J. Acoust. Soc. Jpn. (*E*) **20**, 3 (1999)

JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research

Katunobu Itou,*1 Mikio Yamamoto,*2 Kazuya Takeda,*3 Toshiyuki Takezawa,*4 Tatsuo Matsuoka,*5 Tetsunori Kobayashi,*6 Kiyohiro Shikano,*7 and Shuichi Itahashi*

*1 *Electrotechnical Laboratory,*
1-1-4, Umezono, Tsukuba, 305-8568 Japan
*2 *University of Tsukuba,*
1-1-1, Tennodai, Tsukuba, 305-0006 Japan
*3 *Nagoya University,*
Furo-cho, Chikusa-ku, Nagoya, 464-8603 Japan
*4 *ATR,*
2-2, Hikari-dai, Seika-cho, Soraku-gun, Kyoto, 619-0288 Japan
*5 *NTT,*
2-2-2, Ote-machi, Chiyoda-ku, Tokyo, 100-0004 Japan
*6 *Waseda University,*
3-4-1, Okubo, Shinjuku-ku, Tokyo, 169-8555 Japan
*7 *NAIST*
8916-5, Takayama, Ikoma, 630-0101 Japan

(*Received 3 August 1998*)

In this paper we present the first public Japanese speech corpus for large vocabulary continuous speech recognition (LVCSR) technology, which we have titled JNAS (Japanese Newspaper Article Sentences). We designed it to be comparable to the corpora used in the American and European LVCSR projects. The corpus contains speech recordings (60 h) and their orthographic transcriptions for 306 speakers (153 males and 153 females) reading excerpts from the newspaper's articles and phonetically balanced (PB) sentences. This corpus contains utterances of about 45,000 sentences as a whole with each speaker reading about 150 sentences. JNAS is being distributed on 16 CD-ROMs.

Keywords: LVCSR, Corpus, Speech recognition, Large vocabulary, Assesment

PACS number: 43. 72. Ne

1. INTRODUCTION

In the USA and Europe, effort such as ARPA (NAB)[1] and SQALE[2] have resulted in a large technology push in speaker independent, continuous speech recognition.

In Japan, the Acoustical Society of Japan (ASJ) Continuous Speech Corpora (ASJ-PB)[3] which contain about 10,000 PB sentences, have been widely used as a public resource for LVCSR research. However, we did not have a large text database; the main reason was that Japanese texts were written without spacing between words, and we did not have an adequate automatic word segmentation tool. For this reason, Japanese LVCSR systems were not well developed. Recently, however, progress with morphological analysis systems has enabled automatic segmentation to be used for learning of the statistical language models (SLM), and thus some LVCSR systems have begun to develop.[4]

In Japan, we have been unable to compare different recognition methods and systems, because we did not have any common Japanese speech corpus for LVCSR research. To stimulate Japanese LVCSR research, we designed a Japanese speech corpus for LVCSR technology that is comparable to the corpora used for NAB and SQALE.

In developing the text database, we still have some language-dependent problems with training the language model. The main problem is that we do not have a general rule to separate text into words. One of the other problems is that, because Japanese text consists three character systems (*hiragana, katakana*, and *kanji* (Chinese characters)) there are a lot of variations of spelling.

These problems cause variations between morphological systems (grammar, lexicon, and so on) in Japanese natural language processing (NLP) research. Differences between morphological systems affect the word-frequency lists. For referential comparison, we need to normalize the morphological system or prepare a sharable referential tool as a public standard. We designed and developed the corpus after careful consideration of these points.

In constructing JNAS, the Large Vocabulary Continuous Speech Database Working Group (LVCSD-WG) of the Special Interest Group of Spoken Language Processing (SIG-SLP) of the Information Processing Society of Japan (IPSJ) designed and developed text sets for recording from 1995 to 1997, and the Speech Database Committee of the ASJ developed speech corpus in 1997. We do not plan any formal project for competitive evaluation such as ARPA or SQALE in our community, but we intended the corpus to be used as common reference data.

