Specific Speaker's Japanese Speech Corpus over Long and Short Time Periods

Satoru Tsuge[1], Masami Shishibori[1], Fuji Ren[1],
Kenji Kita[2], and Shingo Kuroiwa[1]

[1] Faculty of Engineering, the University of Tokushima,
2-1, Minami Josanjima, Tokushima, Japan
tsuge@is.tokushima-u.ac.jp

[2] Center for Advanced Information Technology, the University of Tokushima,
2-1, Minami Josanjima, Tokushima, Japan

Abstract. It is known that speech recognition performance varies pending when the utterance was uttered although speakers use a speaker-dependent speech recognition system. This implies that the speech varies even if a specific speaker utters a specific sentence. Hence, we investigate the speech variability of a specific speaker over short and long time periods for getting the stable speech recognition performances. For this investigation, we need a specific speaker's speech corpus which is recorded over long time periods. However, at present, we have not seen such a Japanese speech corpus. Therefore, we have been collecting the Japanese speech corpus for investigating the relationship between intra-speaker speech variability and speech recognition performance. In this paper, first, we introduce our speech corpus. Our corpus consists of six speakers' speech data. Each speaker read specific utterance sets tree times a day, once a week. Using a specific female speaker's speech data in this corpus, we conduct speech recognition experiments for investigating the relationship between intra-speaker speech variability and speech recognition performance. 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 a 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 consider that speech characteristics varies even though the speaker and utterance remain constant. This intra-speaker variability is caused by some factors including emotion and background noise. If the recognition performance is not consistent, then products using speech recognition systems become less useful for the end-user. As the relationship between intra-speaker's speech variability and speech recognition performance is yet unclear, we began to investigate the nature of this relationship.

© A. Gelbukh (Editor)
Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 115-124

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 evaluation 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 and speech recognition performance. In order to examine the intra-speaker's speech variability and its influence on speech recognition performance, we need a new corpus. Consequently, we started collecting some specific speaker's read speech data. Data collection was initiated in October 2002. It is still underway as of December 2005. In this paper, we describe the speech corpus collected by us for investigating intra-speaker's speech variability.

Because the initial speech data which were collected at one recording time was one file, we need to divide one recorded file into separate utterances. This process requires a lot of time and effort. In this paper, we propose an automatic utterance segmentation tool for dividing one recored file into separate utterances.

At present, we have some speech data which has been processed using this segmentation tool. We conducted the speech recognition experiments using these speech data. In this paper, we report phoneme accuracy on each speaking day and at each speaking time.

In the following section, we introduce the our database. In section 3, we propose the automatic utterance segment tool for processing our database. In section 4, we show the experimental conditions and results. In the last section (section 5), we describe the summary of this paper and future works.

2 A Japanese speech corpora

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 difficult to investigate the 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 physical condition of 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 September 2002, we have been collecting speech data for investigating the relationships between the speech variation and speech recognition 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 the 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 Aset.)
  - 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 recoding set differ.

All speakers uttered the common recording set at every recording session. The length of this recording set, which included non-voice 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 DAT link, we copied the recorded speech data from DAT to the computer. At that time, the number of recorded speech data files was one because this data included the non-voice sections and the mistaken sections. 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, after the last recording session in a day, the speaker answers some questions about today's activities and the hours of sleep yesterday.

3 Automatic utterance segment tool

The speaker has to check each utterance each time if we collect the utterances individually. In general, the speech is recorded as one file from recording start to recording end. Hence, there are some noises, non-voiced sections and mistaken sections in this file. For our speech corpus, we have to segment this file into individual utterances and select the useful utterances. However, this process requires a lot of time and effort. In this paper, we propose an automatic utterance segmentation tool for dividing the recorded speech files.

Figure 1 illustrates the flow of the proposed automatic utterance segmentation tool. Below, we explain each part of the proposed method.

