Speech-to-Speech and Speech-to-Text Summarization

Sadaoki Furui, Tomonori Kikuchi, Yousuke Shinnaka
Department of Computer Science
Tokyo Institute of Technology
{furui, kikuchi, shinnaka}@furui.cs.titech.ac.jp

Chiori Hori
Intelligent Communication Laboratory
NTT Communication Science Laboratories
chiori@cslab.kecl.ntt.co.jp

Abstract

This paper presents techniques for speech-to-text and speech-to-speech automatic summarization. For the former case, a two-stage summarization method consisting of important sentence extraction and word-based sentence compaction is investigated. Sentence and word units which maximize the weighted sum of linguistic likelihood, amount of information, confidence measure, and grammatical likelihood of concatenated units are extracted from the speech recognition results. For the latter case, sentences, words and between-filler units are investigated as units to be extracted from original speech and concatenated for producing summaries. These methods are applied to summarization of spontaneous presentations and evaluated by objective and subjective measures. It was confirmed that proposed methods are effective in automatic speech summarization.

1. Introduction

One of the key applications of automatic speech recognition is to transcribe speech documents such as talks, presentations, lectures and broadcast news [1]. Although speech is the most natural and effective method of communication between human beings, it is not easy to quickly review, retrieve and reuse speech documents if they are simply recorded as audio signal. Therefore, transcribing speech is expected to become a crucial capability for the coming IT era. Although high recognition accuracy can be easily obtained for speech read from a text, such as anchor speakers' broadcast news utterances, technological ability for recognizing spontaneous speech is still limited [2]. 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 included in a transcription caused by recognition errors is commonly inevitable. Therefore, an approach in which all words are transcribed is not an effective one for spontaneous speech. Instead, speech summarization which extracts important information and removes redundant and incorrect information is necessary for recognizing spontaneous speech. Efficient speech summarization saves time for reviewing speech documents and improves the efficiency of document retrieval.

Summarization results can be presented by either text or speech. The former method has advantages in that: a) the documents can be easily looked through; b) the part of the documents that is interesting for users can be easily extracted; and c) information extraction and retrieval techniques can be easily applied to the documents. However, it has disadvantages in that: a) wrong information due to speech recognition errors cannot be avoided; and b) prosodic information such as the emotion of speakers that is conveyed only by speech cannot be presented. On the other hand, the latter method does not have such disadvantages and it can preserve all the acoustic information included in the original speech.

Methods for presenting summaries by speech can be classified into two categories; a) presenting simply concatenated speech segments that are extracted from original speech, or b) synthesizing summarization text by using a speech synthesizer. Since state-of-the-art speech synthesizers still cannot produce completely natural speech, the former method can easily produce better quality summarizations, and it does not have the problem of synthesizing wrong messages due to speech recognition errors. The major problem in using the extracted speech segments is how to avoid unnatural noisy sound caused by the concatenation.

This paper investigates automatic speech summarization techniques with the two presentation methods. In both cases, the most appropriate sentences, phrases or word units/segments are automatically extracted from original speech and concatenated to produce a summary. The extracted units cannot be reordered or replaced. Only when the summary is presented by text, transcription is modified into a written editorial article style by certain rules. When the summary is presented by speech, a concatenation-based method is used. Evaluation experiments are performed using spontaneous presentation utterances in the CSJ (Corpus of Spontaneous Japanese) made by the Spontaneous Speech Corpus and Processing project [3].

2. Summarization with text presentation

2.1 Two-stage summarization method

Figure 1 shows the two-stage summarization method consisting of important sentence extraction and sentence compaction [4]. Using speech recognition results, the score for important sentence extraction is calculated for each sentence. After removing all the fillers, a set of relatively important sentences is extracted, and sentence compaction using our proposed method [5, 6] is applied to the set of extracted sentences. The ratios of sentence extraction and compaction are controlled according to a summarization ratio initially determined by the user.

