CROSS-LINGUAL SPEECH RECOGNITION UNDER RUNTIME RESOURCE CONSTRAINTS

Dong Yu, Li Deng, Peng Liu, Jian Wu, Yifan Gong, Alex Acero

Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA
{dongyu, deng, pengliu, jianwu, ygong, alexac}@microsoft.com

ABSTRACT

This paper proposes and compares four cross-lingual and bilingual automatic speech recognition techniques under the constraints of limited memory size and CPU speed. The first three techniques fall into the category of lexicon conversion where each phoneme sequence (PHS) in the *foreign language* (FL) lexicon is mapped into the *native language* (NL) phoneme sequence. The first technique determines the PHS mapping through the international phonetic alphabet (IPA) features; The second and third techniques are data-driven. They determine the mapping by converting the PHS into corresponding context-independent and context-dependent hidden Markov models (HMMs) respectively and searching for the NL PHS with the least Kullback-Leibler divergence (KLD) between the HMMs. The fourth technique falls into the category of acoustic-model (AM) merging where the FL's AM is merged into the NL's AM by mapping each senone in the FL's AM to the senone in the NL's AM with the minimum KLD. We discuss the strengths and limitations of each technique developed, report empirical evaluation results on recognizing English utterances with a Korean recognizer, and demonstrate the high correlation between the average KLD and the word error rate (WER). The results show that the AM merging technique performs the best, achieving 60% relative WER reduction over the IPA-based technique.

***Index Terms* —** Cross-lingual speech recognition, Kullback-Leibler divergence, lexicon conversion, senone mapping, resource constraint

1. INTRODUCTION

In recent years, we have observed an increasing number of deployed automatic speech recognition (ASR) applications in hand-held devices and automobiles where speech modality is shown to be superior to the conventional modalities (e.g., keyboards and stylus) due to the device size or the interaction environments. In many of these ASR applications, users occasionally need to control the system with bilingual commands. For example, a Chinese user may sometimes want to search for an English company using mobile search. To accomplish this task he/she may issue a command with English words mixed in a Chinese utterance or even with a pure English utterance. To support this usage scenario under the constraints of limited memory size and CPU speed new techniques need to be developed to allow the ASR recognizer trained for one language to recognize utterances in a second language (cross-lingual scenario) or in mixed languages (bilingual scenario). We call the language for which the acoustic model (AM) is trained the *native language* (NL) and other languages the *foreign languages* (FLs).

Many multi-lingual and cross-lingual ASR techniques have been proposed and investigated in earlier studies [1][4][6]. However, the goal of the prior arts has been to develop ASR systems for new languages for which less or no training data are available. Our goal of this work differs from the above in that we are interested in conditions where runtime resources (such as memory and CPU speed) rather than training data are limited, the recognition accuracy and speed for the NL not to be sacrificed, and the AM for each language is available or can be trained. We aim to develop the techniques that do not depend on human experts, and are sufficiently general so that they can be applied to different languages with no or very limited manual intervention.

In this paper we propose and compare four techniques under the setting just described. Among the four, the first three techniques fall into the lexicon conversion category where each phoneme sequence (PHS) in the FL lexicon is mapped into the NL PHS: The international phonetic alphabet (IPA) [3] based technique finds the PHS mapping through the IPA features; The data-driven context-independent (CI) and context-dependent (CD) phoneme mapping techniques determine the mapping by converting PHS into corresponding CI-phone and tri-phone hidden Markov models (HMMs) respectively and searching for the NL PHS with the least Kullback-Leibler divergence (KLD) between the HMMs. The fourth technique reported in this paper belongs to the AM-merging category where the FL's AM is merged into the NL's AM by mapping each FL senone to the NL senone with the minimum HMM KLD. We discuss advantages and disadvantages of each technique developed, report empirical results on recognizing English utterances with a Korean recognizer, demonstrate high correlation between the average KLD per phone and the word error rate (WER), and show that the AM-merging technique performs the best with 60% relative WER reduction over the IPA-based technique.

The rest of the paper is organized as follows. In Section 2, we describe the three lexicon conversion techniques and the engineering tricks used to speed up the conversion process. In Section 3, we illustrate the more effective, senone-mapping based AM-merging (AMM) technique and the rationale behind it. We report the experimental results in Section 4 and conclude the paper in Section 5.