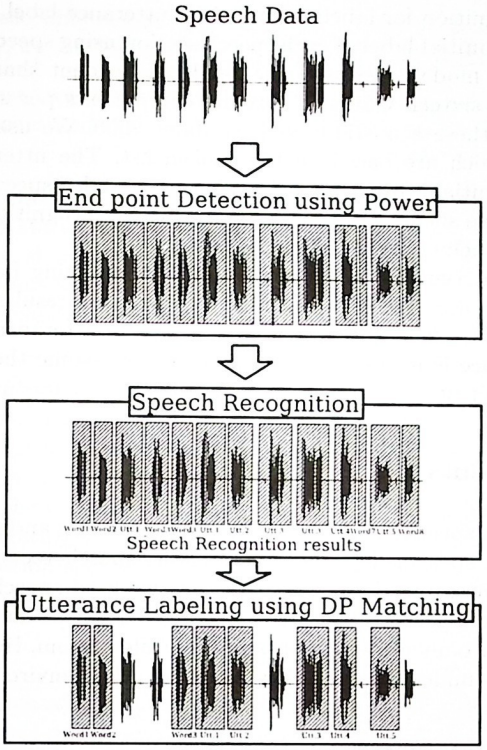

Fig. 1. The flow of the automatic utterance segment tool.

1. End point detection using power
   First, we find the end point of each utterance in the speech data using power. The power of m-th frame ($P(m)$) is calculated as follows,

$$P(m) = \frac{\sum_{n=0}^{T-1}(s((m*S)+n))^2}{T}, \tag{1}$$

   where T, S, and $s(n)$ are the the frame length, the frame shift, and speech signal, respectively. Using this, we detect the end point. In this paper, the frame length and frame shift are set at 5 msec and 2 msec, respectively. If the power of a frame following a silence is bigger than the threshold, this frame is the start of the speech section. Using the threshold selection method[6], we determine the threshold for detecting the end point. After the start section, we find the next silint segment using power then we define the end point of the speech section. However, we ignore sections shorter than 300msec. We added a 1 second silence to speech sections.

2. Speech recognition for labeling the initial utterance label
   We made on initial label of each speech section using speech recognition. For the acoustic models, we used the gender dependent shared-state triphone HMMs with sixteen Gaussian mixture components per state. The number of states of these acoustic models is about 2000. We use a dictionary and grammar which are based on the spoken list. The utterance sections are labeled the initial utterance labels which are included speech list. In addition, we modify the silent period to 0.5 second using recognition result.
3. Utterance labeling using DP matching
   After speech recognition, we conducted DP matching between the spoken list, which is a corrected list, and the recognition results. This process removed the mistaken sections from the separate utterances. In this process, if an utterance is uttered more than once, we assume the last utterance to be the correct utterance.

4 Experiments

For investigating speech recognition variability over long and short time periods, we conducted a speaker-dependent continuous speech recognition experiment using our speech corpus. In this experiment, we used speech data which were downsampled from 48kHz to 16kHz. The collected speech data were recorded in two types of environments: a schoolroom and a silent room. Hence, we conducted two experiments under the condition of the recording environments.

4.1 Experimental Conditions

Training data 502 Japanese phonetically balanced sentences were used for the training for schoolroom recording environment and 503 Japanese phonetically balanced sentences were used for training for the silent room recording environment, respectively. These training sentences were uttered on following days:

- Schoolroom recording environment
  2002.11.12, 19, 26, 2002.12.3, 10, 17, 24, 2003.1.14, 21
- Silent room recording environment
  2004.11.9

Testing data For the testing data, we used 50 kinds of Japanese phonetically balanced sentences. These sentences were uttered three times in a each recording day. These sentences were recorded on the following days:

- Schoolroom recording environment
  2002.11.19, 26, 2002.12.3, 10, 17, 24, 31, 2003.1.7, 14, 21, 28, 2003.2.4, 11, 18, 25, 2003.3.4, 12, 18, 26, 2003.4.2, 7, 14, 21, 28, 2003.5.6, 12, 19, 26, 2003.6.3, 10, 17, 24, 2003.7.1, 8, 15, 22, 2003.8.5, 11, 17, 30, 2003.9.1, 10, 16, 23, 2003.10.03

- Silent room recording environment
  2003.10.10, 17, 24, 31, 200311.07, 15, 21, 28, 2004.1.5, 9, 16, 23, 30, 2004.2.6.

