# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| Title | Speech-to-text and speech-to-speech summarization of spontaneous speech |
| Author | Sadaoki Furui, Tomonori Kikuchi, Yousuke Shinnaka, Chiori Hori |
| Journal/Book name | IEEE Transactions on speech and audio processing, Vol. 12, No. 4, pp. 401-408 |
| 発行日 / Issue date | 2004, 7 |
| 権利情報 / Copyright | (c)2004 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE. |

# Speech-to-Text and Speech-to-Speech Summarization of Spontaneous Speech

Sadaoki Furui, *Fellow, IEEE*, Tomonori Kikuchi, Yousuke Shinnaka, and Chiori Hori, *Member, IEEE*

***Abstract*—This paper presents techniques for speech-to-text and speech-to-speech automatic summarization based on speech unit extraction and concatenation. 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 and concatenated for producing summaries. For the latter case, sentences, words, and between-filler units are investigated as units to be extracted from original speech. These methods are applied to the summarization of unrestricted-domain spontaneous presentations and evaluated by objective and subjective measures. It was confirmed that proposed methods are effective in spontaneous speech summarization.**

***Index Terms*—Presentation, speech recognition, speech summarization, speech-to-speech, speech-to-text, spontaneous speech.**

## I. 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 usually inevitable. 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 expected to save time for reviewing speech documents and improve the efficiency of document retrieval.

Summarization results can be presented by either text or speech. The former method has advantages in that: 1) the documents can be easily looked through; 2) the part of the documents that are interesting for users can be easily extracted; and 3) information extraction and retrieval techniques can be easily applied to the documents. However, it has disadvantages in that wrong information due to speech recognition errors cannot be avoided and prosodic information such as the emotion of speakers conveyed only in 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: 1) presenting simply concatenated speech segments that are extracted from original speech or 2) 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 extracted speech segments is how to avoid unnatural noisy sound caused by the concatenation.

There has been much research in the area of summarizing written language (see [3] for a comprehensive overview). So far, however, very little attention has been given to the question of how to create and evaluate spoken language summarization based on automatically generated transcription from a speech recognizer. One fundamental problem with the summaries produced is that they contain recognition errors and disfluencies. Summarization of dialogues within limited domains has been attempted within the context of the VERBMOBIL project [4]. Zechner and Waibel have investigated how the accuracy of the summaries changes when methods for word error rate reduction are applied in summarizing conversations in television shows [5]. Recent work on spoken language summarization in unrestricted domains has focused almost exclusively on Broadcast News [6], [7]. Koumpis and Renals have investigated the transcription and summarization of voice mail speech [8]. Most of the previous research on spoken language summarization have used relatively long units, such as sentences or speaker turns, as minimal units for summarization.

This paper investigates automatic speech summarization techniques with the two presentation methods in unrestricted

Manuscript received May 6, 2003; revised December 11, 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Julia Hirschberg.

S. Furui, T. Kikuchi, and Y. Shinnaka are with the Department of Computer Science, Tokyo Institute of Technology, Tokyo, 152-8552, Japan (e-mail: furui@furui.cs.titech.ac.jp; kikuchi@furui.cs.titech.ac.jp; shinnaka@furui.cs.titech.ac.jp).

C. Hori is with the Intelligent Communication Laboratory, NTT Communication Science Laboratories, Kyoto 619-0237, Japan (e-mail: chiori@cslab.kecl.ntt.co.jp).

Digital Object Identifier 10.1109/TSA.2004.828699

domains. In both cases, the most appropriate sentences, phrases or word units/segments are automatically extracted from original speech and concatenated to produce a summary under the constraint that 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 waveform concatenation-based method is used.

Although prosodic features such as accent and intonation could be used for selection of important parts, reliable methods for automatic and correct extraction of prosodic features from spontaneous speech and for modeling them have not yet been established. Therefore, in this paper, input speech is automatically recognized and important segments are extracted based only on the textual information.

Evaluation experiments are performed using spontaneous presentation utterances in the Corpus of Spontaneous Japanese (CSJ) made by the Spontaneous Speech Corpus and Processing Project [9]. The project began in 1999 and is being conducted over a five-year period with the following three major targets.

