A JAPANESE NATIONAL PROJECT ON SPONTANEOUS SPEECH CORPUS AND PROCESSING TECHNOLOGY

Sadaoki Furui[*], *Kikuo Maekawa*[#] *and Hitoshi Isahara*[+]

* Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan, furui@ cs.titech.ac.jp
The National Language Research Institute
3-9-14 Nishiga'oka, Kita-ku, Tokyo, 115-8620 Japan, kikuo@kokken.go.jp
+ Communications Research Laboratory
588-2 Iwaoka, Nishi-ku, Kobe, 651-2401 Japan, isahara@crl.go.jp

ABSTRACT

A new national project for raising the technological level of speech recognition and understanding has recently commenced in Japan. This project aims at a) building a large-scale spontaneous speech corpus consisting of roughly 7M words and 800 hours of speech, b) acoustic and linguistic modeling for spontaneous speech understanding and summarization using linguistic as well as para-linguistic information in speech, and c) building a prototype of a spontaneous speech summarization system. The corpus under compilation will contain spontaneously uttered Common Japanese speech and the morphologically annotated transcriptions. Also, segmental and intonation labeling will be provided for a subset of the corpus. The primary application domain of the corpus is speech recognition of spontaneous speech, but it is also planned to become a useful research corpus both for natural language processing and phonetic/linguistic studies.

1. INTRODUCTION

Although high recognition accuracy can be obtained for speech reading a written text or similar by using state-of-the art speech recognition technology, the accuracy becomes miserable for freely spoken spontaneous speech. The principal reason of this problem exists is the fact that acoustic and linguistic models used in speech recognition have been built mainly using written language or speech reading text, while spontaneous speech and written language considerably differ both acoustically and linguistically. Broadening effectively the application of speech recognition thus crucially depends on raising the recognition performance for spontaneous speech.

From this perspective, a Japanese national project on spontaneous speech corpus and processing technology was initiated in 1999. This project aims to build a large-scale spontaneous speech corpus and create spontaneous speech recognition and understanding technology.

2. PROJECT OVERVIEW

The Science and Technology Agency Priority Program (Organized Research Combination System) entitled "Spontaneous Speech: Corpus and Processing Technology" started in 1999 under the supervision of S. Furui. The principal organizations working together to conduct this project are the National Language Research Institute under the Ministry of Education, the Communication Research Laboratory under the Ministry of Posts and Telecommunications, and Tokyo Institute of Technology.

The project will be conducted over a 5-year period in pursuit of the following three major themes as shown in Fig. 1;

1) Building a large-scale spontaneous speech corpus consisting of roughly 7M words with the total speech length of 800 hours. The corpus will primarily consist of monologues such as lectures, presentations, and news commentaries. They will be manually given orthographic and phonetic transcription. One-tenth of the utterances ("Core") will be tagged manually and used for constructing a morphological analysis program for automatically analyzing all of the 800-hour utterances. The Core will also be tagged with para-linguistic information including intonation [1].
2) Acoustic and linguistic modeling for spontaneous speech understanding and summarization using linguistic as well as para-linguistic information in speech.
3) Constructing a prototype of a spontaneous speech summarization system.

The technology created in this project is expected to be applicable to wide areas such as indexing of speech data (broadcast news, etc.) for information extraction and

retrieval, transcription of lectures, preparing minutes of meetings, closed captioning, and aids for the handicapped.

Presentations at various academic meetings, such as the Acoustical Society of Japan (ASJ) meetings, and simulated public speech by non-professional speakers have already been recorded and transcribed in the project. Using these utterances, preliminary recognition experiments are being conducted at several universities participating in the project. At the Tokyo Institute of Technology, for example, preliminary experiments have been conducted using a presentation recorded at an ASJ meeting and one person's talk excerpted from a broadcast political discussion [2].

