INVITED PAPER *Special Section on the 2002 IEICE Excellent Paper Award*

Speech Summarization: An Approach through Word Extraction and a Method for Evaluation*

Chiori HORI†, ***Member and*** **Sadaoki FURUI**†, ***Fellow***

SUMMARY In this paper, we propose a new method of automatic speech summarization for each utterance, where a set of words that maximizes a summarization score is extracted from automatic speech transcriptions. The summarization score indicates the appropriateness of summarized sentences. This extraction is achieved by using a dynamic programming technique according to a target summarization ratio. This ratio is the number of characters/words in the summarized sentence divided by the number of characters/words in the original sentence. The extracted set of words is then connected to build a summarized sentence. The summarization score consists of a word significance measure, linguistic likelihood, and a confidence measure. This paper also proposes a new method of measuring *summarization accuracy* based on a word network expressing manual summarization results. The *summarization accuracy* of each automatic summarization is calculated by comparing it with the most similar word string in the network. Japanese broadcast-news speech, transcribed using a large-vocabulary continuous-speech recognition (LVCSR) system, is summarized and evaluated using our proposed method with 20, 40, 60, 70 and 80% summarization ratios. Experimental results reveal that the proposed method can effectively extract relatively important information by removing redundant or irrelevant information.

key words: *speech summarization, sentence compaction, summarization score, dynamic programming, word network of manual summarization result, summarization accuracy*

1. Introduction

Significant progress has been made with state-of-the-art technology for large-vocabulary continuous-speech recognition (LVCSR). Currently, various applications of LVCSR systems, such as automatic closed captioning [1], meeting/conference summarization [2], and indexing for information retrieval [3], are actively being investigated. However, transcribed speech, which usually includes redundant information, cannot be directly used for captions, indexes and summarization. Practical applications using LVCSR systems require a process of speech summarization, which removes redundant and irrelevant information and extracts relatively important information according to users' requirements, especially for spontaneous speech.

In the closed captioning of broadcast news, the number of words spoken by professional announcers sometimes exceeds the number of words that people can read and understand if they are presented on a TV screen. Meeting/conference summarization would be useful if it could extract relatively important information scattered within the original speech. These goals can be accomplished by reducing the number of words in speech transcriptions according to the target summarization ratio.

Techniques for automatically summarizing written text have been actively pursued throughout the field of natural language processing [4]. One of the main techniques of summarizing written text is the process of extracting important sentences from a document based on keywords or cue phrases. One major difference between text summarization and speech summarization is the fact that transcribed speech is sometimes linguistically incorrect. 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. Removing such redundant and irrelevant information from automatic speech transcriptions is indispensable in the preparation of closed captions, lecture/meeting summaries, and indexes. These problems can be solved by summarizing speech according to the summarization ratio that is required by users. A speech summarization technique that includes both information extraction and skimming technology is required to construct a system that allows archived multimedia to be freely accessed by using large-vocabulary continuous-speech recognition (LVCSR) systems.

In this paper, we propose a new approach to automatically summarizing speech by extracting a set of words from the automatic transcription of each utterance. This extraction is performed according to a target summarization ratio, which is the number of characters/words in the summarized sentence divided by the number of characters/words in the original sentence. The extracted set of words is then concatenated to build a summarized sentence. An utterance is defined as continuous speech containing a meaningful word set. In this paper, each sentence read aloud by anchorpersons in a news broadcast is defined as an utterance.

Automatic summarization should retain the important information that was included in the original utterance under a restricted summarization ratio. In addition, it should not only exclude recognition er-

Manuscript received August 2, 2003.

†The authors are with Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Tokyo, 152–8550 Japan.

*This article was originally published in the IEICE Transactions on Information and Systems (Japanese Edition), vol.J85-D-II, no.2, pp.200–209, February 2002.

rors by the ASR system but also produce understandable sentences. To generate such summarized utterances, our summarization approach focuses on extracting topic words, weighting correct-word concatenations linguistically, and extracting reliable components of speech recognition acoustically as well as linguistically. All possible combinations of words in the original utterance at the target summarization ratio are hypotheses of a summarized utterance. A set of words maximizing a summarization score, indicating the appropriateness of a summarized sentence, is selected from those using a dynamic programming (DP) technique.

The summarization score consists of a word significance measure, a confidence measure, and the linguistic likelihood of summarized sentences. The confidence score indicates the reliability of recognition. The summarization process of extracting a set of words to maximize the summarization score can preserve relatively important words and exclude recognition errors. Simultaneously, words other than the extracted important words, are extracted as linguistically correct. This summarization approach is not simple "keyword extraction", but "sentence generation including keywords", generated by extracting words in the original utterance and concatenating them. In this approach, the summarization ratio can be altered depending on summarization targets, such as providing closed-captioning for multimedia data including speech, preparing minutes of meetings or synopses of speeches, and indexing speech data or multimedia data including speech. In addition, the speech summarization technique enables us to easily find important components of speech data.

In this paper, we also propose a numerical method of evaluating automatic summarizations. These results are compared with manual summarizations generated by humans. The manual summarizations are then merged into a word network, which approximately expresses all the possible correct summarizations, including subjective variations. The word string that is the most similar to the automatic summarization extracted from the word network is considered to be the correct target for automatic summarization. The word accuracy of comparing the summarized sentence with the target word string is defined as *summarization accuracy*, which is a measure of linguistic correctness and retention of the original meanings of the utterance. This method of evaluation takes the subjective variations in manual summarization into consideration.

In this paper, we report an automatic speech summarization generated from automatic transcriptions of spoken Japanese news data, broadcast on TV at summarization ratios of 20%, 40%, 60%, 70% and 80%. We also report on the results of evaluating automatic speech summarizations using *summarization accuracy*. Our experiments indicated that the proposed method can effectively extract relatively important information, and remove redundant and irrelevant information from Japanese news broadcasts at all of the summarization ratios we tested.

