_{n=-\infty}^{\infty} x(n) \frac{\sin(\pi(t - nT_0))}{\pi(t - nT_0)}, \quad (1)$$

$$T_0 = \frac{1}{f_{s_0}}, \quad (2)$$

$$T_1 = \frac{1}{f_{s_1}}, \quad (3)$$

$$y(n) = s_a(nT_1). \quad (4)$$

$y(n)$ is the output signal processed by the proposed method. These processes enable to control F0, F1 and F2 by changing sampling rate of $y(n)$ to f_{s_0}. For example, in case that f_{s_1} is twice as high as f_{s_0}, F0, F1 and F2 of $y(n)$ are half ones. However, these processes have problem that length of utterance varies. Since the ASR system utilizes delta-powers as feature, variable length of utterance may affect ASR performance. In this paper, to improve ASR performance with LC after converting of Lombard features, we adjusted length of utterance to original one by PICOLA [8]. PICOLA is the method to adjust length of utterance by expanding or condensing voiced vowel on the time sequence. Before resampling, we search F0 by using correlation function, and delete repetition parts. Anteroposterior parts are processed by overlap and add. As a result, we obtain $y(n)$ that has the same length of utterance of $x(n)$. We define LC processed by these processes as Convert-Clean speech (CC).

EVALUATION EXPERIMENTS

To evaluate ASR performance with NC, LC and CC, we conducted experiments by using JULIUS [9]. We will confirm that degradation of ASR performance by comparing NC and LC. Also, improvement of ASR performance is confirmed by comparing LC and CC.

Experimental conditions

Table 2 shows experimental conditions. We utilized to recognize 25 ms frame length and 10 ms frame shift, 12-dimentional MFCCs (Mel-Frequency Cepstrum Coefficient), 12-delta-MFCCs and 1-delta-power as features. The evaluation data are phoneme balanced sentences for 40 speakers. We evaluated ASR performance with CC converted by proposed method with the each F0 and formant frequency ascending rate. We converted features based on the ascending rates that are 15 % with F0 for female speakers, 30 % with F0 for male speakers, 16 % with formant frequency for female speakers and 7% with formant frequency for male speakers.

Table 2. Experimental conditions

Decoder	JULIUS 4.1.2
Acoustic model	PTM triphone model
Language model	5,000 vocabulary
Number of speaker	40
Sampling rate	16 kHz
Quantization	16 bits

Experimental Results

Figures 5 and 6 show the ASR performance with NC, LC and CC for female (Left side) and male (Right side). In Fig 5, CC is converted by proposed method with the F0 ascending rate. On the other hand, in Fig 6, CC is converted by proposed method with the formant frequency ascending rate.

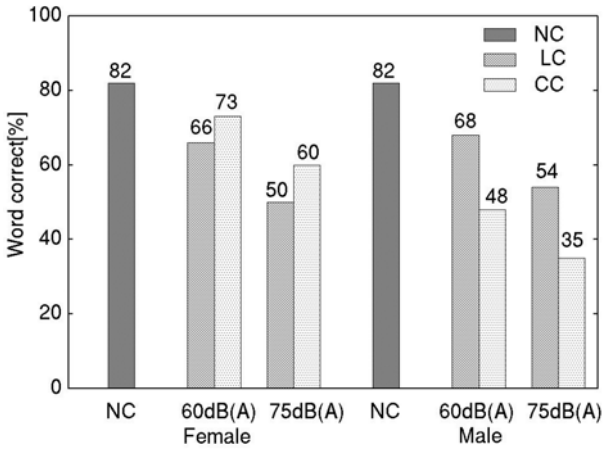

Figure 5. ASR performance with NC, LC and CC by conversion with the F0 ascending rate

Discussion

In case of conversion by proposed method with the F0 ascending rate, ASR performance with CC degraded at 20 % for male speakers in 60 dBA as shown in Fig. 5. Also, ASR performance with CC degraded at 19 % for male speakers in 75 dBA as shown in Fig. 5. From these results, we found that ASR performance with CC degraded in case of conversion by proposed method with the F0 ascending rate. However, as JULIUS doesn't utilize to recognize F0 as feature, we think that this result is true tendency. Conversion with PICOLA and resampling enables to control F0 and formant frequency.