2. LEXICON CONVERSION

A straightforward approach that satisfies the resource constraints is to keep the NL's AM unchanged and convert the FL lexicon into one using the NL phoneme set. That is, for each pronunciation $\theta = \theta_1 \vdash \cdots \vdash \theta_L$ in the FL lexicon represented in the FL phone set $\Sigma_\mathrm{F} = \{\sigma_\mathrm{i} | \mathrm{i} = 1, \cdots, \mathrm{F}\}$ we seek to find the best matched pronunciation

$$\hat{\varphi} = \hat{\varphi}_1 \vdash \cdots \vdash \hat{\varphi}_K = \operatorname*{argmin}_{\varphi} d(\theta, \varphi) \tag{1}$$

represented in the NL phone set $\Sigma_\mathrm{N} = \{\rho_\mathrm{i} | \mathrm{i} = 1, \cdots, \mathrm{N}\}$, where $d(\theta, \varphi)$ is the distance between the two sequences θ and φ. An obvious advantage of this approach is its simplicity since only the lexicon needs to be converted.

The key to this approach is the definition and evaluation of the distance $d(\theta, \varphi)$, and the algorithm to search for the best matched PHS $\hat{\varphi}$. Two sets of techniques have been proposed in the past to estimate the distance: one based on expert knowledge such as the IPA features, and the other based on the likelihood difference [4] or confusion matrix [1] [6]. These previously proposed data-driven techniques require decoding the FL utterances with NL's AM, and are either difficult to be extended to tri-phone pairs or costly in computation. In this section we first describe our baseline - IPA-based technique and then propose two new data-driven techniques based on KLD between HMMs. Compared with the IPA based-approach, the data-driven approach can be easily extended to different languages without expert knowledge or manual work as long as sufficient data are available to estimate the model parameters.

2.1. Lexicon Conversion with IPA

The well established IPA [3] classifies sounds based on knowledge of phonetic characterization of speech sounds. The idea behind the IPA-based lexicon conversion is to find a mapping from the FL phoneme sequences to the NL phoneme sequences based on the IPA features. Since it requires expert knowledge, usually only a limited number of J mapping rules between reasonably short sequences can be specified, i.e.,

$$\theta^j \rightarrow \varphi^j \quad iff\ \varphi^j = \operatorname*{argmin}_{\varphi} d_{IPA}(\theta^j, \varphi), \tag{2}$$

where $j = 1, \cdots, J$ is the index of the mapping rule. To find the best matched NL PHS φ for a long FL PHS θ, we define $d(\theta, \varphi)$ to be ∞ if the mapping from θ to φ is invalid according to the rule set (2) and equals to the minimum number of rules applied to find the mapping otherwise. The best NL PHS φ can be found with the dynamic programming algorithm or a finite state transducer.

The IPA-based technique has three limitations. First, subjective influence may be introduced by the expert when defining the conversion rule set. Second, the IPA symbols are constructed largely based on speech production properties rather than on speech acoustics. Hence, the same symbol in IPA as marked in many databases may be pronounced differently in different languages. Third, the IPA-based conversion rule set can only include short PHS (usually with length one and two) and so cannot represent long-span dependency.

2.2. Lexicon Conversion Using CI-Phone KLD

Given the limitations of the IPA-based technique discussed above, a data-driven approach that defines the distance directly based on speech acoustics becomes attractive, which we present in this and the next subsections.

Note that each phone $\sigma_\mathrm{i} \in \Sigma_\mathrm{F}$ and $\rho_\mathrm{i} \in \Sigma_\mathrm{N}$ can be represented as CI-phone HMMs defined by

$$\begin{aligned} h(\sigma_\mathrm{i}) &= \{\pi_{\sigma_\mathrm{i}}, A_{\sigma_\mathrm{i}}, B_{\sigma_\mathrm{i}}\} \text{ and} \\ h(\rho_\mathrm{i}) &= \{\pi_{\rho_\mathrm{i}}, A_{\rho_\mathrm{i}}, B_{\rho_\mathrm{i}}\} \end{aligned} \tag{3}$$

respectively, where B_{σ_i} and $B_{\rho_\mathrm{i}}$are sets of output distributions, $\pi_{\sigma_\mathrm{i}} = \pi_{\rho_\mathrm{i}} = [1\ 0\ \cdots 0]^T$ are the initial state distributions, and $A_{\sigma_\mathrm{i}}$and A_{ρ_i} are the transition matrices in the form

