STUDY OF INTRA-SPEAKER'S SPEECH VARIABILITY OVER LONG AND SHORT TIME PERIODS FOR SPEECH RECOGNITION

Satoru TSUGE, Masami SHISHIBORI, Kenji KITA, Fuji REN, Shingo KUROIWA

The University of Tokushima
2-1, Minami Josanjima, Tokushima, Japan

ABSTRACT

In this paper, we describe a Japanese speech corpus collected for investigating the speech variability of a specific speaker over short and long time periods and then report the variability of speech recognition performance over short and long time periods. Although speakers use a speaker-dependent speech recognition system, it is known that speech recognition performance varies pending when the utterance was uttered. This is because speech quality varies by occasion even if the speaker and utterance remain constant. However, the relationships between intra-speaker speech variability and speech recognition performance are not clear. Hence, we have been collecting speech data to investigate these relationships since November 2002. In this paper, we introduce our speech corpus and report speech recognition experiments using our corpus. Experimental results show that the variability of recognition performance over different days is larger than variability of recognition performance within a day.

1. INTRODUCTION

Recently, speech recognition systems, such as car navigation systems and cellular phone systems have come into wide use. Although each speaker uses a speaker-dependent speech recognition system, it is known that speech recognition performance varies pending when the utterance was uttered. For this reason, we hypothesized that speech characteristics vary even though the speaker and utterance remain constant. This intra-speaker's speech variability is caused by some factors including emotion, background noise, and so on. If the recognition performance is not consistent, then products using speech recognition systems become less useful for the end-user. However, as the relationships between intra-speaker's speech variability and speech recognition performance are yet unclear, we began to investigate the nature of this relationship.

In the field of speaker identification and verification, it has been reported that speaker verification performance degraded for a standard set of templates after only a few months[1][2]. However, we have not seen this experiment applied to Japanese speech recognition. At present, there are a lot of Japanese speech corpora for studying speech recognition[3][4][5]. However, we have not seen a corpus of Japanese speech data of a specific speaker over a long time period. Hence, we have not been able to investigate the relationships between the intra-speaker's speech variability over time and speech recognition performance. In order to examine the time-related intra-speaker's speech variability and its influence on speech recognition performance, we needed a new corpus. Consequently, we started collecting some specific speakers' read speech data. Data collection was initiated in November 2002. It is still underway as of October 2005.

In this paper, we introduce the speech corpus collected by us for investigating intra-speaker's speech variability over long and short time periods and report the speech recognition experimental results using a part of our speech corpus. In addtion, it has been reported that the speaker adaptation technique, Maximum Likelihood Linear Regression (MLLR), is effective in coping with intra-speaker's variability in speech recognition[6]. Hence, in this paper, we investigate the MLLR's effects on recognition performance over the speaking day and time. In this paper, the speaking time refers to the time when we collected the speech data, such as morning, afternoon, or evening.

2. OUR SPEECH CORPUS

There are a lot of Japanese speech corpora for studying speech recognition[3][4][5]. Most of these speech corpora are designed for studying speaker independent speech recognition systems. Hence, the amount of speech data from one speaker is limited and often collected on one day. Using these speech corpora, it is impossible to investigate a speaker's speech variability over long time periods and relationships between speaker's speech variability and speech recognition performance. In addition, these corpora lack information about speakers, such as the physical condition of the speaker, environmental condition of recording, and so on. To investigate what caused the variability of the speaker's speech, we need this information. Consequently, we began collecting speech data of some specific speakers uttered over a long time period. In our corpus, the speaker fills out a questionnaire which is described in section **2.5** at each recording session. Since November 2002, we have been collecting speech data for investigating the relationships between the speech variation and speech recogni-

tion performance. In this section, we describe our Japanese speech corpus.

2.1. Speakers and recording days

Our corpus consists of six speakers' speech data. The number of male speakers and the number of female speakers are four and two, respectively. Each speaker read utterance sets, described in section **2.3**, three times a day, once a week. The length of each recording was about fifteen minutes.

