論文 / 著書情報
Article / Book Information

Title	Analysis of spectral space reduction in spontaneous speech and its effects on speech recognition performances
Authors	Masanobu Nakamura, Koji Iwano, Sadaoki Furui
Citation	Interspeech2005, Vol. , No. , pp. 3381-3384,
Pub. date	2005, 9
Copyright	(c) 2005 International Speech Communication Association, ISCA
DOI	http://dx.doi.org/

Analysis of Spectral Space Reduction in Spontaneous Speech and Its Effects on Speech Recognition Performances

Masanobu Nakamura, Koji Iwano, and Sadaoki Furui

Department of Computer Science
Tokyo Institute of Technology, Tokyo, 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 broadcast, can be recognized with high accuracy, recognition accuracy 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 speech and read speech using a large-scale spontaneous speech database "Corpus of Spontaneous Japanese (CSJ)". Experimental results show that spontaneous speech can be characterized by reduced size of spectral space in comparison with that of read speech. It has also been found that there is a strong correlation between mean spectral distance between phonemes and phoneme recognition accuracy. This indicates that spectral reduction is one major reason for the decrease of recognition accuracy of spontaneous speech.

1. Introduction

State-of-the-art speech recognition technology can achieve high recognition accuracies for read texts or limited domain spoken interactions. 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 and investigates its contribution to speech recognition performance. Spectral reduction of spontaneous speech in comparison with read speech has already been conducted by several researchers [3]. However, as of yet no research has been conducted using a large spontaneous database nor on the relationships between the spectral reduction and spontaneous speech recognition performance. This paper focuses on the analysis of spectral reduction using cepstral features that are widely used in speech recognition, based on a large-scale spontaneous speech database "Corpus of Spontaneous Japanese (CSJ)"[4].

The paper is constructed as follows. Section 2 describes speech materials used for the analysis. In Section 3, a cepstral feature extraction method for analysis is explained. Section 4 reports the reduction ratio of the cepstral distribution of spontaneous speech in comparison with read speech. Section 5 reports the difference of distances between each pair of phonemes in spontaneous speech and that in read speech. The relationship between the phoneme distances and phoneme recognition performance in various speaking styles is shown in Section 6. Finally, section 7 concludes this paper.

2. Speech materials

Utterances representing four different speaking styles in the CSJ, that is, read speech, academic presentations (AP), extemporaneous presentations (EP), and dialogues, were used in the analysis. AP contains live recordings of academic presentations in nine different academic societies, covering the fields of engineering, social science and humanities, and 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 has a duration of approximately 10 minutes. The read speech contains "reading text" speech reading novels including transcribed dialogues, and "reading transcriptions" speech reading transcription of APs or EPs by the same speakers. The dialogue set is composed of interviews, task oriented dialogues, and free dialogues.

These utterances were digitized by 16 kHz sampling, and segmented by silences with durations of 400 ms or longer. If the length of the segmented unit was shorter than 1 sec, it was merged with the succeeding unit. The segmented utterances are hereafter called "utterance units".

3. Analysis of cepstral features

The mean and variance of cepstrum vectors for each phoneme in various speaking styles were calculated to analyze the spectral characteristics of spontaneous speech. The whole set of 31 Japanese phonemes, consisting of 10 vowels and 21 consonants, are listed in Table 1.

Table 1: Japanese phonemes

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

The mean and variance cepstrum vectors were obtained as follows.

1. 39-dimensional feature vectors, consisting of 12-dimensional MFCC (Mel-frequency cepstrum coefficients), log-energy, and their first and second derivatives, were extracted from utterances using a 25 ms-length

Table 2: Total number of phoneme samples for each speaker and each speaking style.

