論文 / 著書情報
Article / Book Information

論題(和文)	
Title	Toward the Realization of Spontaneous Speech Recognition and Summarization
著者(和文)	古井 貞熙, 篠崎 隆宏
Author	Sadaoki Furui, Chiori Hori, Takahiro Shinozaki
出典(和文)	, Vol. , No. , pp. 1-21
Journal/Book name	Research on Computational Linguistics Conference IV (2001 ROCLING), Vol. , No. , pp. 1-21
発行日 / Issue date	2001,

Proc. Research on Computational Linguistics Conference IV (2001 ROCLING), pp. 1-21 (2001)

Toward the realization of spontaneous speech recognition and summarization

Sadaoki Furui, Chiori Hori and Takahiro Shinozaki

Department of Computer Science

Tokyo Institute of Technology

2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan

furui@cs.titech.ac.jp

Abstract

Although high recognition accuracy can be obtained for speech in the form of reading a written text or similar by using state-of-the art speech recognition technology, the accuracy is quite poor for freely spoken spontaneous speech. From this perspective, a five-year national project for raising the technological level of speech recognition and understanding commenced in Japan in 1999. The project focuses on building a large-scale spontaneous speech corpus and acoustic and linguistic modeling for spontaneous speech recognition and summarization. This paper reports some results of preliminary experiments which have been conducted at Tokyo Institute of Technology. Experimental results show that acoustic and language modeling based on the actual spontaneous speech corpus is far more effective than modeling based on read speech. It was also shown that our proposed automatic speech summarization method could effectively extract relatively important information and remove redundant and irrelevant information.

1. Introduction

Read speech or similar, such as speech reading newspapers and broadcast news utterances made by announcers, can be recognized with a higher than 90% accuracy using the present speech recognition technology. However, the recognition accuracy dramatically declines for spontaneous speech. The principal reason for this is that acoustic and linguistic models used in speech recognition have been built using written language or speech reading text, while spontaneous speech and written language considerably differ both acoustically and linguistically. Broadening the

application of speech recognition thus crucially depends on raising the recognition performance for spontaneous speech.

From this viewpoint, a Japanese national project on spontaneous speech corpus and processing technology was initiated in 1999. This project aims to build a large-scale spontaneous speech corpus and create spontaneous speech recognition and summarization technology.

This paper reports results of preliminary experiments on spontaneous speech recognition and summarization conducted at Tokyo Institute of Technology. Section 2 describes the outline of the national project. Section 3 describes methods and results of experiments on automatic transcription of spontaneous presentation. Section 4 describes our proposed method of automatic speech summarization and its evaluation. Finally some conclusions and perspectives are given in Section 5.

2. Japanese national project on spontaneous speech corpus and processing technology

The Science and Technology Agency Priority Program (Organized Research Combination System) entitled "Spontaneous Speech: Corpus and Processing Technology" started in 1999 under the supervision of S. Furui [1]. The principal organizations working together to conduct this project are National Language Research Institute, Communication Research Laboratory, and Tokyo Institute of Technology.

The project will be conducted over a 5-year period in pursuit of three major themes as shown in Fig. 1:

1) Building a large-scale spontaneous speech corpus consisting of roughly 7M words with a total speech length of 700 hours. Mainly recorded will be monologues such as lectures, presentations, and news commentaries. They will be manually given orthographic and phonetic transcription. One-tenth of the utterances ("Core") will be tagged manually and used for constructing a morphological analysis program for automatically analyzing all of the 700-hour utterances. The Core will also be tagged with para-linguistic information including intonation [2].

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

3) Constructing a prototype of a spontaneous speech summarization system.

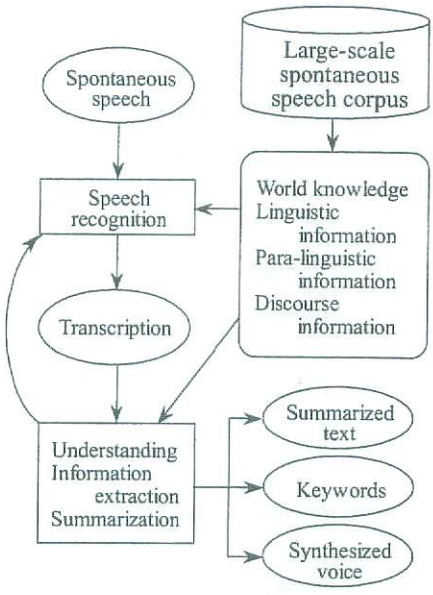

Figure 1. Overview of the national project

The technology created in this project is expected to be applicable to wide areas such as indexing of speech data (broadcast news, etc.) for information extraction and retrieval, transcription of lectures, preparing minutes of meetings, closed captioning, and aids for the handicapped.