2.2. Recording environments

Our corpus was collected in one of the two following types of recording environments,

- Schoolroom
  We used a quiet school room for recording from November 2002 to October 2003.
- Silent room
  We used a silent room for recording from October 2003 to the present.

2.3. Utterance sets

We used two utterance list sets for recording. In this paper, we call these Common recording set and Individual recording set. The contents of each are described below:

- Common recording set
  - Japanese phonetically balanced sentences (The number of sentences is 50. These sentences are called the A set.)
  - Isolated words (The number of words is 10.)
  - Name words (The number of words is 10.)
  - 4 digit strings (The number of items is 10.)
  - Checked sentences (The number of sentences is 16.)
- Individual recording set
  - Japanese phonetically balanced sentences
  - Isolated words
  - 4 digit strings
  - Japanese newspaper sentences

  The items in the individual and common recording set differ.

All speakers uttered the common recording set at every recording session. The length of this recording set, which included non-voiced sections and mistaken sections, is about thirteen minutes. The contents of the individual recording set were different at each recording session.

2.4. Recording file format

We used the head set microphone, Sennheiser HMD410, and the DAT recorder, Sony TCD-D100, for the recording system. The DAT's sampling rate is 48 kHz. Using a DAT link, we copied the recorded speech data from the DAT to the computer. Then, we divided this speech data file to individual speech data. Finally, we resampled the speech data at 16kHz.

2.5. Questionnaire

For investigating the reason of intra-speaker's speech variability, the speaker filled out a questionnaire at every recording session. The contents of the questionnaire are listed below:

- Physical conditions
  - Body temperature
  - Weight
  - Percentage of body fat
  - Pulse rate
  - Blood pressure
  - Feeling or Mood
  - Condition of nose and throat
- Environmental conditions
  - Outdoor temperature
  - Outdoor humidity
  - Temperature in recording room
  - Humidity in recording room
  - Day of recording
  - Time of recording

In addition, when the speaker has finished recording for the day, the speaker answers some questions about today's activities and the hours of sleep yesterday.

3. EXPERIMENT

For investigating speech recognition variability over long and short time periods, we conducted a speaker-dependent continuous speech recognition experiment. In this experiment, we used a female's speech data collected in a schoolroom environment.

3.1. Experimental Conditions

3.1.1. Training data

502 Japanese phonetically balanced sentences were used for the training.These training sentences were uttered on 2002/ 11/12, 19, 26, 2002/ 12/3, 10, 17, 24, 2003/ 1/14, 21.

3.1.2. Testing data

For the testing data, we used 50 kinds of Japanese phonetically balanced sentences, which are a part of the common recording set. These sentences were uttered three times in each recording day. These sentences were recorded from 2002/ 11/19 to 2003/ 10/03. The total number of recording days

Table 1. Schoolroom environment (phoneme accuracy (in %))

	Recording days														
	2002							2003							
	1119	1126	1203	1210	1217	1224	1231	0107	0114	0121	0128	0204	0211	0218	0225
Morning	78.5	75.8	76.3	73.1	76.7	77.2	80.0	75.9	78.1	74.9	74.5	77.4	71.2	70.6	74.6
Afternoon	78.8	78.1	79.2	74.7	77.9	77.4	78.2	74.1	74.1	75.4	74.7	72.8	70.6	70.4	71.9
Evening	77.6	75.9	78.1	76.1	78.4	78.2	76.2	76.6	76.7	77.2	76.0	72.0	74.0	72.2	73.5
Average	78.3	76.6	77.8	74.6	77.7	77.6	78.1	75.5	76.3	75.8	75.0	74.0	71.9	71.1	73.3

	2003														
	0304	0312	0318	0326	0402	0407	0414	0421	0428	0506	0512	0519	0526	0603	0610
Morning	74.9	70.0	72.9	76.2	75.2	77.2	76.7	76.4	74.3	71.5	73.0	67.5	74.7	74.9	75.0
Afternoon	73.0	70.8	75.5	75.0	74.1	71.3	74.1	73.1	73.8	74.8	74.3	66.0	74.5	73.1	71.3
Evening	75.4	72.5	72.9	77.4	75.7	74.9	77.8	76.7	74.5	74.1	73.6	70.0	77.0	75.8	73.9
Average	74.4	71.1	73.8	76.2	75.0	74.5	76.2	75.4	74.2	73.5	73.7	67.8	75.4	74.6	73.4

