IMPROVED MIXED LANGUAGE SPEECH RECOGNITION USING ASYMMETRIC ACOUSTIC MODEL AND LANGUAGE MODEL WITH CODE-SWITCH INVERSION CONSTRAINTS

Ying Li, Pascale Fung

Human Language Technology Center
Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
eewing@ust.hk, pascale@ece.ust.hk

ABSTRACT

We propose an integrated framework for large vocabulary continuous mixed language speech recognition that handles the accent effect in the bilingual acoustic model and the inversion constraint well known to linguists in the language model. Our asymmetric acoustic model with phone set extension improves upon previous work by striking a balance between data and phonetic knowledge. Our language model improves upon previous work by (1) using the inversion constraint to predict code switching points in the mixed language and (2) integrating a code-switch prediction model, a translation model and a reconstruction model together. This integration means that our language model avoids the pitfall of propagated error that could arise from decoupling these steps. Finally, a WFST-based decoder integrates the acoustic models, code-switch language model and a monolingual language model in the matrix language all together. Our system reduces word error rate by 1.88% on a lecture speech corpus and by 2.43% on a lunch conversation corpus, with statistical significance, over the conventional bilingual acoustic model and interpolated language model.

Index Terms— mixed language, multilingual speech recognition

1. INTRODUCTION

Multilingual people often code-switch (CS) mixing two languages in the same sentence (intra-sentential code-switch) or between sentences (inter-sentential code-switch). Linguistic convention defines the principal language in a mixed language sentence as the matrix language (ML), and the embedded language (EL) is the secondary, foreign language [1].

The main challenge in recognizing mixed language speech is the lack of expertly transcribed data for both acoustic and language model training. There are many methods proposed to build bilingual acoustic models for both matrix and embedded languages that range from mapping the pronunciation dictionary to phonetic set combination and acoustic model merging [2, 3, 4, 5]. Bouselmi et al. [6] proposed a simple parallel model for context-independent models. A major weakness in these approaches is that the parameters of acoustic models undergo an irreversible change, and the models lose their ability to cover other (degrees of) accents. Such models do not perform well on matrix language speech, causing overall degradation in performance on mixed language speech [5].

For acoustic modeling, we propose a hybrid knowledge-based and data-driven approach for mapping phones of the matrix language to those of the embedded language, and then carry out a state-level acoustic model reconstruction on these mapped phones.

For language modeling of mixed language speech, a common approach is to interpolate the language models of the matrix and embedded languages, trained separately [2, 3, 4]. This approach allows code-switch anywhere. Vu et al. [7] used language identification techniques to detect the boundaries at which the speaker code-switches and decode the speech segments using the corresponding language model. However, making an early decision on switch points leads to errors being propogated to the next stage. More importantly, linguists have found that code-switch does not occur in positions where the order of the words is inverted between the matrix language and the embedded language [8, 9, 10]. This constraint corresponds to an inversion constraint in statistical machine translation [11, 12]. For the first time, we propose to use this constraint in a code-switch language model and incorporate a CS boundary prediction model, a CS translation model and a reconstruction model in a single weighted finite state transducer framework.

2. ASYMMETRIC ACOUSTIC MODELING

We postulate that mixed language speakers pronounce the embedded language speech in a range of accents and propose a hybrid approach of phonetic mapping using a similarity mea-

sure with an acoustic model reconstruction approach to recognize code-switch speech. The similarity measure is based on the alignment of the canonical transcription of speech in the embedded language and its recognition results using the matrix language recognizer, in other words the distances between the observations of the accented speech in the embedded language and the models of the matrix language. The clustering of the Mandarin and English phones was done using a pure data-driven approach in our previous work [3]. The new approach presented here is a hybrid of rule-based and data-driven methods to strike a balance between data and knowledge since there exist variations between the mixed language speech data and the linguistic knowledge of the matrix language and the embedded language.

In our work, we classify the Mandarin phone set and English phone set into 10 phone classes, each using linguistic knowledge such as the articulatory feature. The phone classes constraint is that each phone in the English data can only be recognized as a Mandarin phone with the same articulatory feature. Our proposed method is as follows:

1. Apply the Mandarin recognizer to the English data to obtain hypothetical Mandarin phonetic transcriptions. Obtain a phonetic confusion matrix between Mandarin and English by aligning the reference English phonetic transcriptions with the time-label information and hypothetical Mandarin phonetic transcriptions;

2. Likelihood ratio tests are used as a confidence measure to obtain phone clustering between the Chinese and English phone sets from 1). Elements of Mandarin initials/finals and English phone sequences with the largest confusion probability can be mapped and removed from the probability matrix. This clustering procedure continues until the Chinese initials/finals and English phones are clustered into a bilingual phone set;

3. For a given pair of bilingual and accented English decision trees of mapped phones, the leaf nodes of individual states of individual phones are merged according to an acoustic distance measure between the trained acoustic models. The new output distribution of the merged mixed language acoustic model is then a linear interpolation of the pre-trained models at the state level.