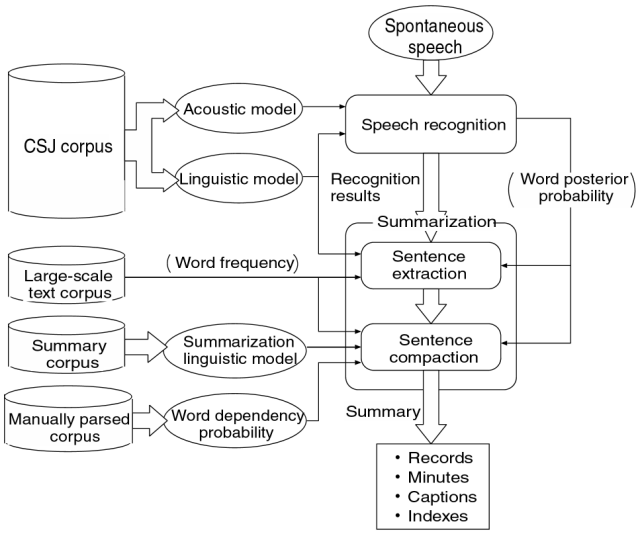

Fig. 1 - Automatic speech summarization system with text presentation.

A. Important sentence extraction

The important sentence extraction is performed according to the following score for each sentence, $W = w_1, w_2, \ldots, w_N$, obtained as a result of speech recognition:

$$S(W) = \frac{1}{N}\sum_{i=1}^{N}\{L(w_i) + \lambda_I I(w_i) + \lambda_C C(w_i)\} \quad (1)$$

where N is the number of words in the sentence W, and $L(w_i)$, $I(w_i)$ and $C(w_i)$ are the linguistic score, the significance score, and the confidence score of word w_i, respectively. The three scores are a subset of the scores originally used in our sentence compaction method and considered to be useful also as measures indicating the appropriateness of including the sentence in the summary. λ_I and λ_C are weighting factors for balancing the scores. Details of the scores are as follows.

Linguistic score: The linguistic score $L(w_i)$ indicates the linguistic likelihood of word strings in the sentence and is measured by n-gram probability:

$$L(w_i) = \log P(w_i \mid \ldots w_{i-1}) \quad (2)$$

In our experiment, trigram probability calculated using transcriptions of presentation utterances in the CSJ consisting of 1.5M morphemes (words) is used. This score de-weights linguistically unnatural word strings caused by recognition errors.

Significance score: The significance score $I(w_i)$ indicates the significance of each word w_i in the sentence and is measured by the amount of information. The amount of information is calculated for content words including nouns, verbs, adjectives and out-of-vocabulary (OOV) words, based on word occurrence in a corpus as shown in (3). A flat score is given to other words.

$$I(w_i) = f_i \log\frac{F_A}{F_i} \quad (3)$$

where f_i is the number of occurrences of w_i in the recognized utterances, F_i is the number of occurrences of w_i in a large-scale corpus, and F_A is the number of all content words in that corpus, that is $\Sigma_i F_i$.

For measuring the significance score, the number of occurrences of 120k kinds of words in a corpus consisting of transcribed presentations (1.5M words), proceedings of 60 presentations, presentation records obtained from WWW (2.1M words), NHK (Japanese broadcast company) broadcast news text (22M words), Mainichi newspaper text (87M words) and text from a speech textbook "Speech Information Processing" (51k words) is calculated. Important keywords are weighted and the words unrelated to the original content, such as recognition errors, are de-weighted by this score.

Confidence score: The confidence score $C(w_i)$ is incorporated to weight acoustically as well as linguistically reliable hypotheses. Specifically, a logarithmic value of a posterior probability for 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 score.

B. Sentence compaction

After removing sentences having relatively low recognition accuracy and/or low significance, the

remaining transcription is automatically modified into a written editorial article style to calculate the score for sentence compaction. All the sentences are combined together, and a linguistic score, a significance score, a confidence score and a word concatenation score are given to each transcribed word. The word concatenation score is incorporated to weight a word concatenation between words with dependency in the transcribed sentences. The dependency is measured by a phrase structure grammar, SDCFG (Stochastic Dependency Context Free Grammar). A set of words that maximizes a weighted sum of these scores is selected according to a given compression ratio and connected to create a summary using a 2-stage dynamic programming (DP) technique. Specifically, each sentence is summarized according to all possible compression ratios, and then the best combination of summarized sentences is determined according to a target total compression ratio.

Ideally, the linguistic score should be calculated using a word concatenation model based on a large-scale summary corpus. Since such a summary corpus is not yet available, the transcribed presentations used to calculate the word trigrams for the important sentence extraction are automatically modified into a written editorial article style and used together with the proceedings of 60 presentations to calculate the trigrams.

The significance score is calculated using the same corpus as that used for calculating the score for important sentence extraction. The word dependency probability is estimated by the Inside-Outside algorithm, using a manually parsed Mainichi newspaper corpus having 4M sentences with 68M words.