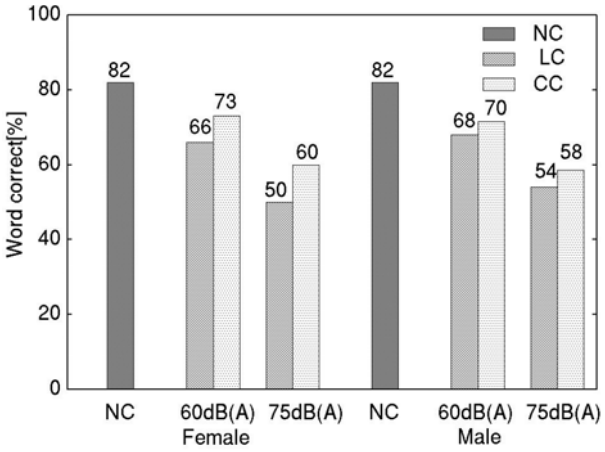

Figure 6. ASR performance with NC, LC and CC by conversion with the formant frequency ascending rate

In case of conversion by proposed method with the ascending rate of F0, ASR performance degraded because the proposed method much decreases formant frequency for male speakers. On the other hand, ASR performance was improved for female speakers both in case of conversion using the F0 and formant frequency ascending rates. We guess that the ascending rates are the same tendency as F0 and formant frequency. In case of conversion by proposed method with the formant frequency ascending rate, ASR performance with CC was improved at 7 % for female speakers and 2 % for male speakers in 60 dBA as shown in Fig. 6. Also, ASR performance with CC was improved at 10 % for female speakers and 4 % for male speakers in 75 dBA as shown in Fig. 6. As a result, ASR performance was improved in case of conversion by proposed method with the formant frequency ascending rate.

CONCLUTIONS

In this paper, we analysed and compared Lombard-clean speech and neutral-clean speech. As a result, we confirmed ascents of the fundamental frequency, the first formant frequency and the second formant frequency compared with neutral-clean speech. We proposed the new approach based on the voice conversion from Lombard-clean speech to neutral-clean speech by using PICOLA and resampling. As a result, in case of conversion by proposed method with the formant frequency ascending rate, we confirmed that ASR performance was improved at 10 % for female speakers and 4 % for male speakers. In the future, we will try to analyse other features such as envelope tilt and much improve the Lombard ASR performance.

ACKNOWLEDGMENTS

This work was partly supported by Grants-in-Aid for Scientific Research funded by Japan's Ministry of Education, Culture, Sports, Science, and Technology.

REFERENCES

1 Norihide Kitaoka et al. "Speech recognition under noisy environments using spectral subtraction with smoothing of time direction" IEICE, vol. J85-D-II, No. 2, pp. 500-508, 2000.

2 Yasunari Obuchi et al. "Development of Evaluation Platform to Improve Speech Recognition Accuracy of Car Navigation Systems in Real Environment" Proc. ASJ, pp. 1-2, 2006.

3 Jean-Claude Junqua et al. "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex" Speech Communication 20, pp. 13-22, 1996.

4 Atsushi Wakao et al. "Variability of Lombard effects under different noise conditions" Proc. ICSLP, vol. 4, pp. 2009-2012, 1996.

5 Tetsuji Ogawa et al. "Influences of Lombard effect on simulation-based assessments of noisy speech recognition for various recognition tasks and noise levels." Proc. ASJ, pp. 195-198, 2007.

6 Hynek Borl et al. "Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environment" Proc. ICASSP, pp. 3937-3940, 2009.

7 Sukeyasu Kanno et al. "Lombard speech recognition based on voiced sound detection and application to the fabric inspection system in factories" IEICE Trans. vol. J85-D-II, No. 5, pp. 851-862, 2002.

8 Naotaka Morita et al. "Time-scale modification algorithm for speech by use of Pointer Interval Control OverLap and Add (PICOLA) and its evaluation." Proc. ASJ, pp. 149-150, 1986.

9 Akinobu. Lee et al. "Recent Development of Open-Source Speech Recognition Engine Julius" Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 131-137, 2009.