論文 / 著書情報
Article / Book Information

Title	Spontaneous Speech Recognition and Summarization
Author	Sadaoki Furui
Journal/Book name	The Second Baltic Conference on HUMAN LANGUAGE TECHNOLOGIES, Vol. , No. , pp. 39-50
発行日 / Issue date	2005, 4

SPONTANEOUS SPEECH RECOGNITION AND SUMMARIZATION

Sadaoki Furui
Tokyo Institute of Technology (Japan)

Abstract

This paper overviews recent progress in the development of corpus-based spontaneous speech recognition technology focusing on various achievements of a Japanese 5-year national project "Spontaneous Speech: Corpus and Processing Technology". Although speech is in almost any situation spontaneous, recognition of spontaneous speech is an area which has only recently emerged in the field of automatic speech recognition. Broadening the application of speech recognition depends crucially on raising recognition performance for spontaneous speech. For this purpose, it is necessary to build large spontaneous speech corpora for constructing acoustic and language models. Because of various spontaneous-speech specific phenomena, recognition of spontaneous speech requires various new techniques. These new techniques include flexible acoustic modeling, sentence boundary detection, pronunciation modeling, acoustic as well as language model adaptation, and automatic summarization. Particularly automatic summarization including indexing, a process which extracts important and reliable parts of the automatic transcription, is expected to play an important role in building various speech archives, speech-based information retrieval systems, and human-computer dialogue systems.

Keywords: spontaneous speech, corpus, recognition, model adaptation, summarization

1. Introduction

Read speech and similar types of speech, e.g. news broadcasts reading a text, can be recognized with accuracy higher than 95%, using the state-of-the-art speech recognition technology. However, recognition accuracy drastically decreases for spontaneous speech (Furui, 2003a; Furui, 2003b). One of the major reasons for this decrease is that acoustic and language models used up until now have generally been built using written language or speech read from a text. Spontaneous speech and speech from written language are very different, both acoustically and linguistically. Spontaneous speech includes filled pauses, repairs, hesitations, repetitions, partial words, and disfluencies. It is quite interesting to note that, although speech is almost always spontaneous, spontaneous speech recognition is a special area that emerged only about 10 years ago within the wider field of automatic speech recognition (e.g. Shinozaki et al., 2001;

Sankar et al., 2002; Gauvain and Lamel; 2003; Evermann et al., 2004; Schwartz et al., 2004). Broadening the application of speech recognition depends crucially on raising the recognition performance for spontaneous speech.

In order to increase recognition performance for spontaneous speech, it is necessary to build acoustic and language models for spontaneous speech. Current methods applying statistical language modeling such as bigrams and trigrams of words or morphemes to spontaneous speech corpus may prove to be inadequate. Our knowledge of the structure of spontaneous speech is currently insufficient to achieve the necessary breakthroughs. Although spontaneous speech effects are quite common in human communication and may increase in human machine discourse as people become more comfortable conversing with machines, modeling of speech disfluencies is only in the initial stage. Since spontaneous speech includes various redundant expressions, recognition of spontaneous speech will require a paradigm shift from simply recognizing speech, where transcribing the spoken words is the primary focus, to understanding where underlying messages of the speaker are extracted.

2. Categories of speech recognition tasks

Speech recognition tasks can be classified into four categories, as shown in Table 1, according to two criteria: whether it is targeting utterances from human to human or human to computer, and whether the utterances have a dialogue or monologue style (Furui, 2003b). The table lists typical tasks for each category.

Table 1. Categorization of speech recognition tasks

	Dialogue	Monologue
Human to human	(Category I) Switchboard, Call Home (Hub 5), meeting, interview	(Category II) Broadcast news (Hub 4), other programs, lecture, presentation, voice mail
Human to machine	(Category III) ATIS, Communicator, information retrieval, reservation	(Category IV) Dictation

Category I targets human-to-human dialogues and includes DARPA-sponsored Switchboard and Call Home (Hub 5) tasks. Speech recognition research in this category aiming to make minutes of meetings (e.g. Janin et al., 2004) has recently started. Waibel et al. have been investigating a meeting browser that observes and tracks meetings for later review and summarization (Waibel and Rogina, 2003). Akita

et al. have investigated techniques for archiving discussions (Akita et al., 2003). In their method, speakers are automatically indexed in an unsupervised way, and speech recognition is performed using the results of indexing. Processing human-human conversational speech under unpredictable recording conditions and vocabularies presents new challenges for spoken language processing.