Presentations at various conferences, such as the Acoustical Society of Japan (ASJ) meetings, and free presentations by voluntary subjects are recorded and transcribed in the project. Using these utterances, preliminary recognition experiments are being conducted at several universities participating in the project. At Tokyo Institute of Technology, for example, experiments have been conducted using a part of the corpus that has so far been built.

3. Automatic transcription of spontaneous presentation

3.1 Recognition task

Presentation speech uttered by 10 male speakers has been used as a test set of speech recognition. Table 1 shows an outline of the test set. The top four presentations in the table were on the subject of speech.

Morphemes (which will be called "words" hereafter in this paper) were used as units for statistical language modeling. For all the following recognition performances, word-based performance is measured. Fillers are counted as words and taken into account in calculating the accuracy.

Table 1: Recognition test set of presentations

ID	Conference name	Length [min]
A22	Acoust. Soc. Jap.	28
A23	Acoust. Soc. Jap.	30
A97	Acoust. Soc. Jap.	12
P25	Phonetics Soc. Jap.	27
J01	Soc. Jap. Linguistics	57
K05	National Lang. Res. Inst.	42
N07	Assoc. Natural Lang. Proc.	15
S05	Assoc. Socioling. Sciences	23
Y01	Spont. Speech Corpus Meeting	14
Y05	Spont. Speech Corpus Meeting	15

3.2 Experimental conditions

Sounds were digitized with 16kHz sampling and 16 bit quantization. They were segmented into utterances using silence periods longer than 500ms. Feature vectors had 25 elements consisting of 12 MFCC, their delta, and the delta log energy. CMS (cepstral mean subtraction) was applied to each utterance. HTK v2.2 was utilized for acoustic modeling and speaker adaptation. Language models were made by the use of CMU SLM Tool Kit v2.05. The Julius v3.1 decoder [3] was used

for speech recognition. Language model weighting used in the decoding process was changed for each combination of acoustic and language model but kept constant over all speech in the test set.

3.3 Corpora

The following two corpora were used for training.

- Corpus of Spontaneous Japanese (CSJ): A part of the corpus completed by the end of December 2000, consisting of approximately 1.5M words of transcriptions, was used. The training set consisted of 610 presentations; 274 academic conference presentations and 336 simulated presentations. The simulated presentations talking about a wide variety of topics including the subjects' experiences in their daily lives were specially recorded for the project.

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

3.4 Language modeling

We built three language models, denoted as SpnL, WebL, and WSpL. Each model consisted of bigrams and reverse trigrams with backing-off. Their vocabulary sizes were all 30k. Table 2 shows the size of the corpus used to build each language model.

Table 2: Corpus size for training each language model

Language model	Corpus size [words]
SpnL	1.5 M
WebL	2 M
WSpL	2+0.06 M

SpnL: Made using the 610 presentations in the CSJ. The speakers had no overlap with those of the test set. Since there were no punctuation marks in the transcription, commas were inserted at silences of 200ms or longer duration.

WebL: Made using the text of our Web corpus.

WSpL: Made by adding whole text of a textbook on speech processing authored by Furui to the Web corpus with equal weighting for task adaptation. The textbook contains about 63k words.

3.5 Acoustic modeling

The following two tied-state triphone HMMs were made. Both models have 2k states and 16 Gaussian mixtures in each state.

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

RdA: The acoustic model made by Information-technology Promotion Agency (IPA) and contained in the CD-ROM "Japanese Dictation Toolkit 99". Approximately 40-hour long read speech uttered by many speakers was used.

3.6 Test-set perplexity and OOV rate

Figure 2 presents test-set perplexity of 3grams and out-of-vocabulary (OOV) rate for each language model. Perplexity of **SpnL** made from the CSJ, is clearly better than that of other models. **WebL** indicates high perplexity and OOV rate. This is because **WebL** is edited as a text and their topics are general. OOV rate of **WSpL** is smaller than that of **WebL** for the four left-hand-side speeches. This shows that task adaptation by adding the textbook worked well. **SpnL** is superior to **WSpL** also in terms of OOV rate.

