Code-switch Language Model with Inversion Constraints for Mixed Language Speech Recognition

Ying Li, Pascale Fung
Human Language Technology Center
Department of Electronic & Computer Engineering, HKUST
eewing@ust.hk, pascale@ee.ust.hk

ABSTRACT
We propose a first ever code-switch language model for mixed language speech recognition that incorporates syntactic constraints by a code-switch boundary prediction model, a code-switch translation model, and a reconstruction model. A WFST-based decoder then recognizes speech by combining an acoustic model, a pronunciation model and the code-switch language model in an integrated approach. Our proposed approach avoids making early decisions on code-switch boundaries and is therefore more robust than previous approaches. Our proposed system using the code-switch language model outperforms a baseline of interpolated language models by a statistically significant 0.91% on a mixed language lecture speech corpus, and 1.25% on a mixed language lunch conversation corpus. Our method also outperforms a language model that permits code-switch at all word boundaries by a statistically significant 1.35% on the lecture speech corpus and 1.69% on the lunch conversation corpus.

KEYWORDS: Code-switch, mixed language, language modeling.

Proceedings of COLING 2012: Technical Papers, pages 1671–1680,
COLING 2012, Mumbai, December 2012.

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). For inter-sentential switches, many researchers use two language models to decode separate sentences (Fugen et al., 2003). However, a sentence that contains two languages poses a more formidable challenge to speech recognition systems, as the same sentence would contain words or phrases belonging to two or more grammatical systems or subsystems (Gumperz, 1982). It is challenging to predict where in the sentence the speaker switches to another language and back, if at all. Some researchers (Lee et al., 2009) use the transcription of the code-switch speech to train a domain-dependent language model. This approach is limited by the small amount of code-switch data available for training.

Code-switch should be distinguished from loanword, which is a word borrowed from one language and incorporated into another language to become part of the lexicon. Code-switch speech is where the speaker actually tries to speak another language in its own grammar. In code-switch, the matrix language is the 'principal' language, where the 'embedded' language is the second language(C. and C., 1993; Coulmas, 1998).

There are two main approaches to recognizing code-switch mixed language speech. One is to detect the boundaries at which the speaker code-switches, then identify the language in the speech segments between the boundaries, and decode the speech segments using the acoustic and language models in the corresponding language(Chan et al., 2005; Shia et al., 2004; Lyu and Lyu, 2008; Vu et al., 2012). For text only code-switch language, Solorio and Liu (2008) used Naive Bayes and Value Feature Interval to classify the hypothesis code-switch points by F-measure and the naturalness rating.

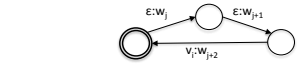
However, this approach requires multiple passes of boundary detection, language identification and speech recognition. The boundaries and language identity of each speech segment are irreversibly determined by the previews pass. Moreover, the speech segment of the embedded language tends to be very short. This poses challenges to the state-of-the-art language identification approaches.

A more holistic way to decode code-switch speech is by using a set of universal acoustic models for both matrix and embedded languages and a language model that permits code-switch (while predicting with probability where CS might occur) and which does not require an early decision of the code-switch points.

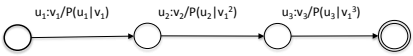
There are many methods proposed to build universal acoustic models for both the matrix and embedded languages that range from mapping the pronunciation dictionary to phonetic set combination and acoustic model merging (Imseng et al., 2011; Li et al., 2011; Bhuvanagiri and Kopparapu, 2010; Zhang et al., 2008). In this paper, we focus on code-switch language modeling.

A common approach is to build a code-switch language model from the language models of the matrix and embedded languages, trained separately from monolingual texts and combined together with linear interpolation (Bhuvanagiri and Kopparapu, 2010; Li et al., 2011; Imseng et al., 2011). This approach does not assume any syntactic constraint.

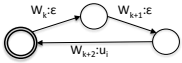
One can also use hand written grammar to constrain the code-switch point, such as in Zhang et al., 2008 or a bilingual dictionary to map the statistical n-grams in one language to the other, such as in Cao et al., 2010. Yeh et al., 2010 used a class-based n-gram language model based on



A portion of the code-switch boundary prediction model



A portion of the code-switch translation model



A portion of the code-switch reconstruction model

Figure 1: Weighted finite state transducers of the code-switch language model

perplexity and part-of-speech tag features. Tsai et al., 2010 proposed to use part-of-speech tags and to use Hakka-Chinese bilingual word mapping in the language model. The code-switch points are deterministic.

