# Speech Summarization using Weighted Finite-State Transducers

*Takaaki Hori, Chiori Hori, and Yasuhiro Minami*

Speech Open Laboratory
NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan
{hori,chiori,minami}@cslab.kecl.ntt.co.jp

## Abstract

This paper proposes an integrated framework to summarize spontaneous speech into written-style compact sentences. Most current speech recognition systems attempt to transcribe whole spoken words correctly. However, recognition results of spontaneous speech are usually difficult to understand, even if the recognition is perfect, because spontaneous speech includes redundant information, and its style is different to that of written sentences. In particular, the style of spoken Japanese is very different to that of the written language. Therefore, techniques to summarize recognition results into readable and compact sentences are indispensable for generating captions or minutes from speech. Our speech summarization includes speech recognition, paraphrasing, and sentence compaction, which are integrated in a single Weighted Finite-State Transducer (WFST). This approach enables the decoder to employ all the knowledge sources in a one-pass search strategy and reduces the search errors, since all the constraints of the models are used from the beginning of the search. We conducted experiments on a 20k-word Japanese lecture speech recognition and summarization task. Our approach yielded improvements in both recognition accuracy and summarization accuracy compared with other approaches that perform speech recognition and summarization separately.

## 1. Introduction

In the past decade, techniques that enable large-vocabulary continuous-speech recognition have been intensively investigated, and they have achieved more than 90% word accuracy for read speech. Currently, spontaneous speech recognition is being investigated as the next target[1][2]. Although state-of-the-art speech recognizers have not yet achieved sufficient accuracy for spontaneous speech, the technique is being expected to be applied to automatic generation of captions, minutes, and so on. However, there is another problem besides the insufficient accuracy. That is, the recognition result is usually difficult to understand, even if the recognition system yields 100% accuracy, because spontaneous speech includes redundant information such as disfluencies, filled pauses, repetitions, repairs, and word fragments. Therefore, speech summarization techniques are required to generate readable and compact sentences from spontaneous speech. In this paper, we define speech summarization as a translation from speech signals to compact written-style sentences, which includes speech recognition, paraphrasing, and compaction.

We have already proposed a speech paraphrasing method using Weighted Finite-State Transducers (WFSTs)[3]. The method translates spontaneous speech into written-style sentences. Such techniques are indispensable in Japanese, because the written style is preferred to the spoken style when making captions or minutes. The style of spoken Japanese is very different from that of the written Japanese in comparison with English. In this paper, we extend our paraphrasing method to speech summarization. Some techniques for speech summarization have been proposed [4]. In [5], dynamic programming was applied to produce understandable summarized sentences by extracting relatively important words with high heuristic likelihood while excluding redundant and irrelevant information. This method can effectively compact transcribed sentences. However, since the summarization is performed as a post-processing of speech recognition, it tends to suffer from recognition errors, and it requires a delay after recognition.

Unlike such methods, our approach is an integrated processing of speech recognition, paraphrasing, and compaction using a single WFST. The WFST is generated by combining WFSTs for those three processes. This framework has two advantages over the separated implementation consisting of speech recognition and the succeeding text processing: One is that the target sentences can be derived almost simultaneously while a human speaks, because the speech can be directly translated into the target sentences frame by frame using a Viterbi search for the integrated network. Therefore, this framework is more effective for on-line applications. The other is that speech recognition accuracy can improve by integrating all the knowledge sources into one single network, meaning that the decoder can choose the best hypothesis under all the constraints. Furthermore, search errors can be reduced by using all the knowledge sources from the beginning of the search, because it is possible to judge whether each hypothesis is promising or not in the early stages of the search. This framework is more effective, especially when recognizing speech with a specific style and translating it into the corresponding general-style sentences, because it is usually difficult to estimate a good language model for a specific style if few data are present, whereas it is relatively easy for the general style. In general, it is difficult to prepare a large corpus of spontaneous speech transcriptions, whereas it is relatively easy to prepare one for written documents. Thus, this framework is suitable for translations such as spontaneous speech to written-style text.