2.2 Evaluation experiments

A. Evaluation set

Three presentations, M74, M35 and M31, in the CSJ by male speakers were summarized at summarization ratios of 70% and 50%. Length and mean word recognition accuracy of each presentation are shown in Table 1. They were manually segmented into sentences before recognition.

Table 1 Evaluation set

Presentation ID	Length [min]	Recognition accuracy [%]
M74	12	70
M35	28	60
M31	27	65

B. Summarization accuracy

To objectively evaluate the summaries, correctly transcribed presentation speech was manually summarized by nine human subjects to create targets. Variations of the manual summarization results were merged into a word network as shown in Fig. 2, which is considered to approximately express all possible correct summaries covering subjective variations. Word accuracy of the summary is then measured in comparison with the closest word string extracted from the word network as the summarization accuracy [5].

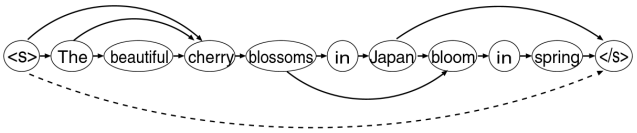

Fig. 2 - Word network made by merging manual summarization results.

C. Evaluation conditions

Summarization was performed under the following nine conditions: single-stage summarization without applying the important sentence extraction (NOS); two-stage summarization using seven kinds of the possible combination of scores for important sentence extraction (*L*, *I*, *C*, *L_I*, *I_C*, *C_L*, *L_I_C*); and, summarization by random word selection. The weighting factors, λ_I and λ_C , were set at optimum values for each experimental condition.

2.3 Evaluation results

A. Summarization accuracy

Results of the evaluation experiments are shown in Figs. 3 and 4. In all the automatic summarization conditions, both one-stage method without sentence extraction and two-stage method including sentence extraction achieve better results than random word selection. In both 70% and 50% summarization conditions, the two-stage method achieves higher summarization accuracy than the one-stage method. The two-stage method is more effective in the condition of the smaller summarization ratio (50%), that is, where there is a higher compression ratio, than in the condition of the larger summarization ratio (70%). In the 50% summarization condition, the two-stage method is effective for all three presentations.

Comparing the three scores for sentence extraction, the significance score (*I*) is more effective than the linguistic score (*L*) and the confidence score (*C*). The

summarization score can be increased by using the combination of two scores (*L_I*, *I_C*, *C_L*), and even more by combining all three scores (*L_I_C*).

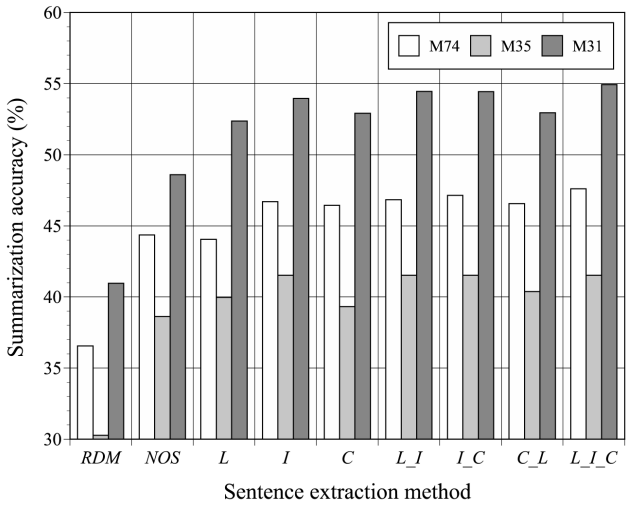

Fig. 3 - Summarization at 50% summarization ratio.

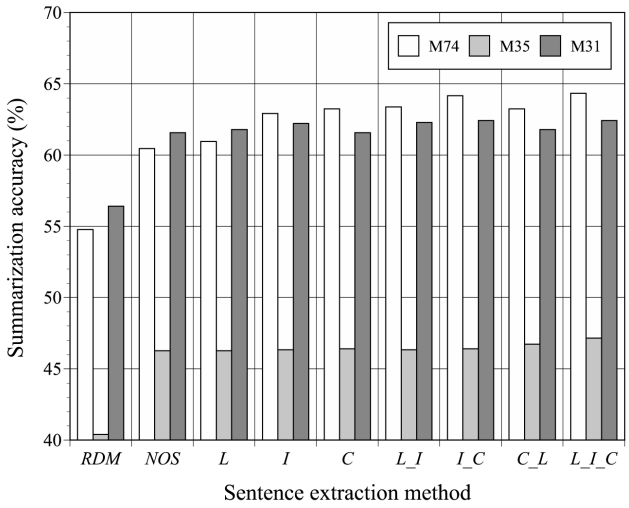

Fig. 4 - Summarization at 70% summarization ratio.