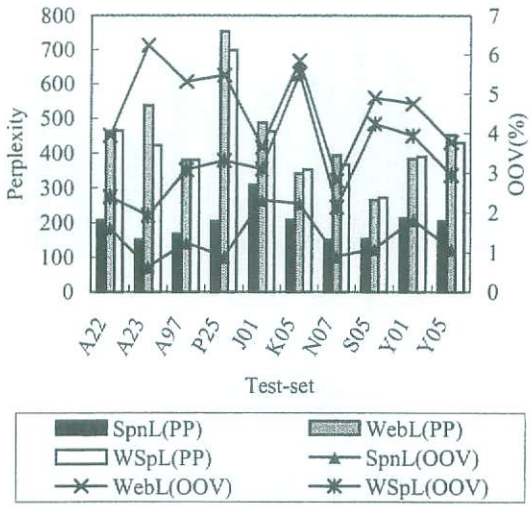

Figure 2. Test-set perplexity (PP) and OOV rate for the three language models.

3.7 Effects of language modeling

Figure 3 shows recognition results for the three language models when **SpnA** is used as the acoustic model. **SpnL** achieves the best results. **WSpL** achieves better results than **WebL**, especially for test sets A22, A23, A97 and P25, reflecting the test-set perplexity and OOV rate reduction. Mean accuracies are 64.5%, 55.2% and 57.3% for **SpnL**, **WebL** and **WSpL**, respectively. A supplementary experiment was performed to analyze the effects of OOV rate and test set perplexity on the accuracy. In this experiment, OOV words were added to the language models as "unknown" class words; 489 words and 710 words were added to **SpnL** and **WSpL**, respectively. Resulting mean word accuracies using **SpnL** and **WSpL** were 66.0% and 60.1%, respectively. These results indicate that OOV rate and test set perplexity are equaly crucial in these models.

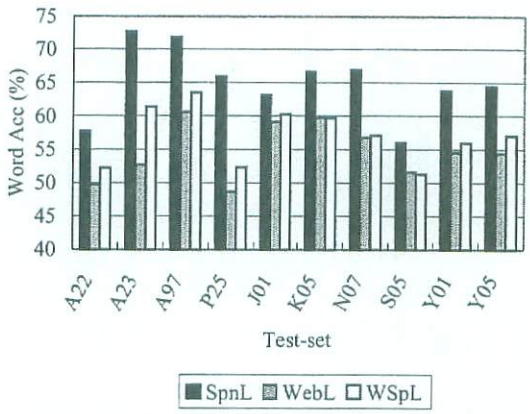

Figure 3. Word accuracy for the three language models.

3.8 Effects of acoustic modeling

The recognition results for **SpnA** and **RdA** when **SpnL** is used as the language model are shown in Fig. 4. Mean accuracies are 64.5% and 53.3% for **SpnA** and **RdA**, respectively. **SpnA** made from the CSJ achieves much better results than **RdA** made from read speech. This is probably because **SpnA** has better coverage of triphones and better matching of acoustic characteristics corresponding to the speaking style. **SpnA** also has better matching of recording conditions with the test set. CSJ and IPA corpora were both recorded using close-talking microphones, but the types of the microphones were different.

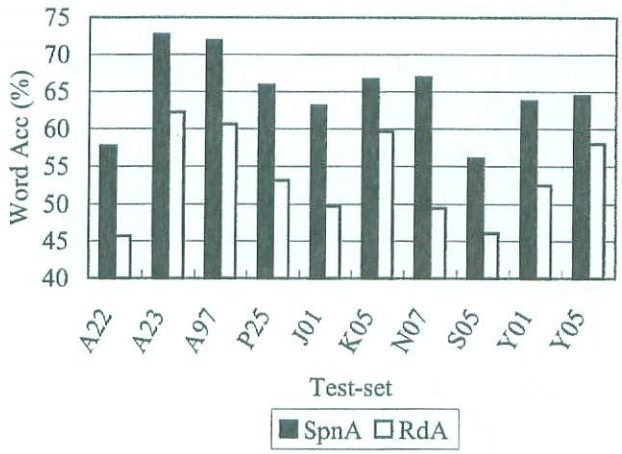

Figure 4. Word accuracy for the two acoustic models.

3.9 Individual differences

As shown in Figs. 3 and 4, the word accuracy largely varies from speaker to speaker. There exist many factors that affect the accuracy of spontaneous speech recognition. These factors include individual voice characteristics and speaking manner including noises like coughs. Although all utterances were recorded using the same close-talking microphones, acoustic conditions still varied according to the recording environment.

Figure 5 presents relationship between speaking rate and word accuracy when **SpnL** and **SpnA** were used as language and acoustic models. The speaking rate was calculated using actual speech periods after removing pauses. 10 dots in the figure correspond to individual speakers. A MMSE line fitted to those dots is also shown in the figure. The correlation coefficient is –0.58. Faster speech generally produces more errors.

