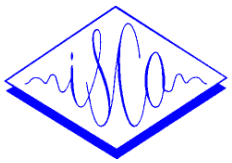


ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13-16, 2003

Recent Advances in Spontaneous Speech Recognition and Understanding

Sadaoki Furui
Tokyo Institute of Technology, Department of Computer Science
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
furui@cs.titech.ac.jp

***Abstract*—How to recognize and understand spontaneous speech is one of the most important issues in state-of-the-art speech recognition technology. In this context, a five-year large-scale national project entitled "Spontaneous Speech: Corpus and Processing Technology" started in Japan in 1999. This paper gives an overview of the project and reports on the major results of experiments that have been conducted so far at Tokyo Institute of Technology, including spontaneous presentation speech recognition, automatic speech summarization, and message-driven speech recognition. The paper also discusses the most important research problems to be solved in order to achieve ultimate spontaneous speech recognition systems.**

I. Introduction

Speech recognition systems are expected to play important roles in an advanced IT society with user-friendly human-machine interfaces [1]. The field of automatic speech recognition has witnessed a number of significant advances in the past 10-20 years, spurred on by advances in signal processing, algorithms, computational architectures, and hardware. These advances include the widespread adoption of a statistical pattern recognition paradigm, a data-driven approach which makes use of a rich set of speech utterances from a large population of speakers, the use of stochastic acoustic and language modeling, and the use of dynamic programming-based search methods [2][3][4].

Read speech and similar types of speech, e.g. that from reading newspapers or from news broadcast, can be recognized with accuracy higher than 90% using the state-of-the-art speech recognition technology. However, recognition accuracy drastically decreases for spontaneous speech. This decrease is due to the fact that the acoustic and linguistic models used have generally been built using written language or speech from written language. Unfortunately spontaneous speech and speech from written language are very different both acoustically and linguistically. Broadening the application of speech recognition thus crucially depends on raising the recognition performance for spontaneous speech. In order to increase the recognition performance for spontaneous speech, it is crucial to build acoustic and language models for spontaneous speech. Our knowledge of the structure of spontaneous speech is currently inadequate to achieve the necessary breakthroughs. Although spontaneous speech effects are quite common in human communication and may be expected to increase in human machine discourse as people become more comfortable conversing with machines, modeling of speech disfluencies is only just beginning. Recognition of spontaneous speech will require a paradigm shift from speech recognition to understanding where underlying messages of the speaker are extracted, instead of transcribing all the spoken words [5].

We can envision a great information revolution on par with the development of writing systems, if we can successfully meet the challenges of speech both as a medium for information access and as itself a source of information. Speech is still the means of communication used first and foremost by humans, and only a small percentage of human communication is written. Automatic speech understanding can add many of the advantages normally associated only with text (random access, sorting, and access at different times and places) to the many benefits of speech. Making this vision a reality will require significant advances.

II. Japanese National Project on Spontaneous Speech Corpus and Processing Technology

For building language models for spontaneous speech, large spontaneous speech corpora are indispensable. In this context, a Science and Technology Agency Priority Program entitled "Spontaneous Speech: Corpus and Processing Technology" started in Japan in 1999 [6]. The project will be conducted over a 5-year period under the following three major themes as shown in Fig. 1.

1) Building a large-scale spontaneous speech corpus, Corpus of Spontaneous Japanese (CSJ), consisting of roughly 7M words with the total speech length of 700 hours. Mainly recorded will be monologues such as lectures, presentations and news commentaries. The recordings will be manually given orthographic and phonetic transcription. One-tenth of the utterances, hereafter referred to as the *Core*, will be tagged manually and used for training a morphological analysis and part-of-speech (POS) tagging program for automatically analyzing all of the 700-hour utterances. The *Core* will also be tagged with para-linguistic information including intonation (see Fig. 2).

2) Acoustic and linguistic modeling for spontaneous speech understanding using linguistic as well as para-linguistic information in speech.

