IMPROVEMENTS IN AUTOMATIC SPEECH SUMMARIZATION AND EVALUATION METHODS

Chiori Hori and Sadaoki Furui

Department of Computer Science
Tokyo Institute of Technology
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552 Japan
e-mail : {chiori,furui}@furui.cs.titech.ac.jp

ABSTRACT

This paper proposes an improved method of summarizing speech in which a confidence measure of a word hypothesis is incorporated in the summarization score and also proposes a new method for evaluating the summarized sentences. The automatically summarized sentences were evaluated based on the precision of extracted keywords and each word string with a certain length in the manual summarizations by human subjects. Japanese broadcast-news speech transcribed using a large-vocabulary continuous-speech recognition (LVCSR) system was summarized using our proposed method. Experimental results show that a confidence score giving a penalty for acoustically as well as linguistically unreliable hypotheses can reduce the meaning alteration of summarizations caused by recognition errors especially when the speech recognition rate is relatively low.

1. INTRODUCTION

Major applications of the LVCSR systems in the near future will include automatic closed captioning and meeting/conference summarization. Since transcribed speech obtained from LVCSR system usually includes some redundant information, a summarization technique is indispensable in order to index and make abstracts for automatically retrieval of speech data and for making closed captioning.

Therefore our future goal is to build a system being designed to extract and output information depending on the desired level of extraction from speech data with a single topic. For example, output can be as simple as keywords, summarization of each utterance for making a closed caption or creating an abstract from speech data.

In the present study, our target is to summarize Japanese broadcast-news speech sentence by sentence for captioning. In our previous speech summarization method, a set of words were extracted from an automatically transcribed sentence according to a target compression ratio so that the word concatenation includes topic words and maintains the meaning of the original utterance as much as possible using a dynamic programming (DP) technique[1].

This paper proposes an improved method of summarizing speech in which a confidence measure of a word hypothesis is incorporated in the summarization score. One of the major differences between text summarization and speech summarization exists in the fact that transcribed speech includes recognition errors. Thus the confidence measure is incorporated to avoid the meaning alteration caused by acoustically as well as linguistically unreliable hypotheses.

To evaluate the automatically summarized sentences, correctly transcribed speech are manually summarized by human subjects and used as correct targets. In consideration of the subjective variation, two measures are proposed as follows. One is the precision of extracted keywords and the other is the precision of each word string with a certain length in the automatically summarized sentences.

The improved method of automatically summarization is described in Section 2, the evaluation methods in Section 3 and the structure of Japanese broadcast-news transcription system in Section 4, and the experimental results are given in Section 5.

2. APPROACH TO SUMMARIZATION OF SPEECH USING A CONFIDENCE MEASURE

To summarize a sentence, we extract a limited number of relatively important words from each sentence so that the number of characters remains in a specified ratio range to the number of characters in the original sentence. The words are extracted using a summarization score consisting of three scores, i.e. a significance score I and a confidence score C of extracted each word, and a linguistic score L of the word concatenation. A set of words that maximizes the summarization score is selected using a dynamic programming (DP) technique. This method is effective in reducing the number of words without losing important information.

The summarization score is calculated as follows. Given a transcription result consisting of N words, $W = w_1, w_2, \ldots, w_N$, the summarization score of the extracted $M (M < N)$ words, $V = v_1, v_2, \ldots, v_M$, is obtained by

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

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

Significance score

A significance score I indicates the relative significance of each word in a sentence. In this paper a topic score based

on a kind of tf/idf measure is used as significance scores for nouns. A flat score is given to words other than nouns.

Linguistic score

A trigram probability $P(v_m|v_{m-2}v_{m-1})$ is used as a linguistic score $L\ (v_m)$.

Confidence score

A confidence score $C\ (v_m)$ is incorporated to give a penalty for acoustically as well as linguistically unreliable hypotheses. Specifically, posterior probability of each transcribed word, that is the ratio of the word hypothesis probability to that of all other hypotheses, is calculated using the word graph obtained by a decoder and used as a confidence measure [2][3].