	Speaker ID	Read speech	AP	EP	Dialogue
Male	M1	7,420	7,371	5,213	9,915
	M2	10,768	10,815	6,000	14,489
	M3	12,118	12,211	8,525	17,616
	M4	23,154	23,208	8,615	19,892
	M5	8,598	8,651	11,518	29,862
Female	F1	12,162	12,071	10,119	25,428
	F2	7,843	7,757	7,206	20,141
	F3	11,383	11,360	4,837	17,044
	F4	8,111	8,038	8,232	20,999
	F5	17,797	17,848	9,598	22,083

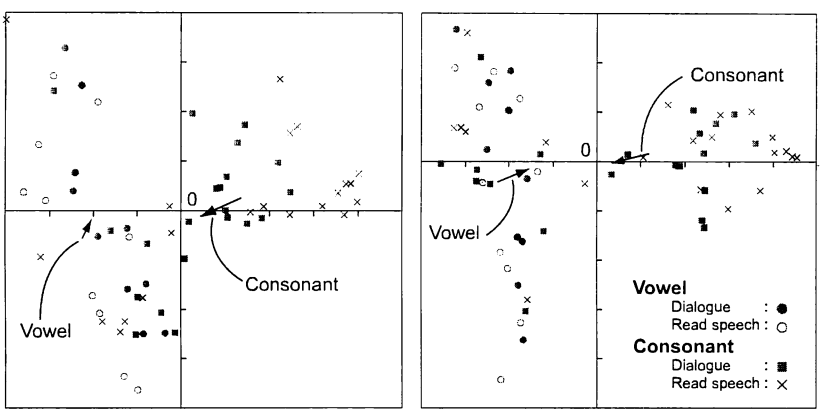

Figure 1: Examples of distributions of mean cepstrum vectors of all the phonemes projected into the 2-dimensional PCA space for dialogue and read speech by two speakers. The arrows indicate the deviations of vowel and consonant centers from the read speech to the dialogue speech.

window shifted every 10 ms. The CMS (cepstral mean subtraction) is applied to each utterance unit.

2. A mono-phone HMM with a single Gaussian mixture was trained using utterances of every combination of phonemes, speakers, and utterance styles. Every HMM had a left-to-right topology with three self-loops.
3. The mean and variance vectors of the 12-dimensional MFCC at the second state of the HMM were extracted for each phoneme and used for the analysis.

4. Reduction ratios of the distributions of phonemes

4.1. Projection into the PCA space

Table 2 shows the total number of phoneme samples used in this experiment for each speaker and each speaking style. Figure 1 shows examples of the distribution of mean cepstrum vectors (12-dimensional MFCC vectors) of all the vowels and consonants, projected into 2-dimensional vector spaces constructed by the Principal Component Analysis (PCA), for the dialogue and read speech by two speakers (left: F5, and right: M5), respectively. Each PCA space was separately built using data from each speaker. These speakers were selected since their voices have relatively large perceptual differences between the two speaking styles. In the figure, x and y axes indicate the first and the second PCA vectors, respectively. The two arrows in each figure indicate deviations of vowel and consonant centers from the read speech to the dialogue speech.

The results clearly show that the distribution of mean cepstrum vectors of dialogue speech is closer to the center of the distribution of all the phonemes than the distribution of read speech. In other words, the size of spectral space for the phonemes in spontaneous speech is smaller compared to that of read speech.

4.2. Reduction ratio

In order to quantitatively analyze the reduction of the distribution of phonemes, Euclidean norms/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 of the distance for spontaneous speech (presentations and dialogues) to that of read speech was calculated for each phoneme as follows.

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

Here $\mu_p(X)$ is the mean vector of a phoneme p uttered with a speaking style X, $\mu_p(R)$ is the mean vector of a phoneme p of read speech, and Av indicates the value averaged over all phonemes.

Results using the mean cepstrum vector of the second state of the HMM with a single Gaussian mixture as the mean vector

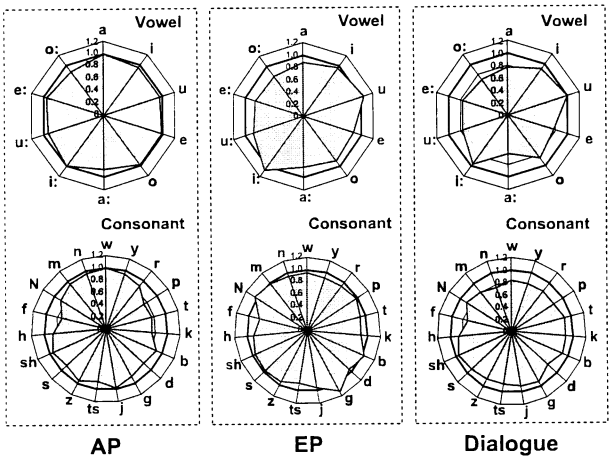

Figure 2: The reduction ratio of the vector norm between each phoneme and the phoneme center in the spontaneous speech to that in the read speech.