Figure 1 shows the construction flow of JNAS. Japanese newspaper articles were morphologically analyzed automatically, and were split to two parts. One part was used for training data of language models. Using the language models, candidate sentences for reading were selected from another part of sentences. The number of sentences first selected was about 40 K. Sentences which had word segmentation errors were rejected, and sentences which had pronunciation tag errors were corrected. From these modified sentences, 150 text sets (about 100 sentences/set) for reading were constructed. The rest of the sentences were kept for spares ('POOL' in Fig. 1). The text sets were handed to the ASJ committee and were checked again by 39 recording sites from the viewpoints of reading difficulty. Problematic sentences were rejected and replaced

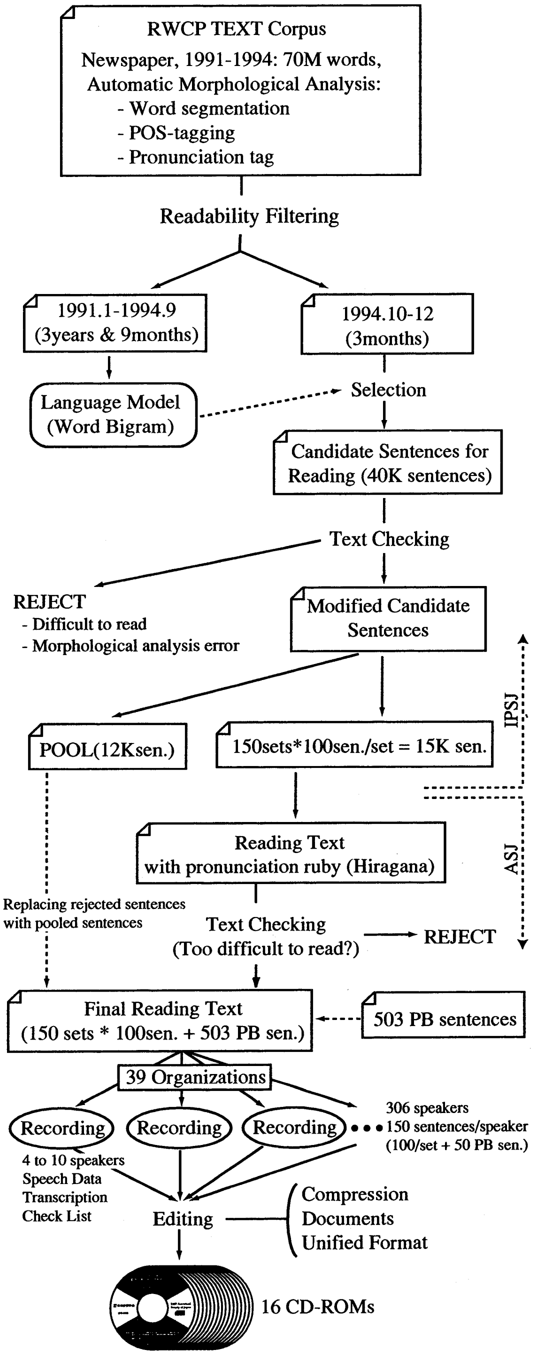

Fig. 1 The construction flow of JNAS.

with spare sentences. The final reading text was made up of 150 sets which consist of about 100 sentences and 50 Phone-balanced sentences each. Each recording site recorded speeches of 4 to 10 speakers. A/D converted data were compressed and edited into 16 CD-ROMs.

In this paper, we describe the details of the specification and development of JNAS.

2. DESIGN AND DEVELOPMENT OF THE CORPUS

2.1 Corpus Structure and Capabilities

The corpus is designed to be comparable to the corpora used in NAB and SQALE project.

The corpus is scalable and built to accommodate variable-sized large vocabularies (5 K, 20 K), variable perplexities (from 0 to 400), and variable sentence lengths (from 5 to 40 words). The training material for the acoustic models amounts to 100 sentences per speaker and over 100 speakers. The speaker adaptation material amounts to 50 PB sentences per speaker. All materials are recorded simultaneously with a standard close-talk microphone and a secondary desktop microphone. The speech data is collected in a "read" speech mode, and an equal number of male and female speakers are chosen. The corpus permits evaluation both with and without "out-of-vocabulary" lexical items and vocabulary open tests.

