Japanese Speech Databases for Robust Speech Recognition

Atsushi Nakamura, Shoichi Matsunaga, Tohru Shimizu,*
Masahiro Tonomura and Yoshinori Sagisaka
ATR Interpreting Telecommunications Research Labs.
2-2 Hikaridai, Seika-Cho, Soraku-Gun, Kyoto, 619-02 JAPAN
* Currently at NTT Human Interface Labs.

ABSTRACT

At ATR, a next-generation speech translation system is under development towards natural trans-language communication. To cope with the various requirements to speech recognition technology for the new system, further research efforts should emphasize the robustness for large vocabulary, speaking variations often found in fast spontaneous speech and speaker variances. These are key problems to be solved not only for speech translation but also for the general use of speech recognition in real environments. In this paper, three large speech databases are designed to cope with these problems in speech recognition and the current status of data collection is reported.

1. Introduction

Since its foundation in 1986, ATR has taken the initiative in building a large database that meets the needs of a variety of studies in speech research. In 1987, we started public release of our first Japanese speech database (ATR-JSDB) which consists of several types of speech materials, such as the isolated words, phonetically balanced sentences, read speech conversations for "Conference registration" task and so on [1]. So far, the ATR-JSDB has been used as a standard Japanese speech database and has greatly contributed to progress in the research of Japanese speech recognition, as well as our success in developing a speech translation system (e.g. [2]).

We are currently developing a next-generation speech translation system towards more natural trans-language communication that requires speech recognition to be much more robust against large vocabulary, speaking variations often found in fast spontaneous speech and speaker variances. And, these are common problems for the general use of speech recognition in real environments. This paper describes three Japanese speech databases which we are now building and are well-considered for the research on the robust speech recognition. In Section 2, the basic concepts for our new speech databases are described with the feature of each database. In Section 3, the task used in the databases are introduced. All speech data are collected and transcribed as shown in Section 4. In Section 5, the current status of data collection are reported and some interim evaluations for the collected data are given. Section 6 introduces our current research using the collected data. And, in Section 7, future plans concerning the databases are discussed.

2. Basic concepts

In order to realize robust speech recognition, taking into account the characteristics of the problems, we are now concentrating on two different research targets: to overcome large vocabulary and high spontaneity, separately (Fig. 1). Two of our new databases correspond to these research targets.

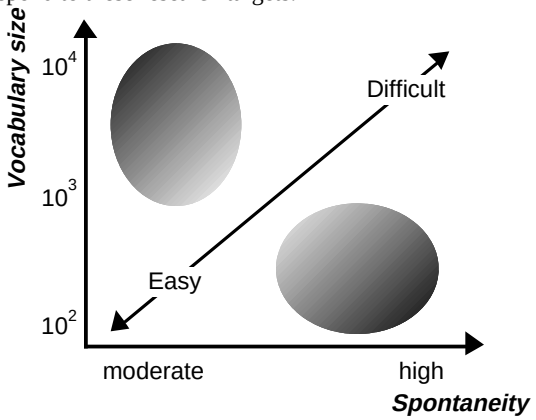


Fig. 1 Two research targets & difficulty of recognition

The first database is designed for *large vocabulary*. At ATR, an integrated speech and language database (SLDB) [3], which comprises roughly 600 bilingual (Japanese-English) dialogues, had previously been constructed as a common material for speech and language research in the development of the speech translation system. Based on the SLDB, our large vocabulary database is designed to be a Japanese monolingual dialogue speech database of moderate spontaneity and fairly large-sized (~10^4) vocabulary.

The second one is designed for *high spontaneity.* In order to build a robust recognition system against the speaking variation found in real fast spontaneous speech, spontaneity of speech in the large vocabulary database mentioned above is not enough. Our spontaneous speech database is designed to be suitable for pursuing highly spontaneous speech recognition, namely, a speech database of high spontaneity and limited (~10^3) vocabulary.

In addition to these two target, the speaking variations caused by speaker variances are also major problem for speech recognition. The third database for the *speaker variances* is designed to cover such variations with a large number of speakers' read speech and dialogue speech.

In the rest of this paper, we refer to these three databases as SDB-L, SDB-S and SDB-I, respectively. Table 1 shows the target feature for each database.

