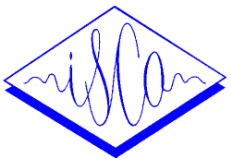


6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

RELATIONSHIP AMONG SPEAKING STYLE, INTER-PHONEME'S DISTANCE AND SPEECH RECOGNITION PERFORMANCE

Kazumasa Yamamoto† *and Seiichi Nakagawa*‡

†Dept. of Electrical & Electronic Eng., Faculty of Engineering, Shinshu University
4-17-1 Wakasato, Nagano, 380-8553, Japan
‡Department of Information and Computer Sciences, Toyohashi University of Technology
1-1 Hibarigaoka, Tenpaku-cho, Toyohashi, 441-8580, Japan
†kyama@sp.shinshu-u.ac.jp, ‡nakagawa@slp.ics.tut.ac.jp

10.21437/ICSLP.2000-405

ABSTRACT

There is a limit of recognition performance for dialogue speech using acoustic models built only with read speech, because various acoustic and linguistic phenomena, which reflect the characteristics of spontaneous speech, are observed in the dialogue speech. In this paper, we investigated the differences of acoustic properties which cause the limit among isolated words, read speech and spontaneous speech. Firstly, the dialogue speech was compared with the read speech through acoustic analyses. Next, the acoustic models were separately built with each of the speech databases. The recognition performance was experimentally evaluated using the acoustic models and the relations of the differences of the performance to those of the acoustic features observed in the analyses were investigated quantitatively. The effectiveness of speaker adaptation was also investigated in the same manner.

1. INTRODUCTION

We have been developing a speech recognition system as a part of a spoken dialogue system. In our previous studies, read speech samples were mainly used for building the acoustic models[1].

In spontaneous speech, various acoustic and linguistic phenomena are observed, such as fast speech rate, neutralization of vowels, restart, filled pause, repair, omission, reverse and so on, but not in read speech. This clearly indicates that there exists a limit of performance on dialogue speech recognition for the acoustic models built only with the read speech data.

For each speaking condition of Broadcast News corpus of DARPA, some experimental results of recognition have been reported [2]. These results showed the differences of recognition performance caused by the differences of speaking style, channel condition and background noise, but the analyses of acoustic properties for each speaking condition have not been carried out. Blaauw investigated phonetic differences between read and spontaneous speech and reported that read syllables were on average longer than spontaneous syllables [3]. Eskénazi also reported the same finding [4]. Kuwabara reported that formant frequencies of individual vowels were largely shifted toward the neutral region in the conventional F_1–F_2 plane for spontaneous speech or fast speech rate [5]. Murakami and Sagayama also compared read speech with spontaneous speech. They reported that compound (or neutral) phone labels increased and phone accuracy rate decreased by 6–10% for spontaneous speech, but the difference of speech rate between speaking styles was small [6].

In this paper, we investigated the differences of acoustic properties among isolated words, read speech and spontaneous speech. Firstly, the dialogue speech was compared with the read speech through acoustic analyses, such as speech rate and inter-phoneme's distance. Next, the acoustic models were separately built with each of the speech databases. The recognition performance was experimentally evaluated using the models, and the relations of the differences of the performance to those of the acoustic features observed in the analyses were investigated quantitatively. The effectiveness of speaker adaptation was also investigated in the same manner.

2. SPEECH MATERIALS AND HMM

Databases used for experiments are as follows:

(a) Isolated words (TM)
The database for isolated spoken words is the Tohoku University – Matsushita Research Ltd. spoken word database uttered by 30 male speakers (212×30 words). The number of training speakers is 25 and 5 speakers are for testing. There are about 18,000 syllables for training.

(b) Read speech (JNAS)
The database for read speech is the Japan Newspaper Article Sentences database uttered by 132 male speakers (about 100 sentences per speaker). The number of training speakers is 123 and 9 for testing. There are about 480,000 syllables for training.

(c) Spontaneous speech / dialogue speech (ATRD)
The database for spontaneous speech is the database of ATR's spoken dialogue on travelers, uttered by 91

This work was supported by the Core Research for Evolutional Science and Technology (CREST).

Table 1: Speech analysis conditions

Sampling frequency	12 kHz
Pre-emphasis	$1 - 0.98z^{-1}$
Hamming window width	21.33 ms (256 points)
Frame period	8 ms (96 points)
LPC analysis order	14-th
Feature parameters	KL-20th *1 $+\Delta, \Delta\Delta$cepstrum (10 dimensions each) $+\Delta, \Delta\Delta$power

*1: Segmental statistics that consist of 4 successive frames (the parameter of each frame is 10 dimensional LPC Mel-cepstrum) and the dimensionality of which is reduced to 20 from 40 by K–L expansion[1].