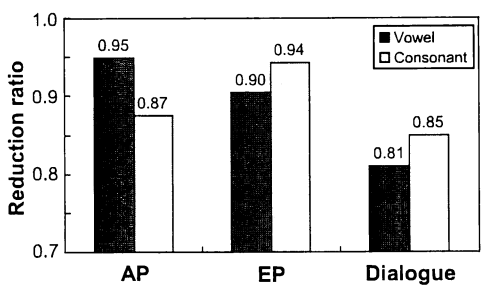

Figure 3: Mean reduction ratios of vowels and consonants for each speaking style.

for each phoneme are shown in Figure 2. The figure shows the reduction ratios $red_p(X)$ averaged over all the speakers, separately for AP, EP, and dialogues. /N:/ and /ch/, which rarely occurred in the utterances listed in Table 2, were not used in this analysis. The condition of $red_p(X) = 1$ is indicated by a thick line. The dialogues include interviews on AP and EP, task dialogues, and free dialogues. 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 3 shows mean reduction ratios for vowels and consonants, respectively, for each speaking style. These results show that the reduction of the distribution of phonemes in the cepstrum domain in comparison with that of read speech is observed for all the speaking styles, and most significantly for dialogue speech.

5. Reduction of distances between phonemes

In the previous section, the reduction of cepstrum space was measured by the ratio of the distance between each phoneme and the phoneme center in spontaneous speech to that in read speech. In this section, the reduction of cepstral distance between each phoneme pair is measured. The Euclidean distance using the mean cepstrum vector of each phoneme and the Mahalanobis distance, which takes into account the variances, were measured. The definition of Mahalanobis distance $D_{ij}(X)$ between phoneme i and j spoken with a speaking style X can be written as follows.

$$D_{ij}(X) = \sqrt{\frac{K\sum_{k=1}^{K}(\mu_{ik}(X) - \mu_{jk}(X))^2}{\sum_{k=1}^{K}\sigma_{ik}^2(X) + \sum_{k=1}^{K}\sigma_{jk}^2(X)}} \quad (2)$$

Where, K is the dimension of an MFCC vector ($K = 12$). $\mu_{ik}(X)$ and $\sigma_{ik}^2(X)$ are the kth dimensional elements of the mean and the variance vector of MFCC for phoneme i uttered with a speaking style X. In the case of the Euclidean distance between phonemes i and j, the denominator in the above formula (2) is set to a constant value.

Five males and five females were randomly selected from the CSJ for this experiment. The total number of phoneme samples for each speaking style was 45,242 (read speech), 80,095 (AP), 55,102 (EP), or 56,583 (dialogues). The read speech set in the CSJ includes various kinds of "reading transcriptions" and "reading novels including dialogues". The dialogue set includes variation of "interview" and "free dialogue". Therefore, speech materials of read speech and dialogues for this experiment were selected so as to represent as many variations of speaking styles as possible. Speech materials of AP and EP were randomly selected from the test-set data of CSJ designed for speech recognition experiments.

Figure 4 shows the cumulative frequency of distances between phonemes for each speaking style. The left-hand side of the figure shows the case using the Euclidean distance, whereas the right-hand side shows the case using the Mahalanobis distance. The x axis indicates the Euclidean or the Mahalanobis distance, and the y axis indicates the cumulative frequency. These results clearly show that the distances between phonemes decrease as the spontaneity of the utterances increases (D $\gg$ EP $>$ AP $\gg$ R). The Wilcoxon's rank order test with a significance level of p-value ≤ 0.01 shows that the distributions of each speaking style are statistically different from each other, except between AP and EP.