	2003															
	0617	0624	0701	0708	0715	0722	0805	0811	0817	0830	0901	0910	0916	0923	1003	Ave
Morning	74.2	75.1	74.7	78.1	75.3	74.8	77.1	76.8	76.0	76.6	74.6	71.4	75.6	74.8	75.3	75.0
Afternoon	71.1	74.4	75.3	77.1	74.2	76.4	78.0	76.9	78.3	76.6	75.1	73.1	77.1	74.7	74.0	74.6
Evening	74.8	75.8	77.1	72.9	73.4	76.0	78.3	78.0	76.1	76.3	76.6	77.0	77.1	78.5	74.2	75.6
Average	73.4	75.1	75.7	76.1	74.3	75.71	77.8	77.2	76.8	76.5	75.4	73.8	76.6	76.0	74.5	75.1

was 47. For the testing set, the total number of utterances was 6,747[1].

3.1.3. Feature vector and acoustic model

The feature vector for the experiment was 25 MFCCs (12 static MFCCs + 12 of their delta MFCCs + one delta-logpower). CMS was performed on this feature vector. In our preliminary experiment, the recognition performance of the speaker-dependent model is lower than that of the speaker-adaptation model. Hence, in this experiment, we used the speaker-adaptation shared-state triphone HMMs with sixteen Gaussian mixture components per state for the acoustic model. The number of the states of this model was 3,000. This speaker-adaptation model was adapted by the MLLR technique using the training data described in *3.1.1*.

3.1.4. Decoder and evaluation

For the decoder, we used the one-pass Viterbi algorithm with the phonotactic constraints of Japanese language expressed. Recognition results are given as phoneme accuracy. We used HTK version 3.2.1 as the acoustic modeling and recognition tools.

We calculated the variance of the recognition accuracy to investigate the variability. To investigate the influence of speaking time, we calculated the variance of the recognition accuracy as in equation (1).

$$V_t = \frac{\sum_{d \in date} (ACC_{d,t} - AVE_t)^2}{N_{date}}, \qquad (1)$$

where, $date$ indicates all speaking days, $ACC_{d,t}$ and AVE_t are the recognition accuracy of speaking day d and speaking time t and the average recognition accuracy of speaking time t. N_{date} is the number of speaking days. To investigate the influence of the speaking day, we also calculated the variance of the recognition accuracy as in equation (2).

$$V = \frac{\sum_{d \in date} \sum_{t \in time} (ACC_{d,t} - AVE_d)^2}{N_{date} * N_{time}}, \qquad (2)$$

where, $time$ indicates the all speaking times. AVE_d indicates the average recognition accuracy of the speaking day d. N_{time} is the number of speaking times in a day.

3.2. Experimental results

Our experimental results show the influence of speaking days and speaking times on recognition accuracy. Table 1 shows the phoneme acuraccy in the schoolroom recording environment. Table 2 shows the variances which were calculated by equations (1) and (2).

Table 1 shows that the recognition performances from 2002/ 11/19 to 2003/ 1/21 are higher than other periods. These are the days when the training data was recorded. Since the training data and the testing data were recorded on the same days, we hypothesize that there are few acoustic mismatches between the training data and the testing data. However, we can see from this table that the recognition performances on other days degraded compared to the training data recording days. Speech variability on different days was greater than the speech variability within a day.

From Table 2, we can see that the variance of the speaking day, V is smaller than the variance of the speaking time,

[1]For recording mistakes, the number of testing sentences is 49 in afternoon on 2003/1/14, morning on 2003/4/28, and morning on 2003/7/1

Table 2. Variance of the recognition accuracy against speaking day and speaking time

	V	V_m	V_a	V_e
schoolroom	1.57	5.58	6.97	3.94

V_m, V_a, V_e. This implies that the variability within a day is smaller than for a set time over our testing period.