B. Effects of the ratio of compression by sentence extraction

Figures 5 and 6 show the summarization accuracy as a function of the ratio of compression by sentence extraction for the total summarization ratios of 50% or 70%. This result indicates that although the best summarization accuracy of each presentation can be obtained at a different ratio of compression by sentence extraction, there is a general tendency where the smaller the summarization ratio becomes, the larger the optimum ratio of compression by sentence extraction becomes. That is, sentence extraction becomes more effective when the summarization ratio gets smaller.

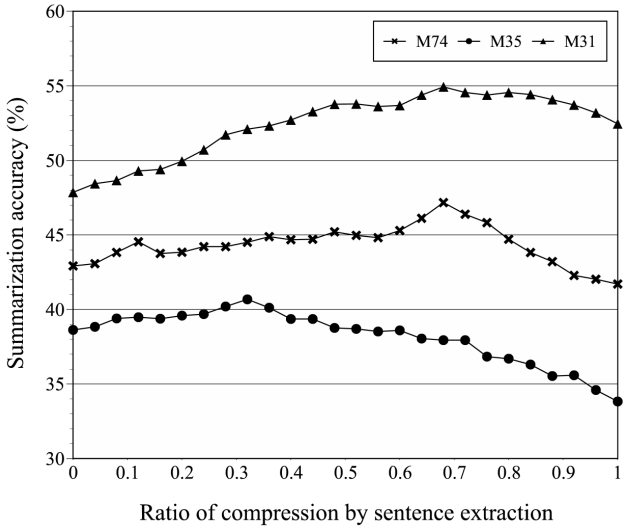

Fig. 5 - Summarization accuracy as a function of the ratio of compression by sentence extraction for the total summarization ratio of 50%.

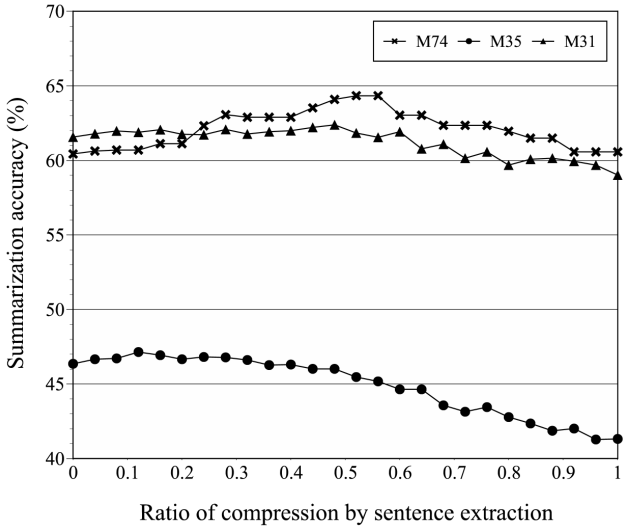

Fig. 6 - Summarization accuracy as a function of the ratio of compression by sentence extraction for the total summarization ratio of 70%.

Comparing results at the right and left ends of the figures, summarization by word extraction, that is, sentence compaction is more effective than sentence extraction for the M35 presentation, which includes a relatively large number of redundant information, such as

disfluencies, fillers, and repetitions. These results indicate that the optimum division of the compression ratio into the two summarization stages needs to be estimated according to the summarization ratio and features of the presentation, such as frequency of disfluencies, fillers and repetitions.

3. Summarization with speech presentation

3.1 Unit selection and concatenation

A. Units for extraction

The following issues need to be addressed in extracting and concatenating speech segments for making summaries:

1) Units for extraction: sentences, phrases, or words;
2) Criteria for measuring the importance of units for extraction; and
3) Concatenation methods for making summary speech.

The following three units are investigated in this paper: sentences, words, and between-filler units. Although prosodic features such as accent and intonation could be used for important part selection, reliable methods for automatically and correctly extracting prosodic features from spontaneous speech and modeling the prosodic features have not yet been established. Therefore, in this paper, input speech is automatically recognized and important segments are extracted based on the same criterion as that used in the previous chapter. Since fillers are automatically detected as the result of recognition, all the fillers are removed before extracting important segments.

