Multilingual acoustic model training based on phoneme set conversion

Yuuki Tachioka
Denso IT Laboratory Tokyo, Japan,
ytachioka@d-itlab.co.jp

***Abstract*—For multilingual automatic speech recognition (ASR), acoustic models are often trained on multiple language datasets. The ASR of a target language can be improved by using auxiliary language data. Generally, universal phoneme sets are used in order to eliminate the differences between phoneme sets. However, this approach is inefficient for languages whose phoneme sets are similar. We propose a more efficient training method based on phoneme set conversion, in which phonemes from the auxiliary language are converted into those of the target language. In addition, we examined the construction of context trees.**

***Index Terms*—multilingual speech recognition, acoustic model, phoneme set, context tree**

I. Introduction

For multilingual automatic speech recognition (ASR), multilingual acoustic models are often trained on multiple language datasets [1]. If models across languages can be shared, training time can be shortened. The use of large-size auxiliary language data can improve the ASR of a target language with small amount of data but differences between the phoneme sets of the languages presents difficulties. Generally, universal phoneme sets, such as the International Phonetic Alphabet [2] or universal character encoding [3], are used in order to eliminate the differences between phoneme sets. However, this approach is inefficient when the phoneme sets of the target and auxiliary languages are similar. Thus, we propose a more efficient training approach based on phoneme set conversion, which converts phonemes from an auxiliary language into similar phonemes in the target language. In this paper, we validated the effectiveness of the proposed method using Korean as the target language and Japanese as the auxiliary language. An acoustic model of the target language was trained on both the target and auxiliary languages after converting phonemes from the auxiliary language into those of the target language. In addition, two types of context trees, separate and shared context were examined.

II. Simultaneous training of multilingual acoustic models with similar phoneme sets

Fig. 1 shows the simultaneous training procedure with phoneme set conversion. Two datasets are used: target language and auxiliary language. First, phoneme sets from auxiliary language are converted into those of the target language. Missing phonemes in the auxiliary language are replaced with similar phonemes from the target language. Second, a shared or separate context tree is constructed. Third, simultaneous training is applied to train acoustic models. Finally, if necessary, fine-tuning is applied to the target language dataset.

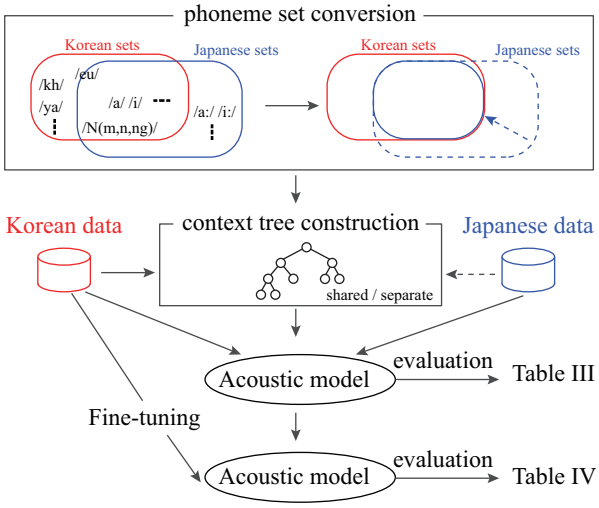

Fig. 1. Simultaneous training procedure on both Korean and Japanese datasets. First step converts Japanese phonemes into Korean phonemes and the second step constructs shared or separate context trees. Afterwards, acoustic models are simultaneously trained on Korean and Japanese. If necessary, acoustic models are fine-tuned only on the Korean dataset.

A. Conversion of phonemes

With the exception of long vowels, Japanese phoneme sets can be converted into Korean phoneme sets because their phoneme sets are similar. The Korean phoneme sets are defined as the zeroth_korean dataset and the Japanese phoneme sets are defined as the Corpus of Spontaneous Japanese (CSJ) dataset. In this paper, Japanese phonemes (J) were converted into Korean phonemes (K) using the conversion rules below. Phonemes (/a/, /i/, /u/, /e/, and /o/) were unchanged except for when /eu/ (K) was used for /u/ (J), as in /s u/, /ts u/, and /z u/. /m/, /ng/, and /n/ (K) were used for /N/ (J) depending on the consecutive phonemes. /j a/, /j u/, and /j o/ (J) were replaced by /ya/, /yu/, and /yo/ (K). Other phonemes /k/, /sh/, /r/, /z/, /f/, and /ts/ (J) were replaced by /kh/, /s/, /l/, /j/, /h/, and /ch/ (K), respectively. Double consonants (J) were substituted with the silent patchim /d2/ (K). Since there are no long vowels in Korean, we used two types of models: the first type models long vowels as one vowel and the second doubles the vowel for each long vowel.

