THE EFFECT OF SPECTRAL SPACE REDUCTION IN SPONTANEOUS SPEECH ON RECOGNITION PERFORMANCES

Masanobu Nakamura, Koji Iwano, and Sadaoki Furui

Tokyo Institute of Technology, Department of Computer Science
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan,
{masa,iwano,furui}@furui.cs.titech.ac.jp

ABSTRACT

Although speech derived from reading texts and similar types of speech, e.g., that from reading newspapers or that from news broadcasts, can be recognized with high accuracy, recognition performance drastically decreases for spontaneous speech. This is due to the fact that spontaneous speech and read speech are significantly different acoustically as well as linguistically. This paper analyzes differences in acoustic features between spontaneous and read speech using a large-scale spontaneous speech database "Corpus of Spontaneous Japanese (CSJ)". Using a linear transformation matrix, experimental results show that spontaneous speech can be characterized by reduced size of spectral space in comparison with that of read speech. These have also clarified that a reduction in the spectral space leads to a reduction in phoneme recognition accuracy. This result indicates that spectral reduction is one major reason for the decrease of recognition accuracy of spontaneous speech.

Index Terms— Corpus of Spontaneous Japanese, Spontaneous speech, MLLR matrix, Spectral space

1. INTRODUCTION

State-of-the-art speech recognition technology can achieve high recognition accuracies on read text or speech uttered clearly. However, the accuracy is still rather poor for spontaneous speech, which is not as well structured acoustically and linguistically as read speech [1, 2]. Analysis of spontaneous speech and clarifying acoustical differences between spontaneous speech and read speech are expected to prove useful for improving spontaneous speech recognition performance.

This paper reports results of our analysis on spectral reduction in spontaneous speech using a linear transformation matrix and investigates its contribution to speech recognition performance. Studies of spectral reduction of spontaneous speech in comparison with read speech have already been conducted by several researchers [3]. Our previous research has also reported the relationships between spectral reduction and spontaneous speech recognition performance [4, 5]. However, as of yet no research has been conducted using a large spontaneous database to investigate the direct effect of spectral reduction on spontaneous speech recognition performance. This paper focuses on the analysis of spectral reduction and its ratio of direct effect on a speech recognition task using the large-scale "Corpus of Spontaneous Japanese (CSJ)"[6].

This paper is structured as follows. Section 2 describes the speech corpus used for the analysis. Section 3 describes the reduction of cepstrum space for spontaneous speech in comparison with read speech using mean Mel-Frequency Cepstrum Coefficients (MFCC) vectors. Section 4 reports the reduction of cepstrum space using a linear transformation matrix, and its effect on spontaneous speech recognition performance. Finally, section 5 concludes this paper.

2. SPEECH CORPUS

Utterances representing four different speaking styles in the CSJ, that is, read speech (R), academic presentations (AP), extemporaneous presentations (EP), and dialogues (D), were used in the analysis. AP contains live recordings of academic presentations from nine different academic societies, covering the fields of engineering, social science and humanities, and therefore is composed of many formal utterances. EP is composed of studio recordings of paid layman speakers' speech on everyday topics like "the most delightful memory of my life" delivered in front of a small audience and in a relatively relaxed atmosphere. Therefore, the speaking style in EP is more informal than in AP. Each presentation had a duration of approximately 10 minutes. The read speech contains speech read from novels that include transcribed dialogues, and transcriptions of AP or EP by the same speakers. The dialogue set is composed of interviews concerning AP and EP, task oriented dialogues, and free dialogues.

Table 1 shows the total duration of speech and number of phoneme samples used in this analysis for each speaking style. There are utterances from five males and five females for each of the four speaking styles. These utterances were segmented by identifying silences with durations of 400 ms or longer. If the length of any segmented unit was shorter than 1 second, it was merged with the succeeding unit. The segmented utterances will hereafter be called "utterance units".

Table 1. Total durations of speech and number of phoneme samples for each speaking style.

Speaking style	Duration(min)	Phoneme samples
R	175	119,354
AP	160	119,330
EP	112	79,863
D	296	195,441

Table 2. Japanese phonemes.

Vowels	/a,i,u,e,o,a:,i:,u:,e:,o:/
Consonants	/w,y,r,p,t,k,b,d,g,j,ts, z,s,sh,h,f,N,m,n/

3. ANALYSIS USING MEAN MFCC VECTORS

3.1. Extraction of cepstral features

The means of the MFCC vectors for each phoneme in various speaking styles were calculated to analyze the spectral characteristics of spontaneous speech. The extraction process was as follows.