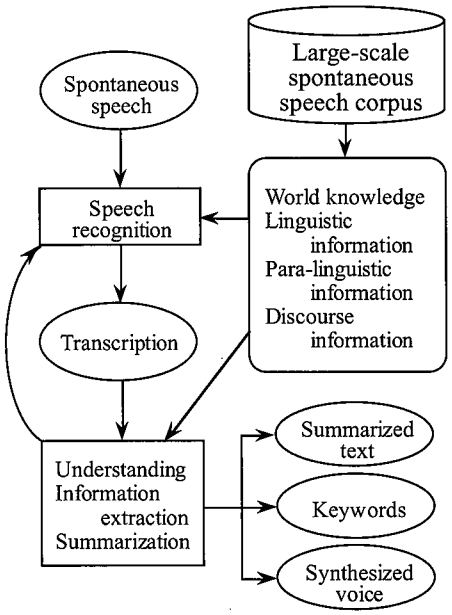

Fig. 1 – Overview of the national project

3. CORPUS DESIGN

3.1. Corpus size

It is very important to have a clear-cut view of the application when we start compiling a corpus. In our project, we will use the corpus mainly for two purposes, 1) construction of the language model for speech recognition for spontaneous speech, and 2) linguistic-phonetic and/or natural language processing studies of spontaneous speech. There is, however, a trade-off between the two purposes: the former requires a large amount of data, while the latter puts more emphasis upon the accuracy and quality of annotations rather than the corpus size. Enlargement of the corpus size and refinement of annotation both result in increased cost of compilation. We try to avoid this problem by applying different annotation strategies to the whole corpus and its substantially smaller subset.

As shown in Table 1, the whole corpus will contain 7M morphemes. Based on our prior experiences, we regarded this to be the minimum requirement for statistical language modeling for speech recognition. Digitized speech, its transcriptions, and morphological annotations of the transcribed speech will be the contents of the whole corpus.

On the other hand, 500K morphemes out of the total of 7M will be in the smaller corpus called the *Core*, to which we will concentrate the cost of annotation. The Core will be provided with segmental and intonation labeling in addition to the above-mentioned contents.

Table 1 - Corpus Size and Contents

	Amount of data	Contents
Whole corpus	7M morphemes (800 hours)	Digitized speech Transcription Morphological annotation
Core	500K morphemes (60 hours)	Segmental labeling Intonation labeling (in addition to the above)

The core has one more *raison d'être* in our project. Although we will provide morphological annotation, namely word boundary and part of speech tagging, for the whole corpus, it is extremely difficult to annotate the corpus of this size by hand. We plan to use the Core as the learning data for the automatic morphological analysis software that we are developing in our project, with which the whole corpus will be annotated. Morphological analyses of the Core will be done manually by the lexicographers at the National Language Research Institute (NLRI).

3.2. Language variety

Selection of speech variety is another vital issue of corpus design. Since it is practically impossible to make a corpus that covers all varieties of a given language, it is desirable as well as necessary to concentrate upon a specific variety. In our project, we concentrate upon spontaneous monologue rather than dialogue. This is because the modeling of dialogue speech for speech recognition is another challenging task, and it is being covered in another national project.

We also concentrate upon a social variety called *Common Japanese* (CJ). In today's Japan, people who are educated at least in high school, speak two varieties: their

native dialect and the CJ. The latter is a variety used in more or less formal situations like business/academic meetings or public lectures in front of an audience. The segmental phonology, syntax, and lexicon of CJ spoken by people who are in their fifties or younger are quite similar to those of Tokyo Japanese. Lexical accent, however, differs considerably reflecting the phonology of the speakers' native dialects even among younger speakers. Therefore we make it our principle not to pay attention to prosody and concentrate our attention on segmental and syntactic characteristics. A speech is classified as CJ, and then stored in our corpus if its segmental, syntactic and lexical characteristics approximate those of Tokyo Japanese. According to our pilot evaluation, out of the total of 289 speeches that we have recorded in the past six months, 278 were evaluated as CJ.

3.3. Sources

Because it is our intention to make the corpus open for researchers, all speech materials in our corpus must be copyright-cleared. We make it a rule to make recordings of those speakers who agreed, with written consent, to provide their speech for our corpus, thereby making it open for academic use.

Currently we are making two different types of data recording: *academic presentations*(AP) and *simulated public speech* (SPS). By AP is meant the live recording of researchers' presentations in various academic meetings and we plan to record at least 300 hours of academic presentations. Currently our recordings are limited to speech related academic societies like the Acoustical Society of Japan, the Phonetic Society of Japan, and the Society for the Study of Japanese Language, but we are planning to correct this bias by enlarging the number of societies. It is also important to note that the distributions of speakers' age and gender are strongly skewed in AP data. Most speakers are male and in their twenties or thirties.

On the other hand, SPS is short speech (mostly 10 to 15 minutes long) spoken specifically for our corpus by paid non-professional speakers. They are instructed to prepare an outline of their talk instead of the completely pre-fixed text. SPS is recorded mostly in the recording studio of the NLRI in front of a small audience. The topic of the speech can vary from speaker to speaker. We plan to make at least 300 hours of SPS in which distributions of speakers in terms of their age and gender are maximally balanced.

We have made recordings of about 80 hours of AP and 70 hours of SPS during the past ten months, since the beginning of our project.