Linguists who study mixed language speech have discovered a typical constraint when speakers switch from one language to another in a sentence(Woolford, 1983; Mahootian, 1993; MacSwan, 1999). They found that code-switch can only occur in positions where "*the order of any two sentence elements, one before and one after the switch, is not excluded in either language*"(Poplack and Sankoff, 1980). This constraint, known as the “equivalence constraint" in linguistics, corresponds to an inversion constraint in statistical machine translation.

In this paper, we propose for the first time to incorporate this syntactic inversion constraint to a statistical code-switch language model. Our CS language model is composed of a CS boundary prediction model, a CS translation model, and a reconstruction model. The prediction model learns from word aligned parallel sentences to give the permissible CS points. The translation model is obtained by logit regression and incorporates syntactic inversion constraints. A *maximum a posterior* framework employs weighted finite state transducers in the process of final decoding, integrating a bilingual acoustic model, a code-switch language model, and a monolingual language model in the matrix language.

The structure of the paper is as follows: Section 2 presents the code-switch language model. The framework of decoding via the weighted finite state transducers is described in Section 3. Section 4 describes the results of the experiments. Section 5 draws the conclusion.

2 Code-switch language modeling

Given a speech utterance, S, with N_S frames, the automatic speech translation system converts S into a word sequence, W_1^M, where M is the total number of words. There are four components in the system, the language model, $P(W_1^M)$; the acoustic models, $P(S_1^{N_S}|W_1^M)$; the pronunciation dictionary; and the decoder.

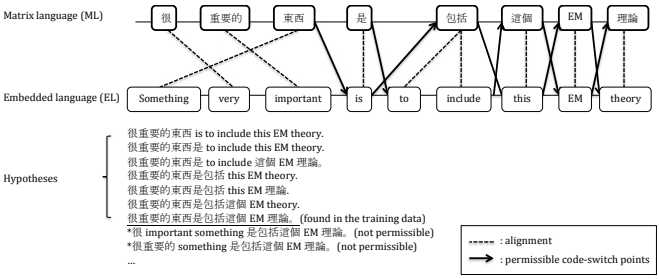


Figure 2: An example of permissible code-switch points

It is difficult to obtain mixed language text data which can be used for training the language model. Instead of directly calculating the probabilities, we use a monolingual language model in the matrix languagetogether with the CS language model.

$$P(W_1^M) = \sum_{w_1^m} P(w_1^m)P(W_1^M|w_1^m) \cong \max_{w_1^m} P(w_1^m)P(W_1^M|w_1^m) \quad (1)$$

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

$$P(W_1^M|w_1^m) \cong \max_{v_1^n, u_1^n, W_1^M} \{P(v_1^n, n|w_1^m) \cdot P(u_1^n|v_1^n, w_1^m) \cdot P(W_1^M|u_1^n, v_1^n, w_1^m)\} \quad (2)$$

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. We assume that w_1^m is a word sequence in the matrix language; the words are segmented into phrases, v_1^n; and u_1^n is a phrase sequence in mixed language.

2.1 Code-switch boundary prediction model

According to the equivalence constraint suggested by linguists(Poplack and Sankoff, 1980), the code cannot occur at the points where the order of the words are inverted between the matrix language and the embedded language. An example of a Mandarin-English mixed language sentence is shown in Figure 2. Code-switch is not allowed between the first three words with syntactic inversion.

Word-aligned parallel sentences in the matrix and the embedded languages are used to constrain at which point code-switch is allowed. We propose to generate bilingual training data as follows:

1) Translate words of the mixed language sentences in the embedded language into the matrix language using a statistical machine translation system;

2) Translate the monolingual sentences from 1) into the embedded language by a statistical machine translation system with inversion transduction grammar constraint (Wu, 1997; Wu and Fung, 2005) to obtain monolingual sentences in the embedded language;

3) Align the pairs of monolingual sentences in the matrix and embedded languages from 1) and 2). In theory, we can generate as many bilingual sentence pairs as possible.