2.2 Japanese Language-Dependent Problems in Statistical Language Model Training

Japanese text is not divided by white space at word boundaries. In Japanese text, we use *kanji* (Chinese orthography) and *kana* (phonetic characters).There are two types of *kana: hiragana* and *katakana*. *Kanji* consists of ideograms. When writing with ideograms, since the objects of linguistic expression are innumerable, there are also an extremely large number of characters. For example, the four years of newspaper text we used, contained more than 5,000 different characters. Many *kanji* have multiple readings: one reading is derived from the Chinese pronunciation, and the other is the "Japanese" reading of the Japanese word that corresponds to the meaning of the Chinese character—what is called *wago*. Many *kanji* have the same Chinese pronunciation. Therefore, it is very hard to disambiguate readings in the "*mixed kana-kanji*" style of writing. Moreover, in the Japanese language, we do not have general rules for what constitutes a single word, and verbs, adjectives, and other inflected words have many inflections.

Therefore, it is difficult to define a word unit, and there are a lot of variations between morphological systems. In making referential comparisons, we need to considered these points carefully.

2.3 Text Selection

First, we discussed which text to select for training and evaluation material. The most important requirement is that the total amount of the text is comparable to the corpora used for NAB and SQALE. To fulfill the requirement, we decided to select a material from news paper article data. As a target of real application of speech recognition, news paper article is not appropriate because, there are grammatical difference between writing style and speaking style in Japanese. However, news paper data was selected because we found no alternative and also selected in NAB and SQALE.

After we discussed which paper to select among some major dailies and a business paper, we decided to use the Mainichi Newspaper, one of the major Japanese dailies, because its copyright permission is most suited to our purpose of releasing the resultant corpus to the public.

2.4 Text Preprocessing

Ideally, the text preprocessing should divide the sentence into words and resolve the ambiguities with all of the readings of the words. This preprocessing is similar to the type that might be used in a text-to-speech system. A text-to-speech systems' preprocessor, however, can only give the readings, and cannot give any grammatical and/or morphological information, such as segmentation of words or the part of speech of the word, which is useful for constructing a language model for speech recognition. In the research community for Japanese natural language processing, a system called "morphological analysis system" is widely used, and the system estimates word segmentation, part of speech, and inflection.

However, estimation of the reading of the word is beyond the ability of the current morphological analysis system, because it is developed for text processing which needs not estimate the reading. Moreover, in Japanese we don't have any standard general rule for the definition of vocabulary, morphological grammar, or a system for parts of speech.

Therefore, we fixed the goal of text preprocessing as analysis of a sentence by a morphological analysis system.

There were no public morphological analysis systems which had the ability to construct a language model, and so we decided to use the morphological tagged corpus of the Mainichi Newspaper which is distributed as the RWCP-Text-Corpus (RWC-DB-TEXT-95-1)[5] by the RWCP (Real World Computing Partnership) as a standard text database for training of the language model.

An example of the RWCP-Text-Corpus is shown in Fig. 2. In the example, each line stands for a morpheme. The first column contains the notation for the morpheme, the second contains a basic form (dictionary form, original form) of the morpheme, and the third contains the ID number of a part of speech (POS).

The RWCP proposed a POS system called THiMCO (Tagset of High quality for Integrated Multi-usage Corpus Openly available to public). In the RWCP-Text-Corpus, THiMCO95 was used. THiMCO95 is a relatively detailed Japanese POS system and contains about 500 parts of speech.

The first step in preprocessing was to extract all of the article paragraphs as fields from the original CD-ROM with RWCP-Text-Corpus. An article has a specific ID number and consists of paragraphs. After extraction, each paragraph was numbered in order automatically (this number was used as the document control number) and collected into a file by month.

Next, paragraphs which had no period were removed automatically for readability filtering. Such paragraphs included poems, recipes, tables, lists, and so on. As another readability filtering, sequences of morphemes between special symbols (for example, round brackets), which were automatically estimated as "unread" expression was removed. Finally, the paragraph was divided into sentences at periods or equivalent symbols. After sentence segmentation, each sentence was numbered in order. Thus, the sentence number was dependent on the morphological analysis system.

2.5 Sentence Selection for Recording

Next it was necessary to divide the text data into a training section and an evaluation section. The most recent three months' data (about 10% of the whole data) were selected for testing, and the rest of the articles covering 45 months (about 90% of the data) were reserved for training. The size of the corpus is shown in Table 1.