1) Building a large-scale spontaneous speech corpus (CSJ) consisting of roughly 7 M words with a total speech length of 700 h. This mainly records monologues such as lectures, presentations and news commentaries. The recordings with low spontaneity, such as those from read text, are excluded from the corpus. The utterances are manually transcribed orthographically and phonetically. One-tenth of them, called Core, are tagged manually and used for training a morphological analysis and part-of-speech (POS) tagging program for automatically analyzing all of the 700-h utterances. The Core is also tagged with para-linguistic information including intonation.
2) Acoustic and language modeling for spontaneous speech understanding using linguistic, as well as para-linguistic, information in speech.
3) Investigating spontaneous speech summarization technology.

## II. Summarization With Text Presentation

### A. Two-Stage Summarization Method

Fig. 1 shows the two-stage summarization method consisting of important sentence extraction and sentence compaction [10]. 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 [11], [12] 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.

Speech summarization has a number of significant challenges that distinguish it from general text summarization. Applying text-based technologies to speech is not always workable and often they are not equipped to capture speech specific phenomena. Speech contains a number of spontaneous effects, which are not present in written language, such as hesitations, false starts, and fillers. Speech is, to some extent, always distorted by ungrammatical and various redundant expressions. Speech is also a continuous phenomenon that comes without unambiguous sentence boundaries. In addition, errors in transcriptions of automatic speech recognition engines can be quite substantial.

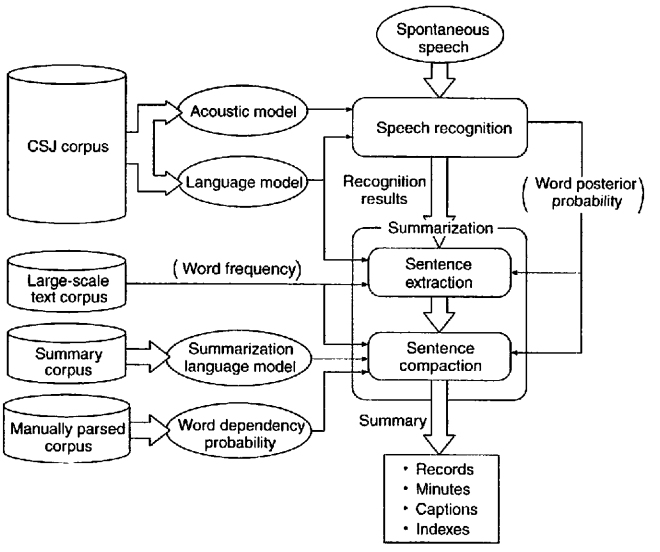



Fig. 1. A two-stage automatic speech summarization system with text presentation.

Sentence extraction methods on which most of the text summarization methods [13] are based cannot cope with the problems of distorted information and redundant expressions in speech. Although several sentence compression methods have also been investigated in text summarization [14], [15], they rely on discourse and grammatical structures of the input text. Therefore, it is difficult to apply them to spontaneous speech with ill-formed structures. The method proposed in this paper is suitable for applying to ill-formed speech recognition results, since it simultaneously uses various statistical features, including a confidence measure of speech recognition results. The principle of the speech-to-text summarization method is also used in the speech-to-speech summarization which will be described in the next section. Speech-to-speech summarization is a comparatively much younger discipline, and has not yet been investigated in the same framework as the speech-to-text summarization.

*1) Important Sentence Extraction:* 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. Although sentence boundaries can be estimated using linguistic and prosodic information [16], they are manually given in the experiments in this paper. 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. $\lambda_I$ and $\lambda_C$ 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

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

In our experiment, trigram probability calculated using transcriptions of presentation utterances in the CSJ consisting of 1.5 M 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 contained in each word 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). The POS information for each word is obtained from the recognition result, since every word in the dictionary is accompanied with a unique POS tag. A flat score is given to other words, and

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

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 120 000 kinds of words is calculated in a corpus consisting of transcribed presentations (1.5 M words), proceedings of 60 presentations, presentation records obtained from the World-Wide Web (WWW) (2.1 M words), NHK (Japanese broadcast company) broadcast news text (22 M words), Mainichi newspaper text (87 M words) and text from a speech textbook "Speech Information Processing" (51 000 words). 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 the posterior probability for each transcribed word, which 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.

