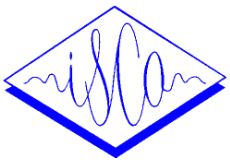

ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13–16, 2003

PARAPHRASING SPONTANEOUS SPEECH USING WEIGHTED FINITE-STATE TRANSDUCERS

Takaaki Hori, Daniel Willett, and Yasuhiro Minami*

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

ABSTRACT

This paper describes an integrated framework to paraphrase spontaneous speech into written-style sentences. Most current speech recognition systems try 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 it has a different style from that of written sentences. Especially, the style of spoken Japanese is much different from that of written one. Therefore, techniques to paraphrase recognition results are indispensable for generating captions or minutes from speech. To realize efficient speech paraphrasing, we attempt to translate spontaneous speech directly into written-style sentences using a Weighted Finite-State Transducer (WFST). This approach enables to use all the knowledge sources in a one-pass search strategy and reduces the search error, since the constraint of the paraphrasing model is used from the beginning of the search. We conducted experiments on a 20k-word Japanese lecture speech recognition and paraphrasing task. Our approach yielded improvements on both recognition accuracy and paraphrasing accuracy compared with other approaches that deal with speech recognition and paraphrasing performed separately.

1. INTRODUCTION

In the past decade, techniques for 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]. However, recognition results of spontaneous speech is usually difficult to read, even if the recognition system yields 100% accuracy, because spontaneous speech includes redundant information such as disfluencies, filled pauses, repetitions, repairs, and word fragments. Furthermore, spontaneous speech has a different style from that of written sentences. The style of spoken Japanese is much different from that of written one in comparison with English. Many verbs, auxiliary verbs, adjectives etc. in spoken sentences change into other words in written sentences. In Japanese, the written style is generally preferred to the spoken style when making captions or minutes from speech. Therefore, it is necessary to paraphrase transcribed speech into readable written-style sentences. We believe that automatic paraphrasing of spoken language will become an important technique for automatic generation of captions, minutes, annotations, and so on.

In this paper, we propose a method to generate readable sentences from spontaneous speech. Our approach is a simultaneous translation of speech into target-style sentences using a single Weighted Finite-State Transducer (WFST) generated by composing two WFSTs which are one for speech recognition and the other for paraphrasing. This framework has two advantages over the separated implementation consisting of speech recognition and the succeeding translation: 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 time-synchronous 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 knowledge sources for translation, since they reinforce linguistic constraints in speech recognition. This 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 due to little data, whereas it is relatively easy for the general style. Thus, this framework is suitable for translations such as spontaneous speech to written-style language, dialectal speech to standard language, and so on.

In this paper, we apply this framework to a simultaneous translation of spontaneous Japanese speech into written

*Currently the author works for TEMIC Speech Processing.

formal Japanese text. In Japanese, the gap between spontaneous speech and written sentence styles is much greater than that of English. In addition, it is difficult to prepare a large corpus of spontaneous speech transcriptions, whereas it is relatively easy to prepare one for the written documents.

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

2. WEIGHTED FINITE-STATE TRANSDUCERS FOR SPEECH RECOGNITION

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

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

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

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 [3][4].

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.

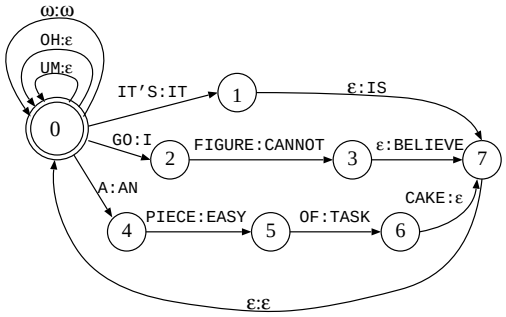

Fig. 1. An example of a substitution WFST

3. SPEECH PARAPHRASING WITH WFSTS

Paraphrasing can be considered as a kind of machine translation. We formulate the speech paraphrasing as a speech-input machine translation[5][6], 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

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

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

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

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

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

$$\simeq \underset{T}{\operatorname{argmax}} \max_{W} P(O|W)P(W|T)P(T). \quad (8)$$

For the translation probability $P(W|T)$, some approximations have been proposed. 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 of ':') substituted

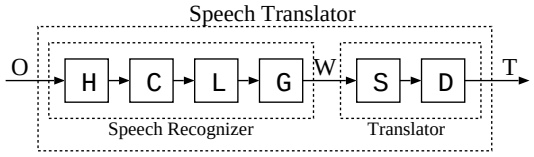

Fig. 2. Cascade of speech-input machine translation

with the other (the right hand of ':') except for "ω:ω" which means any word 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

$$Z = H \circ C \circ L \circ G \circ S \circ D. \tag{10}$$

The process of the speech translation is illustrated by the cascade in Fig. 2. 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.