3) Investigating spontaneous speech summarization technology.

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.

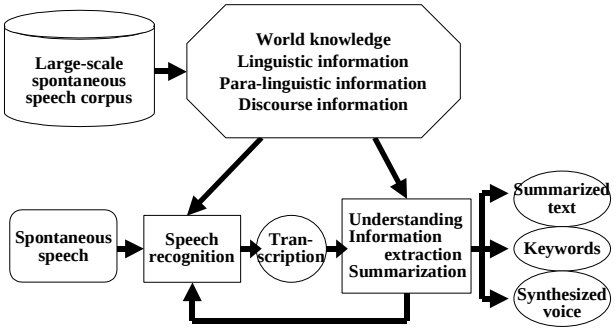


Fig. 1 - Overview of the Japanese national project on spontaneous speech corpus and processing technology.

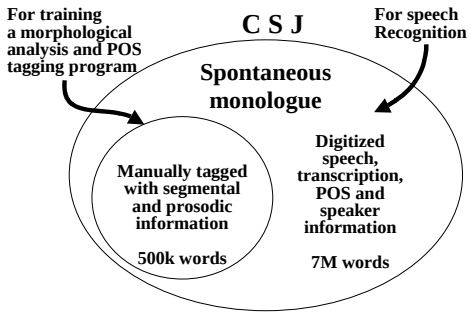


Fig. 2 - Overall design of the *Corpus of Spontaneous Japanese.*

III. Automatic Transcription of Spontaneous Presentations

3.1 Recognition task

Using the CSJ corpus, preliminary recognition experiments are being conducted at Tokyo Institute of Technology as well as at several other universities participating in the project. In this experiment, 4.4 hours of presentation speech uttered by 10 male speakers is used as a test set of speech recognition [7].

The following two corpora are used for training the language and acoustic models.

CSJ: A part of the corpus completed by the end of December 2000, consisting of 610 presentations (approximately 1.5M words of transcriptions), is used.

Web corpus: Transcribed presentations consisting of approximately 76k sentences with 2M words have been collected from the World Wide Web. Spontaneous speech usually includes various filled pauses but they are not included in this presentation corpus. An effort is thus made to add filled pauses to the presentation corpus based on the statistical characteristics of the filled pauses. The topics of the presentations cover wide domains including social issues and memoirs.

The following two language models, denoted as SpnL and WebL, have been constructed. Each model consists of bigrams and reverse trigrams with backing-off. Their vocabulary sizes are 30k words.

SpnL: Made using the 610 presentations in the CSJ. The speakers have no overlap with those of the test set. Since there are no punctuation marks in the transcription, commas are inserted when a silence period of 200ms or longer is encountered.

WebL: Made using the text of our Web corpus.

The following two tied-state triphone HMMs have been made, both having 2k states and 16 Gaussian mixtures in each state.

SpnA: Using 338 presentations in the CSJ uttered by male speakers (approximately 59 hours). The speakers have no overlap with those in the test set.

RdA: Using approximately 40-hours of read speech uttered by many speakers.

3.2 Recognition results

Figure 3 presents the test-set perplexity of trigrams and the out-of-vocabulary (OOV) rate for each language model. The perplexity of **SpnL,** made from the CSJ, is clearly better than that of the web-based model. **WebL** shows high perplexity and OOV rate, since it was edited as a text and their topics are much more diversified than those of the test set.