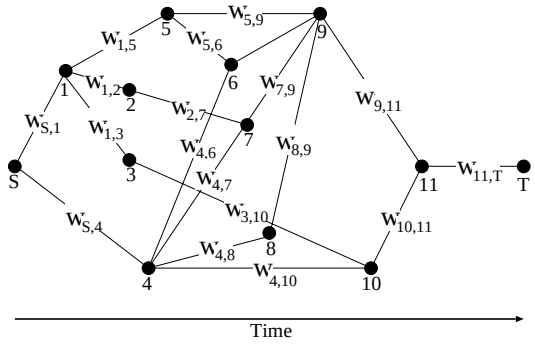

Figure 1: An example of word graph.

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

The posterior probability of a word hypothesis $w_{k,l}$ is given by

$$C(w_{k,l}) = \log \frac{\alpha_k P_{ac}(w_{k,l}) P_{lg}(w_{k,l}) \beta_l}{\mathcal{G}} \quad , \tag{2}$$

k,l	: node number in a word graph ($k < l$)
$w_{k,l}$	: word hypothesis occurred between node k and node l
$C(w_{k,l})$	: log of the posterior probability of $w_{k,l}$
α_k	: forward probability from the beginning node S to node k
β_l	: backward probability from node l to the end node T
$P_{ac}(w_{k,l})$	: acoustic likelihood of $w_{k,l}$
$P_{lg}(w_{k,l})$	: linguistic likelihood of $w_{k,l}$
$\mathcal{G}$	: forward probability from the beginning node S to the end node T

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). A set of words maximizing the summarization score is efficiently selected using a DP technique. Furthermore, in order to evaluate the summarization scores of the summarized sentences obtained from the same original sentence using the various summarizing ratios, a normalization factor was applied to the summarization score[1].

3. EVALUATION METHODS

In our previous experiments the summarization results were evaluated according to the performance of extracting "important words" selected by human subjects from the manual transcriptions and maintaining the meaning of the original speech[1].

In this study to evaluate the automatically summarized sentences, correctly transcribed speech is manually summarized by human subjects according to the target summarization ratio and used as correct targets. In consideration of the subjective variation, the precision of extracted keywords and that of each word string with a certain length in the manual summarizations by human subjects were evaluated as follows.

3.1. Precision of keywords

The precision of extracted keywords corresponds to the coverage of the core information. This measure is calculated as the mean of word significance values defined as percentages of subjects who have selected the words as keywords. The precision of keywords R of the summarization $V = v_1, v_2, \ldots, v_M$ is given as follows.

$$R = \frac{\sum_{m=1}^{M} \frac{c(v_m)}{a}}{M} \quad , \tag{3}$$

a	: number of subjects to make manual summarizations
M	: total number of words in a summarized sentence
v_m	: m-th word in a summarized sentence
$c(v_m)$	: number of subjects extracting v_m

3.2. Precision of word strings

To evaluate linguistic correctness and maintenance of the original meanings of the utterance, the precision of each word string with a certain length in the automatically summarized sentences is defined as how many such word strings are included in at least one of the manual summarizations by human subjects.

The extraction ratio WS_D of each word strings consisting of D words in a summarized sentence $V = v_1, v_2, \ldots, v_M$ is given by

$$WS_D = \frac{\sum_{m=D}^{M} \delta(v_{m-D+1}, \ldots, v_{m-1}, v_m)}{M - D + 1} \quad , \tag{4}$$

where

$$\delta(u_D) = \begin{cases} 1 & if \quad u_D \in U_D \\ 0 & if \quad u_D \notin U_D \end{cases} \tag{5}$$

u_D : each word string consisting of D words
U_D : a set of word strings consisting of D words in all manual summarizations

Note that words occurred in the different location in an original sentence are considered to be a different word even though they are the same words. When D is 1, WS_D indicates the precision of each word and when D is the length of a summarized sentence M, WS_D indicates the precision of the summarized sentence itself.