A relatively new task classified into this category is the MALACH (Multilingual Access to Large spoken ArCHives) project (Oard, 2004). Its goal is to advance the state-of-the-art technology for access to large multilingual collections of spontaneous conversational speech by exploiting an unmatched collection assembled by the Survivors of the Shoah Visual History Foundation (VHF). This collection is indeed a challenging task because of heavily accented, emotional and elderly spontaneous characteristics. Named entity tagging, topic segmentation, and unsupervised topic classification are also being investigated.

Tasks belonging to Category II, which targets recognizing human-to-human monologues, include transcription of broadcast news (Hub 4), news programs, lectures, presentations, and voice mails (e.g. Hirschberg et al., 2001). Speech recognition research in this category has recently become very active. Since the utterances in the Category II are made with the expectation that the audience can correctly understand what is spoken in the one-way communication, they are relatively easier to recognize than the utterances in Category I. If high recognition performance is achieved, a wide range of applications, such as making lecture notes, records of presentations and closed captions, archiving and retrieving these records, and retrieving voice mails, will be realized.

Most of the practical application systems widely used now are classified as Category III, recognizing the utterances in human-computer dialogues, such as in airline information services tasks. DARPA-sponsored projects including ATIS and Communicator have laid the foundations of these systems. Unlike other categories, the systems in the Category III are usually designed and developed after clearly defining the application/task. The machines that we have attempted to design so far are, almost without exception, limited to the simple task of converting a speech signal into a word sequence and then determining, from the word sequence, a meaning that is "understandable". Here, the set of understandable messages is finite in number, each being associated with a particular action (e. g., route a call to a proper destination or issue a buy order for a particular stock). In this limited sense of speech communication, the focus is detection and recognition rather than inference and generation.

Various research has made clear that the utterances spoken by people talking to computers, such as those in Categories III and IV, especially when the speaker is conscious, are acoustically as well as linguistically very different from those spoken to other people, such as those in Categories I and II. One of the typical tasks belonging to Category IV, which targets the recognition of monologues performed when people are talking to a computer, is dictation. Various commercial software for such purposes have been developed. Since the utterances in Category IV are made with the expectation that the utterances will be converted exactly into texts with correct characters, their spontaneity is much lower that that in Category III. In the four categories, spontaneity is considered to be the highest in Category I and the lowest in Category IV.

3. Spontaneous speech corpora

3.1. Issues of corpus construction

The appetite of today's statistical speech processing techniques for training material are well described by the aphorism: "There's no data like more data." Large structured collections of speech and text are essential for progress in speech recognition research. Unlike the traditional approach, in which knowledge of speech behavior is "discovered" and "documented" by human experts, statistical methods provide an automatic procedure to directly "learn" regularities in the speech data. The need for a large set of good training data is, thus, more critical than ever. However, establishing a good speech database for the computer to uncover the characteristics of the signal is not a straightforward process. There are basically two broad issues to be carefully considered: one being the content and its annotation, and the other the collecting mechanism.

The recorded data needs to be verified, labeled, and annotated by people whose knowledge is introduced into the design of the system through its learning process (i.e. via supervised training of the system after the data has been labeled). Labeling and annotation for spontaneous speech can easily become unmanageable. For example: how do we annotate speech repairs and partial words? how do the phonetic transcribers reach a consensus in acoustic-phonetic labels when there is ambiguity? and how do we represent a semantic notion? Errors in labeling and annotation will result in system performance degradation. How to ensure the quality of the annotated results is thus a major concern. Research limited only to automating or creating tools to assist the verification procedure is in itself an interesting subject.

3.2. Corpus of Spontaneous Japanese (CSJ)

In the above-mentioned context, a 5-year Science and Technology Agency Priority Program entitled "Spontaneous Speech: Corpus and Processing Technology" was conducted in Japan from 1999 to 2004 (Furui, 2003a), and a large-scale spontaneous speech corpus, Corpus of Spontaneous Japanese (CSJ), consisting of roughly 7M words with a total speech length of 650 hours was built (Maekawa, 2003; Maekawa et al., 2004).