We conducted experiments on a 20k-word Japanese lecture speech recognition and summarization task. We present the evaluation results and state our conclusions.

## 2. Speech Summarization using WFSTs

We built a spontaneous speech summarization system. This system searches the best summarized result for a given speech input using a one-pass Viterbi algorithm while performing speech

recognition, paraphrasing, and compaction all at once.

### 2.1. Speech Recognition

Continuous speech recognition can be formulated as a problem to find a word sequence $\hat{W}$, such that

$$\hat{W} = \underset{W}{\text{argmax}}\, P(W|O) \quad (1)$$

$$= \underset{W}{\text{argmax}}\, P(O|W)P(W), \quad (2)$$

where $P(O|W)$ is an acoustic probability of speech input $O$ given a word sequence $W$ and $P(W)$ is the language probability of $W$. To estimate these probabilities, a general speech recognition system has phonetic, acoustic and linguistic knowledge sources, which are a pronunciation lexicon, an acoustic model, and a language model, respectively. A speech recognition decoder finds the most likely hypothesis for the input while inquiring such knowledge sources.

Recently, the WFST approach has become a promising alternative formulation to the traditional decoding approach, which offers a unified framework representing various knowledge sources and producing the full search network optimized up to the HMM states [6][7].

WFSTs are finite state networks associating input and output symbols on each arc, which can be weighted with a log probability value. They can represent all of the above mentioned knowledge sources for speech recognition. Furthermore, WFSTs can be combined by using the composition operator, leading to the integration of the underlying knowledge sources into a single input-output relation. An integrated WFST for speech recognition can be composed as

$$R = H \circ C \circ L \circ G, \quad (3)$$

where $H$, $C$, $L$, and $G$ are, for example, a state network of triphone HMMs, a set of connection rules for triphones, a pronunciation lexicon, and a trigram language model, respectively. Here, "$\circ$" represents the composition operator. As a result, decoding with $R$ becomes a one-pass search process using cross-word triphones and trigrams. Once the network is further optimized by proceeding to weighted determinization and minimization, the search efficiency dramatically increases.

### 2.2. Speech Paraphrasing

Paraphrasing can be considered as a kind of machine translation. We formulate the speech paraphrasing as a speech-input machine translation[8][9], where the source language corresponds to spontaneous speech and the target language corresponds to written-style sentences.

The translation of a source language $W$ to a target language can be formulated as the search for a word sequence $\hat{T}$ from a target language, such that

$$\hat{T} = \underset{T}{\text{argmax}}\, P(T|W) \quad (4)$$

$$= \underset{T}{\text{argmax}}\, P(W|T)P(T). \quad (5)$$

If the source language is speech $O$, i.e. speech-input case, the translation can be formulated as the search for $\hat{T}$, such that

$$\hat{T} = \underset{T}{\text{argmax}}\, P(T|O) \quad (6)$$

$$= \underset{T}{\text{argmax}} \sum_{W} P(O|W)P(W|T)P(T) \quad (7)$$

$$\simeq \underset{T}{\text{argmax}}\, \underset{W}{\max}\, P(O|W)P(W|T)P(T). \quad (8)$$

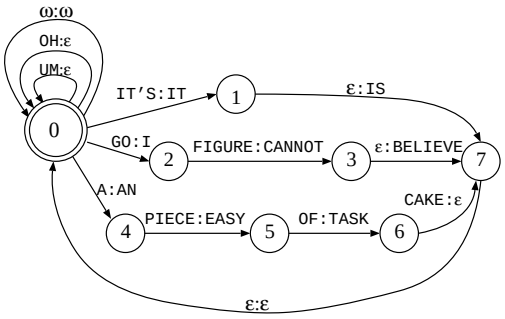


Figure 1: An example of a substitution WFST

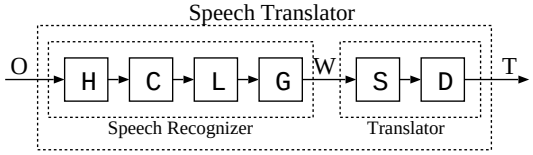


Figure 2: Cascade of speech-input machine translation

Some approximations have been proposed to determine the translation probability $P(W|T)$. In this paper, we assume

$$P(W|T) \approx P_G(W)\delta_S(W,T), \quad (9)$$

where $P_G(W)$ is a prior probability of $W$, given by a language model for speech recognition, and $\delta_S(W,T)$ takes binary 0 or 1 values depending on whether it is possible to substitute $W$ with $T$, which is given by a set of substitution rules of word sequences.