2. Speech Summarization Approach through Word Extraction

Our method of summarizing each utterance is to extract a set of words representing the core information in the original speech from an automatically transcribed sentence according to a summarization ratio, and then to concatenate these to construct a summary. Morphemes are defined as words in this paper.

The summarization score in our approach is defined as a measure indicating the appropriateness of a summarized sentence. A set of words maximizing the summarization score is extracted from the original utterance using a dynamic programming (DP) technique based on a target compression ratio. The summarization score is defined as the sum of word significance score I, the confidence score C of each word in the original sentence, and the linguistic score L of the word string in the summarized sentence.

The summarization score of a set of M words is given by

$$S(V) = \sum_{m=1}^{M} \{I(v_m) + \lambda_L L(v_m) + \lambda_C C(v_m)\}, \tag{1}$$

where λ_L and λ_C are the weighting factors to balance the dynamic ranges of L, I and C.

Given a set of M words, $V = v_1, v_2, \ldots, v_M$, extracted from a transcription consisting of N $(M < N)$ words, $W = w_1, w_2, \ldots, w_N$, the summarization process determines a set of words, $\hat{V}$, that maximizes the summarization score given by Eq. (1). This problem can be solved using the DP technique.

2.1 Word Significance Score

The word significance score $I(v_m)$ indicates the relative significance of each word in the original sentence. The amount of information based on the frequency of each word, given by Eq. (2), is used as the word significance score for topic words.

$$I(w_i) = f_i \log \frac{F_A}{F_i}, \tag{2}$$

where

w_i : a topic word in the transcribed speech,
f_i : the number of occurrences of w_i in the transcription,
F_i : the number of occurrences of w_i in all the training documents, and
F_A : the summation of all F_i in all the training documents $(= \sum_i F_i)$.

The large corpus in the same domain of the target speech to be summarized is used as training data. Our preliminary experiments revealed that test subjects selected mostly nouns as the important words in sentences [5]. The nouns include those acting as part of a verb phrase. Suppose "提案する", which means "propose" in English, is split into a noun for movement and a verb as follows:

提案する/propose →
提案/proposal + する/make

Therefore, the significance score given by Eq. (2) was only assigned to nouns, and a flat score I_{const} was assigned to words other than nouns. To reduce the repetition of words in the summarized sentence, a flat score of I_{const} was assigned to each reappearing noun.

2.2 Linguistic Score

Linguistic score $L(v_m)$ indicates the appropriateness of word strings in a summarized sentence and was measured by the logarithmic value of n-gram probability $P(v_m|\ v_{m-n+1} \dots v_{m-1})$. Trigram probability, $P(v_m|v_{m-2}v_{m-1})$, was applied to the linguistic score in this study.

$$L(v_m) = \log P(v_m|v_{m-2}v_{m-1}) \tag{3}$$

To generate "a sentence including topic words", the significance score extracts topic words and the linguistic score extracts words that were given a flat significance score.

2.3 Confidence Score

We incorporated the confidence score $C(v_m)$ to weight reliable hypotheses acoustically as well as linguistically. Specifically, the posterior probability of each transcribed word, that is, the ratio of word hypothesis probability to that of all other hypotheses, was calculated using a word graph obtained through a decoder and used as a measure of confidence [6], [7].

There is a word graph consisting of nodes and links from beginning node S to end node T in Fig. 1. Nodes represent time boundaries between possible word hypotheses, and the links connecting these nodes represent word hypotheses. Each link is given an acoustic log likelihood and the linguistic log likelihood of a word hypothesis.

Given a word hypothesis $w_{k,l}$ in the word graph, the confidence score, $C(w_{k,l})$, is defined as the posterior probability of $w_{k,l}$ given by

$$C(w_{k,l}) = \log \frac{\alpha_k P_{ac}(w_{k,l}) P_{lg}(w_{k,l}) \beta_l}{\mathcal{G}} \quad , \tag{4}$$

where
k, l : node identifiers in a word graph ($k < l$),
$w_{k,l}$: the word hypothesis occurring between node k and node l,
$C(w_{k,l})$: the log of posterior probability of $w_{k,l}$,

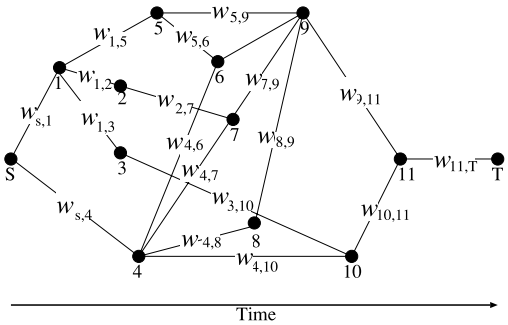

Fig. 1 Example of word graph.

α_k : the forward probability from the beginning node S to node k,
β_l : the backward probability from node l to end node T,
$P_{ac}(w_{k,l})$: the acoustic likelihood of $w_{k,l}$,
$P_{lg}(w_{k,l})$: the linguistic likelihood of $w_{k,l}$, and
$\mathcal{G}$: the forward probability from beginning node S to end node T ($= \alpha_T$).

A word's confidence score indicates the log of the likelihood ratio compared with the other word hypotheses that occur in roughly the same time period on the word graph. Words given higher confidence scores are considered to be reliably recognized.

2.4 Dynamic Programming for Speech Summarization

Given a transcription consisting of N words, $W = w_1, w_2, \dots, w_N$, an algorithm to determine a set of M ($M < N$) words, $V = v_1, v_2, \dots, v_M$, maximizes the summarization score given by Eq. (1).