Mainly recorded are monologues such as academic presentations (AP) and extemporaneous presentations (EP). AP is live recordings of academic presentations in nine different academic societies covering the fields of engineering, social science and humanities. EP is studio recording of paid layman speakers' speech on everyday topics like "the most delightful memory of my life" presented in front of a small audience and in a relatively relaxed atmosphere. The age and gender of EP speakers are more balanced than that of AP speakers. The CSJ also includes some dialogue speech for the purpose of comparison with monologue speech. The recordings were manually given orthographic and phonetic transcription. Spontaneous speech-specific phenomena, such as filled pauses, word fragments, reduced articulation and mispronunciation, as well as non-speech events like laughter and coughing were also carefully tagged.

One-tenth of the utterances, hereafter referred to as the Core, were tagged manually and used for training a morphological analysis and part-of-speech (POS) tagging program (Uchimoto et al., 2003) for automatically analyzing all of the 650-hour utterances (see Fig. 1). The Core consists of 70 APs, 107 EPs, 18 dialogues and 6 read speech files. They were also tagged with para-linguistic/intonation information, dependency-structure, discourse structure, and summarization. For intonation labeling of spontaneous speech, the traditional J_ToBI (Venditti, 1997) was extended to X_JToBI (Maekawa et al., 2002), in which inventories of tonal events as well as break indices were considerably enriched.

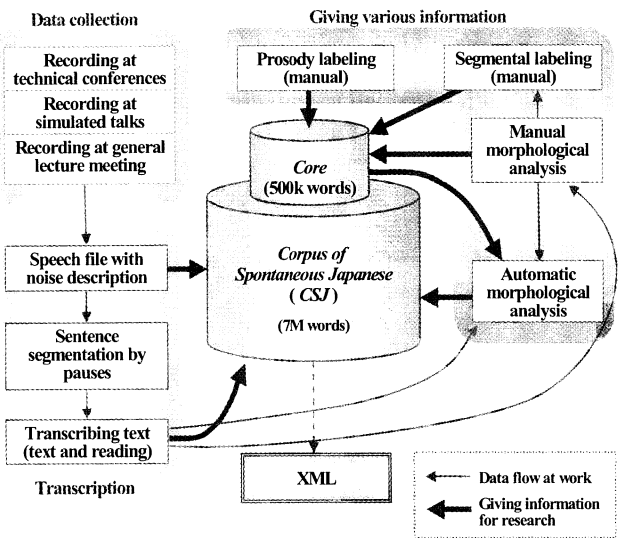

Fig. 1. CSJ corpus construction

4. Progress made in spontaneous speech recognition using the CSJ

4.1. Effectiveness of the CSJ

By constructing acoustic and language models using the CSJ, recognition errors for spontaneous presentation were reduced to roughly half compared to models constructed using read speech and written text (Furui, 2003a; Shinozaki et al., 2001). Increasing the size of training data for acoustic and language models has significantly decreased the recognition error rate (WER: word error rate), and the best WER of 25.3% was obtained when the whole training data set (510 hours, 6.84M words) was used (Ichiba et al.,2004; Furui et al., 2005). When the acoustic model was constructed using the whole training data set and the language model training data size was increased from 1/8 (0.86M words) to 8/8 (6.84M words), the WER, the perplexity and the OOV were relatively reduced by 17%, 19%, and 62%, respectively. On the other hand, when the

language model was made using the whole training data set (6.84M words) and the acoustic model training data was increased from 1/8 (68 hours) to 8/8 (510 hours), the WER was reduced by 6.3%. The WER was almost saturated by using the whole data set.

4.2. Pronunciation variation

In spontaneous speech, pronunciation variation is so diverse that multiple surface form entities are needed for many lexical items. Kawahara et al. (Kawahara et al., 2004) have found that statistical modeling of pronunciation variations integrated with language modeling is effective in suppressing false matching of less frequent entries. They have adopted a trigram model of word-pronunciation entries. Since both orthographic and phonetic transcriptions of the CSJ were made manually for each unit of utterance (sentence), word-based automatic alignment between them was performed to obtain the pronunciation entries for each word. This was incorporated as a post-processor of the morphological analyzer. Heuristic thresholding was applied to eliminate erroneous patterns, in which pronunciation entries whose occurrence probability in each lexical item is lower than a threshold were eliminated. As a result, 30,820 word-pronunciation entries (24,437 distinct words) were obtained, on which a trigram model was trained. Experimental results show that the word-pronunciation trigram model is more effective than simply adding the pronunciation probability in the decoding process.