The new output distribution of the reconstructed model is represented as

$$P'(x|b_j) = \lambda P(x|b_j) + (1-\lambda)\sum_{i=1}^{N} P(x|e_i)P(e_i|b_j) \quad (1)$$

where $P(x|b_j)$ is the output distribution of the pre-trained bilingual model, $P(x|e_i)$ is the output distribution of the accented English model, λ is a linear interpolation coefficient and is the probability of the bilingual model being recognized as itself. In addition, $i = 1, 2, ..., N$ and N is the total number of merged nodes from the accented decision tree; e_i is one possible state from the accented decision tree to be tied to the state in the bilingual decision tree. $P(e_i|b_j)$ is the confusion probability between the bilingual model and accented English model. We use the bilingual model to decode the phone string in the accented English, use the same model to perform forced alignment, and obtain a confusion matrix.

3. CODE-SWITCH LANGUAGE MODELING WITH SYNTACTIC CONSTRAINT

We proposed a code-switch language model, which is the composition of a monolingual language model in the matrix language, a code-switch boundary prediction model, a code-switch translation model and a reconstruction model to avoid propagated error and to incorporate a syntactic constraint of code-switch speech [13].

$$P(W_1^M) = \sum_{w_1^m} P(w_1^m)P(W_1^M|w_1^m) \cong \max_{w_1^m} P(w_1^m)P(W_1^M|w_1^m) \quad (2)$$

where W_1^M is in mixed language, and w_1^m is in the matrix language. The code-switch language model can be modeled as

$$P(W_1^M|w_1^m) \cong \max_{v_1^n, u_1^n, W_1^M} \{P(v_1^n, n|w_1^m) \cdot P(u_1^n|v_1^n, w_1^m) \cdot P(W_1^M|u_1^n, v_1^n, w_1^m)\} \quad (3)$$

where $P(v_1^n, n|w_1^m)$ is the code-switch boundary prediction model, $P(u_1^n|v_1^n, w_1^m)$ is the code-switch translation model, and $P(W_1^M|u_1^n, v_1^n, w_1^m)$ is the reconstruction model. A word sequence in the matrix language w_1^m is segmented into phrases, v_1^n, and u_1^n is a phrase sequence in mixed language.

3.1. Code-switch Boundary Prediction Model

According to linguistics [8, 9, 10], a code-switch can only occur at points where the word order requirements of both the matrix and embedded languages are satisfied. Figure 1 shows an example of a Mandarin-English mixed language sentence. For example, code-switch is not permissible between the first three words with syntactic inversions.

We propose to train the code-switch boundary prediction model on the word-aligned parallel sentences in the matrix and embedded languages. The code-switch boundary prediction model is the probabilities of a sequence of words segmented into a sequence of phrases. We define a phrase as a word or a concatenation of words in which there are one or more inversions of a word-aligned sentence pair.

$$P(v_1^n, n|w_1^m) = \frac{1}{Z_n}\prod_{i=1}^{n} P(v_i) \quad (4)$$

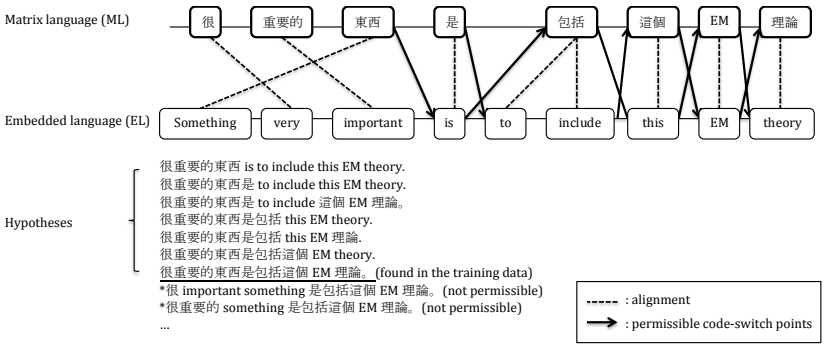

Fig. 1. An example of permissible code-switch points

$$Z_n = \sum_{v_1^n} \prod_{k=1}^{m} P(v_i) \tag{5}$$

where $P(v_i)$ can be approximated by the relative frequency of the i-th phrase.