*2) Sentence Compaction:* After removing relatively less important sentences, the remaining transcription is automatically modified into a written editorial article style to calculate the score for sentence compaction. All the sentences are concatenated while preserving sentence boundaries, and a linguistic score, $L(w_i)$, a significance score $I(w_i)$, and a confidence score $C(w_i)$ are given to each transcribed word. A word concatenation score $T(w_i, w_j)$ for every combination of words within each transcribed sentence is also given to weight a word concatenation between words. This score is a measure of the dependency between two words and is obtained by a phrase structure grammar, stochastic dependency context-free grammar (SDCFG). 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 two-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 4 M sentences with 68 M words. For the details of the SDCFG and dependency scores, readers should refer to [12].

### *B. Evaluation Experiments*

*1) Evaluation Set:* Three presentations, M74, M35, and M31, in the CSJ by male speakers were summarized at summarization ratios of 70% and 50%. The summarization ratio was defined as the ratio of the number of characters in the summaries to that in the recognition results. Table I shows features of the presentations, that is, length, mean word recognition accuracy, number of sentences, number of words, number of fillers, filler ratio, and number of disfluencies including repairs of each presentation. They were manually segmented into sentences before recognition. The table shows that the presentation M35 has a significantly large number of disfluencies and a low recognition accuracy, and M31 has a significantly high filler ratio.

*2) Summarization Accuracy:* To objectively evaluate the summaries, correctly transcribed presentation speech was manually summarized by nine human subjects to create targets. Devising meaningful evaluation criteria and metrics for speech summarization is a problematic issue. Speech does not have explicit sentence boundaries in contrast with text input. Therefore, speech summarization results cannot be evaluated using the F-measure based on sentence units. In addition, since words (morphemes) within sentences are extracted and concatenated in the summarization process, variations of target summaries made by human subjects are much larger than those using the sentence level method. In almost all cases, an "ideal" summary does not exist. For these reasons, 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].

TABLE I
EVALUATION SET

| Presentation ID | M74 | M35 | M31 |
|---|---|---|---|
| Length [min] | 12 | 28 | 27 |
| Recognition accuracy [%] | 71.8 | 55.4 | 69.4 |
| No. of sentences | 110 | 248 | 217 |
| No. of words | 2,179 | 5,337 | 4,518 |
| No. of fillers | 138 | 554 | 633 |
| Filler ratio [%] | 6.3 | 10.4 | 14.0 |
| No. of disfluencies | 35 | 281 | 101 |

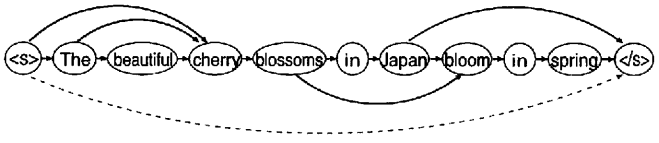



Fig. 2. Word network made by merging manual summarization results.

*3) 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 $\lambda_I$ and $\lambda_C$ were set at optimum values for each experimental condition.

## C. Evaluation Results

*1) Summarization Accuracy:* Results of the evaluation experiments are shown in Figs. 3 and 4. In all the automatic summarization conditions, both the one-stage method without sentence extraction and the two-stage method including sentence extraction achieve better results than random word selection. In both the 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. The two-stage method is especially effective for avoiding one of the problems of the one-stage method, that is, the production of short unreadable and/or incomprehensible sentences.

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 ($L\_I\_C$).

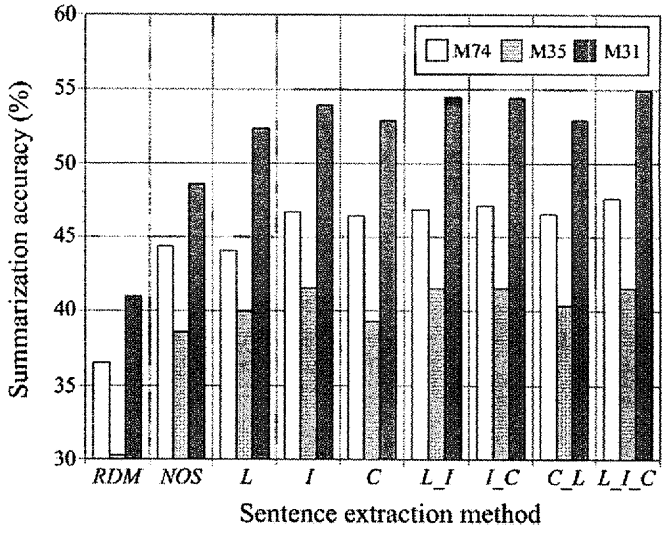



Fig. 3. Results of the summarization with text presentation at 50% summarization ratio.