4.3. Sentence boundary detection

Another difficulty of spontaneous speech recognition is that generally no explicit sentence boundary is given. Therefore, it is impossible to recognize spontaneous speech sentence by sentence. Kawahara et al. developed a decoder in which no sentence boundaries are required (Kawahara et al., 2001). The decoder can handle very long speech with no prior sentence segmentation. Experimental results show that the new decoder performed better than the previous version using sentence boundaries. Based on transcription results and pause lengths, sentence boundaries are automatically determined and punctuation marks are given. Specifically, a linguistic likelihood ratio between a model including sentence boundary and a model without boundary is compared with a threshold and then a decision is made.

4.4. Acoustic model adaptation

Word accuracy varies largely from speaker to speaker. There exist many factors that affect the accuracy of spontaneous speech recognition. They include individual voice characteristics, speaking manners, styles of language (grammar), vocabularies, topics, and noise, such as coughs. Even if utterances are recorded using the same close-talking microphones, acoustic conditions still vary according to the recording environment. A batch-type unsupervised adaptation method has been incorporated to cope with speech variation in the CSJ utterances (Shinozaki et al., 2001). The MLLR method using a binary regression class tree to transform Gaussian mean vectors was employed. The regression class tree was made using a centroid-splitting algorithm.

The actual classes used for transformation were determined at run time according to the amount of data assigned to each class. By applying the adaptation, the error rate was reduced by 15% relative to the speaker independent condition.

4.5. Language model adaptation

Lussier et al. investigated combinations of unsupervised language model adaptation methods for CSJ utterances (Lussier et al., 2004). Data sparsity is a common problem shared by all speech recognition tasks but it is especially acute in the case of spontaneous speech recognition. The method proposed combines information from two readily available sources, clusters of presentations from the training corpus and the transcription hypothesis, to create word-class n-gram models that are then interpolated with a general language model. The interpolation coefficient is estimated based on EM algorithm using a development set. Since this method performs in an offline manner using whole recognition results to suppress influences of local recognition errors, it is more robust against recognition errors than online adaptation methods. Experimental results show that a relative reduction in word error rate of 5-10% is obtained on the CSJ test sets.

4.6. Massively Parallel Decoder-based recognition

Shinozaki et al. have proposed using a combination of cluster-based language models and acoustic models in the framework of a Massively Parallel Decoder (MPD) to cope with the problem of acoustic as well as linguistic variations of presentation utterances (Shinozaki and Furui, 2004). MPD is a parallel decoder that has a large number of decoding units, in which each unit is assigned to each combination of element models. Likelihood values produced by all the decoding units are compared, and the hypothesis having the largest likelihood is selected as the recognition result. The system runs efficiently on a parallel computer, and thus the turnaround time is comparable to the conventional decoder using a single model and processor. In experiments conducted using presentation speeches from the CSJ, two types of cluster models have been investigated: presentation-based cluster models and utterance-based cluster models. It has been confirmed that utterance-based cluster models give significantly lower recognition error rate than presentation-based cluster models in both language and acoustic modeling. It has also been shown that roughly 100 decoding units are sufficient in terms of recognition rate; and, in the best setting, 12% reduction in word error rate was obtained in comparison with the conventional decoder.

5. Spontaneous speech summarization

Spontaneous speech is ill-formed and very different from written text. Spontaneous speech usually includes redundant information such as disfluencies, fillers, repetitions, repairs and word fragments. In addition, irrelevant information caused by recognition errors is usually inevitably included when spontaneous speech is transcribed. Therefore, an approach in which all words are simply transcribed is not an effective one for spontaneous speech. Instead, speech summarization which extracts important

information and removes redundant and incorrect information is ideal for recognizing spontaneous speech. Speech summarization is also expected to reduce time needed for reviewing speech documents and improve the efficiency of document retrieval.