Table 1 Target feature of databases

	SDB-L	SDB-S	SDB-I
Speaking style	Dialogue	Dialogue	Read & Dialogue
Vocabulary size	-10^4	-10^3	-10^5 & -10^3
Spontaneity	Middle	High	Mixed
# of speakers	-10^2	-10^2	-10^4

3. Task design

3.1 "Travel arrangement" task

For SDB-L, a "Travel arrangement" [3] task, that originally had been designed for the SLDB, was adopted. Among several travel scenes in the original task, only dialogues between hotel clerk and customer (room reservation, cancellation, and trouble-shooting) were chosen for SDB-L collection. Each speaker role-plays according to some information sheets.

3.2 "Scheduling appointment" task

For SDB-S and SDB-I, a new task "Scheduling appointment" was designed. In this task, one speaker plans to visit another, and two goals to get cooperatively are given. One goal is to determine the date and the time of the visit according to calenders which schedules are filled in, the other goal is to get or give information concerning the location (transportation, hotels, etc.) according to simple maps. In order to promote spontaneous utterances, sentence-like expressions are omitted from these maps and simple words and iconified symbols are used to give knowledge.

3.3 Read speech task

In addition to dialogue speech, in order to build an acoustic corpus with very broad coverage of words and expressions, we also collect read speech data in SDB-I. The isolated speech of Japanese words and idioms or proverbs with the meaning explanations are collected for each speaker. These words and expressions are assigned to each speaker in the way that maximizes the total number of different words and expressions. Table 2 shows the items of data collected in this task.

Table 2 Data items for read speech task

Items	# utterances per speaker	
Words	Common words	2
	Imported words	2
	Area names	2
Expressions	Idioms / Proverbs	2 & 2 *

* Expression & meaning

4. Data collection and transcription

4.1 Recording conditions

All speech data are recorded with a digital audio tape recorder (DAT) in a quiet environment. Speakers have no facial contact during conversation so that the communication medium is restricted to speech. The details of recording conditions are summarized in Table 3.

Table 3 Recording conditions

Microphone:	Uni-directional (Sanken MU-2C, SONY C-355)
Sampling frequency:	48 kHz
Quantization:	16 bits linear
Noise environment:	Quiet
Facial contact:	None
Dialogue topology:	1-to-1
Language:	Japanese monolingual

4.2 Speakers

Although there are basically no special requirements for speakers' experience, the roles of hotel clerks in the "Travel arrangement" task are exceptionally assigned to persons who have the experience for such kind of occupations in their real life. Practically, the speaker issues are important especially in the collection of SDB-I. For SDB-I, taking into account the regional variances in speech, data collection are carried out at geographically distributed multiple sites in Japan with speakers employed there. Fig. 2 shows the cities where speech data have already been collected and are being collected. The recording sites will be more widely distributed with the collection going ahead.

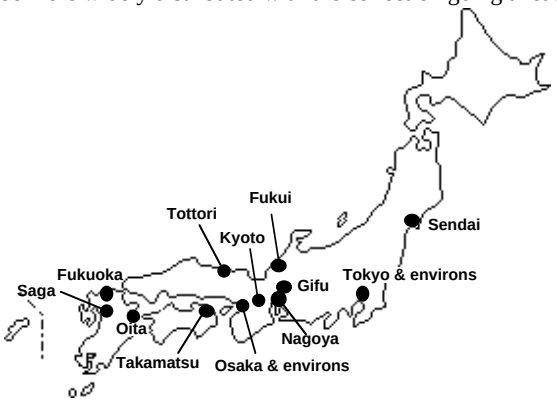


Fig. 2 Current recording sites for SDB-I

4.3 Dialogue control

According to a current realistic image of the speech translation system, speaking turns in dialogues are controlled to refrain from overlapping of two speakers' utterances by giving a ready-prompt to the speaker who has the right to speak. The ready-prompts have been tried to give by three ways: lamp indication, messages on a computer display and hand signals between the speakers. We did not find any big difference between them in the impact on the data

collection efficiency. So, currently the way of giving ready-prompts is determined on a site by site basis.

4.4 Transcription

All collected speech data are transcribed into Japanese text expressions (Japanese text transcription), and then converted into Japanese 26-phoneme expressions with start and end time information for every pause segments (Japanese phoneme transcription). Fig. 3 shows an example of a Japanese phoneme transcription.