4. STRUCTURE OF THE BROADCAST NEWS TRANSCRIPTION SYSTEM

4.1. Acoustic models

The acoustic model involved in the sharable software repository for Japanese large vocabulary continuous speech recognition by IPA was used[4]. The feature vector extracted from speech consists of 12 MFCCs, the delta of their features and the delta of normalized logarithmic power (derivatives). The total number of parameters in each vector is 25. MFCCs were normalized using the CMS (cepstral mean subtraction) method. The phone models used were shared-state triphone HMMs designed using tree-based clustering. The total number of states was 1012, and the number of Gaussian mixture components per state was 8. They were trained using speech reading newspaper by 100 speakers.

4.2. Language models

Broadcast-news manuscripts recorded from August 1992 to May 1996, comprising of approximately 500k sentences consisting of 22M words, were used for constructing language models. The vocabulary size is 20k words.

4.3. Decoder

We used a word-graph-based 2-pass decoder for transcription. In the first pass, frame-synchronous beam search was performed using the above-mentioned HMMs and a bigram language model. A word graph was generated as a result of the first pass. In the second pass, the word graph was rescored using a trigram language model to derive the final transcription that was then used for summarization.

5. EVALUATION EXPERIMENTS

5.1. Evaluation data

Japanese news speech data broadcast on TV in 1996 was used as a test set to evaluate our proposed method. The set consisted of 419 utterances by a female anchor speaker, and was manually segmented into sentences. The out-of-vocabulary (OOV) rate for the 20k word vocabulary is 2.5% and the perplexity for the test set was 54.5. 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, the ratio of the number of characters in the summarized sentences to that in the original sentences, was set to 20, 40, 60, 70 and 80%.

5.2. Language models for summarized sentences

A trigram language model for summarization was built using text from Mainichi newspaper published from 1996 to 1998, comprising of approximately 5.1M sentences consisting of 87M words. We did this because we consider newspaper text to usually be more compact and simpler than broadcast news text and therefore more appropriate for building language models for summarization than broadcast news text. In our previous experiments the automatically summarized sentences using word trigram constructed by newspaper text were much better than those by broadcast-news manuscripts.

5.3. Evaluation results

The automatically summarized utterances without incorporating the confidence measure (REC), that with the confidence measure (CM), and the summarized transcriptions by human (TRS) were evaluated. To set a goal of the automatic summarized sentences, each manual summarization by 25 human subjects (SUB) were evaluated based on the manual summarizations by 24 human subjects except for the evaluated summarization itself. To insure that our method is sound, we considered randomly generated summarizations according to the summarization ratio (RDM) to compare the precisions with those achieved by our proposed methods.

The evaluation results by precision of keywords is shown in Table 1. The better result of CM than that of REC shows the precision of keywords of the summarized speech improved significantly using the confidence score. The difference between TRS and CM is that some of important words in TRS were not included in CM.

Table 1: Evaluation results by precision of keywords

Target ratio	20%	40%	60%	70%	80%
RDM	0.17	0.35	0.54	0.66	0.75
REC	0.20	0.38	0.58	0.69	0.77
CM	0.27	0.41	0.57	0.68	0.75
TRS	0.31	0.43	0.60	0.71	0.78
SUB	0.45	0.56	0.68	0.75	0.80

The evaluation results by precision of word strings in the condition of 70% summarization ratio is shown in Fig.2. The concatenation of words is more constrained when word strings were evaluated on longer word strings. Therefore the word string precisions of all types of summarizations decrease gradually with the length of word strings. The automatic summarization of TRS, REC and CM can maintain the correct word concatenation more often than RDM. The quick decrease of RDM precision indicates the word concatenations of RDM are grammatically and semantically incorrect. However the results of the automatic summarizations cannot reach the performance level of the goal

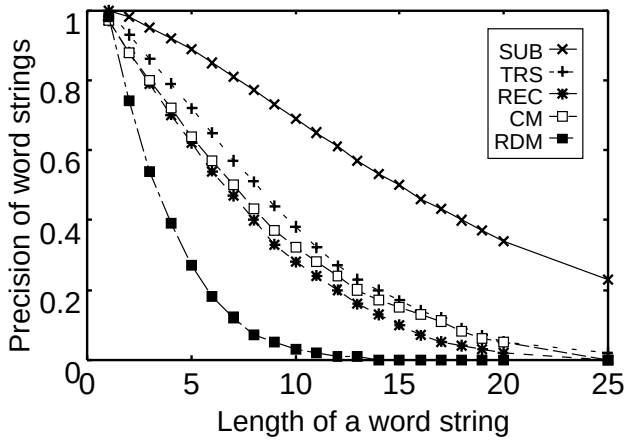

Figure 2: Precision of word strings vs. length of a word string at 70% summarization ratio.

of SUB. The same types of results were also shown in all of the summarizations using the various summarization ratios.