Table 1 Text corpus.

	91/1–94/9	94/10–94/12
# sentence	2,372K	194K
# paragraph	1,438K	138K
# article	282K	21K
# morpheme	65,347K	4,936K
vocabulary size	291K	97K

For classifying sentences, it was necessary to form a language model for selection of the sentences for recording. The first step was to form a word-frequency list (WFL, a frequency-ordered morpheme unigram list) from all of the training text with their morphological information.

To form a WFL, we needed to define a counting unit for word. As we mentioned above, it is reasonable that we treated a morpheme as a word. In this case, we have several choice from many definitions of counting units. We considered the following and other definitions.

1. distinction by notation
2. distinction by basic form
3. distinction by reading

It seems that the third method is better one than any other choice, because we do not need to consider the treating of a word which has multiple readings. However, any sharable adequate text-to-reading translation tool or any large text corpus which has the information of reading is not available.

It seems that the second method is better one from the linguistic viewpoint. The first method is the most simple one. However, in this method, we do not distinguished the morphemes which have the same notation and the different POSs or readings.

Finally, we defined the counting unit as a mor-

notation	basic form	POS ID
認識	認識	1
し	する	63
なけれ	ない	445
ば	ば	422
なら	なる	276
ない	ない	443
。	。	468

Fig. 2 An example of the data of the RWCP-Text-Corpus. The example sentence is "認識しなければならない。" (I have to recognize.)

pheme distinguished by all of the morphological attributes given by RWC Corpus (shown in Fig. 2), because the definition is the most precise and it is advantageous as the reference corpus.

We counted using this counting unit, it yielded a list of 291 K words from all of the learning data. The frequency-weighted word coverage of the WFL is shown in Table 2.

Next, a word bigram language model was constructed to calculate the test-set perplexity of the sentence. The word bigram language model was generated using the CMU SLP Toolkit.[6] The language model was an open vocabulary backoff word bigram which was constructed with the cutoff set to 2, the discount strategy specified as "Good Turing discounting," and a vocabulary size of 20 K words.

Sentences in articles covering three months were classified into 30 categories based on the bigram model. Each category is characterized by the sentence length (2 levels), the vocabulary size (5 levels) and perplexity (3 levels).

We discussed the two methods for design of limited frequent word's vocabulary. The first definition is based on coverage,[4] and the other one is based on size. From a view point to consider statistical evaluation of language modeling, the first method is appropriate to compare with other language system. However, from a view point to consider total performance of system, the first method is not the best because the difference of the vocabulary size affects the performance of the recognition system.

Table 2 Frequency-wedghted word coverage (from the word-frequency list).

Size	Coverage (%)
5K	85.8
8.1K	90.0
20K	95.7
27.6K	97.0
291K	100.0

First of all, Japanese and the European languages are totally different; it is quite difficult to draw comparison. And also, as mentioned above, in Japanese,we don't have any general definition of word unit and general rule of counting words. After discussion, we defined 5 K and 20 K vocabularies to have same size as the vocabularies used for NAB and SQALE.

A statistically-controlled text set consists of 90 sentences (SC-sentences) collected from the 30 categories according to Table 3 and about 10 sentences taken from a few paragraphs which consisted of only the three or more sentences which were satisfied in any category in Table 3. Category's boundary parameters of the sentence length (Table 4) and perplexity (Table 5) are depended on the vocabulary size, because the distributions are sensitive to the vocabulary size.

Five other text sets were "article" sets. An "article" set consisted of three articles. Each article included 10 or more paragraphs that consisted only of sentences classified into any class in Table 3. Each paragraph contained 3–10 sentences. We didn't check for duplication of sentences between "sentence" sets and "article" sets.

Table 3 Distribution of perplexity (pp), sentence length, vocaburaly class for 90 sentences as selected for each speaker.

Sentence Length	Normal			Long		
Perplexity	low pp	normal	high pp	low pp	normal	high pp
MID	2	6	2	1	3	1
MID+	2	6	2	1	3	1
LARGE	4	12	4	2	6	2
LARGE+	2	6	2	1	3	1
LARGE++	2	6	2	1	3	1

MID	=5k voc.
MID+	=5k voc. with one unknown word
LARGE	=20k voc.
LARGE+	=20k voc. with one unknown word
LARGE++	=20k voc. with two unknown words

Table 4 Boundary parameters in sentence length.