The substitution function $\delta_S(W,T)$ can be expressed as a WFST, and an example of a substitution WFST is illustrated in Fig. 1. In the figure, the symbol pair on each arc represents one word (the left-hand side of ':') substituted with the other (the right-hand side of ':') except for "$\omega$:$\omega$," which indicates any word can be substituted with the word itself. The WFST, for example, can substitute a sentence:
"OH, GO FIGURE! IT'S A PIECE OF CAKE,"
with another sentence:
"I CANNOT BELIEVE IT IS AN EASY TASK."

Let $S$ be a WFST of $\delta_S(W,T)$, and $D$ be a WFST of a language model of the target language. The integrated WFST for speech translation can be composed as

$$Z = H \circ C \circ L \circ G \circ S \circ D. \quad (10)$$

The cascade in Fig. 2 illustrates the process of speech translation. Each WFST in the cascade can be optimized, and the resulting WFST in each composition step can also be optimized using weighted determinization and minimization.

### 2.3. Sentence Compaction

To incorporate a sentence compaction mechanism into the integrated WFST $Z$ in Eq. 10, we extend the lexicon transducer $L$. First, we connect a wildcard transducer which accepts an arbitrary phone sequence. The resulting lexicon transducer is illustrated in Fig. 3. The searched path will detour to the wildcard when unreliable utterances are inputted, such as out-of-

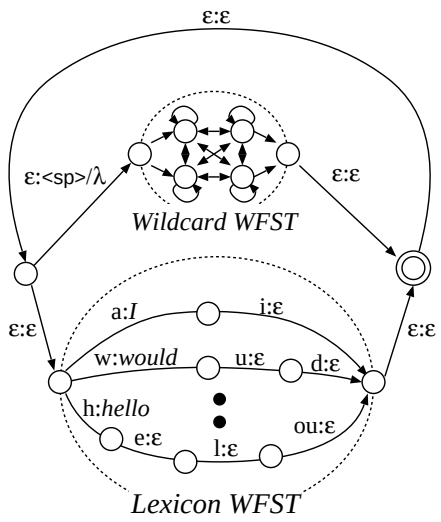


Figure 3: Extended lexicon transducer

vocabulary words, disfluencies, filled pauses, repetitions, repairs, and word fragments. The wildcard transducer outputs a symbol "<sp>," which indicates a phrase boundary. This symbol is introduced to avoid irrelevant connections of words before and after the skipped words (the wildcard). The connections are restricted by the language model $G$, so that each connected point becomes a phrase boundary. However, the language model $G$ needs to be estimated with a modified text corpus including "<sp>."

Secondly, we assign a weight indicating a significance of each word to the first arc of the word in $L$ to extract more important words. We use IDF (Inverse Document Frequency) as the significance measure.

We can control the summarization ratio by changing the penetration weight $\lambda$, which indicates a special weight associated with the arc at the entrance of the wildcard transducer. The summarization ratio is defined as:

$$\text{Summarization Ratio} = \frac{\text{Number of extracted words}}{\text{Number of spoken words}}. \quad (11)$$

If $\lambda$ is large, the summarization ratio will be large, whereas, if $\lambda$ is small, the ratio will be small.

# 3. Experiments

## 3.1. Conditions

We evaluated our summarization system in a 20k-word spontaneous speech recognition and summarization task. The task is based on a corpus of Japanese spontaneous speech [2], mainly consisting of monologues such as lectures, presentations, and news commentaries.