The code-switch boundary prediction model trains the probabilities of a sequence of words segmented into a sequence of phrases from the aligned parallel sentences. A phrase is a word or a concatenation of words in which there exists one or more inversions of an aligned parallel sentence pair in the matrix language and the embedded language.

$$P(v_1^n, n|w_1^m) = \frac{1}{Z_n} \prod_{i=1}^{n} P(v_i) \tag{3}$$

where $P(v_i)$ can be approximated by the relative frequency of the i-th phrase.
$Z_n = \sum_{v_1^n} \prod_{k=1}^{m} P(v_i)$ such that $\sum_{v_1^n} P(v_1^n, n|w_1^m) = 1$.

2.2 Code-switch translation model

Given the permissible code-switch points by the above model, the code-switch translation model is the actual probability of code-switch at these points. We assume that the code-switch translation probability, $P(u_1^i|v_1^i)$, depends on the previous phrase, v_{i-1}.

The code-switch translation probability distribution is specified by probabilities $\pi(\mathbf{x})$ of code-switch and $(1-\pi(\mathbf{x}))$ of not code-switch. $\mathbf{x}$ is an n-tuple containing the conditional probability $P(e|w)$ of code-switch from a word, w, in the matrix language to a word, e, in the embedded language, and reordering probability $\prod_{j=1}^{k} P(r_j|j,k,l)$ of a phrase in the matrix language of length k and a phrase in the embedded language of length m, where r_j denotes that the j-th ML word is aligned to the r_j-th EL word, phrase translation probability $Pr(u|v)$ from a phrase, v, in the matrix language to a phrase, u, in the embedded language and phrase penalty $Pen(v)$. These probabilities are trained from word-aligned bilingual sentences.

The code-switch translation probability is neither linear nor exponential; it changes dramatically near the CS threshold. Thus we propose to use a logit regression model to describe the code-switch translation probability

$$\text{logit}[\pi(\mathbf{x})] = \log(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}) = \alpha + \sum \beta_j x_j \tag{4}$$

where β_j refers to 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 the n-tuple $\mathbf{x}$. The code-switch translation probability

$$\pi(\mathbf{x}) = \frac{\exp(\alpha + \sum \beta_j x_j)}{1 + \exp(\alpha + \sum \beta_j x_j)} \tag{5}$$

$$P(u_i|v_1^i) = \begin{cases} 1 - \pi(\mathbf{x_{i-1}^i}) & u_i = v_i \\ \pi(\mathbf{x_{i-1}^i}) & otherwise \end{cases} \tag{6}$$

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

2.3 Code-switch reconstruction model

The code-switch reconstruction model assigns probabilities to a sequence of mixed language words, W_1^M, given the constraint that the words agree with u_1^n, v_1^n, n, w_1^m

$$P(W_1^M|u_1^n, v_1^n, n, w_1^m) = \prod_{i=1}^{n} P(W_{S_i}^{E_i}|u_i) \tag{7}$$

$$P(W_{S_i}^{E_i}|u_i) = \begin{cases} \frac{1}{Z_i} \prod_{j=S_i}^{E_i} q(W_j) & W_{S_i}^{E_i} = u_i \\ 0 & otherwise \end{cases} \tag{8}$$

$p(W_j)$ 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. $Z_i = \sum_{u_1^n} \prod_{j=S_i}^{E_i} p(W_j)$ is set so that the probabilities sum to unity over possible values of u_i.

3 Decoding via weighted finite state transducers

The decoding of mixed language speech can be considered as searching the weighted finite state network. Suppose S denotes a speech utterance with N_S feature vectors, the recognition result $\hat{W}_1^M$ can be found as

$$\hat{W}_1^M = \arg\max_{W_1^M} P(W_1^M|S_1^{N_S}) \tag{9}$$

By Bayesian rule, the $\hat{W}_1^M$ is in

$$\{\hat{v}_1^n, \hat{u}_1^n, \hat{W}_1^M\} = \arg\max_{v_1^n, u_1^n, W_1^M} P(S_1^{N_S}|W_1^M) P(w_1^m) P(v_1^n, n|w_1^m) P(u_1^n|v_1^n, w_1^m) P(W_1^M|u_1^n, v_1^n, w_1^m) \tag{10}$$

while the possible code-switches are specified by $\hat{v}_1^n$ and $\hat{u}_1^n$.