3.2. Code-switch Translation Model

The code-switch translation model trains the probability of code-switch at the phrase boundaries given by the above model. The code-switch translation probability, $P(u_1^i|v_1^i)$, is assumed to depend on the previous phrase, v_{i-1}. $\pi(\mathbf{x})$ specifies the code-switch translation probability distribution trained from word-aligned bilingual sentences.$\mathbf{x}$ is an n-tuple, which includes the word code-switch probability $P(e|w)$, the reordering probability $\prod_{j=1}^{k} P(r_j|j,k,l)$, the phrase translation probability $Pr(u|v)$ and the phrase penalty $Pen(v)$, where w is an ML word, e is an EL word, k, l are the lengths of phrases in the matrix language and the embedded language, r_j denotes that the j-th word is aligned to the r_j-th EL word, v is an ML phrase, and u is an EL phrase.

We use a logit regression model to describe the code-switch translation probability

$$\text{logit}[\pi(\mathbf{x})] = \log(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}) = \alpha + \sum \beta_j x_j \tag{6}$$

where β_j is the effect of the j-th item in the n-tuple $\mathbf{x}$ on the logit of the code-switch translation probabilities controlling the other items of $\mathbf{x}$. The code-switch translation probability

$$\pi(\mathbf{x}) = \frac{\exp(\alpha + \sum \beta_j x_j)}{1 + \exp(\alpha + \sum \beta_j x_j)} \tag{7}$$

$$P(u_i|v_1^i) = \begin{cases} 1 - \pi(\mathbf{x_{i-1}^i}) & u_i = v_i \\ \pi(\mathbf{x_{i-1}^i}) & otherwise \end{cases} \tag{8}$$

where $\mathbf{x_{i-1}^i}$is the n-tuple of the word alignment probabilities, the reordering probability and the phrase penalty of the $(i-1)$th and ith phrases.

3.3. Code-switch Reconstruction Model

The reconstruction model assigns probabilities to a sequence of mixed language words, W_1^M, given that the words in the sequence are the same as the words of the phrases, u_1^n

$$P(W_1^M|u_1^n, v_1^n, n, w_1^m) = \prod_{i=1}^{n} P(W_{S_i}^{E_i}|u_i) \tag{9}$$

$$P(W_{S_i}^{E_i}|u_i) = \begin{cases} \frac{1}{Z_i} \prod_{j=S_i}^{E_i} q(W_j) & W_{S_i}^{E_i} = u_i \\ 0 & otherwise \end{cases} \tag{10}$$

$$Z_i = \sum_{u_1^n} \prod_{j=S_i}^{E_i} p(W_j) \tag{11}$$

where $p(W_j)$ is the frequency of occurrences of word W_j obtained from the bilingual sentences. $W_{S_i}^{E_i} = u_i$ indicates that the word sequence $W_{S_i}^{E_i}$ is exactly the same as the phrase u_i, S_i is the start of phrase u_i, and E_i is the end of phrase u_i.

4. EXPERIMENTS

The acoustic models used throughout our paper are state-clustered crossword tri-phone HMMs with 16 Gaussian mixture output densities per state. We use the 39 phone set from the CMU dictionary, 21 Mandarin standard initials, 37 Mandarin finals, and 6 zero initials. The pronunciation dictionary is obtained by modifying dictionaries in the matrix and embedded languages using the phone set. The acoustic models

are adapted to the speakers using maximum likelihood linear regression as a baseline. A WFST decoder is used for decoding [14].

4.1. Speech Corpora and Baseline Acoustic Models

We compare our proposed framework to baseline systems consisting ofa bilingual acoustic model adapted to accented speech and an interpolated language model. The bilingual acoustic model is trained from 160 hours of speech from GALE Chinese broadcast conversation, 40 hours of speech from GALE English broadcast conversation, and three hours of in-house nonnative English data.

We evaluate our proposed method on two corpora: (1) a lecture speech corpus of a digital speech processing course recorded at National Taiwan University, and (2) a lunch conversation corpus recorded at the Hong Kong University of Science and Technology. The lecture speech contains 16% embedded English words. 18 hours of the lecture speech is used for adaptation of the acoustic models, 0.9 hours of the speech is used as a development set, and 1037 utterances are used as a test set. The conversation speech contains 22% English words [15]. 127 minutes of the conversation speech is used to adapt acoustic models, 26 minutes of the speech is used as a development set, and 280 utterances are used as a test set.

4.2. Text Corpora and Baseline Language Models

250,000 sentences from digital speech processing conference papers, power point slides and web data are used for language model training and parallel sentence generation for the lecture speech recognition task (LM data 1). 250,000 sentences of the Gale conversational speech transcription are used for language model training and parallel sentence generation for the lunch conversion speech recognition (LM data 2). The baseline language model for the lecture speech recognition is an interpolation of the language model trained from LM data 1 and the language model trained on the transcriptions of the mixed language lecture speech. Another baseline model of the lunch conversations recognition is trained from LM data 2 and interpolated with the language model trained from the transcriptions of the mixed language lunch conversations.