6. Relationship between phoneme distances and phoneme recognition performance

Differences of the size of distribution of between-phoneme distances are expected to be related to the phoneme recognition performance for various speaking styles. This section investigates the relationship between the between-phoneme distances and the phoneme recognition accuracy using utterances by many speakers. Mono-phone HMMs with a single Gaussian mixture for phoneme recognition were trained for each speaking style, using utterances by 100 males and 100 females for AP and 150 males and 150 females for EP. These speakers were randomly selected from the CSJ, and the total number of phoneme samples were approximately two million for AP and EP, respectively. A 38-dimensional feature vector was used as the acoustic feature. The same data as used in Section 5 were used for the evaluation experiment. A phoneme network with di-phone probabilities was used as a language model for recognition. The insertion penalty was optimized for each speaking style.

Figure 5 shows the relationship between the mean phoneme distance and the phoneme recognition accuracy. The left-hand

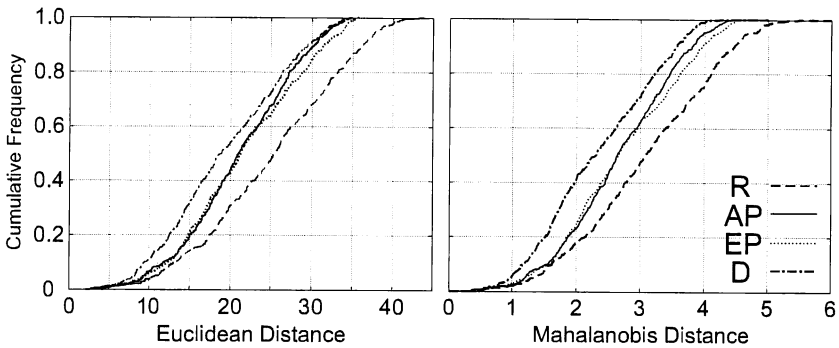

Figure 4: Distribution of distances between phonemes.

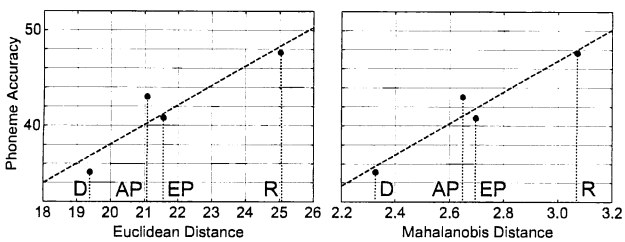

Figure 5: Relationship between phoneme distances and phoneme recognition accuracy.

side of the figure shows the case using Euclidean distance and the right-hand side shows the case using Mahalanobis distance as the distance between phonemes for each speaking style. Correlation coefficients between the mean phoneme distance and the phoneme recognition accuracy are 0.93 in the case using Euclidean distance and 0.97 in the case using Mahalanobis distance. The lines in Figure 5 indicate the regression over the four points. These results clearly show a strong correlation between mean phoneme distance and phoneme accuracy. This means that the phoneme recognition accuracy can be estimated by the mean phoneme distance. That is, the reduction of the Euclidean distances between phonemes is a major factor contributing to the degradation of spontaneous speech recognition accuracy. It can also be concluded that the relationship between the phoneme distance and the phoneme recognition accuracy becomes slightly more significant if the variances of phoneme spectra are also taken into account.

7. Conclusion

This paper reported on research conducted in order to clarify the difference of acoustic features between spontaneous speech and read speech, using utterances with various speaking styles, that is, read speech, academic presentations (AP), extemporaneous presentations (EP), and dialogues, in "Corpus of Spontaneous Japanese (CSJ)". It has been found that the cepstral distribution of spontaneous speech is significantly reduced 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 the more spontaneous the speech, the smaller the distances between phonemes become, and that there is a high correlation between the mean phoneme distance and the phoneme recognition accuracy. 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.

Our future research includes analysis over wider range of spontaneous speech using utterances other than those included in the CSJ. This paper has focused on acoustic properties of spontaneous speech. Obviously, there exist significant differences in linguistic characteristics between spontaneous speech and read speech. Therefore, our future research includes investigating linguistic characteristics of spontaneous speech and their effects on speech recognition performances. 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.

8. 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".

9. References

[1] S. Furui, "Recent advances in spontaneous speech recognition and understanding," *Proc. IEEE Workshop on 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] K. Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," *Proc. IEEE Workshop on SSPR*, Tokyo, pp.7-12, 2003