1. Definition of symbols and variables
   `<s>` : the beginning symbol of a sentence
   `</s>` : the ending symbol of a sentence
   $L(w_n)$: the linguistic score
   $I(w_n)$: the word significance score
   $C(w_n)$: the confidence score
   $s(n)$: the local summarization score of each word

   $$s(n) = I(w_n) + \lambda_L L(w_n) + \lambda_C C(w_n)$$

   Trigram probability is applied to linguistic score. The summarization score of a word is given as

   $$s(k,l,n) = I(w_n) + \lambda_L \log P(w_n|w_k w_l) + \lambda_C C(w_n).$$

   $g(m,l,n)$: the summarization score of a sub-sentence $\langle s \rangle, \dots, w_l, w_n$, consisting of m words, beginning from `<s>`, and ending in w_l, w_n ($0 \leqq l < n \leqq N$)
   $B(m,l,n)$: the back pointer

2. Initialization

$$g(1,0,n) = \begin{cases} I(w_n) + \lambda_L \log P(w_n|\texttt{<s>}) \\ \qquad + \lambda_C C(w_n) \\ \qquad \text{if } 1 \leqq n \leqq (N - M + 1) \\ -\infty \qquad \text{otherwise} \end{cases}$$

3. DP process
   Dynamic programming recursion is applied to each pair of the last two words $(w_l,\ w_n)$ for each sub-sentence hypothesis consisting of m words.

$$\begin{aligned}
&\text{for } m = 2 \text{ to } M\\
&\quad \text{for } n = m \text{ to } N - m + 1\\
&\qquad \text{for } l = m - 1 \text{ to } n - 1\\
&g(m,l,n) = \max_{k<l}\{g(m-1,k,l) + s(k,l,n)\}\\
&B(m,l,n) = \operatorname*{argmax}_{k<l}\{g(m-1,k,l) + s(k,l,n)\}
\end{aligned}$$

4. Select the optimal path
   The best complete hypothesis consisting of M words is determined by selecting the last two words $(w_{\hat{l}}, w_{\hat{n}})$.

$$S(\hat{V}) = \max_{\substack{N-M<n\leqq N\\ N-M-1<l\leqq N-1}} g(M,l,n) + \log P(\texttt{</s>}|w_l w_n)$$

$$(\hat{n},\hat{l}) = \operatorname*{argmax}_{\substack{N-M<n\leqq N\\ N-M-1<l\leqq N-1}} g(M,l,n) + \log P(\texttt{</s>}|w_l w_n)$$

5. Backtracking

$$\begin{aligned}
&\text{for } m = M \text{ to } 1\\
&\quad v_m = w_{\hat{n}}\\
&\quad l' = B(m,\hat{l},\hat{n})\\
&\quad \hat{n} = \hat{l}\\
&\quad \hat{l} = l'
\end{aligned}$$

Figure 2 shows the two-dimensional space for the dynamic programming process. The vertical axis represents the transcription consisting of ten words ($N = 10$), and the horizontal axis represents the summarized sentence consisting of five words ($M = 5$). All possible sets of five words extracted from the ten words are traced by paths from the bottom left corner to the top right corner. The path that maximizes the summarization score has been selected. In Fig. 2, a set of words, $v_1, \ldots, v_5$, maximizing the summarization score is extracted from the transcriptions, w_3, w_4, w_6, w_9 and w_{10}.

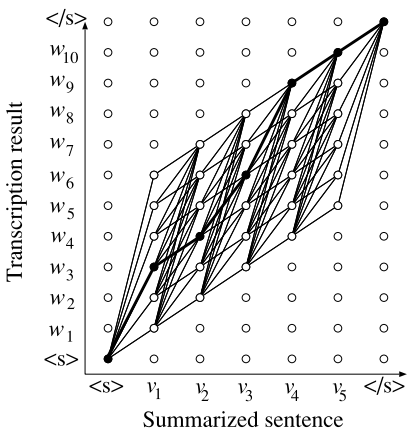

Fig. 2 Example of DP alignment for speech summarization. ($N = 10$, $M = 5$)

Although our process of summarization is carried out based on words, the summarization ratio can be determined based on the number of words or characters. When summarization is calculated based on the number of characters, the number of characters in a summarized sentence changes depending on which words are extracted. Thus the number of words to be extracted in a summarized sentence, M, cannot be determined by the summarization ratio. A summarized sentence that has a maximum number of words under the target summarization ratio is selected from multiple hypotheses consisting of various numbers of words.

3. Evaluation Method for Automatic Summarizations

3.1 Word Network of Manual Summarizations Used for Evaluation

Automatic speech summarizations were compared with correct target summarizations to evaluate our technique. Correctly transcribed speech was manually summarized by human subjects and then used as a correct target to automatically evaluate summarized sentences. The manual summarizations varied among different subjects. These variations bring up the following problems:

1. How to consider all possible correct answers in manual summarization, and
2. How to measure the similarities between evaluated sentences and multiple manual summaries.

If all possible manual summarizations could be collected, the one that was the most similar to the automatic results would be chosen as the correct answer and used for evaluation. However, in real situations, the number of manually summarized sentences that can be collected is limited. The coverage of real answers in the collected manual summaries is unknown. When the coverage is low, the summarizations are compared with inappropriate targets, and the *word accuracy* obtained through such comparisons is inefficient.

In this paper, we propose the concept of *summarization accuracy* to measure global similarity and cope with the coverage problem at the same time. Through this, the manual summarizations are merged into a word network, which approximately expresses all of the possible correct summarizations, including subjective variations. The word sequence in the network that is closest to the evaluation word sequence, is extracted and used to measure the similarity based on word accuracy.

Table 1 shows the manual summarizations generated by six human subjects, while Fig. 3 shows the variations of manual summarizations when merged into a

Table 1 Example of manual summarization by human subjects.