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & a_{S-1,S} \\ 0 & 0 & \cdots & 1 \end{pmatrix} \tag{4}$$

for an S-state left-to-right HMMs. The PHS $\theta = \theta_1 \vdash \cdots \vdash \theta_L$ and $\varphi = \varphi_1 \vdash \cdots \vdash \varphi_K$ can be represented as CI-phone HMMs constructed by concatenating CI-phone HMMs, i.e.,

$$h(\theta) = h(\theta_1) \vdash \cdots \vdash h(\theta_L), \tag{5}$$

$$h(\varphi) = h(\varphi_1) \vdash \cdots \vdash h(\varphi_K). \tag{6}$$

We can thus define the distance between θ and φ as

$$\begin{aligned} d(\theta, \varphi) &= KLD\big(h(\theta), h(\varphi)\big) \\ &= \int P\big(o^{1:t} | h(\theta)\big) log \frac{P\big(o^{1:t} | h(\theta)\big)}{P\big(o^{1:t} | h(\varphi)\big)} do^{1:t} \\ &+ \int P\big(o^{1:t} | h(\varphi)\big) log \frac{P\big(o^{1:t} | h(\varphi)\big)}{P\big(o^{1:t} | h(\theta)\big)} do^{1:t} \end{aligned} \tag{7}$$

where $KLD\big(h(\theta), h(\varphi)\big)$ is the symmetric KLD between two HMMs $h(\theta)$ and $h(\varphi)$ whose upper bound can be estimated efficiently as discussed in [2] even if $h(\theta)$ and $h(\varphi)$ have a different number of Gaussian mixture components or states [5]. More specifically, we can use the approximation of

$$KLD\big(h, \tilde{h}\big) \cong \min_{i,j} \sum_{i,j} \Delta_{i,j} + \emptyset_{i,j}, \tag{8}$$

where $\Delta_{i,j}$ is the symmetric KLD between state i in h and state j in $\tilde{h}$ defined as

$$\Delta_{i,j} = \left[D\left(b_i \parallel \tilde{b}_j\right) + log\frac{a_{i,i}}{\tilde{a}_{j,j}}\right] l_i + \left[D\left(\tilde{b}_j \parallel b_i\right) + log\frac{\tilde{a}_{j,j}}{a_{i,i}}\right]\tilde{l}_j, \qquad (9)$$

$l_i = \left(1 - a_{i,i}\right)^{-1}$ is the expected duration of the *i*-th state in h, and

$$\emptyset_{i,j} = \begin{cases} \frac{l_{i-1} + l_i}{\tilde{l}_j} & \text{state } i \text{ in } h \text{ is an insersion} \\ \frac{\tilde{l}_{j-1} + \tilde{l}_j}{l_i} & \text{state } i \text{ in } h \text{ is a deletion} \\ 0 & \text{otherwise} \end{cases} \qquad (10)$$

compensates for duration differences between HMMs with different state sizes as discussed in [5].

2.3. Lexicon Conversion Using CD-Phone KLD

The technique described in Section 2.2 can be improved by converting PHS into tri-phone HMMs:

$$h_3(\theta) = h(sil - \theta_1 + \theta_2) \vdash \cdots \vdash h(\theta_{L-1} - \theta_L + sil) \qquad (11)$$

and

$$h_3(\varphi) = h(sil - \varphi_1 + \varphi_2) \vdash \cdots \vdash h(\varphi_{K-1} - \varphi_K + sil) \qquad (12)$$

and define the distance

$$d(\theta, \varphi) = KLD\left(h_3(\theta), h_3(\varphi)\right). \qquad (13)$$

Since the number of tri-phones is much greater than the number of CI-phones and since the expansion of the tri-phone sequence is context dependent, the search for the best matched PHS can be time consuming, esp. when converting a large lexicon. To speed up the process, we have developed and used two engineering tricks which we describe now.

First, since the KLD between HMMs is a function of $D\left(b_i \parallel \tilde{b}_j\right)$ and $D\left(\tilde{b}_j \parallel b_i\right)$, we have greatly reduced the computation by caching these values.