According to this formulation, we built a spontaneous speech paraphrasing system. This system searches the best paraphrased result for a given speech input using the one-pass Viterbi algorithm, allowing speech recognition and paraphrasing to be performed in parallel with speaking.

4. EXPERIMENTS

4.1. Conditions

We evaluated our paraphrasing system in a 20k-word spontaneous speech recognition and paraphrasing 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 corpora were prepared for the topic, which were spoken, written, and parallel. The spoken corpus consists of manual transcriptions of 680 lectures. The 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.

As shown in the table, the written text is mostly from the newspaper corpus, which does not include many academic articles, and has a large variety of topics. Hence, it is much broader than the spoken language corpus, which is very topic-focused.

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 a WFST integrating cross-word triphone HMMs and trigrams [4].

We excluded four lectures from training in order to use them for evaluation, which are not included in the spoken corpus. We used scaling factors when composing WFSTs for the source and target language models, and these factors were roughly tuned to minimize the word error rate for each tested lecture.

4.2. Experimental results

We calculated word error rates for speech recognition and paraphrasing in cases of using *separated* and *integrated* WFSTs. Here, "*separated*" means paraphrasing after speech recognition using $H \circ C \circ L \circ G$ and $S \circ D$ separately. The term "*integrated*" means simultaneous translation using a single integrated WFST, Z.

In the *integrated* mode, speech recognition results cannot be observed because output symbols corresponding to recognized hypotheses are lost in the composition process. To evaluate the recognition performance, we used the following WFST,

$$Z' = H \circ C \circ L \circ \text{proj}(G \circ S \circ D), \tag{11}$$

where "proj" indicates the projection operator of a WFST to a WFSA (Weighted Finite-State Acceptor). In our work, the

Table 2. Word error rate [%] in speech recognition

	separated	integrated	reduction[%]
A01M0007	28.2	26.7	5.3
A01M0035	40.0	39.1	2.2
A01M0074	28.2	27.2	3.5
A05M0031	25.4	24.2	4.7
Ave.	31.4	30.1	4.1

Table 3. Word error rate [%] in paraphrasing

	text-input	separated	integrated	reduction
A01M0007	16.8	34.9	33.1	5.2
A01M0035	26.9	52.7	51.5	2.2
A01M0074	15.8	33.1	32.0	3.3
A05M0031	14.2	30.9	29.6	4.2
Ave.	19.3	39.5	38.2	3.2

operation simply substitutes the output symbol of each arc with its input symbol. By decoding with Z', we can obtain the recognition result.

Table 2 shows the word error rate 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 every lecture, the integrated method yielded lower error rates than those of the separated method. The reduction rate varied with the lectures. The reduction in A01M0035 was smaller than that in the other lectures. We conclude that the speaker of A01M0035 had such a high degree of spontaneity that the WFST for paraphrasing did not effectively translate his/her speech.

Table 3 shows the word error rate for each paraphrased lecture, which is calculated by comparing it with the manually paraphrased result made by the same human subject as the one who made the parallel text. For reference, we also attached the error rates in text-input case (i.e. paraphrasing for transcriptions). The reduction rate achieved by the integrated method was comparable with that in the case of speech recognition. Accordingly, we can guess that the reduction was yielded from the improvement in speech recognition. Thus, it is shown that the integrated approach reduces recognition errors and also improves the performance of speech paraphrasing. However, the error rate was higher than that in speech recognition. The difference is caused by the fact that the error rate is not zero in the text-input case, because we used only one sequence as the correctly-paraphrased sentences, though a few expressions should be allowed for the same meaning.

5. CONCLUSIONS

We proposed a spontaneous speech paraphrasing system based on Weighted Finite-State Transducers (WFSTs). This system translates spontaneous speech directly into written-style sentences using a single WFST built by composing two WFSTs, which are one for speech recognition and the other for paraphrasing. Unlike the separated 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 paraphrasing task. Our approach improved accuracies in both speech recognition and paraphrasing. The improvements were not significant, however, this approach has the potential to yield farther improvements by using better paraphrasing models.

The integrated approach looks promising for spoken language processing. In the future, we plan to extend our system to speech summarization and other on-line spoken language systems.

6. ACKNOWLEDGEMENT

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

7. 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, H. Isahara, “A Japanese national project on spontaneous speech corpus and processing technology,” Proc. of ASR2000, pp. 244–248, 2000.

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

[4] 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.

[5] F. Casacuberta, “Finite-state transducers for speech-input translation,” Proc. of ASRU 2001.

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