ORG	日本 で 開かれている 地球 温暖化 防止 会議 (Conference for prevention of global climate change held in Japan)
A	日本 で 開かれている ___ _____ ___ 会議 (Conference held in Japan)
B	日本 で 開かれている ___ _____ 防止 会議 (Conference for prevention held in Japan)
C	日本 で 開かれている ___ 温暖化 防止 会議 (Conference for prevention of climate change held in Japan)
D	日本 で _____ 地球 温暖化 防止 会議 (Conference for prevention of global climate change in Japan)
E	日本 で _____ ___ 温暖化 防止 会議 (Conference for prevention of climate change in Japan)
F	___ __ _____ 地球 温暖化 防止 会議 (Conference for prevention of global climate change)

ORG: original sentence,
A-F: manual summarization by six subjects

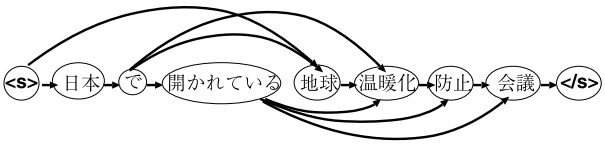

Fig. 3 Example of word network expressing manual summarization.

word network. In this figure, the links connecting words represent the possible word concatenations in the summarized sentence. All word strings connected by links from the initial symbol <s> to </s> in the network can be targets that retain important information and meaning in the original sentence as well as linguistic correctness.

A word string that is most similar to the automatic summarization is extracted from the word network as the target answer for automatic summarization. *Summarization accuracy* given by Eq. (5) for automatic summarization is calculated by comparing the automatic summarization with the extracted sentence.

$$Sum\_acc = \frac{Len - Sub - Ins - Del}{Len} \times 100[\%], \quad (5)$$

where

Sum_acc	: the summarization accuracy,
Sub	: the number of substitutions compared with the target word string,
Ins	: the number of insertions compared with the target word string,
Del	: the number of deletions compared with the target word string, and
Len	: the number of words in the target word string.

The most similar word string to the example summarized sentence "<s> 日本 で 温暖化 会議 </s>" (Conference for climate change in Japan) in the network (Fig. 3) is "<s> 日本 で 温暖化 防止 会議 </s>" (Conference for prevention of climate change). A deletion error of 1 is obtained by comparing these two sentences. The *summarization accuracy* of this automatic summarization is 75%. Such accuracy enables us to evaluate important information extraction, retention of the original meaning, and inappropriateness of summarization simultaneously. However, this measure indicates only the similarities between patterns in automatic summarizations and manual summarizations. *Summarization accuracy* is not likely to match human judgment in reading the automatic summarizations. Since humans understand the meanings of sentences based on their patterns, we can determine a strong correlation coefficient between human judgment and automatic evaluation based on *summarization accuracy*. In addition, human subjective evaluations are varied by psychological conditions. To obtain reliable subjective evaluations, the automatic summarizations should be evaluated by as many human subjects as possible. In comparison with costly and time consuming human judgment, automatic evaluation using *summarization accuracy* is a simple and effective means of numerical evaluation.

4. Evaluation Experiments

4.1 Experimental Conditions

We used spoken Japanese news broadcasts on TV in 1996 as the test data set to evaluate the proposed method. The news delivered by a female anchor was recognized by an automatic speech recognition (ASR) system. The transcription was manually segmented into sentences. The spoken news consisted of 419 utterances. Fifty utterances with a word recognition accuracy above 90%, were automatically summarized and evaluated. The out-of-vocabulary (OOV) rate for a 20,000-word vocabulary was 2.5%, and the perplexity for the test set was 54.5. This test set included one disfluency, thirteen fillers, and no repetitions. These transcriptions were automatically summarized, and the summarization ratios, the ratio of the number of characters in the summarized sentences to that in the original sentences, were set to 20, 40, 60, 70 and 80%.

The automatic transcriptions (RECOG) were summarized with various combinations of word significance scores (I), linguistic scores (L), and confidence scores (C). Seven types of summarization were generated by combining the scores as follows: I, L, C, I_C, L_C, I_L and I_L_C. These automatic summarizations were evaluated using manual summarizations generated by 25 human subjects through word extraction. The weighting factors to balance I, L and C, namely, λ_I, λ_L and λ_C, and the flat score of the word significance score I_{const} were experimentally optimized.

To test the performance of automatic summariza-

tion for transcriptions with 100% word accuracy, the manual transcriptions (TRANS) by humans were automatically summarized. The manual summarization by 25 human subjects of manual transcriptions (SUB) was set as the upper limit for automatic summarization. Each manual summarization was evaluated using all the other 24 manual summarizations as correct summarizations. To insure that our method was sound, Summarization sentences generated by randomly extracting words according to the summarization ratio (RDM) were compared with the automatic summarizations.

4.2 Structure of Transcription System

4.2.1 Feature Extraction

Sounds were digitized at 16-kHz sampling and 16-bit quantization. The frame width was 25 ms and the frame shift was 10 ms. Each feature vector extracted from speech consisted of 12 MFCCs, their delta features (derivatives), and a delta feature of normalized logarithmic power. There were a total of 25 parameters in each vector. Cepstral coefficients were normalized with the CMS (cepstral mean subtraction) method.

4.3 Acoustic Models

The acoustic models used were shared-state triphone HMMs designed using tree-based clustering. The total number of states was 1,012, and the number of Gaussian mixture components per state was 8. A speaker-dependent model was trained based on the maximum likelihood method. There were a total of 985 training utterances, and the total length of the training data was approximately 2 hours. The initial model was a female model provided by IPA [8].

4.3.1 Language Models

A bigram and a trigram were used for the language models. Manuscripts prepared for news broadcasts from July 1992 to May 1996, containing approximately 500,000 sentences consisting of 22,000,000 words, were used for constructing the language models. The vocabulary was 20,000 words in size. Since Japanese sentences are written without spaces between words, the broadcast-news manuscripts were segmented into words by using the JUMAN, morphological analyzer [9]. In addition, the morphological analyzer tagged words with POS and pronunciation simultaneously. The n-gram language models were calculated using these tuples consisting of word, POS and pronunciation.

4.3.2 Decoder

A word-graph-based 2-pass decoder transcribed speech [10]. A frame-synchronous beam search was performed in the first pass using the previously mentioned HMMs and the bigram language model. Context dependency between phonetics was also considered. A word graph was generated as a result of the first pass. In the second pass, the word graph was rescored with the trigram language model. Since each word entry was tagged with POS and pronunciation in our Japanese LVCSR system, the recognition results obtained with our system were words appended to POS and pronunciation.