Sentence units: The method described in 2.1A is applied to the recognition results to extract important sentences. Since sentences are basic linguistic as well as acoustic units, it is easy to maintain acoustical smoothness by using sentences as units, and therefore the concatenated speech sounds natural. However, since the units are relatively long, they tend to include unnecessary words. Since fillers are automatically removed even if they are included within sentences as described above, the sentences are cut and shortened at the position of fillers.

Word units: Word sets are extracted and concatenated by applying the method described in 2.1B to the recognition results. Although this method has the advantage in that important parts can be precisely extracted in small units, it is prone to cause acoustical discontinuity since many small units of speech need to be concatenated. Therefore, summarization speech made by this method sometimes sounds unnatural.

Between-filler units: Speech segments between fillers as well as sentence boundaries are extracted using speech recognition results. The same method as that used for extracting sentence units is applied to evaluate these units. These units are introduced as intermediate units between sentences and words, in anticipation of both reasonably precise extraction of important parts and naturalness of speech with acoustic continuity.

B. Unit concatenation

Units for building summarization speech are extracted from original speech by using segmentation boundaries obtained from speech recognition results. When the units are concatenated at the inside of sentences, it may produce noise due to the difference of amplitudes of the speech waveforms. In order to avoid this problem, amplitudes of approximately 20ms length at the unit boundaries are gradually attenuated before the concatenation. Since this causes an impression of increasing the speaking rate and thus creates an unnatural sound, a short pause is inserted. The length of the pause is controlled between 50 and 100ms empirically according to the concatenation conditions. Each summarization speech which has been made by this method is hereafter referred to as "summarization speech sentence", and the text corresponding to its speech period is referred to as "summarization text sentence".

The summarization speech sentences are further concatenated to create a summarized speech for the whole presentation. Speech waveforms at sentence boundaries are gradually attenuated and pauses are inserted between the sentences in the same way as the unit concatenation within sentences. Short and long pauses with 200ms and 700ms lengths are used as the pauses between sentences. Long pauses are inserted after sentence ending expressions, and otherwise short pauses are used. In the case of summarization by word-unit concatenation, long pauses are always used, since many sentences terminate with nouns and need relatively long pauses to make them sound natural.

3.2 Evaluation experiments

A. Experimental conditions

The three presentations, M74, M35 and M31, were automatically summarized with the summarization ratio of 50%. Summarization accuracies for the three presentations using sentence units, between-filler units, and word units are given in Table 2. Manual summaries

made by nine human subjects were used for the evaluation.

Table 2 - Summarization accuracy for the three presentations using the three summarization units

	M74	M35	M31	Average
Word units	49.6%	37.6%	50.0%	45.7%
Between-filler units	44.7%	37.5%	46.9%	43.0%
Sentence units	45.5%	37.6%	53.4%	45.5%

Subjective evaluation by 11 subjects was performed in terms of ease of understanding and appropriateness as summarization speech with five levels; 1: very bad, 2: bad, 3: normal, 4: good, and 5: very good. The subjects were instructed to read the transcriptions of the presentations and understand the contents before hearing the summarization speech.

B. Evaluation results and discussion

Figures 7 and 8 show the evaluation results. Averaging over the three presentations, the sentence units show the best whereas the word units show the worst results. For the two presentations, M74 and M35, the between-filler units achieve almost the same results as the sentence units. The reason why the word units which show slightly better summarization accuracy in Table 2 also show the worst subjective evaluation results here is unnatural sound due to the concatenation of short speech units. Relatively large number of fillers included in the presentation M31 produced many short units when using between-filler units. This is the reason why between-filler units show worse subjective results than the sentence units for M31.

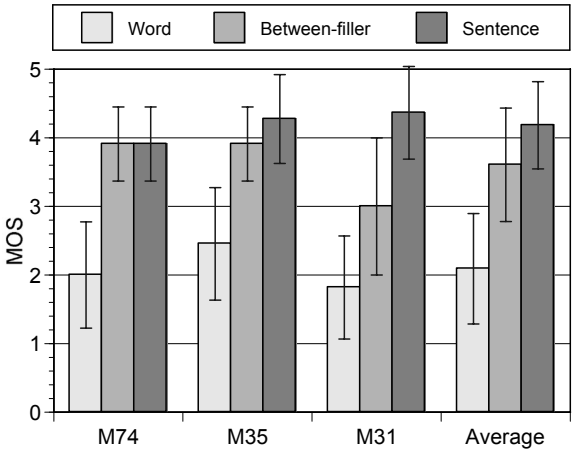

Fig. 7 - Evaluation results in terms of the ease of understanding.