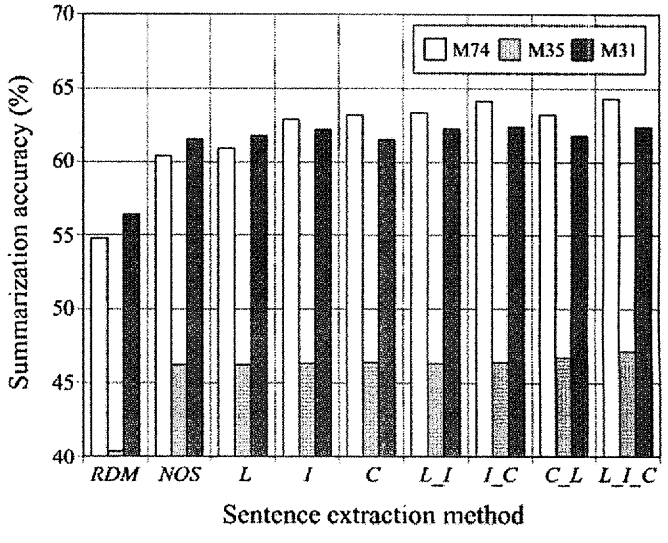



Fig. 4. Results of the summarization with text presentation at 70% summarization ratio.

The differences are, however, statistically insignificant in these experiments, due to the limited size of the data.

*2) Effects of the Ratio of Compression by Sentence Extraction:* Figs. 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%. The left and right ends of the figures correspond to summarizations by only sentence compaction and sentence extraction, respectively. These results indicate 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.

Comparing results at the left and right ends of the figures, summarization by word extraction (i.e., sentence compaction) is more effective than sentence extraction for the M35 presentation. This presentation includes a relatively large amount of redundant information, such as disfluencies and repairs, and has a significantly low recognition accuracy. These results indicate that the optimum division of the compression ratio into the two summarization stages needs to be estimated according to the specific summarization ratio and features of the presentation in question, such as frequency of disfluencies.

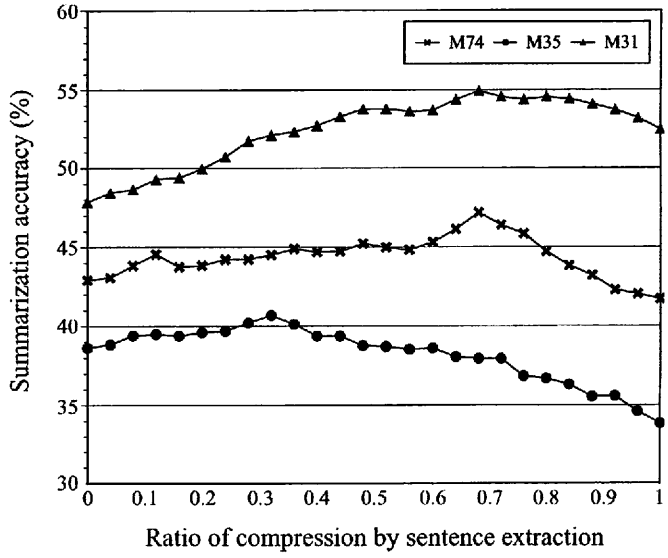



Fig. 5. Summarization accuracy as a function of the ratio of compression by sentence extraction for the total summarization ratio of 50%.