Vocaburary class	Normal	Long
MID	6～19	20～39
LARGE	6～29	30～39

Table 5 Boundary parameters in perplexity.

Vocaburary class	Perplexity
MID	$0 \leq L < 40 \leq M < 85 \leq H < 400$
LARGE	$0 \leq L < 70 \leq M < 130 \leq H < 400$

2.6 Recording

The speech data were recorded in collaboration with 39 sites, so the recording conditions and AD conversion characteristics, including low-pass filter characteristics, were not unified. Each recording site collected data sets for 4–10 speakers (equal numbers of male and female speakers chosen). Each speaker read one set (about 100 sentences) from SC sentence sets, and one subset (about 50 sentences) from the ATR PB sentence sets.[7]

These PB-sentences were chosen by ATR Interpreting Telephony Research Laboratories. Entropy was calculated based on the clusters of two phonemes (120 CV's, 227 VC's and 55 VV's, making 402 clusters in all) and three phonemes (69 CVC's where C is an unvoiced consonant, 18 CVC's where C is a nasal consonant and 136 VCV's where C is a semivowel, making 223 clusters in all) on the assumption that they occur independently. 10,196 original sample sentences were extracted at random from newspapers, magazines, novels, letters, textbooks, *etc*. Of these, 503 PB sentences were chosen to maximize the entropy. They were sorted so that each set of 50 sentences was also phonetically balanced.

From the 150 SC sentence sets, 138 sets were read by both one male speaker and one female speaker, 4 sets were read by both of two male speakers, 4 sets were read by both of two female speakers, 2 sets were read by one male speaker, and 2 sets were read by one female speaker. At each of the recording sites, all of the speakers read the same PB sentence subset.

The utterances were recorded with two microphones simultaneously: a standard close-talk microphone (Sennheiser HMD410/HMD25-1 or equivalent) and a desktop microphone which was selected independently at each site (Sanken, Sony, and similar). The two versions of the data were stored in separate files.

Reading text was printed out to papers and the speakers read the text. Only kanji characters in reading sentences had ruby (readings) for easy to read. They were automatically generated from morphologically analyzed sentences with readings using 'diff' command of UNIX which may use dynamic programming.

Each utterance was checked at each recording site. In the prompting text, each word was given a single reading. However, in Japanese, there are words which have several readings (*i.e.*, "Japan" has two readings: *nihon*, and *nippon*). An orthographic transcription was made at each recording site. No changes to the content of the newspaper articles were permitted under the copyright permission, and the orthographic transcription was not modified for any other errors or variations. A list of these errors was collected in the check list file at each recording site.

3. SPEECH CORPUS: JNAS

The data described here was compiled into 16 CD-ROMs and titled JNAS (Japanese Newspaper Article Sentences). 8 discs (Vol. 1 through Vol. 8) contain the close-talk microphone data, and the others (Vol. 9 through Vol. 16) contain the desktop microphone data. Additional components are the prompting text, the modified transcription in three styles: *kanji*-text with ruby (readings), *katakana*, and *romaji*, original recorded text, check report files,

Table 6 The specifications of JNAS.

#Reading text sets	Newspaper	155sets (16, 176 sen.) (abous 100 sen./set)
	PB sen.	10sets (503 sen.) (about 50 sen./set)
#Speakers		306 (153 fe/males)
#Utteraces	Newspaper	31,938
	PB sen.	15,372
Recorded time of newspaper sentences		215,247 s (about 59 h 47 m)
#Recording site		39
Microphone	headset	common
	desktop	inconsistent

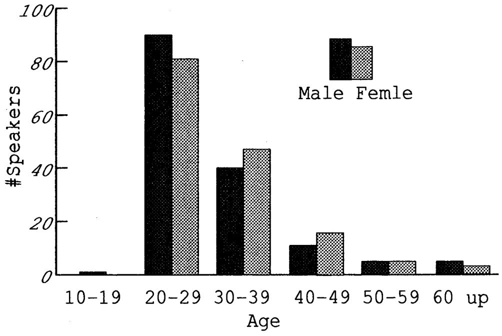

Fig. 3 Speaker's age distribution.

and the bigram language model. The speech waves were digitized at a 16 kHz sampling frequency and quantized at 16 bits. They were stored with the NIST SPHERE headers in the compressed format, using the "shorten" compression technique implemented in the NIST SPHERE PACKAGE. About 18 Gigabytes speech data were compressed to about 9 Gigabytes by the "shorten."