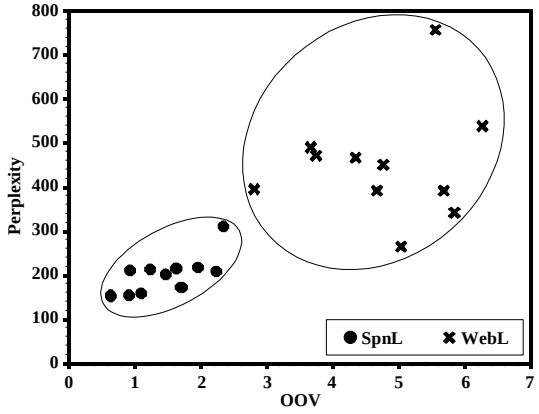
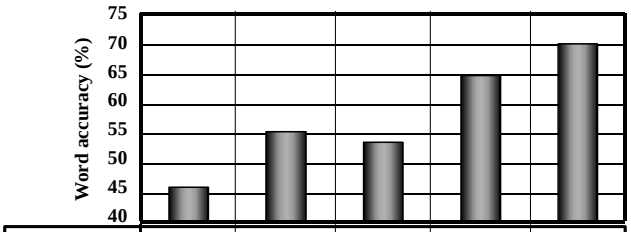


Fig. 3 - Test-set perplexity (PP) and OOV rate for the three language models.

Figure 4 shows recognition results for the combinations of the two language models, **SpnL** and

WebL, and the two acoustic models, **SpnA** and **RdA**. Fillers are counted as words and included in calculating the accuracy. It is clearly shown that **SpnL** achieves much better results than **WebL**, and **SpnA** gives much better results than **RdA**. These results indicate that it is crucial to make language models from a spontaneous speech corpus to adequately recognize spontaneous speech. It is also suggested that acoustic models made from CSJ have better coverage of triphones and better matching of acoustic characteristics corresponding to the speaking style and also have better matching of recording conditions with the test set. The mean accuracy for the combination of **SpnL** and **SpnA** is 65.3%.



Acous. model	RdA	SpnA	RdA	SpnA	SpnA
Ling. model	WebL	WebL	SpnL	SpnL	SpnL
Speaker adapt.	w/o	w/o	w/o	w/o	with

Fig. 4 - Word accuracy for each combination of models.

The 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 and noise like coughs. Although all utterances were recorded using the same close-talking microphones, acoustic conditions still varied according to the recording environment. A batch-type unsupervised adaptation method has been incorporated to cope with the speech variation. The MLLR method [8] using a binary regression class tree to transform Gaussian mean vectors is employed. The regression class tree is made using a centroid-splitting algorithm. The actual classes used for transformation are determined at run time according to the amount of data assigned to each class. By applying the adaptation, the error rate is reduced by 15% relative to the speaker independent case, and the accuracy is raised to 70.5% as shown in Fig. 4.

3.3 Analysis on individual differences

Individual differences in spontaneous presentation speech recognition performances have been analyzed using 10 minutes from each presentation given by 51 male speakers, for a total of 510 minutes [9]. Seven kinds of speaker attributes have been considered in the analysis. They are word accuracy (Acc), averaged acoustic frame likelihood (AL), speaking rate (SR), word perplexity (PP), out of vocabulary rate (OR), filled pause rate (FR) and repair rate (RR). The speaking rate defined as the number of phonemes per second and the averaged acoustic frame likelihood are calculated using the result of forced alignment of the reference tri-phone labels after removing pause periods. The word perplexity is calculated using trigrams, in which prediction of out of vocabulary words is not included. The filled pause rate and the repair rate are the number of filled pauses and repairs divided by the number of words, respectively.

Figure 5 shows correlation between the seven attributes. This result indicates that the attributes having real correlation with the accuracy are speaking rate, out of vocabulary rate, and repair rate.

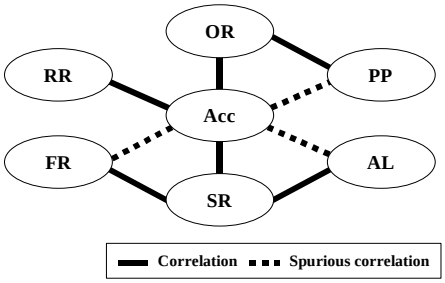


Fig. 5 – Correlation between various attributes; Acc: word accuracy, OR: out of vocabulary rate, RR: repair rate, FR: filled pause rate, SR: speaking rate, AL: averaged acoustic frame likelihood, PP: word perplexity.