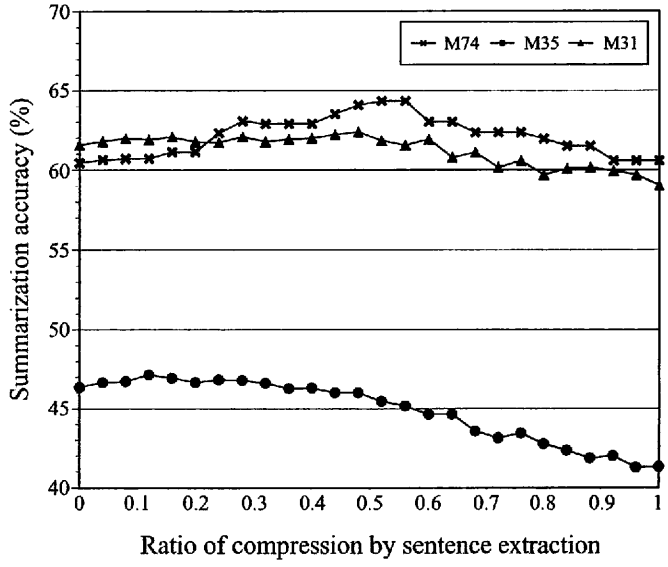



Fig. 6. Summarization accuracy as a function of the ratio of compression by sentence extraction for the total summarization ratio of 70%.

## III. Summarization With Speech Presentation

### A. Unit Selection and Concatenation

*1) 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.
3) Concatenation methods for making summary speech.

The following three units are investigated in this paper: sentences, words, and between-filler units. All the fillers automatically detected as the result of recognition are removed before extracting important segments.

**Sentence units**: The method described in Section II-A.1 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 Section II-A.2 to the recognition results. Although this method has an advantage in that important parts can be precisely extracted in small units, it tends 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.

*2) 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 a difference of amplitudes of the speech waveforms. In order to avoid this problem, amplitudes of approximately 20-ms length at the unit boundaries are gradually attenuated before the concatenation. Since this causes an impression of

TABLE II
SUMMARIZATION ACCURACY AND NUMBER OF UNITS FOR THE THREE KINDS OF SUMMARIZATION UNITS

| Presentation ID | | M74 | M35 | M31 | Average |
|---|---|---|---|---|---|
| Length [min] | | 12 | 28 | 27 | - |
| Word units | Summarization acc. | 49.6% | 37.6% | 50.0% | 45.7% |
| | No. of units | 2,311 | 5,180 | 4,850 | - |
| Between-filler units | Summarization acc. | 44.7% | 37.5% | 46.9% | 43.0% |
| | No. of units | 215 | 478 | 693 | - |
| | No. of fillers | 190 | 432 | 614 | - |
| Sentence units | Summarization acc. | 45.5% | 37.6% | 53.4% | 45.5% |
| | No. of units | 86 | 212 | 208 | - |

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 100 ms 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 200- and 700-ms lengths are used as pauses between sentences. Long pauses are inserted after sentence ending expressions, 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.

### B. Evaluation Experiments

*1) Experimental Conditions:* The three presentations, M74, M35, and M31, were automatically summarized with a summarization ratio of 50%. Summarization accuracies for the three presentations using sentence units, between-filler units, and word units, are given in Table II. Manual summaries made by nine human subjects were used for the evaluation. The table also shows the number of automatically detected units in each condition. For the case of using the between-filler units, the number of detected fillers is also shown.

Using the summarization text sentences, speech segments were extracted and concatenated to build summarization speech, and subjective evaluation by 11 subjects was performed in terms of ease of understanding and appropriateness as a summarization 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.

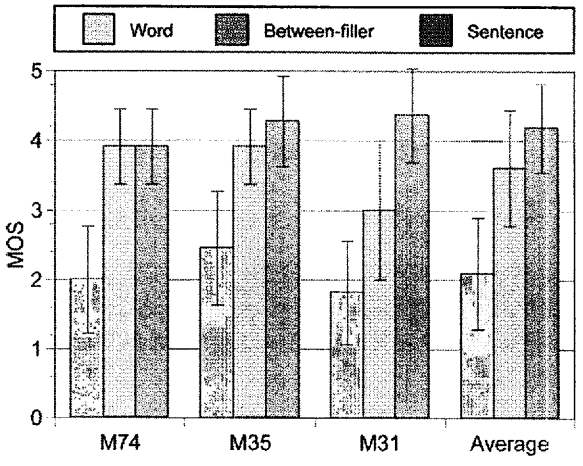



Fig. 7. Evaluation results for the summarization with speech presentation in terms of the ease of understanding.