The CD-ROMs have been released to the public. Table 6 shows the final specification of JNAS and Fig. 3 shows the age distribution of speakers of JNAS. Since JNAS is a relatively large corpus, it is not error free. At this time, we know some errors such as reading error, A/D conversion error, disfluencies, lack offiles, and so on. We plan to maintain these error information about JNAS on the WWW (http://www.milab.is.tsukuba.ac.jp/jnas/).

4. CONCLUSION AND FUTURE WORKS

The JNAS corpus and its components have been designed and developed for LVCSR research by the joint efforts of the LVCSD-WG IPSJ and the Speech Database Committee of ASJ.

The Speech Database Committee of ASJ are now selecting text sets for referential evaluation. It plans a training set that contains 100 speakers × 100 sentences and an evaluation set that contains 25 speakers × 4 sentences.

To promote both research of component technologies and development of systems for LVCSR, we have recognized the necessity of a sharable software repository which includes recognition engines, acoustic models and language models. Thus, we are developing a Japanese Dictation ToolKit,[8] sponsored by the Information-Technology Promotion Agency (IPA), in Japan. In the project, we will also develop a tool to normalize the differences between morphological analysis systems.

ACKNOWLEDGEMENTS

The prompting texts and the bigram language models for the Mainichi Newspaper article sentences were prepared by Akinori Ito (Yamagata Univ.), Takehito Utsuro (NAIST), Tatsuya Kawahara (Kyoto Univ.), Toru Shimizu (KDD), Masafumi Tamoto, Kazuhiro Arai (NTT), and Nobuaki Minematsu (TUT). We used the NIST SPHERE package to attach headers to the wave files and for the "shorten" compression technique used to reduce the number of CD-ROMs. The NIST SPHERE package was implemented by the Spoken Natural Language processing group, National Institute of Standards and Technology, U. S. A. The 'shorten' compression technique was developed by Tony Robinson at Cambridge University and SoftSound Limited, UK. The speech data was collected by the efforts of many volunteers at the 39 research institutes.

We would like to thank all of the above groups and individuals.

REFERENCES

1) D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," Proc. DARPA Speech & Natural Language Workshop, 357-361 (1992).

2) H. J. M. Steeneken and V. Leeuwen, "Multilingual assessment of speaker independent large vocabulary speech recognition system: the SQALE-project," EUROSPEECH 95, 1271-1274 (1995).

3) S. Hayamizu, S. Itahashi, T. Kobayashi, and T. Takezawa, "Design and creation of speech and text corpora of dialogue," IEICE Trans. Inf. Syst. **E76-D**(1), 17-22 (1993).

4) T. Matsuoka, K. Ohtsuki, T. Mori, and S. Furui, "Japanese large-vocabulary continuous speech recognition using a business-newspaper corpus," Proc. ICSLP 96, 22-25 (1996).

5) K. Hasida, H. Isahara, T. Tokunaga, M. Hashimoto, S. Ogino, W. Kashino, J. Toyoura, and H. Takahashi, "The rwc text databases," Proc. LREC '98, 457-461, May (1998).

6) R. Rosenfeld, "The CMU Statistical Language Modeling ToolKit and its use in the 1994 ARPA CSR Evaluation," Proc. ARPA Spoken Language Systems Technology Workshop, 47-50, Jan. (1995).

7) K. Iso, T. Watanabe, and H. Kuwabara, "Design of Japanese sentence list for a speech database," Proc. Spring Meet. Acoust. Soc. Jpn., 89-90 (1988) (in

Japanese).

8) T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano, "Common platform of Japanese large vocabulary continuous speech recognizer assessment—proposal and initial results—, "Proc. Oriental-COCOSDA Workshop, 117–122 (1998).