1. The speech waveforms were converted to 12-dimensional MFCC vectors using a 25 ms-length window shifted every 10 ms. The MFCC vectors were augmented with log-energy, and their first and second derivatives, to give 39-dimensional feature vectors. Cepstral mean subtraction (CMS) was applied to each utterance unit.
2. A mono-phone hidden Markov model (HMM) with a single Gaussian mixture was trained using utterances from every combination of phonemes, speakers, and speaking styles. Every HMM had a left-to-right topology with three self-loops.
3. The mean vectors of the 12-dimensional MFCC at the second state of the HMM were extracted for each phoneme and used for the analysis.

The set of 29 Japanese phonemes used in this analysis, consisted of 10 vowels and 19 consonants, is listed in table 2. A few extremely rare phonemes, which occurred a very small number of times in the utterances, were not included in the analysis.

3.2. Reduction ratio

In order to quantitatively analyze the reduction of the distribution of phonemes, Euclidean distances between the mean vector of each phoneme and the center of the distribution of all phonemes, that is the vector averaged over all the phonemes, were calculated, and the ratio $red_p(X)$ of the distance for spontaneous speech (presentations and dialogues) to that of read speech was calculated for each phoneme as follows.

$$red_p(X) = \frac{||\boldsymbol{\mu}_p(X) - \mathrm{Av}[\boldsymbol{\mu}_p(X)]||}{||\boldsymbol{\mu}_p(R) - \mathrm{Av}[\boldsymbol{\mu}_p(R)]||} \quad (1)$$

Here $\boldsymbol{\mu}_p(X)$ is the mean vector of a phoneme p uttered with a speaking style X. X takes a value of AP, EP, or D, corresponding to academic presentations, extemporaneous presentations, and dialogues, respectively. $\boldsymbol{\mu}_p(R)$ is the mean vector of a phoneme p of read speech, and Av indicates the averaged value.

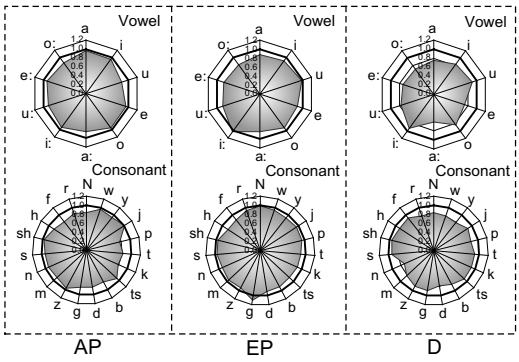

Fig. 1. Relative reduction ratio of the vector distance between each phoneme and the phoneme center for various speaking styles. A larger shaded area indicates that individual phonemes are more easily distinguishable.

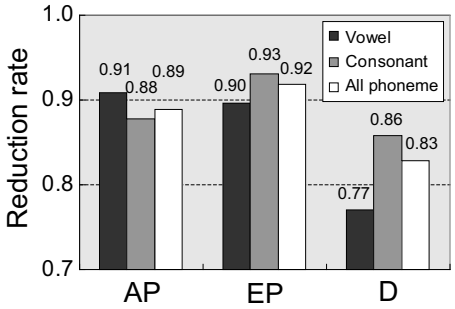

Fig. 2. The mean reduction ratios for vowels, consonants, and all phonemes for each speaking style.

3.3. Result of reduction ratio

Figure 1 shows the reduction ratios $red_p(X)$ averaged over all the speakers, separately for AP, EP, and dialogues. The case when $red_p(X) = 1$ is indicated by a thick line. Results in the figure show the reduction of the cepstrum space for almost all the phonemes in the three speaking styles, and this is most significant in dialogue utterances.

Figure 2 shows the mean reduction ratios for vowels, consonants, and all phonemes, respectively, for each speaking style. These results show that there is a reduction of the distribution of phonemes in the cepstrum domain in comparison with that of read speech for all the speaking styles, and the reduction is most significant for dialogue speech.

4. ANALYSIS USING A LINEAR TRANSFORMATION

4.1. MLLR matrix

In Section 3, the results showed that the distribution of the mean MFCC vectors for spontaneous speech reduces in com-

parison with that for read speech. In this section, the transformation from the cepstrum space for read speech to that for spontaneous speech is modeled by a linear transformation which consists of linear reduction and rotation of the space. The ratio of the contribution of the linear reduction of the cepstrum space to that of the whole transformation is calculated. A maximum likelihood linear regression (MLLR) transformation is used for the linear transformation. A MLLR matrix is made for the whole set of phonemes. The case when the MLLR matrix is the unit matrix means that the size of the cepstrum space for read and spontaneous speech is the same.

