# LANGUAGE MODEL ADAPTATION USING WFST-BASED SPEAKING-STYLE TRANSLATION

*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,willett,minami}@cslab.kecl.ntt.co.jp

## ABSTRACT

This paper describes a new approach to language model adaptation for speech recognition based on the statistical framework of speech translation. The main idea of this approach is to compose a weighted finite-state transducer (WFST) that translates sentence styles from in-domain to out-of-domain. It enables to integrate language models of different styles of speaking or dialects and even of different vocabularies. The WFST is built by combining in-domain and out-of-domain models through the translation, while each model and the translation itself is expressed as a WFST. We apply this technique to building language models for spontaneous speech recognition using large written-style corpora. We conducted experiments on a 20k-word Japanese spontaneous speech recognition task. With a small in-domain corpus, a 2.9% absolute improvement in word error rate is achieved over the in-domain model.

## 1. INTRODUCTION

Language models have become indispensable for large-vocabulary continuous-speech recognition. These models, which are usually $n$-gram models, provide prior probabilities of hypothesized sentences to disambiguate their acoustical similarities. In order to build an $n$-gram model, text corpora are necessary to estimate the conditional word probabilities in the context of the preceding $n$-1 words. Since such statistical models are deeply dependent on data, it is desirable to use a large corpus in the same domain as that of the spoken sentences to be recognized. However, it is generally expensive to make a large corpus, especially from spontaneous speech, for a specific domain.

Language model adaptation techniques can be used to improve the reliability of models using large corpora similar to the target domain. For example, adaptation is achieved by weighted linear or non-linear interpolation of an in-domain model and out-of-domain models [1][2]. The weight for each model is determined based on the similarity to the target domain.

However, even finding a small, but sufficiently large corpus for a specific target domain is not easy. Therefore, we focus on corpora in which the styles of sentences are different. For example, the style of Japanese spoken sentences is much different from that of written sentences in comparison with English, even if the topic is the same. Many Japanese verbs, auxiliary verbs, adjectives etc. in written sentences usually change into other words in spoken sentences. In traditional adaptation techniques, this mismatch of styles obstructs effective adaptation, because those methods combine just probabilities of models made from different style corpora. If such mismatch were canceled, language models for spontaneous speech recognition would be easily built using rich written texts on various topics, which are much easier available than spontaneous speech corpora.

To solve this problem, we utilize a framework of speech-input machine translation based on weighted finite-state transducers (WFSTs) [5][6]. In this framework, knowledge sources for speech recognition and translation are integrated into a single WFST. Hence, this allows the incorporation of knowledge about the mapping of words from one language into another.

In this paper, we present a WFST that translates a sentence style for effective adaptation. The WFST is built by combining in-domain and out-of-domain models through a translation model, where each model is expressed as a WFST. The resulting WFST can be considered as an adapted language model. This adaptation technique can incorporate vocabulary and its statistics of different-style corpora into one adapted model. Consequently, the adapted model is well supported by large out-of-domain corpora.

We conducted experiments on a 20k-word Japanese lecture speech recognition 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

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

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

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 [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. "$\circ$" 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.

## 3. LANGUAGE MODEL ADAPTATION BY SPEAKING-STYLE TRANSLATION

### 3.1. Statistical framework of speech translation

Our language model adaptation technique is based on speech-input machine translation. First, we outline the general framework of statistical translation.

The translation of a source language $W$ to a target language can be formulated as the search for a word sequence

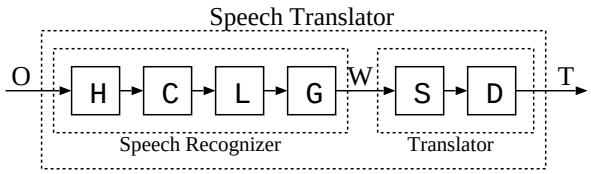


**Fig. 1**. Cascade of speech-input machine translation

$\hat{T}$ 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

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

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

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

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.

Let $S$ be a WFST that substitutes word sequences according to $\delta_S(W,T)$ 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 process of the speech translation is illustrated as the cascade in Fig. 1.

### 3.2. Principle of the adaptation approach

Suppose $W$ is a sentence in the target domain and $T$ is a sentence whose style is different from that of $W$ but whose meaning is the same as $W$. In this case, the Eq. (8) represents sentence-style translation.

We can simply modify Eq. (8) to get the speech recognition result instead of the translation result:

$$\hat{W} = \underset{W}{\text{argmax}}\, P(O|W) \max_{T} P(W|T)P(T). \quad (11)$$

When we consider this equation as a formulation of speech recognition, the language model probability for speech recognition can be interpreted as

$$P^A(W) = \max_T P(W|T)P(T) \quad (12)$$

$$\approx \max_T P_G(W)\delta_S(W,T)P_D(T), \quad (13)$$

where $P_D(T)$ is a prior probability of $T$, given by a language model for speech translation. Thus $P^A(W)$ contains constraints not only in the target domain ($W$) but also in the other-style domain ($T$). Actually, the following equation including weights $\lambda_G$ and $\lambda_D$ is used.

$$P^A(W) = \max_T P_G(W)^{\lambda_G}\delta_S(W,T)P_D(T)^{\lambda_D} \quad (14)$$

This is something similar to log-linear interpolation [2] which is a common adaptation technique with the difference of having a module for translating from the one to the other domain. Therefore, the model of $P^A(W)$ can be interpreted as an adapted language model that is composed of a model in the target domain and a second one in a different domain. Consequently, the adapted model can be supported by a large corpus through translation if there is a set of translation rules and a large corpus whose sentence style does not need to be the same as the target.