4.4 Training Data for Summarization Models

4.4.1 Word Significance Model

The same manuscripts used for building language models in speech recognition were used to calculate the word significance measure for summarization. The recurrence frequency of each word in the training data was used for the significance score.

4.4.2 Language Model for Summarization

The language models for summarization should represent word strings in the summarized sentences. However, there is no large corpus consisting of summarized sentences that permits the language models for summarization to be calculated. The topics in newspapers and broadcast news are the same. Newspaper texts are usually more compact and simpler than broadcast news text in terms of the number of modifiers, the abbreviation of postpositions, and the termination of sentences with noun/noun-phrase. Therefore, newspaper texts can be used to calculate language models to summarize broadcast news. A trigram language model for summarization was built using text from the Mainichi newspaper published from 1996 to 1998. The vocabulary used for the recognition system was also used in the language model for summarization.

The newspaper text had fewer morphemes than the broadcast news manuscripts as follows:

broadcast-news text : 44 morphemes/sentence
newspaper text : 17 morphemes/sentence

4.4.3 Confidence Measure

The confidence measure for each word in the one best set of a recognition result was calculated using the word graph obtained in the first pass by the recognition system. As the confidence measure, we used the log of the posterior probability of each word in the word graph calculated using the acoustic and linguistic likelihoods.

5. Summarization Results

Table 2 lists the automatic summarization results,

Table 2 Summarization results for manual and automatic transcriptions.

TRANS	ジュネーブで開かれている地球温暖化対策の国際会議で日本政府は西暦二千年以降先進各国が ＧＤＰに応じた二酸化炭素の排出削減に努めるという新たな国際目標を日本としては今回初めて提案することを決めました (The Japanese government for the first time decided to propose a new international goal that every advanced country try to reduce CO_2 emission according to its own GDP after A.D. 2000 at the International conference for prevention of global climate change held in Geneva)
80%	ジュネーブで開かれている地球温暖化対策の国際会議で日本政府は＿＿＿＿＿先進各国が＿＿＿ ＧＤＰに＿＿＿二酸化炭素の排出削減＿＿＿＿＿＿＿＿＿＿目標を日本としては今回初めて提案することを決めました (The Japanese government decided to propose a goal that every advanced country reduce CO_2 emission its own GDP at the International conference for prevention of global climate change held in Geneva)
70%	ジュネーブで開かれている地球温暖化対策の国際会議で日本政府は＿＿＿＿＿先進各国が＿＿＿ ＿＿＿＿＿二酸化炭素の排出削減＿＿＿＿＿＿＿＿＿＿目標を日本として＿＿＿＿＿提案することを決めました (The Japanese government decided to propose a goal that every advanced country reduce CO_2 emission at the International conference for prevention of global climate change held in Geneva)
60%	ジュネーブで開かれている地球温暖化対策の国際会議で日本政府は＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿ ＿＿＿＿＿二酸化炭素の排出削減＿＿＿＿＿＿＿＿＿＿目標を日本として＿＿＿＿＿提案することを決めました (The Japanese government decided to propose a goal for reducing CO_2 emission at the International conference for prevention of global climate change held in Geneva)
40%	＿＿＿＿＿＿＿＿＿＿地球温暖化対策＿＿＿会議で日本政府は＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿ ＿＿＿＿＿二酸化炭素の排出削減＿＿＿＿＿＿＿＿＿＿目標を＿＿＿＿＿＿＿＿＿＿提案することを決め＿＿＿ (The Japanese government decided to propose a goal for reducing CO_2 emission at the International conference for prevention of global climate change)
20%	＿＿ ＿＿＿＿＿二酸化炭素の排出削減＿＿＿＿＿＿＿＿＿＿目標を＿＿＿＿＿＿＿＿＿＿提案すること＿＿＿＿＿＿ (Proposing a goal for reducing CO_2 emission)
RECOG	[年] で開かれている [月いう] 温暖化対策の国際会議で日本政府は西暦二千年以降先進各国がＧＤＰ 国内総生産に応じた二酸化炭素の排出削減に努めるという新たな国際目標を日本としては今回初めて提案することを決めました (The Japanese government for the first time decided to propose a new international goal that every advanced country try to reduce CO_2 emission according to its own GDP after A.D. 2000 at the International conference for prevention of [moon says] climate change held in [year])
80%	＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿温暖化対策の国際会議で日本政府は＿＿＿＿＿＿＿先進各国が＿＿＿＿＿ ＿＿＿＿＿＿＿＿二酸化炭素の排出削減に努めるという新たな国際目標を日本としては今回初めて提案することを決めました (The Japanese government for the first time decided to propose a new international goal that every advanced country try to reduce CO_2 emission at the International conference for prevention of climate change)
70%	＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿温暖化対策の国際会議で日本政府は＿＿＿＿＿＿＿先進各国が＿＿＿＿＿ ＿＿＿＿＿＿＿＿二酸化炭素の排出削減に努めるという＿＿＿＿＿＿＿日本としては今回初めて提案することを決めました (The Japanese government for the first time decided to propose that every advanced country try to reduce CO_2 emission at the International conference for prevention of climate change)
60%	＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿温暖化対策の国際会議で日本政府は＿＿＿＿＿＿＿先進各国が＿＿＿＿＿ ＿＿＿＿＿＿＿＿二酸化炭素の排出削減に努めるという＿＿＿＿＿目標＿日本として＿＿＿＿＿提案することを決めました (The Japanese government decided to propose a goal that every advanced country try to reduce CO_2 emission at the International conference for prevention of climate change)
40%	＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿温暖化対策の国際会議で日本＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿ ＿＿＿＿＿＿＿＿二酸化炭素の排出削減＿＿＿＿＿＿＿＿＿＿目標を＿＿＿＿＿＿＿＿＿＿提案することを決めました (Japan decided to propose the goal for reducing CO_2 emission at the International conference for prevention of climate change)
20%	＿＿ ＿＿＿＿＿＿＿＿二酸化炭素の排出削減＿＿＿＿＿＿＿＿＿＿目標＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿ (A goal for reducing CO_2 emission)

percent: summarization ratio, [ ]: recognition error, TRANS: manual transcriptions, RECOG: automatic recognition,
( ): Translations from Japanese to English.

which indicate that this method can extract relatively important information and generate understandable sentences according to the target summarization ratio. In addition, the automatic summarization of automatic transcriptions in Table 2 excludes recognition errors.