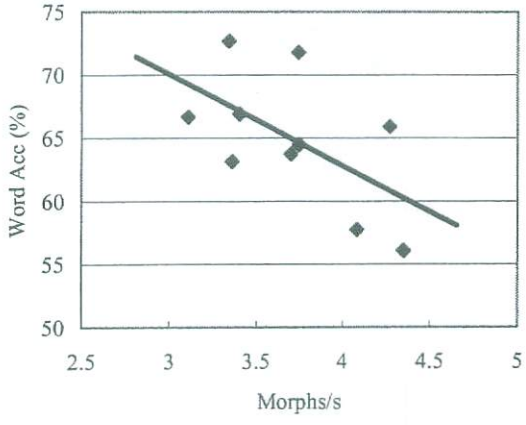

Figure 5. Speaking rate vs. word accuracy.

Figures 6 and 7 respectively show the effects of the frequencies of fillers and repairs on word accuracy. The recognition conditions were the same as those for Fig. 5. There is a general tendency that the more frequently the filler and/or the repair occurs, the more recognition error occurs.

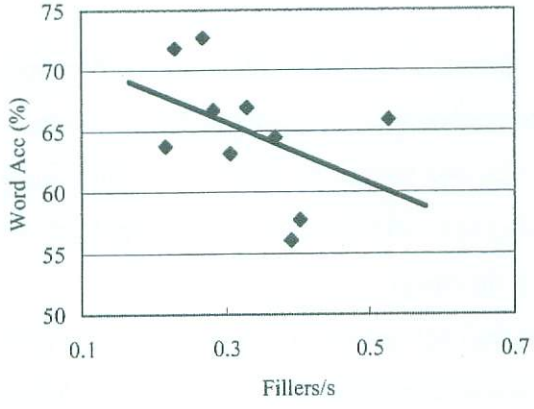

Figure 6. Filler frequency vs. word accuracy.

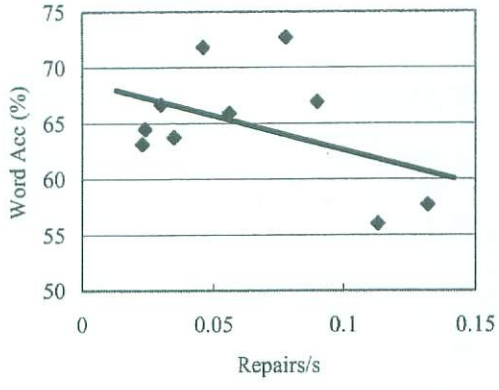

Figure 7. Repair frequency vs. word accuracy.

3.10 Unsupervised adaptation

A batch-type unsupervised adaptation method was incorporated to cope with the speech variation due to speakers and recording environment. We applied the MLLR method using a binary regression class tree to transform Gaussian mean vectors. The regression class tree was made using a centroid-splitting algorithm. The actual classes used for transformation were determined on run time according to the amount of data assigned to each class [4].

The adaptation was performed based on recognition results and no confidence measure was applied. The following steps were performed:

1. Making a regression class tree having 64 leaf nodes for the **SpnA** phone model.

2. Recognizing the test-set utterances using the **SpnA** as a speaker independent model.

3. Applying the MLLR adaptation based on the recognition result for each utterance to make a speaker adaptive model.

4. Re-recognizing the test-set utterances using the speaker adaptive model.

5. Iterating the adaptation process using the resulting transcription.

Figure 8 presents the effect of the adaptation when **SpnL** was used as the language model. "SpnA" indicates the baseline condition. "mllr" indicates the result without iterations and "mllr-i" indicates the results after one iteration of adaptation. The single step of MLLR improved word accuracy by 2 - 6 % absolute, and the second adaptation step further improved accuracy by 1% on average. By applying the two steps of MLLR, error rate was reduced by 15% relative to the speaker independent case.

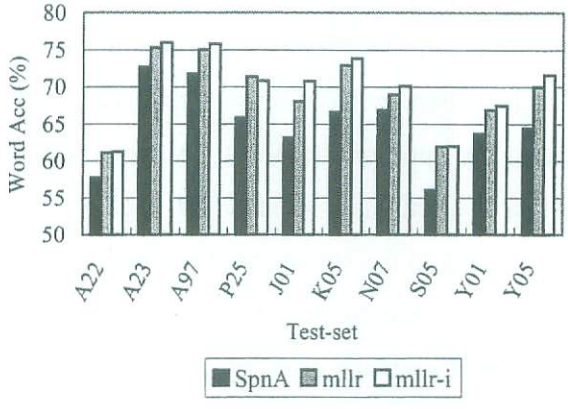

Figure 8. Results of unsupervised adaptation.