4.3. Experimental Results

Table 1 shows the word error rates (WER) of the experiments on the mixed language lecture speech and lunch conversations. The asymmetric acoustic model outperforms the bilingual acoustic models by 2.8% on the lecture speech data and 4.03% on the lunch conversation data. Compared to the adapted acoustic models, the asymmetric acoustic model gives about 1.04% word error rate reduction on the lecture speech data and 1.29% word error rate reduction on the lunch conversation data. Moreover, the code-switch language model (CSLM) reduces the word error rate by 0.84% on the lecture speech data and 1.14% on the lunch conversation data. All the WER reductions are statistically significant at 99%.

Table 1. *Our proposed system outperforms the baselines in terms of WER*

	Lecture speech	Lunch conversations
BilingualAM +InterpolatedLM	37.53%	50.23%
AdaptedAM +InterpolatedLM	35.77%	47.49%
AsymmAM +InterpolatedLM	34.73%	46.20%
AsymmAM +Code-switchLM	**33.89%**	**45.06%**

5. CONCLUSIONS

In this paper, we propose an integrated mixed language speech recognition framework that incorporates asymmetric acoustic models and the language model with syntactic inversion constraints. Our language model is composed of a code-switch prediction model, which learns from word-aligned parallel sentences to give the permissible CS points, a translation model, which is obtained by logit regression and incorporates syntactic inversion constraints, and a reconstruction model in an integrated way. A maximum a posterior framework employs weighted finite state transducers in the process of final decoding, integrating the asymmetric acoustic models, a code-switch language model, and a monolingual language model in the matrix language. We tested our system on two tasks, in mixed language lecture speech recognition and in mixed language lunch conversation. Our system reduces word error rate in a baseline of the adapted acoustic models and the interpolated language model by 1.88% in the first task and by 2.43% in the second task. Our model also outperforms another baseline, that of the asymmetric acoustic models and the interpolated language model, by 0.84% in the first task and by 1.14% in the second task. All results are statistically significant. In addition, our method reduces word error rates for both the matrix language and the embedded language.

6. ACKNOWLEDGMENTS

This work was partially supported by grant number RGF 612211 of the Hong Kong Research Grant Council.

7. REFERENCES

[1] F. Coulmas, *The handbook of sociolinguistics*, vol. 4, Wiley-Blackwell, 1998.

[2] D. Imseng, H. Bourlard, M. Magimai-Doss, and J. Dines, "Language dependent universal phoneme posterior estimation for mixed language speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5012–5015.

[3] Y. Li, P. Fung, P. Xu, and Y. Liu, "Asymmetric acoustic modeling of mixed language speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5004–5007.

[4] K. Bhuvanagiri and S. Kopparapu, "An approach to mixed language automatic speech recognition," in *Oriental COCOSDA, Kathmandu, Nepal*, 2010.

[5] Q. Zhang, J. Pan, and Y. Yan, "Mandarin-English bilingual speech recognition for real world music retrieval," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4253–4256.

[6] G. Bouselmi, D. Fohr, and I. Illina, "Combined acoustic and pronunciation modelling for non-native speech recognition," *arXiv preprint arXiv:0711.0811*, 2007.

[7] N.T. Vu, D.C. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E.S. Chng, T. Schultz, and H. Li, "A first speech recognition system for mandarin-english code-switch conversational speech," 2012.

[8] E. Woolford, "Bilingual code-switching and syntactic theory," *Linguistic Inquiry*, vol. 14, no. 3, pp. 520–536, 1983.

[9] J. MacSwan, "13 code-switching and grammatical theory," *The Handbook of Bilingualism and Multilingualism*, p. 323, 2012.

[10] S. Poplack and D. Sankoff, "A formal grammar for code-switching," *Papers in Linguistics: International Journal of Human Communication*, vol. 14, pp. 3–45, 1980.

[11] D. Wu, "Stochastic inversion transduction grammars and bilingual parsing of parallel corpora," *Computational linguistics*, vol. 23, no. 3, pp. 377–403, 1997.

[12] D. Wu and P. Fung, "Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora," *Natural Language Processing–IJCNLP 2005*, pp. 257–268, 2005.

[13] Ying Li and Pascale Fung, "Code-switch language model with inversion constraints for mixed language speech recognition," .

[14] P. Dixon, T. Oonishi, K. Iwano, and S. Furui, "Recent development of wfst-based speech recognition decoder," in *Asia-Pacific Signal and Information Processing Association 2009 Annual Summit and Conference*, 2009, pp. 138–147.

[15] Ying Li, Yue Yu, and Pascale Fung, "A mandarin-english code-switching corpus," .