VARIABILITY OF LOMBARD EFFECTS UNDER DIFFERENT NOISE CONDITIONS

Atsushi WAKAO, Kazuya TAKEDA and Fumitada ITAKURA

Graduate School of Engineering, Nagoya University
Furo-cho 1, Chikusa-ku, Nagoya, 464-01 JAPAN
takeda@nuee.nagoya-u.ac.jp

ABSTRACT

In this paper, variability of Lombard speech under different noise conditions and an adaptation method to the different Lombard speech are discussed. For this purpose, various kinds of Lombard speech are recorded under different conditions of noise injected into a earphone with controlled feedback of voice. First, DTW word recognition experiments using clean speech as a reference are performed to show that the higher the noise level becomes the more seriously the utterance is affected. Second, linear transformation of the cepstral feature vector is tested to show that when given enough (more than 100 words) training data, the transformation matrix can be correctly learned for each of the noise conditions. Interpolation of the transfer matrix is then proposed in order to reduce the adaptation parameter and number of training samples. We show, finally, that five words are enough for the learning interpolated transformation matrix for unknown noise conditions.

1. INTRODUCTION

The degradation of recognition performance under noisy conditions prevent many important speech recognition applications, e.g. aircraft control and factory automation, from being widely utilized. Therefore, not only enhancement of noisy speech, but also dealing with the unusual effort of utterance under noisy conditions, i.e. the Lombard effect, is still one of the most important issues in speech recognition. As many studies have reported, presence of the noise affects the speech characteristics in various ways such as power increasing, shift of formant frequency, change in formant band width and spectral tilt, etc [1],[2]. In order to deal with such Lombard effects, many recognition methods have been proposed, as well as investigations from the standpoints of production and perception.

One important idea of those proposed methods is to transform the reference utterance to that affected by Lombard effect [3],[4]. In order to implement such an idea, it is needed to learn transformation parameters before the recognition process starts. Which means that if some variability exists in the way masking noise affects the utterance, the transformation does not work well except for the trained noise condition. However, neither variability of Lombard effects nor the method to realize robust transformation have been investigated well.

The purposes of this study are 1) to clarify the variability of Lombard effect under different noise conditions, and 2) propose a new transformation method which can adapt to new noise conditions easily. In Section 2, recording conditions of Lombard speech will be described. In Section 3, the baseline recognition experiment will be performed to clarify the different characteristics of Lombard speech under different noise conditions. In Section 4, effectiveness and problem of the adaptation method to Lombard effect, which uses linear transformation of cepstral parameters will be discussed. Finally, in Section 5, interpolation of transfer matrices will be proposed for adapting the transformation to unknown noise conditions.

2. EXPERIMENTAL SETUP

2.1. Recording Lombard Speech

Speech materials are recorded in a sound proof room (background noise level was 30 dBA) where 12 different noise conditions (11 noise sounds and a silent condition) are provided to speakers through earphones. The utterance being recorded is fed back to the speaker also through earphones so that the speech level at the ear position is as natural as possible.

The 11 different noise sounds used were; seven band limited (0-4 kHz) white noises of the level `52`, `62`, `68`, `74`, `80`, `86` and `92` dBA; four differently band limited (`0-1`, `1-2`, `2-4`, `4-8`, `(kHz)`) white noises of the level `80 (dBA)`.

Under presence of each noise sound, each of three female speakers uttered 120 Japanese city names twice. Thus, including one session without noise (`clean`), 2880 word utterances are recorded for each speaker. All utterances are recorded by DAT and then digitized in 16-bit by a sampling rate of 12,000 Hz after being filtered by an anti-aliasing filter.

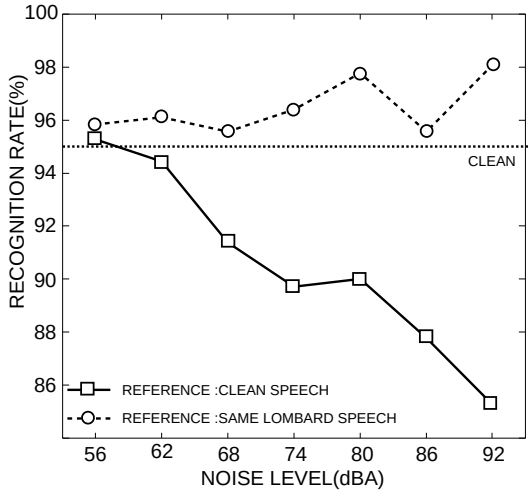
2.2. Recognition System

LPC cepstral parameters were calculated under the conditions listed in Table 1 and DTW word recognizer was adopted as recognition system. The first utterance is used as a reference and the second one is used for a test utterance, respectively. In order to avoid the acoustic difference between sessions due to long time recording, cepstral coefficients averaged over all utterances in a same noise condition