The following equation has been obtained as a result of linear regression model of the word accuracy with the six presentation attributes.

$$Acc = 0.12\,AL - 0.88\,SR - 0.020\,PP - 2.2\,OR + 0.32\,FR - 3.0\,RR + 95 \quad (1)$$

In the equation, the regression coefficient for the repair rate is -3.0 and the coefficient for the out of vocabulary rate is -2.2. This means that a 1% increase of the repair rate or the out of vocabulary rate respectively corresponds to a 3.0% or 2.2% decrease of the word accuracy. This is probably because a single recognition error caused by a repair or an out of vocabulary word triggers secondary errors due to the linguistic constraints. The determination coefficients of the multiple linear regression is 0.48, which is significant at 1% level. This means that roughly half of the variance of the word accuracy can be explained by the model.

Normalized representation of the regression analysis, in which the variables are normalized in terms of the mean and variance before the analysis in order to show the effects of explaining variables on the word accuracy, indicates that coefficients of the speaking rate, the out of vocabulary rate and the repair rate are relatively large.

IV. Automatic Speech Summarization and Evaluation

4.1 Sentence compaction-based summarization

Currently various new applications of LVCSR systems,

such as automatic closed captioning, making minutes of meetings and conferences, and summarizing and indexing of speech documents for information retrieval, are actively being investigated. Transcribed speech usually includes not only redundant information such as disfluencies, filled pauses, repetitions, repairs and word fragments, but also irrelevant information caused by recognition errors. Therefore, especially for spontaneous speech, practical applications using speech recognizer require a process of speech summarization which removes redundant and irrelevant information and extracts relatively important information corresponding to users' requirements. Speech summarization producing understandable and compact sentences from original utterances can be considered as a kind of speech understanding.

A method for automatically summarizing speech based on sentence compaction has been investigated [10]. The method can be applied to the summarization of each sentence/utterance and also to a set of multiple sentences. The basic idea of this method is to extract a set of words maximizing a summarization score from an automatically transcribed sentence according to a target compression ratio (Fig. 6). This method aims to effectively reduce the number of words by removing redundant and irrelevant information without losing relatively important information. The summarization score indicating the appropriateness of a summarized sentence consists of a word significance score I as well as a confidence score C for each word of the original sentence, a linguistic score L for the word string in the summarized sentence, and a word concatenation score T_r. The word concatenation score indicates a word concatenation probability determined by a dependency structure in the original sentence given by a Stochastic Dependency Context Free Grammar (SDCFG). The total score is maximized using a dynamic programming (DP) technique.

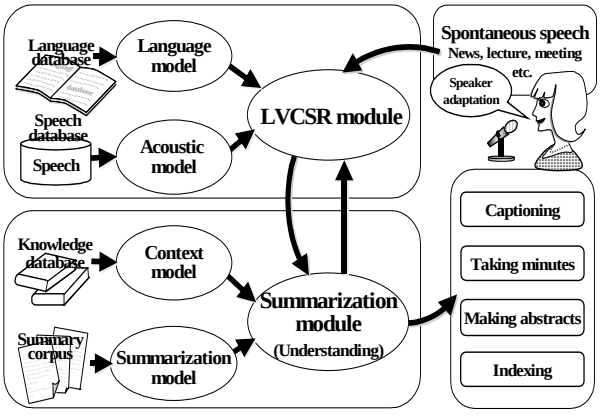


Fig. 6 – Automatic speech summarization system.