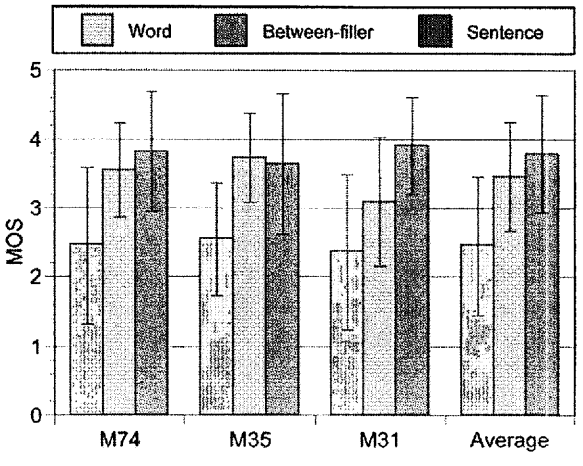



Fig. 8. Evaluation results for the summarization with speech presentation in terms of the appropriateness as a summary.

*2) Evaluation Results and Discussion:* Figs. 7 and 8 show the evaluation results. Averaging over the three presentations, the sentence units show the best results whereas the word units

show the worst. 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 II also show the worst subjective evaluation results here is because of unnatural sound due to the concatenation of short speech units. The relatively large number of fillers included in the presentation M31 produced many short units when the between-filler unit method was applied. This is the reason why between-filler units show worse subjective results than the sentence units for M31.

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.

## IV. 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, 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 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, the speech-to-speech summarization, three kinds of units—sentences, words, and between-filler units—were investigated as 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 the summarization ratio becomes smaller.

As stated in the introduction, 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 a similar task, in that they both map an abundance of information to a (much) smaller piece to be presented to the user [17]. Therefore, speech summarization research will help the advancement of QA systems using speech documents. By condensing important points of long presentations and lectures, speech-to-speech summarization can provide the listener with a valuable means for absorbing much information in a much shorter time.

Future research includes evaluation by a large number of presentations at various summarization ratios including smaller ratios, investigation of other information/features for important unit extraction, methods for automatically segmenting a presentation into sentence units [16], 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," in *Proc. ICASSP2001*, vol. 1, Salt Lake City, UT, 2001, pp. 13–16.

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

[3] I. Mani and M. T. Maybury, Eds., *Advances in Automatic Text Summarization*. Cambridge, MA: MIT Press, 1999.

[4] J. Alexandersson and P. Poller, "Toward multilingual protocol generation for spontaneous dialogues," in *Proc. INLG-98*, Niagara-on-the-lake, Canada, 1998.

[5] K. Zechner and A. Waibel, "Minimizing word error rate in textual summaries of spoken language," in *Proc. NAACL*, Seattle, WA, 2000.

[6] J. S. Garofolo, E. M. Voorhees, C. G. P. Auzanne, and V. M. Stanford, "Spoken document retrieval: 1998 evaluation and investigation of new metrics," in *Proc. ESCA Workshop: Accessing Information in Spoken Audio*, Cambridge, MA, 1999, pp. 1–7.

[7] R. Valenza, T. Robinson, M. Hickey, and R. Tucker, "Summarization of spoken audio through information extraction," in *Proc. ISCA Workshop on Accessing Information in Spoken Audio*, Cambridge, MA, 1999, pp. 111–116.

[8] K. Koumpis and S. Renals, "Transcription and summarization of voicemail speech," in *Proc. ICSLP 2000*, 2000, pp. 688–691.

[9] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proc. LREC2000*, Athens, Greece, 2000, pp. 947–952.

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

[11] C. Hori and S. Furui, "Advances in automatic speech summarization," in *Proc. Eurospeech 2001*, 2001, pp. 1771–1774.

[12] C. Hori, S. Furui, R. Malkin, H. Yu, and A. Waibel, "A statistical approach to automatic speech summarization," *EURASIP J. Appl. Signal Processing*, pp. 128–139, 2003.

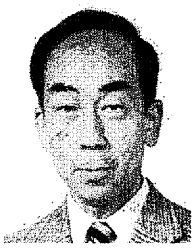
[13] K. Knight and D. Marcu, "Summarization beyond sentence extraction: A probabilistic approach to sentence compression," *Artific. Intell.*, vol. 139, pp. 91–107, 2002.

[14] H. Daume III and D. Marcu, "A noisy-channel model for document compression," in *Proc. ACL-2002*, Philadelphia, PA, 2002, pp. 449–456.