Speech summarization has a number of significant challenges that distinguish it from general text summarization. Applying text-based technologies (Mani and Maybury, 1999) to speech is not always workable and often they are not equipped to capture speech specific phenomena (Kolluru et al, 2003; Christensen et al., 2003). One fundamental problem with the speech summarization is that they contain speech recognition errors and disfluencies. We have proposed a two-stage summarization method consisting of important sentence extraction and word-based sentence compaction, as shown in Fig. 2 (Hori and Furui, 2003; Hori et al., 2003; Kikuchi et al., 2003). After removing all the fillers based on speech recognition results, a set of relatively important sentences is extracted, and sentence compaction is applied to the set of extracted sentences. The ratio of sentence extraction and compaction is controlled according to a summarization ratio initially determined by the user. Sentence and word units are extracted from the speech recognition results and concatenated for producing summaries so that they maximize the weighted sum of linguistic likelihood, amount of information, confidence measure, and grammatical likelihood of concatenated units. The proposed method has been applied to summarization of broadcast news utterances as well as unrestricted-domain spontaneous presentations and has been evaluated by objective and subjective measures. It has been confirmed that the proposed method is effective in both English and Japanese speech summarization.

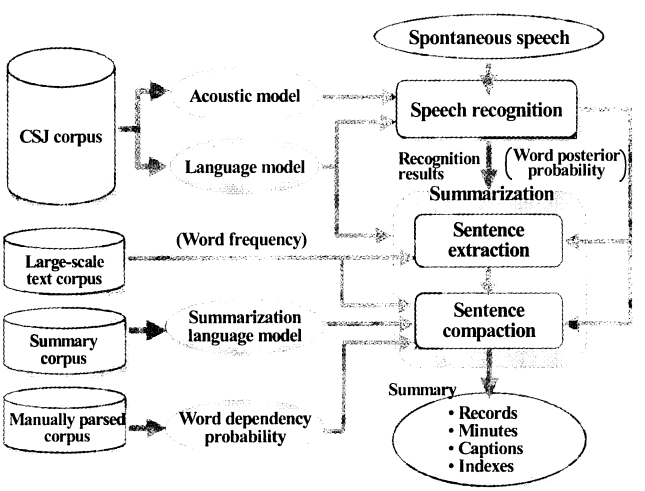

Fig. 2. A two-stage automatic speech summarization system

Speech summarization technology can be applied to any kind of speech document and is expected to play an important role in building various speech archives including broadcast news, lectures, presentations, and interviews. Summarization and question answering (QA) perform similar tasks, in that they both map an abundance of

information to a (much) smaller piece which is then presented to the user. Therefore, speech summarization research will help the advancement of QA systems using speech documents. By condensing important points of long presentations and lectures and presenting them in a summary speech, the system can provide the listener with a valuable means for absorbing more information in a much shorter time.

6. Conclusions and future research

Since speech is the most natural and effective method of communication between human beings, various important speech documents, including lectures, presentations, meeting records and broadcast news, are produced everyday. However, it is not easy to quickly review, retrieve, selectively disseminate, and reuse these speech documents, if they are simply recorded as audio signal. Therefore, automatically transcribing speech using speech recognition technology is a crucial aspect of creating knowledge resources from speech.

Speech recognition technology is expected to be applicable not only to indexing of speech data (lectures, broadcast news, etc.) for information extraction and retrieval, but also to closed captioning and aids for the handicapped. Broadening these applications depends crucially on raising the recognition performance of spontaneous speech. Various spontaneous speech corpora and processing technologies have recently been created under several recent projects. However, how to incorporate filled pauses, repairs, hesitations, repetitions, partial words, and disfluencies still poses a big challenge in spontaneous speech recognition.

The large-scale spontaneous speech corpus, CSJ (Corpus of Spontaneous Japanese) will be stored with XML format in a large-scale database system developed by the COE (Center of Excellence) program "Framework for Systematization and Application of Large-scale Knowledge Resources" at Tokyo Institute of Technology so that the general population can easily access and use it for research purposes (Furui, 2004). Since the recognition accuracy for spontaneous speech is still rather low, the collection of the corpus will be continued in the COE program in order to increase coverage of variations in spontaneous speech.