Table 2: Speech rate as syllable duration

Speech database	Number of frames per syllable			Speech rate [syllables/sec] (average)
	mean	variance	peak	
TM	16.7	42.5	14	7.5
JNAS	14.1	29.1	13	8.9
ATRD	17.3	125.0	11	7.2

male speakers. The number of speakers for training is 82 and 9 for testing. There are about 22,500 syllables for training.

Speech analysis conditions are shown in Table 1.

The syllable-based HMMs having full covariance matrices consist of 5 states with 4 output distributions with duration control. There are 114 Japanese syllable HMMs. Each distribution per state is represented by 4 mixture Gaussian densities. Since there were not enough data for training from scratch for TM and ATRD, the HMMs for them are adapted by the MAP estimation method from the HMMs of JNAS.

3. ACOUSTIC DIFFERENCES BETWEEN READ AND DIALOGUE SPEECH

3.1. Speech rate

We calculated the speech rate of each speaking style as syllable duration by Viterbi alignment using the acoustic models.

Figure 1 illustrates the distribution of syllable duration of all syllables and Table 2 shows the speech rate for various speaking styles on the measure of the number of syllables per second. We can see that the rate for isolated words is low. Also, the variance for spontaneous speech is very large, although the rate is almost the same as that for isolated words (the rate is the same in terms of mean, but the peak or middle point in Fig. 1 is shorter than those of TM and JNAS). We guess this large deviation causes the difficulty of spontaneous speech recognition.

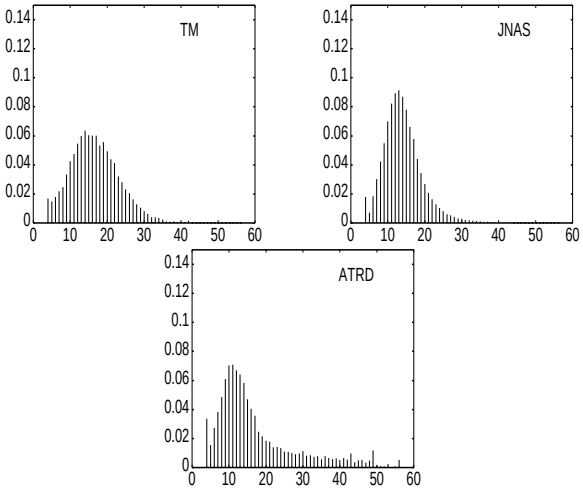


Figure 1: The distribution of syllable duration
upper left: TM, upper right: JNAS, lower: ATRD
horizontal axis: number of frames
vertical axis: frequency of occurrences

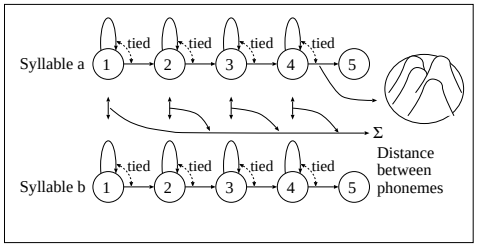


Figure 2: Distance between phonemes

3.2. Distance between phonemes

As described previously, each HMM has 5 states with 4 output distributions (last states do not have any distributions). Almost all Japanese syllables are composed of a preceding consonant and a following vowel, i.e. CV. Therefore we assumed that the first two states of HMM correspond to a consonant and the third and fourth states correspond to a vowel. As shown in Fig. 2, we calculated the distance between phonemes by using Bhattacharyya distance, that is,

$$D(a,b) = \frac{1}{M}\sum_{i=1}^{M} \min_{j,k} BD\{P_a(S_a^i, j), P_b(S_b^i, k)\} \qquad (1)$$

$$BD(P_a, P_b) = \frac{1}{8}(\mu_a - \mu_b)\left\{\frac{\Sigma_a + \Sigma_b}{2}\right\}^{-1}(\mu_a - \mu_b)^t + \frac{1}{2}\log\left(\frac{|(\Sigma_a + \Sigma_b)/2|}{|\Sigma_a|^{\frac{1}{2}}|\Sigma_b|^{\frac{1}{2}}}\right) \qquad (2)$$