Given a transcription result consisting of N words, $W=w_1,w_2,\ldots,w_N$, the summarization is performed by extracting a set of M ($M<N$) words, $V=v_1,v_2,\ldots,v_M$, which maximizes the summarization score given by eq.(2).

$$S(V) = \sum_{m=1}^{M} \{ L(v_m \mid \cdots v_{m-1}) + \lambda_I I(v_m) + \lambda_C C(v_m) + \lambda_T T_r(v_{m-1}, v_m) \} \quad (2)$$

where λ_I, λ_C and λ_T are weighting factors for balancing among L, I, C and T_r.

(a) Word significance score

The word significance score I indicates the relative significance of each word in the original sentence. Since the most important words conveying meanings are nouns and verbs, the amount of information based on the frequency of each word is given to each noun and verb as the word significance score, and a flat score is given to other words. To reduce the repetition of words in the summarized sentence, a flat score is also given to each reappearing noun and verb.

(b) Linguistic score

The linguistic score $L(v_m \mid \ldots v_{m-1})$ measured by a trigram probability $P(v_m \mid v_{m-2}, v_{m-1})$ indicates the appropriateness of word strings in a summarized sentence.

(c) Word confidence score

The confidence score $C(v_m)$ is incorporated to weight acoustically as well as linguistically reliable recognition results. Specifically, a posterior probability of each transcribed word, that is the ratio of a word hypothesis probability to that of all other hypotheses, is calculated using a word graph obtained by a decoder and used as a confidence measure.

(d) Word concatenation score

Suppose “the beautiful cherry blossoms in Japan” is summarized as “the beautiful Japan”. The latter phrase is a grammatically correct but semantically incorrect summarization. Since the above linguistic score is not powerful enough to avoid such a problem, a word concatenation score $T_r(v_{m-1}, v_m)$ is incorporated to give a penalty for a concatenation between words with no dependency in the original sentence.

Word concatenation in a summarized sentence is restricted by the dependency structure in the original sentence as exemplified in Fig. 7. The word at the beginning of an arrow is named “modifier” and the word at the end of the arrow is named “head” respectively. The dependency grammar consists of both “right-headed” dependency indicated by right arrows and “left-headed” dependency indicated by left arrows as shown in the figure. The dependencies can be written as phrase structure grammar, DCFG (Dependency Context Free Grammar).

Since the dependencies between words are usually ambiguous, whether dependencies exist or not between

words is given by probabilities that one word is modified by others based on the SDCFG. The word dependency probability is a posterior probability estimated by the Inside-Outside probabilities obtained using a manually parsed corpus. In the SDCFG, only the number of non-terminal symbols is determined and all combinations of rules are applied recursively. The non-terminal symbol has no specific function such as a noun phrase. Even if transcription results by a speech recognizer are ill-formed, the dependency structure can be robustly estimated by the SDCFG. Since Japanese sentences have only "right-headed" dependencies, word concatenation score for Japanese is more compact than English. In addition, the word dependency structure in each phrase is deterministic and can be represented by a regular grammar.

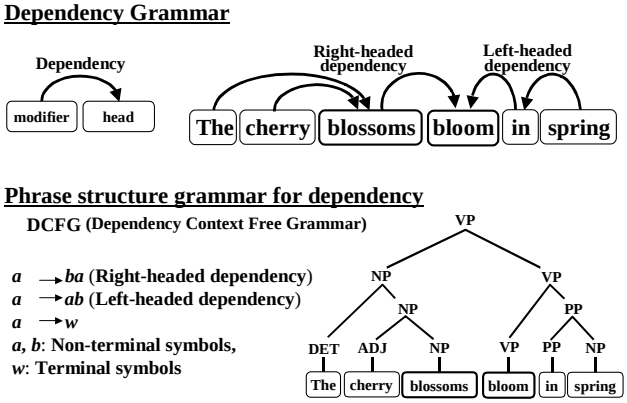


Fig. 7 – Dependency structure.