The word string precisions in the condition of 20% summarization ratio is shown in Fig.3. In order to evaluate the efficiency of the confidence score for the recognition results with lower accuracy, 10 utterances with relatively lower accuracy in the test set (called "difficult test set") were separately evaluated. The upper figure shows all the test set and the lower one shows difficult test set results. The results show that the confidence measure can improve the automatic summarization especially when the speech recognition rate is relatively low.

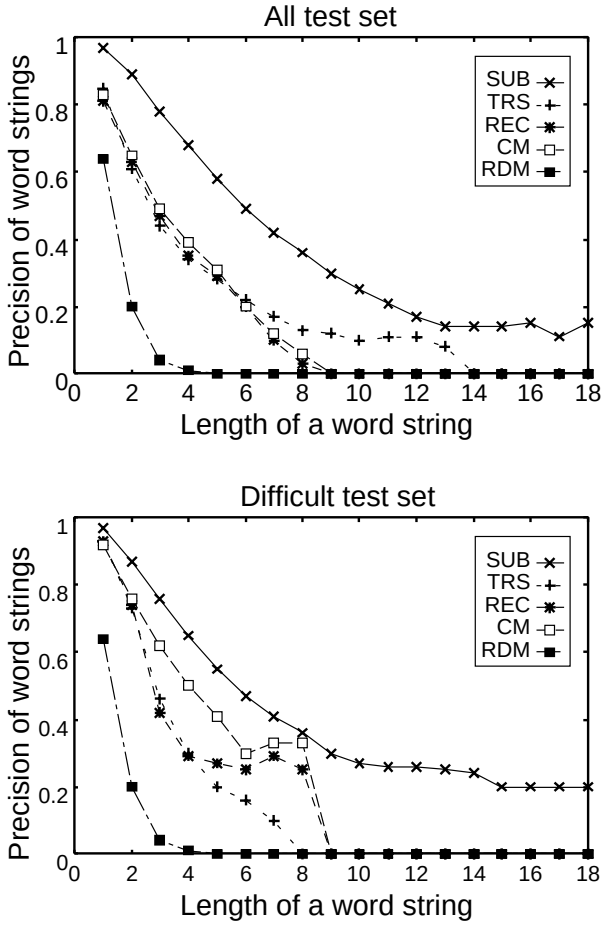

Figure 3: Precision of word strings vs. length of a word string at 20% summarization ratio.

6. CONCLUSIONS

An improved method of automatically summarizing broadcast news speech based on topic words, linguistic likelihood and a confidence measure, facilitated by a dynamic programming technique, has been proposed. In addition to evaluate the automatically summarized sentences, we proposed two measures, i.e. the precision of extracted keywords and that of each word string with a certain length in the automatically summarized sentences using the manual summarizations by human subjects.

Experimental results show that a confidence score giving a penalty for acoustically as well as linguistically unreliable hypotheses can reduce the meaning alteration of summarizations caused by recognition errors especially when the speech recognition rate is relatively low. These results clearly show the effectiveness of the proposed summarization.

Future research includes summarization of a set of sentences with one topic and further making abstracts of monologues such as lectures.

REFERENCES

[1] C. Hori and S. Furui, "Automatic speech summarization based on word significance and linguistic likelihood" Proc. ICASSP2000, Istanbul, Vol.3, pp.1579-1582(2000).

[2] T. Kemp and T. Schaaf, "Estimating confidence using word lattices" Proc. 5th Eurospeech, Rhodes, Vol.2, pp.827-830 (1997).

[3] 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).

[4] T. Kawahara et al., "Sharable software repository for Japanese large vocabulary continuous speech recognition", Proc. ICSLP, pp.3257-3260 (1998).