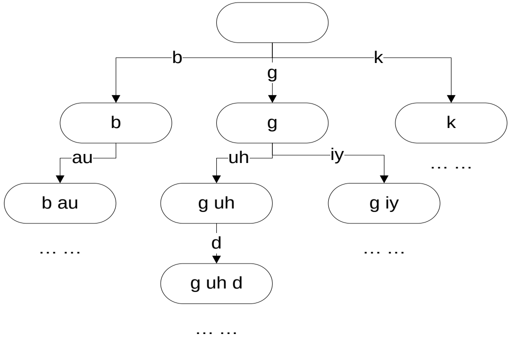

Fig. 1. An example of the lexicon pre-fix tree

Second, since many pronunciations share sub-sequences, we can speed up the conversion process by caching the mapping results for those sub-sequences. The trick we used is to construct a prefix-tree based on the FL lexicon, where each arc is a FL phoneme and each node is associated with a partial pronunciation following the path from the root. Fig.1 shows an example of the tree. To do the conversion, we start from the root and find the best-N NL phone sequences for each node with the depth-first strategy till all nodes are processed. Each time a new node is processed, the results from its parent is used as the starting point.

3. ACOUSTIC MODEL MERGING

Although the CD-phone KLD based approach can perform better than the CI-phone based approach as demonstrated in the experimental results shown in Section 4, there is still room for improvement since the possible target tri-phoneme sequences in the CD-phone KLD based approach are constrained by the possible NL PHS candidates. If we allow for concatenation of any tri-phones (i.e., the left-phone of the tri-phone does not need to be the same as the right-phone of the previous tri-phone) we may obtain a better matched concatenated HMM than what can be achieved with the constrained tri-phone sequence. Pushing this line of thinking further, if we allow for any concatenation of states (or senones) instead of multi-state tri-phone HMMs, we should be able to find even better matched concatenated HMM. Unfortunately, this cannot be achieved in the lexicon conversion framework since the HMM state sequence found is very unlikely to be associated with any valid NL PHS.

Let's examine what the best matched HMM would look like. Using (8), (9), (10) and noticing $\emptyset_{i,j} \geq 0$ and

$$\left(\frac{1}{1 - a_{i,i}} - \frac{1}{1 - \tilde{a}_{i,i}}\right) log\frac{a_{i,i}}{\tilde{a}_{i,i}} \geq 0 \qquad (14)$$

we obtain:

$$\begin{aligned} KLD\left(h, \tilde{h}\right) &\geq \min_i \sum_i \Delta_{i,i} \\ &\geq \min_i \sum_i D\left(b_i \parallel \tilde{b}_i\right) l_i + D\left(\tilde{b}_i \parallel b_i\right)\tilde{l}_i \qquad (15) \\ &= \sum_i \min_i \left(D\left(b_i \parallel \tilde{b}_i\right) l_i + D\left(\tilde{b}_i \parallel b_i\right)\tilde{l}_i\right). \end{aligned}$$

Eq. (15) indicates that we can find the best matched concatenated NL HMM by searching for the most similar NL senone for each FL senone and keeping the same transition matrix. This suggests that we can merge the FL's AM (including phone set and decision tree) into the NL's AM with shared NL senones. Merging the AM this way will only slightly increase the AM size and will not affect the decoding speed since the number of senones is unchanged and it is the senone number that dominates the size of the AM and the speed in evaluating the HMMs. Note that when merging the AMs we need to rename the FL phones in order to avoid name conflicts.

The AM merging technique just described has many additional advantages. First, finding the best matched NL's senone for each FL's senone can be efficiently done. Second, the FL lexicon can be directly used without lexicon conversion. Third, the NL and FL letter-to-sound (LTS) rules can be integrated into one making it easy to introduce new FL words in the applications. However, as a weakness, the AM merging approach requires modification of the AM and so cannot be easily extended to additional FLs once the AM is fixed.

4. EXPERIMENTAL RESULTS

The techniques we proposed and discussed above can be applied to both cross-lingual and mixed-lingual ASR. In this paper we focus on cross-lingual ASR and evaluate these techniques by recognizing English utterances with a Korean recognizer.

The 33-dimension feature used in the experiments consists of 11-dimension MFCC and its first and second order derivatives. The English and Korean tri-phone models were trained with about hundred hours of speech and contain 2300 and 2100 senones, respectively. Each senone in the AM has two to nine, and on average five, Gaussian mixture components. The AMs were further compressed using the sub-space coding algorithm [7] to reduce the size and to speed up the decoding.