3.3. Speaking day and time adaptation

For coping with intra-speaker's speech variability against speaking day and time, we try to adapt the acoustic model used in the previous experiment, to each speaking day and time. Experimental conditions were the same as in section **3.1**.

Figure 1 shows the average phoneme accuracy in a day. In this figure, the "baseline" indicates phoneme accuracy of the baseline model which was used in the previous experiment. The "current" indicates the phoneme accuracy of the model which was adapted using the testing data, i.e., closed test condition. The "pre-time" indicates the phoneme accuracy of the model which was adapted using the 50 sentences uttered previous speaking session, i.e., if the afternoon speech were recognized, the acoustic model was adapted using the 50 sentence uttered on morning in same day[2]. The "last-week" indicates the phoneme accuracy of the model which was adapted using the 50 sentences uttered at the same speaking session the last week.

This figure shows that the phoneme accuracy of the pre-time line is higher than that of the last-week line. The last-week line shows results similar to the baseline. From these results, we can see that an intra-speaker's speech variability within a day is smaller than that over different days. In addition, we can see from this figure that the phoneme accuracy of the baseline, last-week, and pre-time line show almost the same results as when the training data were uttered, from 2002/11/19 to 2003/1/21.

4. SUMMARY

In this paper, we described a Japanese speech corpus for investigating speech variability in a specific speaker over long and short time periods. This corpus has been collected by us since November 2002. The data collection is still ongoing. We have collected utterances from six speakers. Each speaker spoke three times a day once a week.

Using a part of our speech corpus, we conducted speaker-dependent speech recognition experiments. Experimental results show that recognition performance degraded when there are acoustic mismatches between testing and training data collected over different periods.

[2]In this experiment, when a morning test set was tested, we used 50 sentences uttered on that testing day's evening for adaptation data

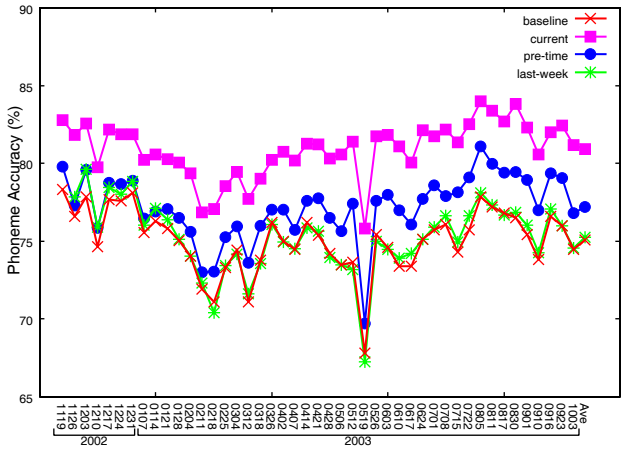

Fig. 1. Average phoneme accuracy in a day (in %)

In the future, we will continue to collect speech data and enhance our speech corpus. We will conduct speech recognition experiments using other speaker's speech data and investigate the influence of recording period on recognition performance.

5. ACKNOWLEDGMENT

This research has been partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), 15700163 and the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 17300065 & 17300036.

6. REFERENCES

[1] T. Matsui, T. Nishitani, and S. Furui, "A study of model and a priori threshold updating in speaker verification," *IEICE (D-II)*, Vol. J81-DII, pp. 268–276, 1998, (in Japanese).

[2] S. Hayakawa, K. Takeda, and F. Itakura, "A speaker verification method which can control false acceptance rate," *IEICE (D-II)*, Vol. J82-DII, pp. 22212–2220, 1999, (in Japanese).

[3] K. Ito, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus," *Proc. ICSLP*, pp. 3261–3264, 1998.

[4] A. Nakamura, S. Matsunaga, T. Shimizu, M. Tonomura, and Y. Sagisaka, "Japanese speech databases for robust speech recognition," *Proc. ICSLP*, pp. 2199–2202, 1996.

[5] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," *Proc. of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*, pp. 7–2, 2003.

[6] B. Li, K. Hirose, and N. Minemtsu, "Robust speech recognition using inter-speaker and intra-speaker adaptation," *Proc. ICSLP*, pp. 1397–1400, 2002.