For the testing set, we used 6,747 and 2,099 utterances in the schoolroom recorded environment and the silent room recorded environment, respectively[3].

Feature vector and acoustic model The feature vector for the experiment was 25 MFCCs (12 static MFCCs + 12 of their deltas + one delta-logpower). For the acoustic model, shared-state triphone HMMs with sixteen Gaussian mixture components per state were trained. We set the number of states at about 270. We used the same conditions in both recording environments.

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 use HTK version 3.2.1[7] as acoustic modeling and recognition tools.

We calculated the variance of recognition accuracy for investigating the variability. For investigating the influence of speaking time, we calculated the variance of the recognition accuracy in equation (2).

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

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 time, we also calculated the variance of the recognition accuracy in equation (3).

$$V = \frac{\sum_{d \in date} \sum_{t \in time} (ACC_{d,t} - AVE_d)^2}{N_{date} * N_{time}}, \tag{3}$$

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.

4.2 Experimental results

In this section, we present the experimental results in order to investigate the influence of speaking days and speaking time. Tables 1 and 2 show the speech recognition performance on the schoolroom recording environment and on silent room recording environment, respectively. Table 3 shows the variances which were calculated by equation (2) and equation (3).

[3] For recording mistakes, the number of testing sentences is 49 in afternoon on 2003.1.14, morning on 2003.4.28, morning on 2003.7.1, and evening on 2003.11.28

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

	Recording days											
	2002							2003				
	1119	1126	1203	1210	1217	1224	1231	0107	0114	0121	0128	0204
Morning	78.5	76.5	76.4	73.8	75.6	77.3	78.6	75.0	75.1	72.9	72.4	74.4
Afternoon	78.1	76.6	81.6	73.8	76.7	76.3	75.9	72.6	73.3	72.7	72.7	70.9
Evening	76.9	75.6	76.6	75.5	77.4	77.5	76.3	75.2	74.0	75.3	75.0	69.4
Average	77.9	76.2	78.2	74.4	76.6	77.0	76.9	74.3	74.1	73.6	73.4	71.6
	2003											
	0211	0218	0225	0304	0312	0318	0326	0402	0407	0414	0421	0428
Morning	69.5	67.4	71.6	73.6	67.7	71.4	72.9	74.3	74.7	73.8	75.4	71.9
Afternoon	67.9	68.4	70.2	71.2	71.1	73.4	71.7	74.1	68.7	71.2	71.5	71.4
Evening	71.4	69.6	71.5	73.4	70.7	73.0	75.1	73.6	72.4	74.5	73.5	72.8
Average	69.6	68.5	71.1	72.7	69.8	72.6	73.2	74.0	71.9	73.1	73.5	72.0
	2003											
	0506	0512	0519	0526	0603	0610	0617	0624	0701	0708	0715	0722
Morning	69.4	72.7	62.5	72.5	76.2	75.0	73.9	76.2	75.6	77.1	73.3	74.9
Afternoon	72.5	71.7	64.4	74.2	73.2	70.3	71.1	75.0	74.5	76.2	73.2	75.5
Evening	73.5	72.6	68.6	74.0	74.9	73.9	73.0	75.7	75.9	74.9	75.0	74.5
Average	71.8	72.3	65.1	73.6	74.8	73.1	72.7	75.6	75.3	76.1	73.8	75.0
	2003									Average		
	0805	0811	0817	0830	0901	0910	0916	0923	1003			
Morning	76.3	73.4	75.4	75.6	74.5	70.6	74.3	73.8	73.1	73.7		
Afternoon	74.7	74.6	75.4	71.8	73.1	72.8	74.7	71.8	70.9	73.0		
Evening	76.1	76.7	73.9	73.6	74.8	75.4	73.9	74.1	74.7	74.1		
Average	75.7	74.9	74.9	73.7	74.1	73.0	74.3	73.2	72.9	73.6		

Comparison of recognition performances

Table 1 shows that the recognition performances from 2002.11.19 to 2003.1.21 are higher than other days. These are the days when the training data was recorded. Because the recording days of the training data and the testing data were the same, we consider 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. I believe that the speech variability on different days was greater than the speech variability within a day.