The acoustic models $P(S_1^{N_S}|W_1^M)$, the pronunciation dictionary, the matrix language model $P(w_1^m)$, the code-switch boundary prediction model $P(v_1^n, n|w_1^m)$, the code-switch translation model $P(u_1^n|v_1^n, w_1^m)$ and the reconstruction model $P(W_1^M|u_1^n, v_1^n, w_1^m)$ are implemented as weighted finite state transducers and composed into a search network.

A lattice of the hypothesis speech recognition result is generated. When each token in the lattice transits from the end of one word to the start of the next word, a word insertion penalty is added. The insertion penalties of Chinese words and English words are separately trained from the development data. Lattice rescoring is proposed to adjust the word penalty of recognition results.

4 Experimental setup

This section briefly describes the data resources and feature analysis that were used for all the experiments used to evaluate the presented approach in the paper.

The acoustic features used in our experiments consist of 39 components (13MFCC,13ΔMFCC,13$\Delta\Delta$MFCC using subtraction of the cepstral mean), which are

analyzed at a 10msec frame rate with a 25msec window size. The acoustic models used throughout our paper are state-clustered crossword tri-phone HMMs with 16 Gaussian mixture output densities per state. The phone set consists of 21 standard Mandarin initials, 37 toneless Mandarin finals, 6 zero initials and 6 English extended phones. The pronunciation dictionary is obtained by modified dictionaries in the matrix and embedded languages using the phone set. The acoustic models are adapted to the speakers using maximum likelihood linear regression. The WFST decoder is used for decoding.

4.1 Corpora

We evaluate our proposed method on two mixed language speech corpora of different speaking styles, namely a lecture speech corpus and a lunch conversation corpus. The audio data is sampled at 16kHz.

About 20 hours of lecture speech of a digital speech processing course recorded at National Taiwan University are separated into three sets. Eighteen hours of the speech is used for adaptation of the acoustic models, and 0.9 hours of the speech is used as a development set. The testing set contains one hour of 1037 utterances. The lecture is given in Mandarin by a single speaker with 16% embedded English words.

The lunch conversation speech was recorded at the Hong Kong University of Science and Technology from a single speaker. The speech is highly spontaneous and the topics are wide ranging. The total length of the conversations is 163 minutes and 127 minutes are used to adapt acoustic models. The development set contains 26 minutes of the speech. Ten minutes of the speech is used for testing. There are 14762 Chinese words and 4280 English words. The percentage of the embedded English words is 22%.

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).

4.2 Langauge models

The baseline language model (Model_IP) 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 (Model_IP) of the lunch conversations is trained from LM data 2 and interpolated with the language model trained from the transcriptions of the mixed language conversations.

Our proposed language model (Model_CS) is constructed from combining the monolingual language model, the code-switch boundary prediction model, the code-switch translation model and the reconstruction model.

5 Experimental Results

Table 1 shows the word or character error rates of experiments on the mixed language lecture speech. The baseline interpolated language model reduces the overall word error rate by 0.49%. However, it degrades the recognition results of English phrases. Our proposed method outperforms the interpolated language model by 0.45% (Mandarin CER), 1.86% (English WER)

and 0.89% (overall).

Table 1: *Word/character error rate (%) of the lecture speech*

	Mandarin (ML)	English (EL)	Overall
Allow code-switch any where	35.16	43.54	36.55
Interpolated LM	34.35	44.36	36.09
Proposed method	**33.90**	**41.68**	**35.2**

The results of experiments on the mixed language lunch conversation speech are shown in Table 2. The baseline interpolated language model gives a 0.8% character error rate reduction on the Mandarin phrases and a 0.44% overall word error rate reduction, but degrades the performance on the English phrases. On the other hand, our proposed CS language model reduces the character error rate of Mandarin phrases by 0.41%, the word error rate of English phrases by 1.98% and overall word error rate by 1.25%. All the character error rate and word error rate reductions are statistically significant at 99%.