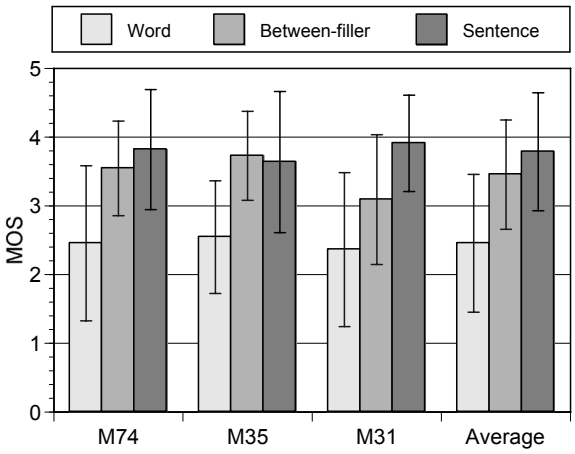

Fig. 8 - Evaluation results in terms of the appropriateness as a summary.

If the summarization ratio is set lower than 50%, between-filler units are expected to achieve better results than sentence units, since sentence units cannot remove redundant expressions within sentences.

4. Conclusion

In this paper, we have presented techniques for compaction-based automatic speech summarization and evaluation results for summarizing spontaneous presentations. The summarization results are presented by either text or speech. In the former case, that is the speech-to-test summarization, we proposed a two-stage automatic speech summarization method consisting of important sentence extraction and word-based sentence compaction. In this method, inadequate sentences including recognition errors and less important information are automatically removed before sentence compaction. It was confirmed that in spontaneous presentation speech summarization at 70% and 50% summarization ratios, combining sentence extraction with sentence compaction is effective; this method achieves better summarization performance than our previous one-stage method. It was also confirmed that three scores, the linguistic score, the word significance score and the word confidence score, are effective for extracting important sentences. The two-stage method is effective for avoiding one of the problems of the one-stage method, that is, the production of short unreadable and/or incomprehensible sentences. The best division for the summarization ratio into the ratios of sentence extraction and sentence compaction depends on the summarization ratio and features of presentation utterances.

For the case of presenting summaries by speech, that is the speech-to-speech summarization, three kinds of units,

sentences, words, and between-filler units, were investigated as the units to be extracted from original speech and concatenated to produce the summaries. A set of units is automatically extracted using the same measures used in the speech-to-text summarization, and the speech segments corresponding to the extracted units are concatenated to produce the summaries. Amplitudes of speech waveforms at the boundaries are gradually attenuated and pauses are inserted before concatenation to avoid acoustic discontinuity. Subjective evaluation results for the 50% summarization ratio indicated that sentence units achieve the best subjective evaluation score. Between-filler units are expected to achieve good performance when summarization ratio becomes smaller.

Future research includes evaluation of the usefulness of other information/features for important unit extraction, investigation of methods for automatically segmenting a presentation into sentence units, those methods' effects on summarization accuracy, and automatic optimization of the division of compression ratio into the two summarization stages according to the summarization ratio and features of the presentation.

Acknowledgment

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

References

[1] S. Furui, K. Iwano, C. Hori, T. Shinozaki, Y. Saito and S. Tamura, "Ubiquitous speech processing," Proc. ICASSP2001, Salt Lake City, vol. 1, pp. 13-16 (2001)

[2] S. Furui, "Recent advances in spontaneous speech recognition and understanding," Proc. ISCA-IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, MMO1 (2003)

[3] K. Maekawa, H. Koiso, S. Furui, H. Isahara, "Spontaneous speech corpus of Japanese," Proc. LREC2000, Athens, pp. 947-952 (2000)

[4] T. Kikuchi, S. Furui and C. Hori, "Two-stage automatic speech summarization by sentence extraction and compaction," Proc. ISCA-IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, TAP10 (2003)

[5] C. Hori and S. Furui, "Advances in automatic speech summarization," Proc. Eurospeech 2001, pp. 1771-1774 (2001)

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