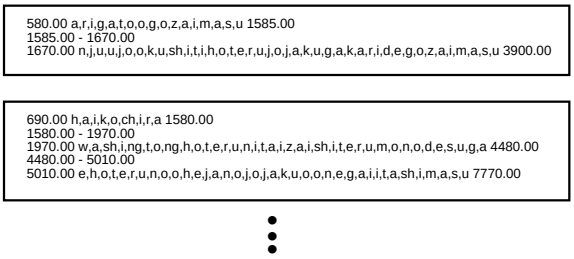


Fig. 3 Japanese phoneme transcription

5. Current status & data evaluations

5.1 Current status

Table 4 shows the current status of data collection and transcription at the end of first quarter 1996. As for the "Travel arrangement" corpus, by putting the SDB-L and the SLDB together, more than 1,200 dialogues have already been collected. As for SDB-I, since a new dialogue features two new speakers, the number of speakers is always twice the number of dialogues.

Table 4 Current status of collection and transcription

	SDB-L	SDB-S	SDB-I	SLDB
Start date	Jan. '94	Jul. '95	Feb. '95	Oct. '93
Collected (dialogues)	655	417	573	618
% transcribed	91.8 %	75.0 %	77.9 %	79.7 %
Speaker participants	215	39	1146	71

5.2 Quantitative evaluation on acoustic & linguistic feature

The acoustic and linguistic features are analyzed on the SDB-L.Table 5 shows the examination results of speaking rate and phoneme boundary uncertainty in comparison with the read speech conversations in the ATR-JSDB. We found that the speech of the SDB-L are in 30 % higher speaking rate than the ATR-JSDB and that each phoneme there can more easily be uncertain than the ATR-JSDB. These results suggest that the speech data collected in this project are much more difficult to recognize than the ATR-JSDB.

Table 5 Speaking rate and phoneme boundary uncertainty

	SDB-L		ATR-JSDB	
	Result	Examined sample size	Result	Examined sample size
Speaking rate	8.91 mora/sec	15,573.1 sec	6.86 mora/sec	9,545.3 sec
Phoneme boundary uncertainty	25.8 %	71,656 phonemes	22.6 %	17,454 phonemes

* Examined samples are different by items

Table 6 shows the frequencies of filled pauses and hesitations, that are typical linguistic phenomena found in spontaneous speech, in comparison with the bilingual dialogues (indirect, through human interpretation) for the same task in the SLDB. We can find that the dialogue styles (e.g. direct or indirect) have great effect on the speaking manner.

Table 6 Filled pauses and hesitations

	SDB-L		SLDB	
	Result	Examined sample size	Result	Examined sample size
Filled pauses	0.75 times/utter.	5,575 utterances	0.41 times/utter.	5,860 utterances
Hesitations	0.073 times/utter.	5,575 utterances	0.018 times/utter.	5,860 utterances

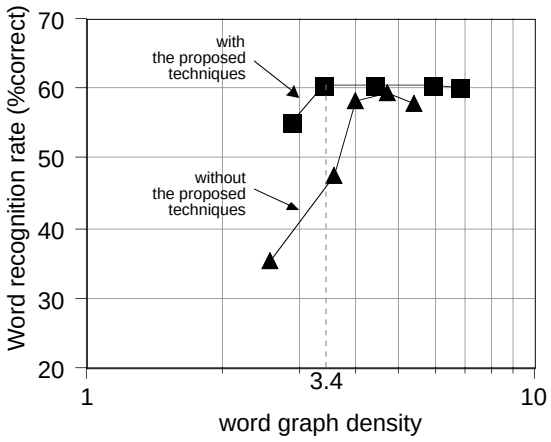
6. Current research using the collected data

Current versions of the collected data are being used for several experiments in our laboratories. In [4], continuous speech recognition experiments for the "Travel arrangement" task were carried out on SDB-L and SLDB with a speech recognizer that features class bigram constraints and word graph output. It was designed and developed to be a part of our new speech translation system. The class bigram was generated by a variable-order N-gram procedure [5] from 828 dialogues (330,513 words) from SDB-L and SLDB. State-shared context dependent HMMs (HMnet [6]) were used as acoustic models. The acoustic models were adapted to the target speaking style and the target speaker with the transfer vector field smoothing (VFS [7]). The lexicon consisted of 6,635 words and the test set had 7 dialogues with word perplexity of 49.6. The following two approximation techniques have been proposed there.