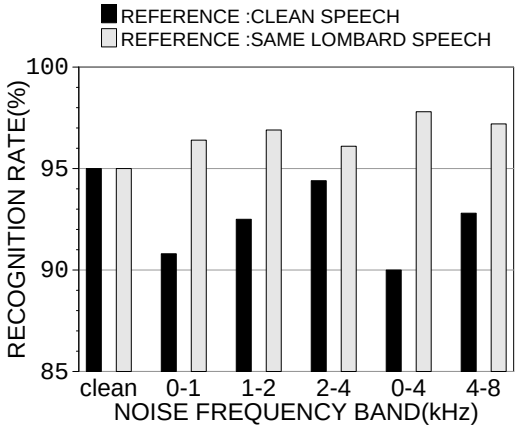
Table 1: Speech analysis conditions.

Sampling frequency	12 kHz
Window type	Hamming
Frame Length	20 msec
Frame period	10 msec
LPC order	14
Cepstral order	20

are subtracted before recognition. The baseline recognition accuracy of using clean speech for both reference and test utterance was 95.0 %.



(a) across the level.



(b) across the band.

Figure 1: Recognition accuracy when Lombard speech under various noise conditions are recognized using two different references uttered under clean and the same noise condition with the test utternaces. The baseline recognition accuracy of using clean speech for both reference and test utterance was 95.0 %.

3. LOMBARD SPEECH UNDER DIFFERENT CONDITIONS

In this section, we will discuss the differences among Lombard speech under different noise conditions from the viewpoint of speech recognition accuracy. As for the first experiment, using clean speech as reference, Lombard speech under 11 noise conditions were recognized. The results are illustrated in Figure 1(a) and (b) for across noise level and noise band, respectively. In both figures, the recognition accuracies of using Lombard speech under the same conditions as the reference are also illustrated.

In Figure 1 (a), it can be seen that the recognition accuracy decreases as the noise level increases down to about 10 % below of the clean-to-clean case. From this result it is clarified that the higher the noise level becomes the more seriously the utterance affected.

In Figure 1 (b), same as the case of changing noise level, the recognition accuracy varies across the spectral band of presented noise, although the range of that is not so wide. In both cases, higher recognition accuracy is obtained than that of the clean-to-clean case when the Lombard speech under the same condition are used for both test and reference utterances. The intelligibility of Lombard speech are discussed in [1], [2] based on perceptual experiment. In [1], Summers et al. concluded the improvement of intelligibility although Junqua reported the speaker and utterance dependency of the improvement in [2]. Our results clarifies, in the view point of pattern matching based speech recognition, the Lombard reflex improves the discriminability of confusing words.