We can see from Table 2 that the recognition performances are consistent compared to the schoolroom recording environments. We consider that the speakers became used to speaking utterances. Hence, the acoustic feature vectors vary little in testing periods.

Compared to Table 1 and 2, we can see that the phoneme accuracies of the silent room environment are lower than those of the schoolroom environment. We suppose that the reason is that the period between testing and training data in the silent room recording environment is 7 months.

On the other hand, we can see from Table 1 that the recognition performances of 2003.5.19 are lower than that of other days. Hence, we investigated the questionnaire of 2003.5.19. This questionnaire showed that the speaker caught a cold. From this result, we can confirm that the physical condition influences the recognition performance. We will investigate the detail of this result and the relationships between the emotion and recognition performance from the questionnaires.

Table 2. Silent room environment (phoneme accuracy (in %))

	Recording days										
	2003									2004	
	1010	1017	1024	1031	1107	1115	1121	1128	0105	0109	0116
Morning	72.0	70.6	68.5	68.4	71.7	67.9	70.6	70.6	71.5	70.7	72.9
Afternoon	72.2	73.4	67.3	70.8	72.0	73.5	73.2	70.6	70.8	73.3	73.6
Evening	72.1	70.2	75.1	72.9	72.3	73.2	70.7	74.6	71.4	72.0	73.7
Average	72.1	71.4	70.3	70.7	72.0	71.5	71.5	71.9	71.2	72.0	73.4

	2003			
	0123	0130	0206	Average
Morning	72.1	69.4	68.7	70.4
Afternoon	72.3	70.9	70.2	71.7
Evening	73.7	70.4	72.1	72.4
Average	72.7	70.2	70.3	71.5

Table 3. Variance of the recognition accuracy against speaking day and speaking time

recording environment	V	V_m	V_a	V_e
schoolroom	1.60	8.88	8.36	4.01
silent room	2.32	2.91	2.12	2.38

Comparison of variances

From Table 3, we also see that the difference between the variance of the speaking time in the silent room (V_m, V_a, V_e) is smaller than in the schoolroom. The training data for the silent room was collected in one day although the training data for the schoolroom was collected in two months. Hence, the acoustic model for the schoolroom may be able to be trained using the variation of the speech resulting from a longer collection period.

5 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 October 2002. The data collection is still ongoing. We have collected six speakers' utterances. Each speaker spoke three times in a day once a week. In addition, we proposed an automatic utterance segmentation tool for dividing one recorded file to individual useful utterances.

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.

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 the recognition performance.

6 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.

References

1. Matsui, T., Nishitani, T. , and Furui, S.: "A study of model and a priori threshold updating in speaker verification," *IEICE (D-II)*, Vol. J81-DII, No. 2, pp. 268–276, 1998, (in Japanese).
2. Hayakawa, S., Takeda, K., and Itakura, F.: "A speaker verification method which can control false acceptance rate," *IEICE (D-II)*, Vol. J82-D-II, No. 12, pp. 22212–2220, 1999, (in Japanese).
3. Ito, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T., Shikano, K., and Itahashi, S.: "The design of the newspaper-based japanese large vocabulary continuous speech recognition corpus," *Proc. ICSLP*, pp. 3261–3264, 1998.
4. Nakamura, A., Matsunaga, S., Shimizu, T., Tonomura, M., and Sagisaka, Y.: "Japanese speech databases for robust speech recognition," in *Proc. ICSLP*, pp. 2199–2202, 1996.
5. Maekawa, K.: "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. Otsu, N.: "A threshold selection method from gray-level histograms," *IEEE Trans. Sys., Man, and Cybernetics*, Vol. SMC-9, No.1, pp. 62–66, 1979.
7. "HTK: Hidden Markov Model Toolkit," http://htk.eng.cam.ac.uk/.