Table 3: Average inter-phoneme distances

Acoustic model	Between vowels		Between consonants	
	mean	var.	mean	var.
TM	5.32	5.75	5.09	5.80
JNAS	3.63	2.38	3.71	4.01
ATRD	2.62	0.43	4.21	4.99

Table 4: Average inter-phoneme distances for speaker adapted models (average for 9 speakers)

Adaptation sentences	Between vowels		Between consonants	
	mean	var.	mean	var.
0	3.63	2.38	3.71	4.01
10	4.03	2.70	3.92	3.74
30	4.65	3.91	4.23	3.90
50	4.93	4.62	4.41	4.07
70	5.14	5.12	4.54	4.21
100	5.40	5.96	4.73	4.44

S_a^i : i-th state of syllable a
$P(S_a^i, j)$: j-th mixture component for S_a^i
$BD(P_a, P_b)$: Bhattacharyya distance between two normal distributions, P_a and P_b
μ_a : mean vector for P_a
Σ_a : covariance matrix for P_a
M : $M = 4$ for syllable consisting of a vowel and $M = 2$ for consonant

We calculated the distance between consonants using syllables having the same vowel. For example, the distance between /m/ and /n/ was calculated from the averaged distance of /ma/ and /na/, /mi/ and /ni/, /mu/ and /nu/, /me/ and /ne/, and /mo/ and /no/.

Table 3 summarizes the distances between phonemes for various speaking styles. As expected, the distance between vowels for isolated words is the largest and for read speech is the next largest. The distance for spontaneous speech is the smallest, especially for vowels, because spontaneous speech was not uttered clearly for each syllable and had a variety of speech rates. The variance for spontaneous speech is also small. We think these acoustic properties cause the degradation of recognition performance for spontaneous speech.

Next, we adapted HMMs for read speech (JNAS) to each speaker by using 10, 30, 50, 70 and 100 sentences. The distance for adapted HMMs averaged for 9 speakers is shown in Table 4. From this result, we found that the distance became larger by using more adaptation sentences, indicating that the speaker adaptation was useful for improving the speech recognition performance.

4. SPEECH RECOGNITION PERFORMANCE

We compared continuous speech recognition rates (without language models) for various speaking styles. The results in syllable recognition are shown in Tables 5 (a) for the closed test and (b) for the open test. In the tables, "Acc." denotes the accuracy, "Corr." is the correct rate of recognition and "Subst.", "Ins." and "Del." denote substitution, insertion and deletion error rate, respectively.

Table 5: Continuous syllable recognition results [%]

(a) CLOSE test

Model	Subst.	Ins.	Del.	Corr.	Acc.
TM	5.4	9.8	0.6	94.0	84.2
JNAS	16.5	9.9	2.7	80.9	71.0
ATRD	14.8	18.0	4.6	80.6	62.6

(b) OPEN test

Model	Subst.	Ins.	Del.	Corr.	Acc.
TM	18.8	18.8	4.1	77.1	58.4
JNAS	19.2	5.5	3.5	77.3	71.9
ATRD	25.0	20.2	7.9	67.1	47.0

Table 6: Continuous syllable recognition results for speaker adapted models on OPEN test [%]

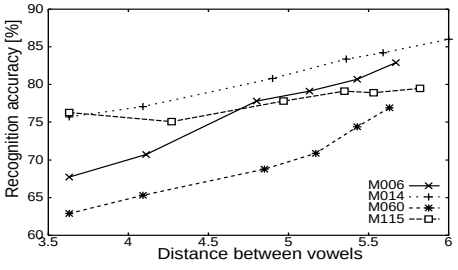
Adaptation sentences	Subst.	Ins.	Del.	Corr.	Acc.
0	19.2	5.5	3.5	77.3	71.9
10	16.9	8.5	2.5	80.7	72.2
30	13.8	7.8	2.3	83.9	76.2
50	12.6	7.3	2.2	85.2	78.0
70	11.6	6.9	2.1	86.2	79.4
100	10.9	6.5	2.0	87.1	80.6

For the closed test (for the training speakers), we got high recognition accuracy for words spoken in isolation (TM). It seems that recognition rates are proportional to inter-phoneme's distances described in the previous section. However, for the open test (for the open speakers), there are irregular results, such as those for isolated words. We guess this is caused by insufficient training samples for TM. For both test sets, we got poor accuracy for spontaneous speech (ATRD), mainly caused by the variety of speech rates and smaller distances between phonemes in spontaneous speech.

Next, we investigated speech recognition rates of adapted acoustic models for read speech (JNAS). The results in syllable recognition are shown in Table 6. We found that a speaker adaptation technique was useful for improving the speech recognition performance, and that recognition rates were proportional to distances between phonemes.