Table 2: *Word/character error rate (%) of the lunch conversations*

	Mandarin (ML)	English (EL)	Overall
Allow code-switch any where	47.20	49.14	47.63
Interpolated LM	46.40	49.98	47.19
Proposed method	**45.99**	**48.01**	**45.94**

Conclusions

In this paper, we propose a first ever statistical language model of code-switch speech that incorporates syntactic inversion constraints that have been found in that kind of speech. Our language model is composed of a code-switch prediction model, a translation model and a reconstruction model. A WFST-based decoder integrates this code-switch language model with an acoustic model and a monolingual language model in the matrix language for the final decoding. We tested our system on two tasks in mixed language lecture speech recognition, with 16% English words in Chinese sentences; and in mixed language lunch conversation, with 22% English words in Chinese sentences. Our system reduces word error rate in a baseline of the interpolated language model by 0.91% in the first task, and by 1.25% in the second task. Our model also outperforms another baseline, that of allowing code-switch at all points by 1.35% in the first task, and by 1.69% in the second task. All results are statistically significant. In addition, our method reduces error rates for both the matrix language and the embedded language.

Acknowledgments

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

References

Bhuvanagiri, K. and Kopparapu, S. (2010). An approach to mixed language automatic speech recognition. In *Oriental COCOSDA, Kathmandu, Nepal*.

C., M.-S. and C., M. (1993). *Duelling languages: Grammatical structure in codeswitching*. Clarendon Press Oxford.

Cao, H., Ching, P., Lee, T., and Yeung, Y. (2010). Semantics-based language modeling for cantonese-english code-mixing speech recognition. In *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*, pages 246–250. IEEE.

Chan, J., Ching, P., Lee, T., and Meng, H. (2005). Detection of Language Boundary in Code-switching utterances by Bi-phone Probabilities. In *Chinese Spoken Language Processing, 2004 International Symposium on*, pages 293–296. IEEE.

Coulmas, F. (1998). *The handbook of sociolinguistics*, volume 4. Wiley-Blackwell.

Fugen, C., Stuker, S., Soltau, H., Metze, F., and Schultz, T. (2003). Efficient handling of multilingual language models. In *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pages 441–446. IEEE.

Gumperz, J. (1982). *Discourse strategies*, volume 1. Cambridge Univ Pr.

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

Lee, H., Tang, Y., Tang, H., and Lee, L. (2009). Spoken term detection from bilingual spontaneous speech using code-switched lattice-based structures for words and subword units. In *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, pages 410–415. IEEE.

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

Lyu, D. and Lyu, R. (2008). Language identification on code-switching utterances using multiple cues. In *Ninth Annual Conference of the International Speech Communication Association*.

MacSwan, J. (1999). *A minimalist approach to intrasentential code switching*. Routledge.

Mahootian, S. (1993). *A null theory of codeswitching*. PhD thesis, Northwestern University.

Poplack, S. and Sankoff, D. (1980). A formal grammar for code-switching. *Papers in Linguistics: International Journal of Human Communication*, 14:3–45.

Shia, C., Chiu, Y., Hsieh, J., and Wu, C. (2004). Language boundary detection and identification of mixed-language speech based on map estimation. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 1, pages I–381. IEEE.

Solorio, T. and Liu, Y. (2008). Learning to predict code-switching points. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 973–981. Association for Computational Linguistics.

Tsai, T., Chiang, C., Yu, H., Lo, L., Wang, Y., and Chen, S. (2010). A study on hakka and mixed hakka-mandarin speech recognition. In *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*, pages 199–204. IEEE.

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

Woolford, E. (1983). Bilingual code-switching and syntactic theory. *Linguistic Inquiry*, 14(3):520–536.

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

Wu, D. and Fung, P. (2005). Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. *Natural Language Processing–IJCNLP 2005*, pages 257–268.

Yeh, C., Huang, C., Sun, L., and Lee, L. (2010). An integrated framework for transcribing mandarin-english code-mixed lectures with improved acoustic and language modeling. In *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*, pages 214–219. IEEE.

Zhang, Q., Pan, J., and Yan, Y. (2008). Mandarin-English bilingual speech recognition for real world music retrieval. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4253–4256. IEEE.