Spectral analysis using various styles of utterances in the CSJ shows that the spectral distribution/difference of phonemes is significantly reduced in spontaneous speech compared to read speech (Furui et al., 2005). This is considered as one of the reasons for the difficulty of spontaneous speech recognition. It has also been observed that speaking rates of both vowels and consonants in spontaneous speech are significantly faster than those in read speech. In our previous experiment for recognizing spontaneous presentations, it was found that speaking rate was variable even within a sentence and therefore it was effective to model local speaking rate variation using stochastic models such as dynamic Bayesian networks (Shinozaki and Furui, 2003).

Although it is quite obvious that human beings effectively use prosodic features in speech recognition, how to use them in automatic speech recognition is still difficult. This is mainly because prosodic features are difficult to extract automatically and correctly from speech signal and difficult to model due to their dynamic natures. This is especially true for spontaneous speech.

This paper has focused on corpus-based spontaneous speech recognition issues mainly from the viewpoint of human-to-human monologue speech processing (Category II). Most of the issues discussed in this paper, however, are expected to be applicable to another important category, human-computer dialogue interaction (Category III).

Acknowledgments

The author would like to thank all members of the "Spontaneous Speech" project as well as students at Tokyo Institute of Technology and Kyoto University who contributed to building the large-scale corpus and creating the new technologies described in this paper.

References

Akita, Y., Nishida, M. and Kawahara, T., 2003. Automatic transcription of discussions using unsupervised speaker indexing. In: Proc. IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, pp. 79-82.

Christensen, H., Gotoh, Y., Kolluru, B. and Renals, S., 2003. Are extractive text summarization techniques portable to broadcast news," In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, St. Thomas, pp. 489-494.

Evermann, G.. et al., 2004. Development of the 2003 CU-HTK conversational telephone speech transcription system. In: Proc. IEEE ICASSP, Montreal, pp. I-249-252.

Furui, S., 2003a. Recent advances in spontaneous speech recognition and understanding. In: Proc. IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, pp. 1-6.

Furui, S., 2003b. Toward spontaneous speech recognition and understanding. In: Pattern Recognition in Speech and Language Processing. Chou, W., Juang, B.-H. (Eds.), CRC Press, New York, pp. 191-227.

Furui, S., 2004. Overview of the 21st century COE program "Framework for Systematization and Application of Large-scale Knowledge Resources". In: Proc. International Symposium on Large-scale Knowledge Resources, Tokyo, pp. 1-8.

Furui, S., Nakamura, M., Ichiba, T. and Iwano, K., 2005. Analysis and recognition of spontaneous speech using Corpus of Spontaneous Japanese. In: Speech Communication (to be published)

Gauvain, J.-L. and Lamel, L., 2003. Large vocabulary speech recognition based on statistical methods. In: Pattern Recognition in Speech and Language Processing. Chou, W., Juang, B.-H.(Eds.), CRC Press, New York, pp. 149-189.

Hori, C. and Furui, S., 2003. A new approach to automatic speech summarization," In: IEEE Trans. Multimedia, pp. 368-378.

Hori, C., Furui, S., Malkin, R., Yu, H. and Waibel, A., 2003. A statistical approach to automatic speech summarization," In: EURASIP Journal on Applied Signal Processing, pp. 128-139.

Hirschberg, J., Bacchiani, M., Hindle, D., Isenhour, P., Rosenberg, A., Stark, L., Stead, L., Whittaker, S. and Zamchick, G., 2001. SCANMail: Browsing and

searching speech data by content. In: Proc. Eurospeech 2001, Aalborg, pp. 2377-2380.

Ichiba, T., Iwano, K. and Furui, S., 2004. (in Japanese) Relationships between training data size and recognition accuracy in spontaneous speech recognition. In: Proc. Acoustical Society of Japan Fall Meeting, 2-1-9.

Janin, A., Ang, J., Bhagat, S., Dhillon, R., Edwards, J., Macias-Guarasa, J., Morgan, N., Peskin, B., Shriberg, E., Stolcke, A., Wooters, C. and Wrede, B., 2004. The ICSI meeting project: Resources and research. In: Proc. NIST ICASSP 2004 Meeting Recognition Workshop, Montreal.