The spontaneous speech data for each speaker was divided into two parts. Two-thirds of the data was used for estimating the MLLR matrix, and the remaining one-third was used for the evaluation of the experiment in Section 4.3. In this section, HMM acoustic models with 16 Gaussian mixtures are trained for each speaker on read speech, and are then adapted with MLLR transformations for use with spontaneous speech.

4.2. Results of analysis using MLLR matrices

Figure 3 shows MLLR matrices averaged over all speakers used for transformation from the cepstrum space for read speech to that for AP, EP, and dialogue (left: AP, center: EP, right: dialogue). In this figure, it is observed that the values of the diagonal elements in the MLLR matrices for all speaking styles are significantly greater than values of the other elements, and that values of the diagonal elements are less than one for all speaking styles. Therefore, it is clarified that the cepstrum space for spontaneous speech is reduced in comparison with that of read speech.

The mean values averaged over all the diagonal elements composed of MLLR matrices for AP, EP, and D in Figure 3 are 0.79, 0.82, and 0.74, respectively. This result is matched with the previous result in Section 3 that the reduction of the cepstrum space for dialogue speech is more significant than that for presentations.

4.3. Effect of cepstrum space reduction to recognition performance

In Section 4.2, MLLR matrices were used for transformation from read speech to spontaneous speech in the cepstrum space. The analysis clarified that the cepstrum space for spontaneous speech is reduced in comparison with that of read speech. The strong diagonal elements in the MLLR matrix mean that the transformation from read speech to spontaneous speech is more linear reduction than rotation of the cepstrum space.

In this section, the effect of the reduction in cepstrum space on speech recognition performance is investigated using the diagonal elements in the MLLR matrices.

4.3.1. Experimental conditions

In this experiment, phoneme recognition accuracy is evaluated through the recognition experiments. The acoustic models for each speaker for the recognition experiment is made as follows:

1. The MLLR matrices are calculated using the data for transformations as described in Section 4.1.
2. The acoustic model for read speech for each speaker is adapted to one for spontaneous speech using the MLLR matrix.

The data for evaluation described in Section 4.1 is used for testing the adapted acoustic models. In the recognition experiments, 38-dimensional MFCC vectors are used as acoustic characteristics, which leads to a 38×38 MLLR matrix.

Three different MLLR matrices are made for comparison with each other in this experiment. The first matrix is the unit matrix I. The second one is a MLLR matrix T_X which is calculated by a MLLR transformation. The third one is a diagonal MLLR matrix $T_X^{(diag)}$, which is calculated by a MLLR transformation under the assumption the matrix is diagonal. The full MLLR matrix T_X can perform reduction and rotation transformations, whilst the diagonal matrix $T_X^{(diag)}$ can only perform reductions.

Acoustic models adapted from the model for read speech to that for spontaneous speech using three different regression matrices $I, T_X, T_X^{(diag)}$ are called **BASE**, **FULL**, and **DIAG**, respectively. These three acoustic models are evaluated by considering their accuracy on phoneme recognition. The number of Gaussian mixtures for the acoustic models for read and spontaneous speech was optimized in advance, and was found to be 16 Gaussian mixtures. A phoneme network with di-phone probabilities was used as a language model for recognition. The insertion penalty was optimized for each speaking style.

4.3.2. Results of recognition experiments

Figure 4 shows phoneme accuracies using acoustic models **BASE**, **FULL**, and **DIAG**, respectively. It is observed that phoneme accuracy of **DIAG** is greater than that of **BASE** for all speaking styles. Therefore, it is clarified that the reduction of the cepstrum space contributes to the decrease of speech recognition performance for spontaneous speech when compared with read speech.

The effect in recognition performance due to reduction in the cepstrum space for speaking style X in spontaneous speech is expressed as the "effect ratio *eff*(X)":

$$\mathit{eff}(X) = \frac{acc_D(X) - acc_B(X)}{acc_F(X) - acc_B(X)} \tag{2}$$

$acc_B(X)$, $acc_D(X)$, and $acc_F(X)$ show phoneme accuracies calculated with the data for speaking style X by evaluating