4.2 Evaluation

To automatically evaluate summarized sentences, correctly transcribed speech utterances are manually summarized by human subjects and used as correct targets. The manual summarization results are merged into a word network which approximately expresses all possible correct summarization including subjective variations. A summarization accuracy of automatic summarization is calculated using the word network. A word string extracted from the word network that is the most similar to the automatic summarization result is considered as a correct word string corresponding to the automatic summarization. The accuracy, comparing the summarized sentence with the word string extracted from the network, is used as an indicator of the "summarization accuracy", measuring the linguistic correctness and maintenance of the original meanings of the utterance.

In the English speech case, CNN-TV broadcast news utterances recorded in 1996 and provided by NIST as a test set of Topic Detection and Tracking (TDT) were tagged by the Brilltagger and used to evaluate the proposed method. Five news articles, consisting of 25 utterances in average, were transcribed by the JANUS [11] speech recognition system. The mean word recognition accuracy was roughly 80%.

In the Japanese speech case, NHK-TV broadcast news utterances recorded in 1996 were used to evaluate the proposed method. 50 utterances with word recognition accuracy above 90%, which was the average rate over the 50 utterances, were selected and used for the evaluation.

Experimental results show that the proposed method can effectively extract relatively important information and remove redundant and irrelevant information from Japanese as well as English news speech [12].

4.3 Combination with important sentence extraction

The sentence compaction-based method has a disadvantage that when multiple spontaneous utterances including many recognition errors and disfluencies are summarized with a high compression ratio (a small summarization ratio), the summary sometimes includes unnatural, incomplete sentences consisting of a small number of words, and it becomes difficult to read. To solve this problem, we have recently proposed a new two-stage summarization method, consisting of important sentence extraction and sentence compaction [13]. In the new method, relatively well-structured and important sentences including important information and fewer speech recognition errors are extracted, and sentence compaction is applied to the set of extracted sentences.

Figure 8 shows the two-stage summarization method. The important sentence extraction is performed according to the weighted sum of the word significance score, the linguistic score and the word confidence score averaged over each sentence. The ratio of sentence extraction and compaction are controlled according to a summarization ratio given by the user.

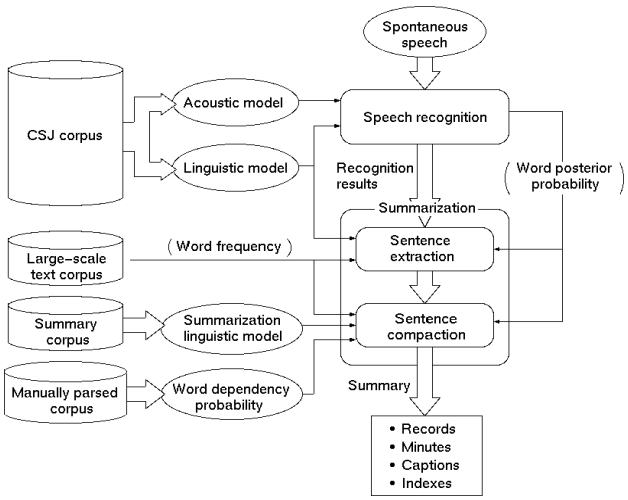


Fig. 8 – Automatic summarization based on the combination of important sentence extraction and sentence compaction.

One of the presentations in the CSJ by a male speaker having a length of roughly 12 minutes was summarized at

summarization ratios of 70% and 50%. The word recognition accuracy of this presentation is 70% on average. Experimental results show that, in both summarization conditions, the two-stage method achieves higher summarization accuracy than the previous one-stage method. It was also found that, when the summarization ratio becomes smaller, it is better to remove more sentences and then apply the sentence compaction technique to the remaining sentences to achieve better results. Comparing the three scores for the sentence extraction, the significance score or the confidence score is more effective than the linguistic score.