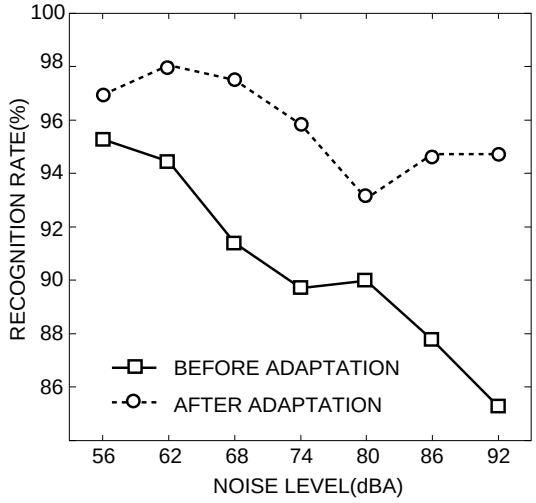
4. ADAPTATION TO DIFFERENT CONDITIONS

4.1. Cepstral Linear Transformation

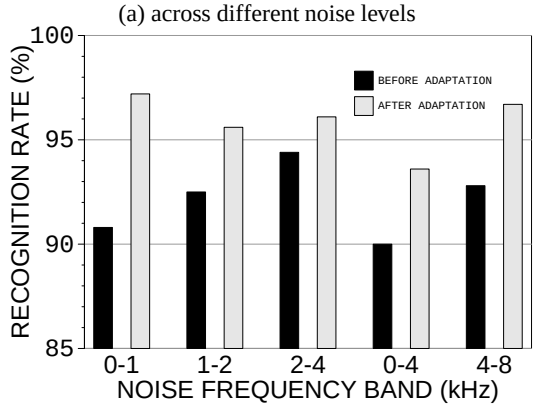
In [4], Mokbel et al. proposed a compensation method of noisy speech through linear transformation of feature vector to that of clean speech. Since the acoustic manifestations of the Lombard effect is not so simple, such general class of transformations as linear transformation is expected to also work better for Lombard speech compensation than additive model [3] or frequency warping. On the other hand, in general, a learning transfer matrix needs a large amount of training data. In this section, we discuss the effectiveness and problems of the linear transformation for Lombard effect normalization.

The linear adaptation method used in this section is the same as proposed in [4] except for the following two points; 1) adopting the transformation for normalizing Lombard effect only, i.e. not for noise contamunated speech, and 2) prepare the transfer matrix for each phoneme. Thus, the adaptation procedure can be summarized as follows.

1: Find phoneme boundary: Since reference utterances are manually segmented into phoneme segments, phoneme boundaries in the Lombard utterances can be decided by taking the optimal matching path in DTW.



(a) across different noise levels



(b) across different noise bands.

Figure 2: Recognition accuracy before and after adaptation through linear transformation of cepstral vector.

2: Calculate parameters: For each phoneme that appeared in the training utterances, transfer matrix A of transformation;

$$y = Ax$$

is calculated so as to minimize MS error between Lombard cepstrum vector y and transformed clean vector Ax.

3: Realign the phoneme boundary: For further refinement of transformation parameters, phoneme boundaries are realigned by executing DTW using a transformed cepstral vector sequence of clean speech as the new reference, until convergence.

The results of adaptation is illustrated in Figure 2, in which recognition accuracy after adaptation is averaged over three speakers across the noise conditions. In the experiment, clean utterance was used as the original reference and all of 120 words are used for training each transfer matrices. In all conditions, recognition accuracies are improved more than 2 %, which is 50 % to 70 % reduction of error rate. From the results, it can be concluded that linear transformation in cepstral domain is a promising method of adapting Lombard speech with given enough training data.

4.2. Training Vocabulary

It becomes, therefore, an important issue on how the adaptation works when the training data is limited. In order to answer the question, we have performed the same experiment as above with limiting training utterances to 20, 40, 60, 80 and 100 words. This series of experiments is performed only on the noise condition of 86 dBA, 0-4kHz. Training utterance for each experiment are randomly selected and recognition accuracy shown in Figure 4 is obtained by averaging the total of five times repetition for each experiment.

As shown in the figure, improvement of recognition accuracy is proportional to the number of training utterances. The improving curve, however, is concave and about 100 training words were required to achieve 50 % reduction of error.

5. INTERPOLATION FOR ROBUST ADAPTATION

5.1. Interpolating Transformation Matrix

The experimental results of the above section claimed that effective parameter reduction is required in order to utilize linear transformation as an adaptation method to Lombard reflex. As such method, we propose an interpolation of transformation matrices, which is

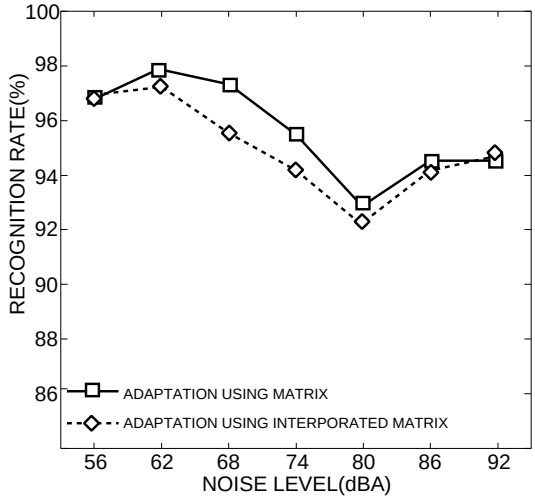
$$y = \{\lambda A' + (1-\lambda)I\}x,$$

where λ is an interpolation parameter, A' is the transformation matrix of most the difficult condition, i.e. 90 dBA in this experiment, and I is an identity matrix.