The 50 utterances recognized by the ASR system included 69 recognition errors other than substitution errors for kana such as 「等」to「など」that did not change the meaning of sentences. When recognition errors changing the meanings of sentences were extracted into automatic summarizations, the summarization could not maintain the meaning of expressions users had uttered. Table 3 lists the number of word errors and the number of sentences including word errors in automatic summarization. The table indicates

Table 3 Number of word errors and summarized sentences including word errors.

summarization ratio	RDM	SUM
100%	69 (50)	
80%	36 (17)	12 (8)
70%	31 (16)	5 (5)
60%	25 (15)	3 (3)
40%	18 (13)	2 (2)
20%	8 (7)	3 (3)

100% is obtained directly from the recognition results.
():number of sentences including recognition errors

that the automatic summarization approach proposed can alleviate summarization errors caused by extracting recognition errors into summarization results.

6. Evaluation Results Using Summarization Accuracy

Table 4 shows examples of automatic summarization and the corresponding target extracted from a manual summarization word network. Examples of the automatic summarizations of automatic transcriptions are listed in Table 2.

The *summarization accuracies* of automatic summarization are shown in Figs. 4 to 8. These results indicate that our proposed technique of automatic speech summarization is significantly more effective than random word selection (RDM) under all of the summarization ratios. The linguistic score, L can improve the *summarization accuracy* of automatic summarization for recognition results (REC) under every summarization ratio. In comparison with all other combinations of scores, the combination of the significance score and the linguistic score, I_L, and the combination of the significance score, the linguistic score and the confidence score, I_L_C, achieved the highest *summarization accuracy* with summarization ratios less than 40% and more than 60%, respectively.

Since automatic summarization was carried out using recognition with more than 90% word accuracy in this study, the confidence score C contributed only to

Table 4 Example of evaluation results based on a manual summarization word network.

TRANS	ジュネーブで開かれている地球温暖化対策の国際会議で日本政府は西暦二千年以降先進各国が ＧＤＰに応じた二酸化炭素の排出削減に努めるという新たな国際目標を日本としては今回初めて提案することを決めました (The Japanese government for the first time decided to propose a new international goal that every advanced country try to reduce CO_2 emission according to its own GDP after A.D. 2000 at the International conference for prevention of global climate change held in Geneva)
RECOG	[年] で開かれている [月いう] 温暖化対策の国際会議で日本政府は西暦二千年以降先進各国がＧＤＰ 国内総生産に応じた二酸化炭素の排出削減に努めるという新たな国際目標を日本としては今回初めて提案することを決めました (The Japanese government for the first time decided to propose a new international goal that every advanced country try to reduce CO_2 emission according to its own GDP after A.D. 2000 at the International conference for prevention of [moon says] climate change held in [year])
80%	地球 温暖 化 対策 の 国際 会議 で 日本 政府 は ＧＤＰ に 応じた 二酸化炭素 の 排出 削減 に 努める と いう 新たな 国際 目標 を 日本 と して は 今回 初めて 提案 する こと を 決め ました (The Japanese government for the first time decided to propose a new international goal for trying to reduce CO_2 emission according to its own GDP at the International conference for prevention of global climate change)
	DEL 温暖 化 対策 の 国際 会議 で 日本 政府 は <先進> <各国> <が> 二酸化炭素 の 排出 削減 に 努める と いう 新たな 国際 目標 を 日本 と して は 今回 初めて 提案 する こと を 決め ました (The Japanese government for the first time decided to propose a new international goal that every advanced country try to reduce CO_2 emission at the International conference for prevention of climate change)
70%	温暖 化 対策 の 国際 会議 で 日本 政府 は ＧＤＰ に 応じた 二酸化炭素 の 排出 削減 に 努める と いう 目標 を 日本 と して は 今回 初めて 提案 する こと を 決め ました (The Japanese government for the first time decided to propose a goal for trying to reduce CO_2 emission according to its own GDP at the International conference for prevention of climate change)
	温暖 化 対策 の 国際 会議 で 日本 政府 は <先進> <各国> <が> 二酸化炭素 の 排出 削減 に 努める と いう DEL DEL 日本 と して は 今回 初めて 提案 する こと を 決め ました (The Japanese government for the first time decided to propose that every advanced country try to reduce CO_2 emission at the International conference for prevention of climate change)
60%	温暖 化 対策 の 国際 会議 で 日本 政府 は 先進 各国 が 二酸化炭素 の 排出 削減 に 努める と いう 目標 INS INS INS 提案 する こと を 決め ました (The Japanese government decided to propose a goal that every advanced country try to reduce CO_2 emission at the International conference for prevention of climate change)
	温暖 化 対策 の 国際 会議 で 日本 政府 は 先進 各国 が 二酸化炭素 の 排出 削減 に 努める と いう 目標 日本 と して 提案 する こと を 決め ました (The Japanese government decided to propose a goal that every advanced country try to reduce CO_2 emission at the International conference for prevention of climate change)
40%	温暖 化 対策 の 国際 会議 で 日本 政府 二酸化炭素 の 排出 削減 INS を 提案 する こと を 決め ました (The Japanese government decided to propose reducing CO_2 emission at the International conference for prevention of climate change)
	温暖 化 対策 の 国際 会議 で 日本 DEL 二酸化炭素 の 排出 削減 目標 を 提案 する こと を 決め ました (Japan decided to propose a goal for reducing CO_2 emission at the International conference for prevention of climate change)
20%	二酸化炭素 の 排出 削減 目標 提案 (Proposing a goal for reducing CO_2 emission)
	二酸化炭素 の 排出 削減 目標 DEL (A goal for reducing CO_2 emission)

TRANS: manual transcriptions, RECOG: automatic recognition, percent: summarization ratio,
upper: correct target summarization, lower: automatic summarization results,
[ ]: recognition error, < > substitution error, INS: insertion error, DEL: deletion error.