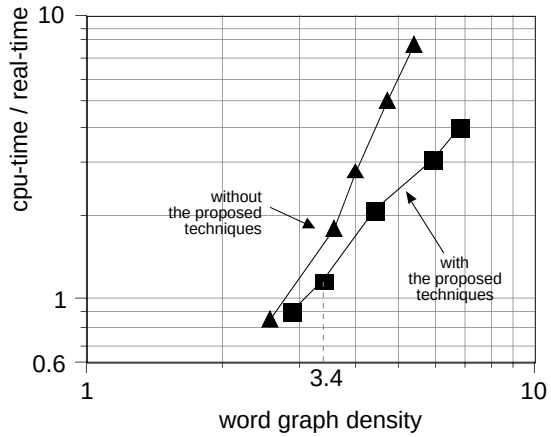
- *Cross-word context approximation*
  Hypotheses that have the same word, the same HMM-state and the same HMM-state end-time are merged if they have the same phoneme context at the end of the preceding word.
- *Language model score interpolation*
  Approximate language scores for HMM-states are given by interpolation using the language scores for phonemes.

Fig. 3 shows experimental results. Here, the word graph density that indicates the size of a word graph is defined as follows.

$$\textit{Word graph density} = \frac{\textit{Number of words in the word graph}}{\textit{Number of spoken words}}$$



(a) Word recognition rate
(for the path giving maximum score)



(b) Cpu time requirement

Fig. 3 Results of continuous speech recognition

The proposed techniques reduced cpu-time requirement with slight improvement in the recognition rate and realized almost real time processing at a word graph density of 3.4, where the recognition rate saturates, on HP 9000/735 workstation.

Other research focusing on high spontaneity or speaker variances using SDB-L or SDB-I are also planned.

7. Future plans

The data collection described above is still ongoing. As shown in Fig. 4, during the first quarter of 1998, the collected data will form huge Japanese dialogue speech corpora. In addition to ATR-JSDB [1], they are expected to be used widely as large-scale speech databases which are well-considered for the research on the robust speech recognition including the speech translation research, and we are planning to start public release of them.

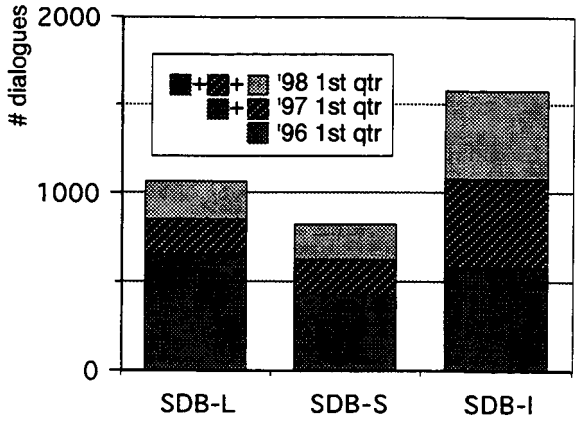


Fig. 4 Future plans for data collection

ACKNOWLEDGMENTS

Speech data collection and transcription described in this paper were carried out cooperatively with NEC Corp., Mitsubishi Electric Corp., Sharp Corp., and INTER GROUP Corp. ATR is grateful to them for their superb efforts. The authors would like to thank Dr. Y. Yamazaki for his continuous support. We would also like to thank all members of ATR Interpreting Telecommunications Research Laboratories for their cooperative efforts in collection and transcription of speech data.

REFERENCES

[1] Sagisaka et. al., "A large-scale Japanese speech database," Proc. of ICSLP '90, pp. 1089-1092 (1990)

[2] Yamazaki, "ATR research activities on speech translation," Denshi Tokyo, vol. 33, pp. 109-114 (1995)

[3] Morimoto et. al., "Speech and language database for speech translation research," Proc. of ICSLP '94, pp. 1791-1794 (1994)

[4] Shimizu et. al., "Spontaneous dialogue speech recognition using cross-word context constrained word graphs," Proc. of ICASSP '96 (1996)

[5] Masataki et. al., "Variable order N-gram generation by word class splitting and consecutive word grouping," Proc. of ICASSP '96 (1996)

[6] Takami et al., "A successive state splitting algorithm for efficient allophone modeling," Proc. of ICASSP 92, pp. 573-576 (1992).

[7] Ohkura et. al., "Speaker adaptation based on transfer vector field smoothing with continuous mixture density," Proc. of ICSLP'92, pp. 369-372 (1992)