Kawahara, T., Kitade, T., Shitaoka, K. and Nanjo, H., 2004. Efficient access to lecture audio archives through spoken language processing. In: Proc. Special Workshop in Maui (SWIM).

Kawahara, T., Nanjo, H. and Furui, S. 2001. Automatic transcription of spontaneous lecture speech, In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, Madonna di Campiglio, Italy.

Kikuchi, T., Furui, S. and Hori, C., 2003. Two-stage automatic speech summarization by sentence extraction and compaction. In: Proc. ISCA-IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, TAP10.

Kolluru, B., Christensen, H., Gotoh, Y. and Renals, S., 2003. Exploring the style-technique interaction in extractive summarization of broadcast news," In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, St. Thomas, pp. 495-500.

Lussier, L., Whittaker, E. W. D. and Furui, S., 2004. Combinations of language model adaptation methods applied to spontaneous speech. In: Proc. Third Spontaneous Speech Science & Technology Workshop, Tokyo, pp. 73-78.

Maekawa, K., 2003. Corpus of spontaneous Japanese: its design and evaluation. In: Proc. IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, pp. 7-12.

Maekawa, K., Kikuchi, H., Igarashi, Y. and Venditti, J., 2002. X-JtoBI: an extended J_ToBI for spontaneous speech. In: Proc. ICSLP 2002, Denver, pp. 1545-1548.

Maekawa, K., Kikuchi, H. and Tsukahara, W., 2004. Corpus of spontaneous Japanese: design, annotation and XML representation. In: Proc. International Symposium on Large-scale Knowledge Resources, Tokyo, pp. 19-24.

Mani, I. and Maybury, M.T. (Eds.), 1999. Advances in automatic text summarization," MIT Press, Cambridge, MA.

Oard, D. W., 2004. Transforming access to the spoken word. In: Proc. International Symposium on Large-scale Knowledge Resources, Tokyo, pp. 57-59.

Sankar, A., Gadde, V. R. R., Stolcke, A. and Weng, F., 2002. Improved modeling and efficiency for automatic transcription of broadcast news. In: Speech Communication, 37, pp. 133-158.

Schwartz, R. et al., 2004. Speech recognition in multiple languages and domains: the 2003 BBN/LIMSI EARS system. In: Proc. IEEE ICASSP, Montreal, pp. III-753-756.

Shinozaki, T. and Furui, S., 2003. Hidden mode HMM using Bayesian network for modeling speaking rate fluctuation. In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, St. Thomas, pp. 417-422.

Shinozaki, T. and Furui, S., 2004. Spontaneous speech recognition using a massively

parallel decoder. In: Proc. Interspeech-ICSLP, Jeju, Korea, 3, pp. 1705-1708.
Shinozaki, T., Hori, C. and Furui, S., 2001. Towards automatic transcription of spontaneous presentations. In: Proc. Eurospeech2001, Aalborg, Denmark, pp. 491-494.
Uchimoto, K., Nobata, C., Yamada, A., Sekine, S. and Isahara, H., 2003. Morphological analysis of the Corpus of Spontaneous Japanese. In: Proc. IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, pp. 159-162.
Venditti, J., 1997. Japanese ToBI labeling guidelines. In: OSU Working Papers in Linguistics, 50, pp. 127-162.
Waibel, A. and Rogina, I., 2003. Advances on ISL's lecture and meeting trackers. In: Proc. IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, pp. 127-130.

SADAOKI FURUI is currently a Professor at Tokyo Institute of Technology, Department of Computer Science. He is engaged in a wide range of research on speech analysis, speech recognition, speaker recognition, speech synthesis, and multimodal human-computer interaction and has authored or coauthored over 400 published articles. He is a Fellow of the IEEE, the Acoustical Society of America and the Institute of Electronics, Information and Communication Engineers of Japan (IEICE). He is President of the International Speech Communication Association (ISCA). He has served as President of the Acoustical Society of Japan (ASJ) and the Permanent Council for International Conferences on Spoken Language Processing (PC-ICSLP). He has also served as a Board of Governor of the IEEE Signal Processing Society, and Editor-in-Chief of Speech Communication as well as the Transaction of the IEICE. He has received numerous awards from IEEE, IEICE, ASJ and the Japanese Minister of Science and Technology. E-mail: furui@cs.titech.ac.jp.