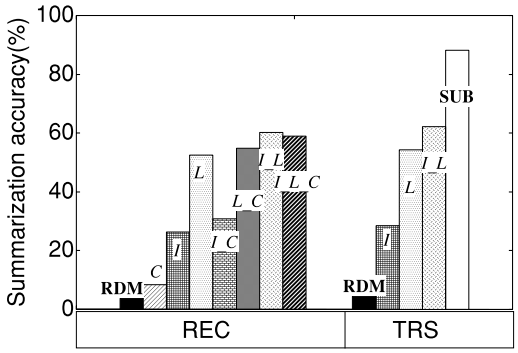

Fig. 4 Summarization at 20% summarization ratio. REC: summarization of recognition, TRS: summarization of manual transcription, RDM: random word selection, C: confidence score, I: significance score, L: linguistic score, I_C, L_C, I_L: combination of 2 scores, I_L_C: combination of all scores, SUB: subjective summarization.

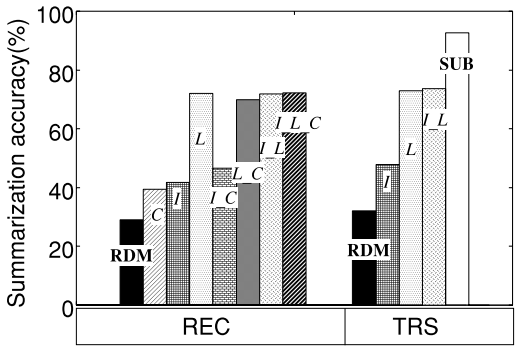

Fig. 5 Summarization at 40% summarization ratio. C: confidence score, I: significance score, L: linguistic score, I_C, L_C, I_L: combination of 2 scores, I_L_C: combination of all scores.

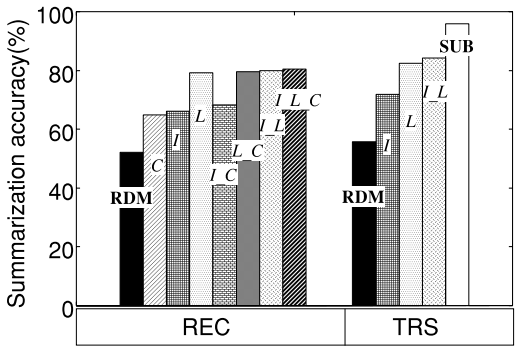

Fig. 6 Summarization at 60% summarization ratio. C: confidence score, I: significance score, L: linguistic score, I_C, L_C, I_L: combination of 2 scores, I_L_C: combination of all scores.

automatic summarization based on a higher summarization ratio in which the probability of extracting recognition errors increases.

The *summarization accuracy* of automatic summarization for manual transcription (TRS) was improved by the linguistic score, L. The best summarization accuracy was achieved by combining the significance score and the linguistic score, I_L. However, automatic summarization could not summarize as accurately as manual summarization.

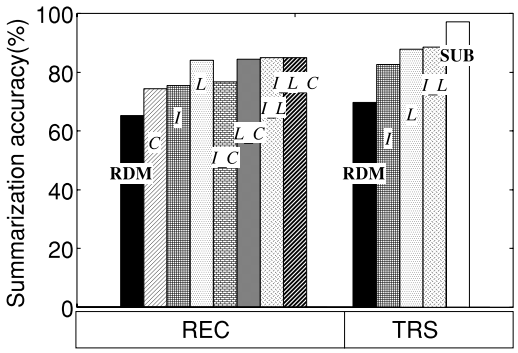

Fig. 7 Summarization at 70% summarization ratio. C: confidence score, I: significance score, L: linguistic score, I_C, L_C, I_L: combination of 2 scores, I_L_C: combination of all scores.

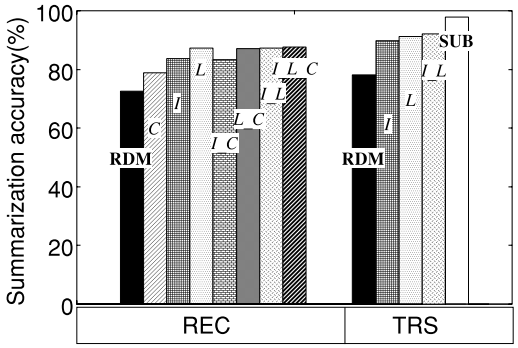

Fig. 8 Summarization at 80% summarization ratio. C: confidence score, I: significance score, L: linguistic score, I_C, L_C, I_L: combination of 2 scores, I_L_C: combination of all scores.

The linguistic score, L, is more effective in increasing the summarization accuracy of both REC and TRS than the significance score, I. Humans generate summarized sentences consisting of important topic words as well as other words that are essential for retaining the original meaning. Therefore, these manual summarizations are not a string of topic words. Automatic summarization generated by using the higher significance score, I, extracted mainly topic words. As a result, these were difficult to understand as sentences, and their summarization accuracy decreased.

In comparison with automatic summarization using only the significance score, I, more well-formed sentences could be generated as summarizations when the linguistic score was weighted more. Therefore, the summarization accuracy of the automatic summarization using only the linguistic score, L, is higher than that using only the significance score, I. Combining the significance and linguistic scores can generate relatively well-formed sentences that include more important topic words and thus accomplish the highest summarization accuracy.

There is a decline in the performance of automatic summarization with a lower summarization ratio. When the summarization ratio is lower, summarization should be generated by extracting fewer topic words. Topic words should be detected more accurately to summarize well with a lower summarization ratio. However, the significance score calculated using all of the data in the prepared manuscripts of broadcast-news is general and consequently does not always detect topic words well. This problem can be solved by adapting the significance score to the test set.