3.11 Discussion

This chapter reported experimental results for recognizing spontaneous presentation speech. Language models based on a spontaneous speech corpus and Web corpus were compared in terms of test-set perplexity, OOV rate, and word (morpheme) accuracy. Two acoustic models made by spontaneous speech and read speech were also compared. Both comparisons showed that models made from spontaneous speech were superior to models based on read speech. It was revealed that the recognition accuracy had a wide speaker-to-speaker variability. Correlation between word

accuracy and speaking rate, filler and repair frequency was observed. When linguistic and acoustic models made form spontaneous speech were used, an average word recognition accuracy of 64.5% was achieved. This performance improved to 70.0% with the help of unsupervised MLLR adaptation for the acoustic model.

Since word accuracy for this task is still very low, further improvement is required for building application systems. Future research issues include a) how to transcribe and annotate spontaneous speech, b) how to build filled pause models, c) how to incorporate repairs, hesitations, repetitions, partial words, and disfluencies, and d) how to adapt to speaking styles and topics of presentations.

4. Automatic speech summarization and evaluation

4.1 Summarization of each sentence utterance

Our proposed method to summarize speech, sentence by sentence, extracts a set of words maximizing a summarization score from an automatically transcribed sentence according to a target compression ratio. This method aims to effectively reduce the number of words by removing redundant and irrelevant information without losing relatively important information. The summarization score indicating the appropriateness of a summarized sentence consists of a word significance score I as well as a confidence score C for each word of the original sentence, a linguistic score L for the word string in the summarized sentence [5][6], and a word concatenation score T_r. The word concatenation score indicates a word concatenation probability determined by a dependency structure in the original sentence given by SDCFG [7]. The total score is maximized using a dynamic programming (DP) technique [5][6].

Given a transcription result consisting of N words, $W=w_1,w_2,\ldots,w_N$, the summarization is performed by extracting a set of M ($M<N$) words, $V=v_1,v_2,\ldots,v_M$, which maximizes the summarization score given by eq.(1).

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

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

(a) Word significance score

The word significance score I indicates the relative significance of each word in the original sentence. The amount of information based on the frequency of each word is used as the word significance score for each noun. A flat score is given to words other than nouns. To reduce the repetition of words in the summarized sentence, a flat score is also given to each reappearing noun.

(b) Linguistic score

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

(c) Word confidence score

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

(d) Word concatenation score

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

(e) Word concatenation rule

Word concatenation in a summarized sentence is restricted by the dependency structure in the original sentence as exemplified in Fig.9. Whereas our experiments are conducted for Japanese, the example is shown in English for the sake of explanation. The word concatenation rule augments the words modified by many other words in the same phrase of the original sentence, such as the "blossoms" in Fig.9, so that they remain in the summarized sentence even when the number of words extracted for summarization decreases (intra-phrase rule). The word concatenation rule also gives a score to the concatenation of words in separate phrases in the original sentence based on the dependency structure of the phrases (inter-phrase rule).

Since the dependency structure within a phrase is deterministic, the word concatenation probability between words with dependency within a phrase of the original sentence is set to 1 and that between words without dependency is set to 0. On the other hand, since the dependency structure between phrases is ambiguous, the word concatenation probability between words in different phrases is determined by a probability that one phrase is modified by others based on the SDCFG as follows.

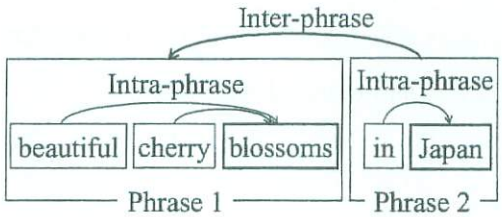

Figure 9. Word concatenation rule.

(f) Computation of word concatenation score

Suppose a sentence consists of H phrases, $P_1,\ldots,P_H$. When the kth word, w_k, belongs to a phrase $P_{h(w_k)}$ and the lth word, w_l, belongs to a phrase $P_{h(w_l)}$, the word concatenation score of w_k and w_l in the same phrase ($h(w_k)=h(w_l)$) is defined using the intra-phrase word concatenation rule($R(w_k,w_l)=0,1$). On the other hand, the word concatenation score w_k and w_l in the different phrases ($h(w_k)<h(w_l)$) is defined using the probability that $P_{h(w_k)}$ and $P_{h(w_l)}$ have the dependency structure. A word concatenation score $T_r(w_k, w_l)$ is defined as a logarithmic value of the word concatenation probability as shown in eq.(2).

$$T_r(w_k, w_l) = \begin{cases} \log \sum_{i=1}^{h(w_k)} \sum_{j=h(w_l)}^{H} \sum_{\alpha,\beta} g(\alpha \rightarrow \beta\alpha; i, h(w_k), j) & \text{if } h(w_k) < h(w_l) \\ \log R(w_k, w_l) & \text{if } h(w_k) = h(w_l) \end{cases} \quad (2)$$

where α, β are nonterminal symbols of SDCFG.