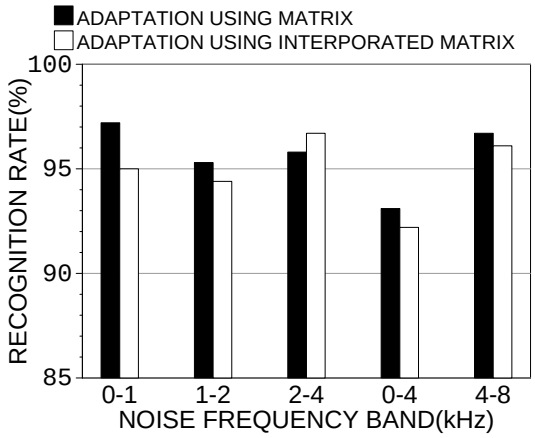
The performance of the transformation generated by the interpolating matrix is demonstrated in Figure 5, in which two transformation matrices; 1) fully trained (thick line) and 2) calculated using trained interpolation parameter (dashed line) are compared. As no significant difference can be seen in the figure, it is confirmed that only one parameter per phoneme is enough for the cepstral transformation.

5.2. Reduction of Training Data

In order to confirm the effectiveness of reducing the number of parameters, finally, an adaptation experiment with limiting training data is performed based on proposed interpolation method using 86 dBA of 1-4 kHz as unknown condition and 90 dBA matrix as the most difficult condition (A'). This time, 1 to 100 training data are randomly selected for training λ. As in the previous experiment, averaged accuracy of five repetitions is shown together with the previous results in Figure 4. As shown in the figure, about 50 % reduction of error can be achieved with about 5 words for training which clarify the effectiveness of reducing the number of parameters in the adaptation scheme.



(a) across different noise levels



(b) across different noise bands.

Figure 3: Recognition accuracy after linear transformation of cepstral vector using interploated transformation matrix $\lambda A' + (1-\lambda)I$ where A' matrix is trained under the condition of `90 dBA`, `0-4 kHz` is used for both experiments.

6. SUMMARY

In this paper, we discussed variability of Lombard effect under different noise conditions and proposed a robust adaptation method to deal with the variability. In the first experiment, Lombard speech under various noise conditions are recognized using clean speech as reference. The results clarified that 1) recognition accuracy varies from 96 to 86 % up to the noise condition, 2) the higher noise level is the lower recognition accuracy obtained, and 3) discriminability of the Lombard speech is superior to the normal speech. In the second experiment, we clarified the effectiveness of the cepstrum domain linear transformation as an adaptation method to Lombard reflex, also showing that the the method required more than 100 utterances for promising result. In the third experiment, we showed that the interpolating transfer matrix is an effective method of reducing data for training linear transformation parameters and thus robust adaptation of Lombard reflex can be achieved with changes of ambient noise conditions.

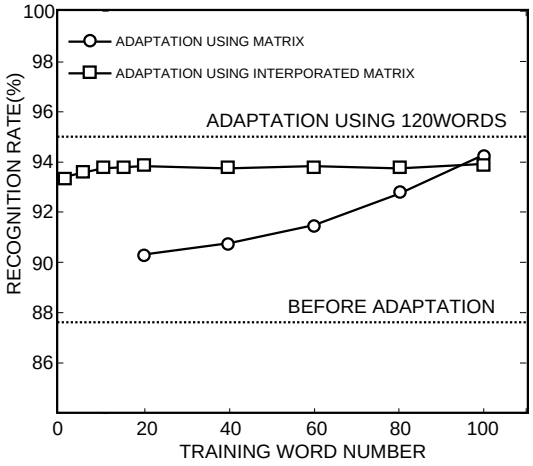


Figure 4: Recognition accuracy of linear transformation across the number of training words. Circles and boxes indicate the each case of training full matrices and training interpolation parameters. The test utterance is uttered under `86 dbA`, `0-4 kHz` condition and A' matrix is trained under the condition of `90 dBA`, `0-4 kHz`.

REFERENCES

1. Summer, W.V., Pisoni, D.B., Bernacki, R.H., Pedlow, R.I. and Stokes, M.A. “Effects of noise on speech production: acoustic and perceptual analyses”, *J.Acoust.Soc.Am.*, 93-1, 917-928, 1988
2. Junqua, J.C. “The Lombard reflex and its role on human listeners and automatic speech recognizers”, *J.Acoust.Soc.Am.*, 85-2, 849-900, 1989
3. Chen, Y. “Cepstral domain talker stress compensation for robust speech recognition”, *IEEE Trans. ASSP*, 36-4, 433-439, 1988
4. Mokbel, C.E. and Chollet, G.F.A “Word recognition in the car - speech enhancement / spectral transformations-”, *Proc. ICASSP91* 925-928, 1991