3.4. Degree of spontaneity

When we started data collection, we made it a principle to exclude speeches that were completely pre-fixed. As we went on with live recordings of AP, however, it turned out that spontaneity of speech could vary even within one presentation. For example, spontaneity was reduced considerably when a talker was reading a manuscript prepared for the presentation, but even in this type of well-prepared presentation, degree of spontaneity was not always low throughout the presentation. It suddenly increased when the talker made a digression or found mistakes in the manuscript. Judging from this experience, we no longer pay much attention to the level of preparation of the presentation. Consequently, our corpus involves samples that hardly will be classified as spontaneous speech if we interpret the term very rigorously. Rough estimation suggests that about five percent of the whole corpus will be of this sort.

We believe, however, inclusion of prepared speech does not deteriorate our corpus. Even the least spontaneous samples recorded so far are distinctively more spontaneous than typical read speech, such as professional announcers' news reading, in that the prepared speeches in our corpus contain many fillers and disfluencies.

Spontaneity is a matter of more or less, and we need a wide range of samples differing in spontaneity in order to know what really are the phonetic and/or linguistic characteristics of speech spontaneity. To help conduct this sort of study, all speech samples in our corpus will be provided with subjectively evaluated indices of spontaneity, ranging from 0 (=completely prepared-to-be-read) to 5 (= completely spontaneous). Table 2 is the tentative list of information that will be provided for each sample.

Table 2 - Information about speakers and speech samples that will be involved in the corpus

Speaker information	Age and gender Birth place and past and present living places. Education level Existence of prepared manuscript etc.
Speech and recording information	Recording date and place Recording equipment used Evaluation of spontaneity Evaluation of fluency Evaluation of proficiency Description of voice quality characteristics (if any) Description of noise sources (if any) etc.

3.5. Recording

All speech samples are recorded using a head-worn unidirectional condenser microphone and a digital tape recorder (DAT) and downsampled to 16kHz sampling frequency and 16 bit quantization before being stored in the corpus. Although maximum care is taken to make good recordings, the recording condition varies considerably from one recording to another in live recordings of AP, due mainly to the difference of the room acoustics of conference sites. The condition of SPS is generally much better than that of AP, because they are recorded mostly in the recording studio of the NLRI.

Video recordings are also made. Videos can be quite helpful in the speech transcription work, because we can get much information about the content of speech by checking the viewgraphs used in presentations. They are also useful to know the reasons of sudden interruptions (floor littered with viewgraphs or sudden-death of a presentation computer) or the interaction between the speakers and the chairperson who may intervene in the presentation. Unfortunately, however, recorded videos will not be involved in our corpus, because the usage of videos is restricted to the transcription work in the copyright-clearance contract. People are generally very reluctant to give consent to open usage of their video images.

4. TRANSCRIPTIONS

4.1. Orthographic and phonetic transcriptions

Our transcription format was devised to satisfy as much as possible the requirements both from speech recognition and natural language processing studies, which can be incompatible at times. The transcription file contains two different kinds of transcriptions; the transcription that we call *orthographic* which is written using both Kanji (Chinese logographs) and Kana (Japanese syllabary) and the transcription called *phonetic* which is written in Kana only. Orthographic transcription will be used in language modeling for speech recognition. It will be used also in automatic morphological analyses; morphological analysis programs for Japanese text require that input text be written in Kanji and Kana because the boundary between Kanji and Kana provides precious information about word boundary (Japanese orthography does not use blank spaces to indicate word boundary).

Though we call it orthographic, our orthographic transcription is different from the standard orthography of Japanese in some respects. Most importantly, our orthographic transcription does not allow any variation of word-to-letter correspondence that characterizes the standard orthography. Phonetic transcription is needed basically to show the readings of Kanji strings (which can often have more than two readings). At the same time, this transcription shows as precisely as possible within the limit of syllable letters, the actual pronunciation as it appears in the recorded speech.

4.2. Filler, disfluency and noise

Fillers, or filled pauses, are among the most eminent indicators of spontaneity of speech. They are marked by a tag (F) in both orthographic and phonetic transcriptions. A typical filler is a prolonged monophthong accompanied with flat pitch, but short conjunctions are recognized as fillers when they are accompanied, typically, by prolongation of the last vowels and flat pitch. Disfluency, another indicator of spontaneity, is marked using tags (D) or (W). (D) is used typically to mark cases where speakers pronounced a word, or fragment of it, and corrected it later, while (W) marks the cases without correction, *i.e.* cases where speakers are not aware of the mispronunciation. (C) indicates a split word. This tag is used when a word is split by a pause longer than 200 ms and accordingly belongs to two separate utterance units. Table 3 shows the tentative list of tags used in our corpus.