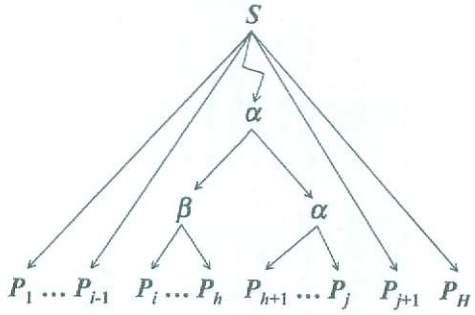

Figure 10. Inside-Outside probability

$g(\alpha \rightarrow \beta\alpha;\ i,\ h,\ j)$ is a posterior probability that the rule of $\alpha \rightarrow \beta\alpha$ is applied and then $P_i \ldots P_h$ is derived from β and $P_{h+1} \ldots P_j$ is derived from α, when a sentence is derived from the initial symbol S as shown in Fig. 10. The posterior probability is estimated using the Inside-Outside probability.

4.2 Summarization of multiple utterances with consistent meanings

Our proposed automatic speech summarization technique for each sentence can be extended to summarize a set of multiple utterances (sentences) having consistent meanings by combining a rule which is applied at sentence boundaries. As a result, the original sentences including many informative words are preserved and the sentences including few informative words are deleted or shortened. This summarization technique can be considered as a combination of the summarization method extracting important sentences investigated in the field of natural language processing and our sentence-by-sentence summarization method.

Given a transcription result consisting of J utterances, $S_1,\ldots,S_J$ ($S_j = w_{j1},w_{j2},\ldots,w_{jN_j}$) the summarization is performed by extracting a set of M ($M < \Sigma_j N_j$) words, $V = v_1,v_2,\ldots,v_M$, which maximizes the summarization score given by eq.(1). Figure 11 illustrates the DP process for summarizing multiple utterances.

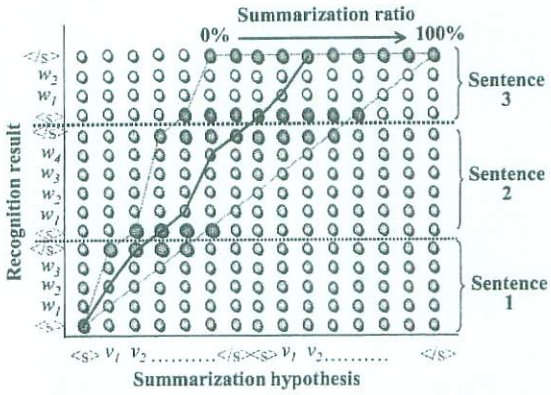

Figure 11. An example of DP process for summarization of multiple utterances.

4.3 Evaluation

(a) Word network of manual summarization results for evaluation

To automatically evaluate summarized sentences, correctly transcribed speech is manually summarized by human subjects and compared to the automatically summarized sentences. The manual summarization results are merged into a word network, and the word accuracy of automatic summarization is calculated using the word network. The network approximates all possible correct summarization including subjective variations. The word accuracy based on the word string that is the most similar to the automatic summarization result extracted from the word network, "summarization accuracy", is used to measure both linguistic correctness of the summarization and maintenance of original meanings of the utterance.

(b) Evaluation data

Since the recognition accuracy for the presentations is not yet high enough, Japanese TV broadcast news utterances recorded in 1996 were used to evaluate our proposed method. 50 utterances with word recognition accuracy above 90%, which was the average rate over the 50 utterances, were selected and used for the evaluation. The summarization ratio was set to 70%. In addition, 5 news articles consisting of 5 sentences each were summarized using the summarization technique for multiple utterances at 30% summarization ratio.

(c) Training data for summarization models

Broadcast-news manuscripts recorded from August 1992 to May 1996, comprising of approximately

500k sentences with 22M words, were used both for building a language model in speech recognition and calculating the word significance measure for summarization.