[15] C.-Y. Lin and E. Hovy, "From single to multi-document summarization: A prototype system and its evaluation," in *Proc. ACL-2002*, Philadelphia, PA, 2002, pp. 457–464.

[16] M. Hirohata, Y. Shinnaka, and S. Furui, "A study on important sentence extraction methods using SVD for automatic speech summarization," in *Proc. Acoustical Society of Japan Autumn Meeting*, Nagoya, Japan, 2003.

[17] K. Zechner, "Spoken language condensation in the 21st Century," in *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 1989–1992.



**Sadaoki Furui** (F'93) is a Professor at the Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan. 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 350 published articles. From 1978 to 1979, he served on staff at the Acoustics Research Department of Bell Laboratories, Murray Hill, NJ, as a Visiting Researcher, working on speaker verification. He is the author of Digital Speech Processing, Synthesis, and Recognition (New York: Marcel Dekker, 1989; revised, 2000), Digital Speech Processing (Tokai, Japan: Tokai Univ. Press, 1985), Acoustics and Speech Processing (Tokyo, Japan: Kindai-Kagaku-Sha, 1992) in Japanese, and Speech Information Processing (Tokyo, Japan: Morikita, 1998). He edited (with M. M. Sondhi) Advances in Speech Signal Processing (New York: Marcel Dekker, 1992). He has translated into Japanese Fundamentals of Speech Recognition (Tokyo, Japan: NTT Advanced Technology, 1995), authored by L. R. Rabiner and B.-H. Juang, and Vector Quantization and Signal Compression (Tokyo, Japan: Corona-sha, 1998), authored by A. Gersho and R. M. Gray.

Dr. Furui is a Fellow of the Acoustical Society of America and the Institute of Electronics, Information and Communication Engineers of Japan (IEICE). He is President of the Acoustical Society of Japan (ASJ), the International Speech Communication Association (ISCA), and the Permanent Council for International Conferences on Spoken Language Processing (PC-ICSLP). He is a Board of Governor of the IEEE Signal Processing Society (SPS) and in 1993, he served as an IEEE SPS Distinguished Lecturer. He has served on the IEEE Technical Committee on Speech and MMSP and on numerous IEEE Conference organizing committees. He is Editor-in-Chief of the *Transactions of the IEICE*. He is also an Editorial Board member of *Speech Communication*, the *Journal of Computer Speech and Language*, and the *Journal of Digital Signal Processing*. He has received numerous awards, including: the Yonezawa Prize and the Paper Awards from the IEICE (1975, 1988, 1993, and 2003); the Sato Paper Award from the ASJ (1985 and 1987); the Senior Award from the IEEE Acoustics, Speech, and Signal Processing Society (1989); the Achievement Award from the Minister of Science and Technology of Japan (1989); the Technical Achievement Award and the Book Award from the IEICE (1990 and 2003); the Mira Paul Memorial Award from the AFECT of India (2001).

**Tomonori Kikuchi** received the B. E. and M. E. degrees in computer science from Tokyo Institute of Technology, Tokyo, Japan, in 2001 and 2003, respectively.

He has been with Japan Patent Office, Tokyo, Japan, since 2003.

**Yousuke Shinnaka** received the B. E. degree in electrical and electronic engineering from Tokyo Institute of Technology, Tokyo, Japan, in 2003. He is currently pursuing the M.S. degree at Tokyo Institute of Technology.

**Chiori Hori** (M'02) received the B.E. and M.E. degrees in electrical and information engineering from Yamagata University, Yonezawa, Japan, in 1994 and 1997, respectively, and the Ph.D. degree from the Graduate School of Information Science and Engineering, Tokyo Institute of Technology (TITECH), Tokyo, Japan, in 2002.

From April 1997 to March 1999, she was a Research Associate with the Faculty of Literature and Social Sciences, Yamagata University. She is currently a Researcher with NTT Communication Science Laboratories (CS Labs), Nippon Telegraph and Telephone Corporation (NTT), Kyoto, Japan, which she joined in 2002.

Dr. Hori is a member of the Acoustical Society of Japan (ASJ), the Institute of Electronics, Information and Communication Engineers of Japan (IEICE), and the Information Processing Society of Japan (IPSJ). She received the Paper Award from the IEICE in 2002 for her work on speech summarization.