However, our adaptation technique needs a set of translation rules. Although this seems to be a drawback in the task of language model adaptation, strict rules are not necessarily needed because $P^A(W)$ can be calculated by the product of $P_G(W)$ and $P_D(W)$ by setting $\delta_S(W,W)$ to 1 when $W$ cannot be translated into any sequence by the defined rules. This should however only be the case for $W$ with a defined $P_D(W)$, which means that $W$ is also a part of $T$ and does not need to be mapped to another word sequence. Therefore, at least the effect of the traditional adaptation technique can be expected even for cases that require no translation.

The adapted model can be expressed as a WFST composed of WFSTs for $P_G(W)$, $\delta_S(W,T)$, and $P_D(T)$:

$$G^A = \text{proj}(G \circ S \circ D), \quad (15)$$

where "proj" indicates a projection operator of a WFST to a WFSA (Weighted Finite-State Acceptor). The operation in our work simply substitutes the output symbol of each arc with its input symbol. Finally, the integrated WFST including the adapted language model is composed as

$$R^A = H \circ C \circ L \circ G^A. \quad (16)$$

## 4. EXPERIMENTS

### 4.1. Conditions

We evaluated our adaptation method in a 20k-word spontaneous speech recognition task. The task is based on a corpus of Japanese spontaneous speech [7], most of which are 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 of 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 (6 lectures) and its manual translations into written language.

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 compared to the spoken language corpus, which is very focused in topic. Therefore, language model adaptation was performed under relatively difficult conditions.

**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 3000 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].

### 4.2. Experimental results

We tested three types of language models: in-domain models, adapted models with translation (proposed), and adapted models with no translation. The in-domain models were trained using only the spoken corpus (the manual transcriptions). The "with no translation" means to set $\delta_S(W,T) = 1$ if $W = T$, and $\delta_S(W,T) = 0$ if $W \neq T$. In

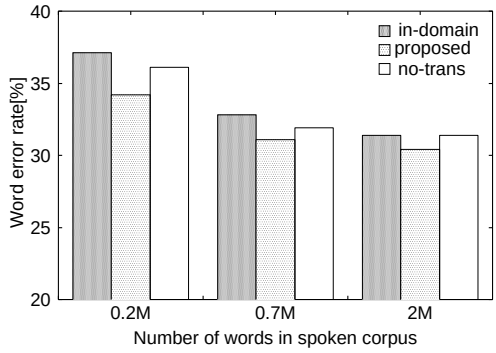


**Fig. 2**. Word error rate in various sizes of the spoken corpus

this case, the vocabulary of $P_D(\cdot)$ is made identical to that of $P_G(\cdot)$, which is the union of the two vocabularies.

We excluded four lectures from training in order to use them for evaluation, which are not included in the spoken corpus. The weights for the model combination were tuned to minimize word error rate for each test lecture.

Figure 2 shows average word error rate for the four lectures with different sizes of the spoken corpus of 0.2 M, 0.7 M, and 2 M words (M= $\times 10^6$). In any of these cases, the proposed model yielded a lower error rate than those of the in-domain model and the no-translation model. When the size is smaller, the improvement achieved with this adapted model is more significant. The proposed method shows a 2.9% absolute improvement over the in-domain model for the case of the 0.2 M words. However, the no-translation model had little effect in comparison with the proposed method. The primary reason for this is the mismatch of styles.

Table 2 shows the word error rate for each lecture when the number of words in the spoken corpus is 2 M words. A01M0007, A01M0035, A01M0074, and A05M0031 represent lecture IDs, and their lengths are 30, 28, 12, and 27 minutes, respectively. The reduction rate by the proposed method varied with the lectures. The reduction in A01M0035 was smaller than those of the other lectures. We conclude that the speaker of A01M0035 had such a high degree of spontaneity that the translation model hardly decreased the mismatch of styles.

**Table 2**. Word error rate [%] for each lecture

| | in-domain | proposed | reduction rate[%] |
|---|---|---|---|
| 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 |

## 5. CONCLUSIONS

We proposed a new adaptation method of language models based on speaking-style translation using weighted finite-state transducers. It has the potential of integrating language models of different vocabularies that represent different styles of speaking or dialects. Compared to conventional language model adaptation, it allows the incorporation of knowledge about the mapping of words from the one domain (style, dialect) into another. The approach was applied to adaptation for Japanese spontaneous speech using a small in-domain spoken corpus and a large out-of-domain written corpus. We conducted experiments on a 20k-word Japanese spontaneous speech recognition task. With a small in-domain corpus, a 2.9% absolute improvement in word error rate was achieved over the in-domain model. For further investigation, the technique needs to be evaluated on a wider variety of tasks.

## 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] R. Kneser and V. Steinbiss, “On the dynamic adaptation of stochastic language models,” Proc. of ICASSP’93, vol. 2, pp. 586–589, 1993.

[2] D. Klakow, “Log-liniear interpolation of language models,” Proc. of ICSLP’98, pp. 1695–1698, 1998.

[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.

[7] T. Shinozaki, C. Hori, and S. Furui, “Towards automatic transcription of spontaneous presentations,” Proc. of Eurospeech 2001, vol. 1, pp. 491–494, 2001.