A trigram language model for summarization was built using texts from the Mainichi newspaper published from 1996 to 1998, comprising of 5.1M sentences with 87M words. The newspaper text is usually more compact and simpler than broadcast news text and therefore more appropriate for building language models for summarization. Our previous experiments confirmed that the automatically summarized sentences using word trigram based on newspaper text were much better than those based on broadcast news manuscripts [5].

SDCFG for word concatenation score was built using text from the manually parsed corpus of the Mainichi newspaper published from 1996 to 1998, comprising of approximately 4M sentences with 68M words. The number of non-terminal symbols was 100.

(d) Evaluation results

Manual transcription (TRS) and automatic transcription (REC) were summarized. In the summarization of REC, the following score conditions were compared;

- Confidence score (CM)
- Significance score (SIG)
- Linguistic score (3gram)
- SIG + 3gram
- SIG + 3gram + CM
- SIG + 3gram + CM + SDCFG (Word concatenation score)

In the summarization of TRS, since there is no recognition error, the conditions including CM were not tried.

To set the upper limit of the automatic summarization, manual summarization by human subjects for manual transcription (TRS_SUB) was performed. The results were evaluated using all other manual summarization results as correct summarization. In addition, as the upper bound of automatic speech summarization for transcription including speech recognition errors, manual summarization of automatically transcribed utterances was also evaluated (REC_SUB). To ensure that our method is sound, we made randomly generated summarization sentences (RDM) according to the summarization ratio and compared them with those obtained by our proposed methods.

Figure 12 shows results of utterance summarization at 70% summarization ratio and Fig. 13

shows those of summarizing articles having multiple sentences at 30% summarization ratio. These results show that our proposed automatic speech summarization technique is significantly more effective than RDM. The better result using the word concatenation score compared with that without using the word confidence score (CM) shows that the summarization accuracy is improved by the confidence score. The method using the word concatenation score (SDCFG) can reduce meaning alteration compared to the methods that are not using this score.

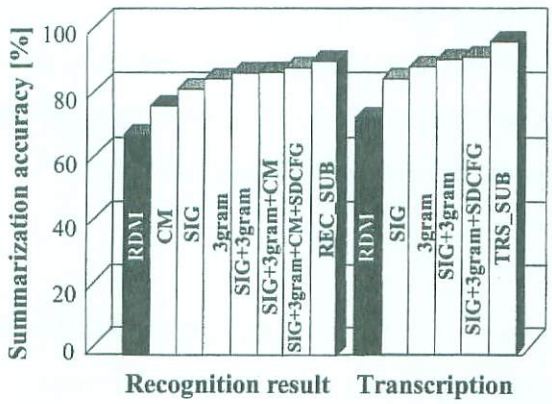

Figure 12. Each utterance summarizations at 70% summarization ratio

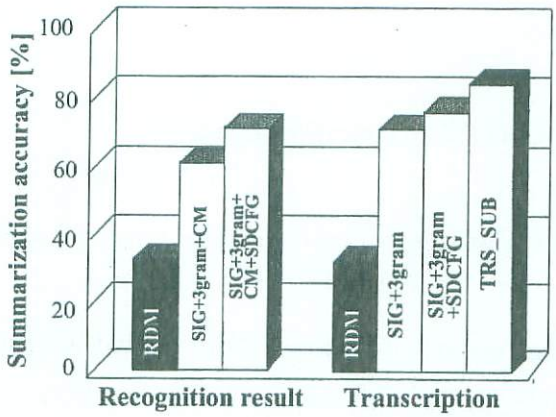

Figure 13. Article summarizations at 30% summarization ratio

4.4 Discussion

An automatic speech summarization method based on a word significance score, linguistic likelihood, a word confidence measure, and a word concatenation probability has been proposed. A

word set maximizing the total score was extracted by using the dynamic programming technique and connected to build a summarized sentence. We proposed a new method for measuring the summarization accuracy based on a word network constructed using manual summarization results.

Single utterance and multiple utterances with consistent meanings of Japanese broadcast news speech were summarized using the proposed method. Experimental results show that the proposed method can effectively extract relatively important information and remove redundant and irrelevant information. A confidence score giving a penalty for acoustically as well as linguistically unreliable words could reduce the meaning alteration of summarization caused by recognition errors. A word concatenation score giving a penalty for a concatenation between words with no dependency in the original sentence could also reduce the meaning alteration of the summarization.

In this study, newspaper texts were used for training linguistic models for summarization. If we can use a summarization model constructed by using a manual summarization corpus, the automatic summarization performance will be greatly improved. Our next step is to summarize presentations recorded in the national project. Future research includes task-dependent evaluation from the viewpoint of how much the original meaning is maintained in the summarization results based on the performance of information retrieval. Future research also includes applying our method to other languages such as English.