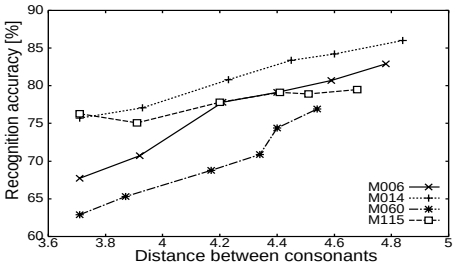
5. RELATIONSHIP BETWEEN RECOGNITION PERFORMANCE AND INTER-PHONEME'S DISTANCE OR LOG LIKELIHOOD

Figures 3 and 4 illustrate the relationship between inter-phoneme's distance and recognition performance for speaker adapted models (typical 4 speakers), for distance between



average of correlation coeffs. for each speaker = 0.967
correlation coefficient for all speakers = 0.829
(In the case of ATRD, 0.987 and 0.796, respectively)

Figure 3: Relationship between inter-vowel distance and recognition performance for speaker adaptation models



average of correlation coeffs. for each speaker = 0.975
correlation coefficient for all speakers = 0.844
(In the case of ATRD, 0.983 and 0.827, respectively)

Figure 4: Relationship between inter-consonant distance and recognition performance for speaker adaptation models

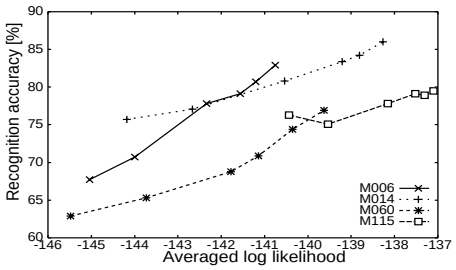
vowels and distance between consonants, respectively. The figures show that the recognition performance depends on the speaker, even if the same acoustic models are used (or, inter-phoneme distances are the same). So, strictly speaking, only the averaged performance is meaningful in the comparison. From these figures, there are very strong correlations between inter-phoneme's distance and the recognition performance. In other words, the recognition rates are approximately proportional to distances between phonemes.

Figure 5 show the relationship between averaged log likelihood of recognition results and recognition performance. There are also very strong correlations between the log likelihood per frame and the recognition performance. The recognition rates are also approximately proportional to the log likelihood.

6. CONCLUSION

In this paper, we investigated the relationship between acoustic differences of speaking style and recognition performance for isolated words, read and dialogue speech.

We found that the speech rate of dialogue speech was faster



average of correlation coeffs. for each speaker = 0.976
correlation coefficient for all speakers = 0.769
(In the case of ATRD, 0.996 and 0.626, respectively)

Figure 5: Relationship between log likelihood and recognition performance for speaker adaptation models

on average and dialogue speech had a larger variance of speech rate than read speech. We also found that the inter-phoneme's distance in dialogue speech was smaller than that in read speech and the phonemes in dialogue speech were more easily confused. In the continuous syllable recognition experiments, more insertion and deletion errors occurred in dialogue speech. This result was caused by the variety of speech rate and the small inter-phoneme's distance.

As future works, we will study a compensation method for variety of speech rate and lexical expression for variety of utterances for spontaneous speech.

7. REFERENCES

[1] K. Hanai, K. Yamamoto, N. Minematsu, S. Nakagawa: Continuous speech recognition using segmental unit input HMMs with a mixture of probability density functions and context dependency, Proc. ICSLP-98, pp.2935–2938, 1998.

[2] D.S. Palett, J.G. Fiscus, J.S. Garofolo, A. Martin and M. Przybocki: 1998 Broadcast news benchmark test results: English and non-English word error rate performance measures, Proc. DARPA Broadcast News Workshop, pp.5–12, 1999.

[3] E. Blaauw: Phonetic differences between read and spontaneous speech, Proc. ICSLP-92, pp.751–754, 1992.

[4] M. Eskénazi: Changing speech styles: strategies in read speech and casual and careful spontaneous speech, Proc. ICSLP-92, pp.755–758, 1992.

[5] H. Kuwabara: Acoustic and perceptual properties of phonemes in continuous speech as a function of speaking rate, Proc. Eurospeech-97, pp.1003–1006, 1997.

[6] J. Murakami and S. Sagayama: A discussion of acoustic problems in spontaneous speech recognition, IEICE Trans., Vol.78-D-II, No.12, pp.1741–1749, 1995. (in Japanese)