Table 3 - Tentative list of tags used in transcription

Symbol	Meaning
(F)	Fillers.
(D)	Disfluency.
(W;)	Mispronounced word. Supposed-to-be correct form is shown after the semicolon.
(?,)	Uncertainty in perception. Multiple candidate words are shown separated by comma if necessary.
(Q)	Metalinguistic expression.
(Laugh) (Cough) (Yawn) (Breath)	Non-verbal vocal events that co-occur with speech such as laughter, cough, yawn, and breath.
<Laugh> <Cough> <Breath> <Lip>	Non-verbal vocal events that do not co-occur with speech. <Lip> means lip noise.
<P>	Pause.
<H>	Prolonged word-final vowel that functions as a filler.
<FV>	Uncertainty of phonetic quality of vowels used as filler.
<C>	Split word. Use when a word is split by a pause and belongs to two separate utterance units.

4.3. Morphological annotations

As noted earlier, we have two different schemata of annotation for the whole corpus and the Core. All speech samples in our corpus will be analyzed in terms of word boundaries and parts of speech. A big problem of morphological analysis is that there is no widely-agreed-upon definition of word in Japanese. This is partly because Japanese orthography does not have the custom of showing word boundaries by blank spaces, but more fundamentally, this is a reflection of the linguistic characteristics of Japanese morphology which allow quite free word-formation. Moreover, criteria of word recognition can be different depending on the purposes of morphological analyses. This makes our analyses more complicated. As a general tendency, speech recognition prefers to recognize longer units as word while natural language processing and linguistics prefer shorter units. In our morphological analysis, we plan to reconcile this problem by providing two different analyses based on two different working definitions of word, namely longer and shorter words.

5. CONCLUSION

This paper introduced a 5-year Japanese national project on spontaneous speech corpus and processing technology started in 1999. The project will be conducted toward realizing three major themes; 1) Building a large-scale spontaneous speech corpus consisting of roughly 7M words having a total speech length of 800 hours, 2) Acoustic and linguistic modeling for spontaneous speech understanding and summarization using linguistic as well as para-linguistic information in speech, and 3) Constructing a prototype of a spontaneous speech summarization system.

We still have so many short term as well as long term research issues, such as a) how to transcribe spontaneous speech, b) how to apply morphological analysis to the transcribed spontaneous speech, c) how to build precise and yet general filled pause models, d) how to incorporate repairs, hesitations, repetitions, partial words, and disfluency, and e) how to adapt the language models to each task. It is also important to investigate the method of building acoustic models fit to spontaneous speech.

After finishing the compilation, we plan to make the corpus available free-of-charge for academic purposes. We also plan to make part of the corpus open for monitoring even before the end of compilation. The corpus, which we tentatively call the *Corpus of Spontaneous Japanese*, will be the world's first of this type, and it is our wish that it is used by many people, both in scientific and technological fields, both at home and abroad. We welcome comments on our project and would like to exchange the ideas and experiences of corpus compilation with colleagues who are working on similar projects.

In the processing of spontaneous speech, a paradigm shift from speech recognition to understanding, where underlying messages of the speaker, namely meaning/ context that the speaker intends to convey, are extracted, instead of transcribing all the spoken words, will be indispensable. Speech summarization, which is one of the main targets of the national project, is considered to be one of the variations of speech understanding [3]. Speech summarization will be also applicable to a range of applications, such as preparing minutes of meetings, close captioning of broadcast news, and presenting information in news-on-demand systems.

ACKNOWLEDGMENT

The authors wish to express their thanks to all the members of the national project for their great effort for building the speech corpus and fruitful discussions for creating new ideas for building language and acoustic models for spontaneous speech recognition and understanding.

REFERENCES

[1] K. Maekawa, H. Koiso, S. Furui and H. Isahara: "Spontaneous speech corpus of Japanese", Proc. 2nd International Conference on Language Resources and Evaluation, Athens, Greece, pp. 947-952 (2000)

[2] S. Furui, K. Maekawa, H. Isahara, T. Shinozaki and T. Ohdaira: "Toward the realization of spontaneous speech recognition – Introduction of a Japanese priority program and preliminary results –", Proc. ICSLP2000, Beijing (2000)

[3] C. Hori and S. Furui: "Automatic speech summarization based on word significance and linguistic likelihood", Proc. ICASSP2000, Istanbul, Turkey, pp. 1579-1582 (2000)