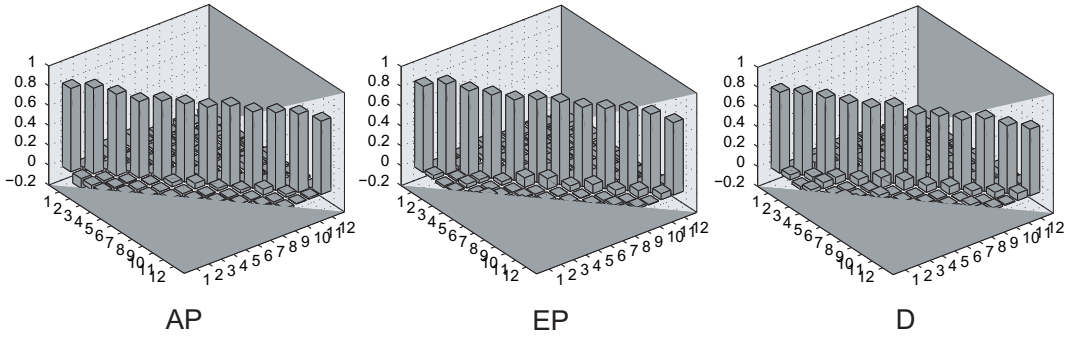

Fig. 3. MLLR matrix components averaged over five males and five females for AP, EP, and dialogue.

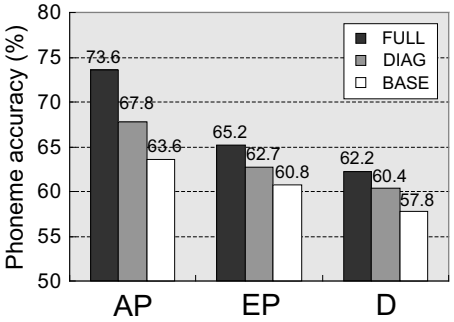

Fig. 4. Phoneme accuracies using acoustic model of **BASE**, **DIAG**, and **FULL**.

the acoustic models of **BASE**, **DIAG**, and **FULL**, respectively. $eff(X) = 1$ means that the decrease in recognition performance is only due to a reduction in the cepstrum space. The effect ratio for AP, EP, and dialogues is 0.42, 0.43, and 0.59, respectively. This result clarifies that about a half of the decrease in phoneme recognition accuracy is due to reduction in the cepstrum space, and the reduction is most significant for dialogues uttered spontaneously.

5. CONCLUSIONS

In order to increase recognition accuracy for spontaneous speech, the difference in acoustic features between spontaneous speech and read speech has been analyzed using utterances with speaking styles of read speech, academic presentations, extemporaneous presentations, and dialogues, that occur in the "Corpus of Spontaneous Japanese". It has been found by using linear transformation matrices that the cepstral distribution of spontaneous speech reduces significantly in comparison with that of read speech. Although this was true for all the spontaneous speech analyzed in this paper, that is, AP, EP, and dialogues, the reduction was most significant for dialogues, which are obviously more spontaneous than the other styles. It has also been found that about a half of the decrease in recognition performances is due to a reduction in the cepstrum space.

In summary, spontaneous speech can be characterized by the reduction of spectral space in comparison with that of read speech, and this is one of the major factors contributing to the decrease in recognition accuracy. The reduction in the feature space could be explained by neutralization of vocal-tract shapes in spontaneous speech.

Our future research will include analysis with an increased number of speakers. Although we have clarified spectral reduction and its effects on spontaneous speech recognition, it is not yet clear how we can use these results for improving recognition performances. Creating methods for adapting acoustic models to spontaneous speech based on the results obtained in this research is also one of our future targets.

6. ACKNOWLEDGEMENTS

This research was supported by the Science and Technology Agency Priority Program "Spontaneous Speech: Corpus and Processing Technology" and the 21st Century COE Program "Framework for Systematization and Application of Large-scale Knowledge Resources".

7. REFERENCES

[1] S. Furui, "Recent advances in spontaneous speech recognition and understanding," *Proc. IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, Tokyo, pp.1-6, 2003.

[2] S. Furui, "Toward spontaneous speech recognition and understanding," Pattern Recognition in Speech and Language Processing, W. Chou, B.-H. Juang (Eds.), CRC Press, New York, pp.191-227, 2003.

[3] R.J.J.H. van Son, L.C.W. Pols, "An acoustic description of consonant reduction," *Speech Communication*, vol.28, no.2, pp.125-140, 1999.

[4] S. Furui, M. Nakamura, T. Ichiba, and K. Iwano, "Analysis and recognition of spontaneous speech using *Corpus of Spontaneous Japanese*," *Speech Communication*, vol.47, no.1-2, pp.208-219, 2005.

[5] M. Nakamura, K. Iwano, and S. Furui, "Acoustic and linguistic characterization of spontaneous speech", *Proc. Speech Recognition and Intrinsic Variation(SRIV2006)*, Toulouse, France, pp.3-8, 2006.

[6] K. Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," *Proc. IEEE Workshop on SSPR*, Tokyo, pp.7-12, 2003.