978-1-7281-9802-6/20/$31.00 ©2020 IEEE

TABLE I
DETAILS OF ZEROTH_KOREAN CORPUS AND CSJ.

	data length	number of speakers
training data		
zeroth_korean	51.6 hours	105 speakers
CSJ	239 hours	986 speakers
evaluation data		
zeroth_korean	1.2 hour	10 speakers

TABLE II
BASELINE WER[%] OF ZEROTH_KOREAN CORPUS.

	tri-gram		four-gram
	small (tgs)	large (tgl)	large (fgl)
baseline	17.25	10.60	10.15

B. Construction of context trees

Constructing context trees based only on the target language may be more effective because the importance of discrimination between phonemes is dependent on the language. In this paper, we constructed shared context trees for Japanese and Korean and a separate context tree for Korean only.

C. Fine-tuning by only using target language

We validated the effectiveness of fine-tuning the target language dataset after simultaneously training multilingual datasets. Simultaneous training increased the amount of training data for overlapping phonemes in both languages but degraded the discrimination of phonemes specific to the target language. The introduction of fine-tuning can mitigate this degradation.

III. EXPERIMENTS (ZEROTH_KOREAN DATASET+CSJ)

A. Setup

Table I shows the details of the available datasets. For Korean speech recognition, the zeroth_korean dataset[1] was used. This dataset consisted of training and evaluation data. An nnet3-type model was trained on the basis of the Kaldi toolkit recipe[2]. The parameters were the same as that of the attached script. In addition, CSJ was used for additional training data. Three types of Korean n-gram language models were used for testing: small-size tri-gram (tgs), large-size tri-gram (tgl), and large-size four-gram (fgl). Performance was evaluated in terms of word error rate (WER).

B. Results and discussions

Table II shows the baseline WER when the acoustic models were only trained on the zeroth_korean dataset. A large-scale language model improved the performance. Table III shows the WER when additional Japanese training data were used. Simultaneously training both Korean and Japanese without fine-tuning degraded the performance compared with the baseline. We expected speaker adaptation to improve because the number of training speakers was improved but such an improvement was not observed.

[1]https://github.com/goodatlas/zeroth
[2]https://kaldi-asr.org

TABLE III
WER[%] OF ZEROTH_KOREAN CORPUS WHEN KOREAN AND JAPANESE MODELS WERE SIMULTANEOUSLY TRAINED WITHOUT FINE-TUNING.

	tri-gram		four-gram
long vowels	small (tgs)	large (tgl)	large (fgl)
separate context tree for Korean			
double	20.08	11.77	11.43
ignore	**18.99**	11.46	11.16
shared context tree for both Korean and Japanese			
double	20.96	11.76	11.27
ignore	19.45	**11.03**	**10.48**

TABLE IV
WER[%] OF ZEROTH_KOREAN CORPUS WHEN FINE-TUNING WAS APPLIED TO KOREAN AFTER SIMULTANEOUS TRAINING.

	tri-gram		four-gram
long vowels	small (tgs)	large (tgl)	large (fgl)
separate context tree for Korean			
double	16.11	**9.88**	**9.42**
ignore	**15.94**	9.90	9.48
shared context tree for both Korean and Japanese			
double	16.12	10.09	9.73
ignore	16.36	10.00	9.83

Table IV shows the WER when fine-tuning was applied after simultaneous training. Fine-tuning on the target dataset after simultaneous training improved the performance compared with the baseline, showing that fine-tuning is essential.

Though long vowels were treated two different ways, it did not affect performance except for the small-size language model (tgs). The performance of the separate context tree for Korean was more optimal than that of the shared tree. This indicates the importance of maintaining the context tree of the target language because the discrimination between phonemes is dependent on language and the key to discriminating them is lost when sharing context trees.

IV. CONCLUSION

Phoneme set conversion was proposed for efficient simultaneous multilingual acoustic model training when phoneme sets are similar across languages. Experiments showed that the WER of the target language improved 0.7–1.3 points by using an additional language dataset and fine-tuning the models in the target language dataset.

REFERENCES

[1] S. Hara and H. Nishizaki, "Acoustic modeling with a shared phoneme set for multilingual speech recognition without code-switching," in *Proc. APSIPA*, 2017, pp. 1–4.
[2] S. Tong, P. N. Garner, and H. Bourlard, "Multilingual training and cross-lingual adaptation on CTC-based acoustic model," *Speech Communication*, vol. 104, 2017.
[3] S. Watanabe, T. Hori, and J. R. Hershey, "Language independent end-to-end architecture for joint language and speech recognition," in *Proceedings of ASRU*, 12 2017, pp. 265–271.

978-1-7281-9802-6/20/$31.00 ©2020 IEEE