V. Conclusion

Although high recognition accuracy can be obtained using state-of-the art speech recognition technology for speech in the form of reading a written text or similar, the accuracy is quite poor for freely spoken spontaneous speech. To challenge this problem, a five-year national project for raising the technological level of speech recognition and understanding commenced in Japan in 1999. The project focuses on building a large-scale spontaneous speech corpus together with acoustic and linguistic modeling for spontaneous speech recognition and summarization. Experimental results show that acoustic and language modeling based on the actual spontaneous speech corpus is far more effective than modeling based on read speech. It is also shown that the proposed automatic speech summarization method effectively extracts relatively important information and removes redundant and irrelevant information.

Since the recognition accuracy for spontaneous speech is still rather low, it is imperative to continue the collection of a large corpus of spontaneous speech and use it for building language and acoustic models. Future research issues include: a) how to transcribe and annotate 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 disfluencies; e) how to adapt the language models to each task; f) how to adapt to speaking styles and topics of presentations; and g) how to build acoustic models that fit spontaneous speech.

Speech summarization will be applicable to a range of applications, such as making abstracts of presentations, preparing minutes of meetings and voicemails, close captioning of broadcast news, and presenting information in news-on-demand systems. Future research includes: a) evaluation using more test data with a manual summary; b) task-dependent evaluation from the viewpoint of how much the original meaning is maintained in the summarization results based on the IR performance; c) investigation of other useful information/features for important sentence extraction; and d) automatic optimization of the division of compression ratio into the two summarization stages.

A paradigm shift from speech recognition to understanding, where the underlying messages of the speaker, i.e., meaning/content that the speaker intended to convey, are extracted, instead of simply transcribing all the spoken words, will be indispensable.

Acknowledgment

The author would like to thank NHK (Japan Broadcasting Corporation) for providing us with the broadcast news database.

References

[1] B.-H. Juang and S. Furui, "Automatic recognition and understanding of spoken language – A first step towards natural human-machine communication," *Proc. IEEE*, 88, 8, pp. 1142-1165, 2000

[2] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, New Jersey, Prentice-Hall, Inc., 1993

[3] S. Furui, *Digital Speech Processing, Synthesis, and Recognition, 2nd Edition*, New York, Marcel Dekker, 2000

[4] H. Ney, "Corpus-based statistical methods in speech and language processing," in *Corpus-based Methods in Language and Speech Processing, Young, S. and Bloothooft, G. Ed.*, pp. 1-26, 1997

[5] B. H. Juang, "From speech recognition to understanding: Shifting paradigm to achieve natural human-machine communication," *Proc. 16th ICA and 135th Meeting ASA*, pp. 617-618, 1998

[6] Furui, S., Maekawa, K., Isahara, H., Shinozaki, T. and Ohdaira, T., "Toward the realization of spontaneous speech recognition – Introduction of a Japanese priority program and preliminary results –," *Proc. Int. Conf. Spoken Language Processing*, Beijing, pp. III-518-521, 2000

[7] T. Shinozaki, C. Hori, and S. Furui, "Towards automatic transcription of spontaneous presentations," *Proc. EUROSPEECH*, Aalborg, Vol. 1, pp.491-494, 2001.

[8] C. J. Legetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, 9, pp. 171-185, 1995

[9] T. Shinozaki and S, Furui, "Analysis on individual differences in automatic transcription of spontaneous presentations," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Orlando, pp. I-729-732, 2002

[10] C. Hori and S. Furui, "Automatic speech summarization based on word significance and linguistic likelihood," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Istanbul, pp. 1579-1582, 2000

[11] Z. Klaus, "Automatic generation of concise summaries of spoken dialogues in unrestricted domains," *Proc. SIGIR2001*, New Orleans, 2001

[12] C. Hori, S. Furui, R. Malkin, H. Yu and A. Waibel, "Automatic speech summarization applied to English broadcast news speech," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Orlando, pp. I-9-12, 2002

[13] T. Kikuchi, S. Furui and C. Hori, "Automatic speech summarization based on sentence extraction and compaction," Technical Report of IEICE, SP2002-158, 2002 (in Japanese)