The target topic was limited to lectures in academic fields. Three types of text corpora were prepared for the topic, which were spoken, written, and parallel. The spoken corpus consists of manual transcriptions of 680 lectures, which written corpus consists of newspaper text from one year, World Wide Web (WWW) text, and automatically translated text of the manual transcriptions. The parallel corpus consists of a subset of the manual transcriptions (six lectures) and its manually paraphrased text rendered into written language by a human subject.

The automatically translated text was generated from the manual transcriptions using the WFST $S \circ D'$, where $S$ was constructed with the substitution rules extracted from the parallel corpus, and $D'$ was the trigram language model trained with only the newspaper text and the WWW text. The corpora are summarized in Table 1.

Table 1: Text corpora for experiments

| type | text set | #words | purpose | |
|---|---|---|---|---|
| spoken | Manual transcription | 2 M | $G$ | |
| written | Newspaper | 35 M | $D'$ | $D$ |
| | WWW | 1.8 M | | |
| | Auto-translation | 1.9 M | | |
| parallel | Spoken-written parallel text | 30K | $S$ | |

The speeches were digitized with 16-kHz sampling and 16-bit quantization. Feature vectors had 25 elements consisting of 12 MFCC, their delta, and delta log energy. Tied-state triphone HMMs with 3,000 states and 16 Gaussians per state were made by using 338 lectures in the corpus uttered by male speakers (approximately 59 hours). Decoding was performed by a one-pass Viterbi search for WFSTs [7].

## 3.2. Evaluation method for speech summarization

To automatically evaluate summarized sentences, correctly transcribed speech is manually summarized by human subjects and used as the correct target. The manual summarization results are merged into a word network that approximately expresses all possible correct summarization, including subjective variations. The summarization accuracy of automatic summarization is calculated using the word network [4]. The word string that is the most similar to the automatic summarization result extracted from the word network is considered as a correct answer for the automatic summarization. The similarity is measured based on the word accuracy. The best accuracy, comparing the summarized sentence with the set of words extracted from the network, is used as a measure of linguistic correctness and maintenance of the original meanings of the utterance (summarization accuracy).

We excluded four lectures from training in order to use them for evaluation, and these are not included in the spoken corpus. To make the word network, the transcriptions of the test lectures are first translated into written-style text, and then summarized by 9 human subjects.

## 3.3. Experimental results

Table 2 shows word accuracy in speech recognition for each lecture, where A01M0007, A01M0035, A01M0074, and A05M0031 represent lecture IDs, and their lengths are 30, 28, 12, and 27 minutes, respectively. In the table, "*baseline*" indicates results for the WFST $R$ (Eq. 3), while "*integrated*" indicates results for the WFST $Z$ (Eq. 10). In every lecture, the *integrated* method yielded higher accuracies than those of the baseline method.

In the *integrated* mode, speech recognition results are not observed because the WFST $Z$ does not output recognized hypotheses, but written-style sentences. However, we can easily obtain recognition results by using the following WFST instead.

$$Z' = H \circ C \circ L \circ \text{proj}(G \circ S \circ D), \quad (12)$$

Table 2: Word accuracy [%] in speech recognition

| lecture ID | baseline | integrated |
|---|---|---|
| A01M0007 | 71.8 | 73.3 |
| A01M0035 | 60.0 | 60.9 |
| A01M0074 | 71.8 | 72.8 |
| A05M0031 | 74.6 | 75.8 |
| Ave. | 68.6 | 69.9 |

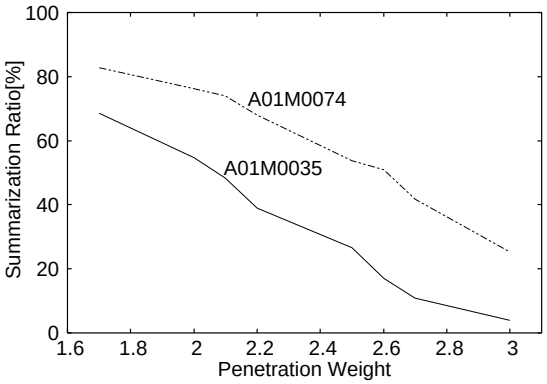


Figure 4: Summarization ratio

where "proj" indicates the projection operator of a WFST to a WFSA (Weighted Finite-State Acceptor). In our work, the operation simply substitutes the output symbol of each arc with its input symbol. In this first evaluation, the lexicon transducer $L$ did not include a wildcard; that is, the sentence compaction had not yet been performed.