5. Conclusions

This paper first introduced a 5-year Japanese national project on spontaneous speech corpus and processing technology started in 1999, and then reported preliminary experimental results for recognizing and summarizing spontaneous speech performed at Tokyo Institute of Technology. The project is being conducted toward realizing three major themes: 1) building a large-scale spontaneous speech corpus, 2) acoustic and linguistic modeling for spontaneous speech understanding and summarization, and 3) constructing a prototype of a spontaneous speech summarization system.

The preliminary recognition experiments have been performed using 10 speakers' presentation utterances of approximately 4.5 hours. Recognition results show that acoustic and language modeling based on an actual spontaneous speech corpus is far more effective than conventional modeling based on read speech. The recognition accuracy has a wide speaker-to-speaker variability according to the speaking rate, the number of fillers, the number of repairs, etc. It was confirmed that unsupervised speaker adaptation of acoustic models was effective to improve the recognition

accuracy. Since the recognition accuracy for spontaneous speech is, however, still rather low, it is imperative to continue collecting a large corpus of spontaneous speech and use it for building language and acoustic models and challenging various research issues.

Future research issues include: a) how to transcribe spontaneous speech; b) how to apply morphological analysis to the transcribed spontaneous speech; c) how to build precise and yet general filled pause models; d) how to incorporate repairs, hesitations, repetitions, partial words, and disfluencies; e) how to adapt the language models to each task; and f) how to build acoustic models that fit spontaneous speech.

Indispensable in the processing of spontaneous speech will be a paradigm shift from speech recognition to understanding, where the underlying messages of the speaker, namely the meaning/context that the speaker intends to convey, are extracted instead of transcribing all of the spoken words. Speech summarization, which is one of the main targets of the national project, is considered to be one of the variations of fostering speech understanding. Speech summarization will also be applicable to a range of applications, such as preparing minutes of meetings, close captioning of broadcast news, and presenting information in news-on-demand systems.

In our proposed speech summarization method, a set of words maximizing a summarization score is extracted from automatically transcribed speech. This extraction is performed according to a target compression ratio using the dynamic programming technique. The extracted set of words is then connected to build a summarized sentence. The summarization score consists of a word significance measure, a confidence measure, linguistic likelihood, and a word concatenation probability which is determined by a dependency structure in the original speech. Japanese broadcast news speech transcribed using a LVCSR system was summarized and evaluated in comparison with manual summarization by human subjects. It was shown that the summarization method could effectively extract relatively important information and remove redundant and irrelevant information.

Our future research includes applying the summarization method to recognition results of presentations recorded in our national project. We are also planning to evaluate our summarization method by applying it to other languages such as English.

Acknowledgment

The authors wish to express their gratitude to the members of the "Spontaneous Speech" national project for constructing the corpus and for fruitful discussions. The authors wish to express their

thanks to Professor Tatsuya Kawahara at Kyoto University for his contribution to the national project and for several valuable comments and fruitful discussions related to the preliminary recognition experiments reported in this paper. The authors would also like to thank NHK (Japan Broadcasting Corporation) for providing us with the broadcast news database.

References

[1] S. Furui, K. Maekwa, H. Isahara, T. Shinozaki and T. Ohdaira, "Toward the realization of spontaneous speech recognition – Introduction of a Japanese Priority Program and preliminary results", Proc. ICSLP, Beijing, pp. 518-521 (2000).

[2] K. Maekawa, H. Koiso, S. Furui and H. Isahara, "Spontaneous speech corpus of Japanese," Proc. 2nd International Conference on Language Resources and Evaluation, Athens, Greece, pp. 947-952 (2000).

[3] A. Lee, T. Kawahara and S. Doshita, "An Efficient Two-pass Search Algorithm using Word Trellis Index", Proc. ICSLP, Australia, pp.1831-1834 (1998).

[4] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev and P. Woodland, "The HTK Book, Version 2.2", Entropic Ltd (1999).

[5] C. Hori and S. Furui, "Automatic speech summarization based on word significance and linguistic likelihood", Proc. ICASSP2000, Istanbul, pp.1579-1582 (2000).

[6] C. Hori and S. Furui, "Improvements in automatic speech summarization and evaluation methods", Proc. ICSLP2000, Beijing, pp.IV-326-329 (2000).

[7] A. Ito, C. Hori, M. Katoh and M. Kohda, "Language modeling by stochastic dependency grammar for Japanese speech recognition", Proc. ICSLP2000, Beijing, pp.I-246-249 (2000).