Table 1: Comparisons of the average KLD per phone, WER, and RWERR for different techniques

Technique	KLD Per Phone	WER	RWERR
IPA Mapping	260.6	62.1%	0% (baseline)
CI-KLD	222.3	51.7%	16.7%
CD-KLD	146.5	44.4%	28.5%
AM Merging	41.7	25.0%	59.7%
Bound	0	18.8%	69.7%

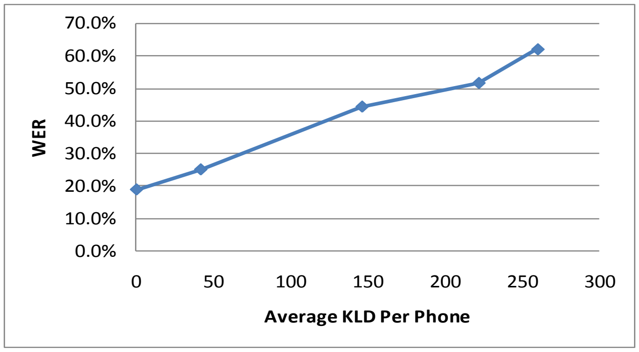

Fig. 2. The relationship between average KLD and WER

The English test set used in the experiment contains 14K words. If we replace the Korean tri-phone model with the English tri-phone model in the system and use the English lexicon directly we get 18.8% WER, which is the lower bound we can obtain for this test set using this compact AM. Table 1 summarizes the average KLD per phone, WER, and relative WER reduction (RWERR) for different techniques. We can clearly observe from the table that the shared senone AM merging technique achieved the best result with 25.0% WER or 59.7% RWERR over the baseline - IPA lexicon conversion technique. Fig. 2 shows the relationship between the average KLD and the WER and indicates that the average KLD is highly correlated with the WER. This confirms that HMM KLD is an appropriate objective function for optimization.

5. SUMMARY AND CONCLUSION

In this paper we have proposed and compared four techniques for cross-lingual and mixed-lingual ASR under run-time resource constraints. We have shown that we can gradually reduce the WER as we reduce the KLD between the mapped NL's PHS and the FL's PHS by relaxing constraints on possible phoneme sequences. We have demonstrated that the shared-senone AM merging technique achieved the best result with 59.7% relative WER reduction over the baseline IPA approach. If additional senones are allowed to be included in the NL's AM, the KLD measure can also be used to determine what senones to add. Our technique can be easily extended to recognizing accented FL utterances, where we need an additional step of adapting the FL's AM with accented FL utterances before applying the AM-merging technique described in this paper.

We believe that the behavior of the bilingual ASR will be highly correlated to the cross-lingual ASR result we have obtained. This hypothesis will be tested in our future work by collecting and testing on bilingual data sets.

6. ACKNOWLEDGEMENTS

We would like to thank Dr. Patrick Nguyen, Dr. Ye Tian, Lesley Carmichael, and Shahid Durrani at Microsoft Corporation for valuable discussions and helps.

REFERENCES

[1] O. Andersen, P. Dalsgaard, and W. Barry, "Data-driven Identification of Poly- and Mono-phonemes for Four European languages," Proc. *EUROSPEECH 1993*, pp. 759-762.

[2] M. N. Do, "Fast Approximation of Kullback Leibler Distance for Dependence Trees and Hidden Markov Models," *IEEE Signal Processing Letters*, vol. 10, pp. 115-118, Apr. 2003.

[3] International Phonetic Association, "Report on the 1989 Kiel convention," *Journal of the International Phonetic Association*, vol. 19(2), pp. 67-82, 1989.

[4] J. Kohler, "Multi-lingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities of Sounds," Proc. *ICSLP 1996*, pp. 2195-2198.

[5] P. Liu, F. K. Soong, and J.-L. Zhou, "Divergence-Based Similarity Measure for Spoken Document Retrieval," Proc. *ICASSP 2007*, vol. IV, pp. 89-92.

[6] A. Zgank, B. Imperl, F. T. Johansen, Z. Kacic, B. Horvat, "Cross-lingual Speech recognition with Multilingual Acoustic Models Based on Agglomerative and Tree-Based Triphone Clustering," Proc. *EUROSPEECH 2001*.

[7] E. Bocchieri and B. K. -W. Mak, "Subspace distribution clustering hidden Markov model," *IEEE Trans. Speech and Audio Proc.*, vol. 9, no. 3, pp. 264-275, 2001.