We then investigated the changes of the summarization ratio when the penetration weight $\lambda$ varied. Fig. 4 shows relationships between the penetration weight and the summarization ratio for two lectures. These results mean that the summarization ratio can be controlled by the penetration weight, although the relationship changes for each lecture. The property seems to depend on recognition accuracy. Further investigation is necessary for confirmation.

Finally we compared our integrated summarization with the post-processing method [5]. In the post-processing mode, the speech signal was recognized using the transducer $R$, and the recognition result was paraphrased into written-style text using $S \circ D$. The resulting text was summarized using the Dynamic Programming technique according to the word significance score and the linguistic likelihood. Dependency structures of the original sentences were not considered, an idea introduced in [5]. Table 3 shows the summarization accuracies for the two lectures. In every lecture and summarization ratio, the integrated method yielded higher accuracies than those of the post-processing method. Accordingly, we can suppose that the improvement was yielded from the improvement in speech recognition. Thus, it is shown that our integrated approach reduces recognition errors and also improves the performance of speech summarization.

## 4. Conclusions

We proposed a spontaneous speech summarization system based on Weighted Finite-State Transducers (WFSTs). This system translates spontaneous speech directly into written-style compact sentences using a single WFST built by combining WFSTs for speech recognition, paraphrasing, and compaction. Unlike the post-processing method, the integrated method allows the incorporation of knowledge about the paraphrasing to improve speech recognition.

We conducted experiments on a 20k-word Japanese spontaneous speech recognition and summarization task. Our approach improved accuracies in speech recognition and summarization. The improvements were not significant; however, this approach has the potential to yield further improvements by incorporating better translation models.

Table 3: Summarization Accuracy [%]

| | Summarization Ratio | | | |
|---|---|---|---|---|
| | 50% | | 70% | |
| lecture ID | post-process | integrated | post-process | integrated |
| A01M0035 | 22.9 | 25.7 | 35.4 | 35.9 |
| A01M0074 | 35.7 | 39.8 | 54.2 | 55.9 |

## 5. Acknowledgement

We thank the Japanese Science and Technology Agency Priority Program, "Spontaneous Speech: Corpus and Processing Technology," for providing speech data and transcriptions.

## 6. References

[1] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," Proc. of ICASSP'92, vol. 1, pp. 517–520, 1992.

[2] S. Furui, K. Maekawa, and H. Isahara, "A Japanese national project on spontaneous speech corpus and processing technology," Proc. of ASR2000, pp. 244–248, 2000.

[3] T. Hori, D. Willett, and Y. Minami, "Paraphrasing spontaneous speech using weighted finite-state transducers," To appear in Proc. of Spontaneous Speech Processing and Recognition 2003.

[4] C. Hori, "A study on statistical methods for automatic speech summarization," Doctoral dissertation, Tokyo Institute of Technology, 2002.

[5] C. Hori, and S. Furui, "Automatic speech summarization applied to English broadcast news speech," Proc. of ICASSP2002, vol. 1, pp. 9–12, 2002.

[6] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," Proc. of ASR2000, pp. 97–106, 2000.

[7] D. Willett, E. McDermott, Y. Minami, and S. Katagiri, "Time and memory efficient Viterbi decoding for LVCSR using a precompiled search network," Proc. of Eurospeech 2001, vol. 2, pp. 847–850, 2001.

[8] F. Casacuberta, "Finite-state transducers for speech-input translation," Proc. of ASRU 2001.

[9] S. Bangalore and G. Riccardi, "A finite-state approach to machine translation," Proc. of ASRU 2001.