7. Conclusions

This paper proposed an automatic method of speech summarization through word extraction using a DP technique based on the word significance score, linguistic likelihood, and the word confidence measure. This paper also proposed an automatic method evaluating the summarization achieved by extracting words from original sentences using *summarization accuracy* based on a word network produced by merging manual summarizations.

The evaluation results revealed that the proposed automatic summarization method can generate relatively well-formed sentences that include more important topic words while minimizing the extraction of recognition errors.

Our future research will include making abstracts from speech consisting of multiple utterances, such as complete news stories and lectures, and practical applications of screening to extract more reliable components from automatic transcriptions at lower recognition accuracy.

Although we proposed a new method of evaluating automatic summarization by comparing it with manual summarizations, a problem remains where all possible summarized sentences cannot always be collected by a limited number of humans. Summarizations obtained from ill-formed speech are sometimes linguistically incorrect but semantically correct and understandable. *Summarization accuracy* is too strict in these cases. Our future research will include task-dependent evaluation methods such as information retrieval. Performance needs to be evaluated from the viewpoint of how much the original meaning, which is indispensable to accomplish the target tasks, is maintained in the summarization results.

Acknowledgments

The authors would like to thank NHK (Japan Broadcasting Corporation) for providing us with the broadcast-news database and the Language Media Laboratory of Kyoto University for providing *Kyoto Univ. Corpus*.

References

[1] T. Imai, A. Kobayashi, S. Sato, and A. Ando, "Broadcast news transcription system with a progressive 2-pass decoder," IEICE Technical Report, SP99-129, Dec. 1999.
[2] T. Shinozaki, Y. Saito, C. Hori, and S. Furui, "Toward spontaneous speech recognition," IEICE Technical Report, SP2000-96, Dec. 2000.
[3] R. Valenza, T. Robinson, M. Hickey, and R. Tucker, "Summarization of spoken audio through information extraction," Proc. ESCA Workshop on Accessing Information in Spoken Audio, pp.111–116, 2000.
[4] I. Manu and M. Maubury, Advances in Automatic Text Summarization, The MIT Press, 1999.
[5] A. Iwasaki and S. Furui, "Topic extraction from broadcast-news speech," Proc. Autumn Meet. Acoust. Soc. Jpn., vol.1, 1-1-14, pp.27–28, 1998.
[6] T. Kemp and T. Schaaf, "Estimating confidence using word lattices," Proc. 5th Eurospeech, vol.2, pp.827–830, Rhodes, 1997.
[7] V. Valtchev, J.J. Odel, P.C. Woodland, and S.J. Young, "MMIE training of large vocabulary recognition systems," Speech Communication, vol.22, pp.303–314, 1997.
[8] "Japanese Dictation Toolkit," IPA (Information-technology Promotion Agency), Japan, http://www.lang.astem.or.jp/dictation-tk/
[9] S. Kurohashi, T. Nakamura, Y. Matsumoto, and M. Nagao, "Improvements of Japanese morphological analyzer JUMAN," Proc. Int. Workshop on Sharable Natural Language Resources, Nara, Aug. 1994.
[10] T. Hori, N. Oka, M. Katoh, A. Ito, and M. Kohda, "A study on a phoneme-graph-based hypothesis restriction for large vocabulary continuous speech recognition," Trans. IPSJ, vol.40, no.4, pp.1365–1373, 1999.

Chiori Hori received B.E. and M.E. degrees in electrical and information engineering from Yamagata University, Yonezawa, Japan, in 1994 and 1997, respectively. In April 1999, she started the doctoral course in the Graduate School of Information Science and Engineering at Tokyo Institute of Technology (TITECH), Tokyo, Japan, and received the Ph.D. degree in March 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).

Sadaoki Furui is currently a Professor with the Department of Computer Science, Tokyo Institute of Technology (TITECH), Japan. He is engaged in a wide range of research on speech analysis, speech recognition, speaker recognition, speech synthesis, and multimodal human-computer interaction. From 1978 to 1979, he served on the staff of the Acoustics Research Department, Bell Laboratories, Murray Hill, NJ, as a Visiting Researcher working on speaker verification. 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 is the author of *Digital Speech Processing, Synthesis, and Recognition* (New York: Marcel Dekker, 1989; revised 2000), in English, *Digital Speech Processing* (Tokyo, Japan: Tokai University Press, 1985), in Japanese, *Acoustics and Speech Processing* (Tokyo, Japan: Kindai-Kagaku-Sha, 1992), in Japanese, and *Speech Information Processing* (Tokyo, Japan: Morikita, 1998), in Japanese. He edited *Advances in Speech Signal Processing* (New York: Marcel Dekker, 1992) jointly with Dr. M.M. Sondhi. He has translated into Japanese *Fundamentals of Speech Recognition*, authored by Dr. L.R. Rabiner and Dr. B.H. Juang (Tokyo, Japan: NTT Advanced Technology, 1995) and *Vector Quantization and Signal Compression*, authored by Dr. A. Gersho and Dr. R.M. Gray (Tokyo, Japan: Corona-sha, 1998). Dr. Furui is a Fellow of the Acoustical Society of America. 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 on the Board of Governors of the IEEE Signal Processing Society (SPS). He has served on the IEEE Technical Committees on Speech and MMSP and on numerous IEEE conference organizing committees. He received the Yonezawa Prize and the Paper Award from the IEICE in 1975, 1988, and 1993, and the Sato Paper Award from the ASJ in 1985 and 1987. He received the Senior Award from the IEEE ASSP Society in 1989 and the Achievement Award from the Minister of Science and Technology, Japan, also in 1989. He has received the Book Award from the IEICE in 1990. He has also received the Mira Paul Memorial Award from the AFECT, India, in 2001. In 1993, he served as an IEEE SPS Distinguished Lecturer.