# A FIRST SPEECH RECOGNITION SYSTEM FOR MANDARIN-ENGLISH CODE-SWITCH CONVERSATIONAL SPEECH

Ngoc Thang Vu[1], Dau-Cheng Lyu[2], Jochen Weiner[1], Dominic Telaar[1], Tim Schlippe[1], Fabian Blaicher[1]
*Eng-Siong Chng[2], Tanja Schultz[1], Haizhou Li[2]*

[1]Cognitive Systems Lab, Institute for Anthropomatics, Karlsruhe Institute of Technology (KIT)
[2]School of Computer Engineering, Nanyang Technological University, Singapore

thang.vu@kit.edu, dclyu@ntu.edu.sg

## ABSTRACT

This paper presents first steps toward a large vocabulary continuous speech recognition system (LVCSR) for conversational Mandarin-English code-switching (CS) speech. We applied state-of-the-art techniques such as speaker adaptive and discriminative training to build the first baseline system on the SEAME corpus [1] (South East Asia Mandarin-English). For acoustic modeling, we applied different phone merging approaches based on the International Phonetic Alphabet (IPA) and Bhattacharyya distance in combination with discriminative training to improve accuracy. On language model level, we investigated statistical machine translation (SMT) - based text generation approaches for building code-switching language models. Furthermore, we integrated the provided information from a language identification system (LID) into the decoding process by using a multi-stream approach. Our best 2-pass system achieves a Mixed Error Rate (MER) of 36.6% on the SEAME development set.

***Index Terms—*** code-switching, multilingual speech recognition

## 1. INTRODUCTION

Code-switching speech is defined as speech which contains more than one language within an utterance and is a common phenomenon in many multilingual countries [2]. This paper introduces our first automatic speech recognition system (ASR) for code-switch conversational speech on the SEAME corpus [1] (South East Asia Mandarin-English). The task of building an ASR system on a code-switch corpus imposes several challenges: Since language model training data at code-switch points are very scarce it is very difficult to reliably estimate the probability of word sequences where code-switching appears. Another challenge are co-articulation effects between phones at code-switches. Additionally, only a small amount of data for the task of recognizing spontaneous conversational code-switching speech was provided.

To address the former challenge, [3] [4] applied class-based language models using POS information. Further studies explored the use of translation- and semantic-based LMs [6] to improve the probability of infrequent and unseen code-switches. The latter problem was tackled in [3] [4] [5] where speaker adaptation and phone sharing between languages were investigated. Additionally, in [7] [8] monolingual acoustic models were used in combination with language identification to recognize code-switch sentences. Except for [3], who experimented on lecture speech, mostly read speech corpora were examined so far.

To overcome the effect of co-articulation at code-switch points and to make better use of our limited training resources we investigate two approaches for phone merging in combination with discriminative training. On language model level, we apply different SMT-based methods to generate artificial code-switch texts. Furthermore, we integrate information from a language identification system (LID) into the decoding process by using the multi stream approach to improve the accuracy.

## 2. SEAME CORPUS

SEAME is a conversational Mandarin-English code-switching speech corpus recorded from Singaporean and Malaysian speakers [1]. The corpus is designed for multiple research purposes which include language boundary detection, language identification studies and multilingual LVCSR systems. Hence, a word-level manual transcription with language boundary alignment is provided. To take regional language variations into account, we collected data from two countries: Singapore and Malaysia. As the corpus was developed for spontaneous code-switching speech research, our recordings consist of interviews and conversations without prepared transcription. The interview scenario featured two speakers, an interviewer who asked questions and an interviewee who answered them. Only the interviewees speech was recorded. Recordings of conversational speech consist of speech from two speakers. All speech was recorded in a quiet recording room using close-talk microphones. The audio was sampled at 16 kHz with a resolution of 16 bit. Compared to [1], we extended the corpus to about 63 hours

of audio data. Considering the particular speaking styles in Singapore and Malaysia, we classify transcribed words into four categories for language identification research: English and Mandarin words, Silence, and Others (discourse particle, other languages, and hesitations). The ratio of Mandarin, English, Silence and Others is 44%, 26%, 21%, and 7% respectively. The average number of code-switches within each utterance is 2.6 when counting only switches between Mandarin and English. The corpus contains 9,210 unique English and 7,471 unique Mandarin words. The duration of monolingual segments is very short: More than 82% English and 73% Mandarin segments are less than 1 second long while the average duration of English and Mandarin segments is only 0.67 seconds and 0.81 seconds, respectively. Further details and analysis on the 25-hrs corpus can be found in [1]. We divide the corpus into three sets (training, development and test set) and distribute the data based on several criteria (e.g. gender, speaking style, ratio of Singaporean and Malaysian speakers, ratio of the four categories, and the duration in each set). Table 1 lists the statistics of the SEAME corpus in these three sets. As performance measure for our systems we adopted the Mixed Error Rate (MER) which applies word error rates for English and character error rates for Mandarin. The presented MER is the weighted average over all English and Mandarin portions of the speech recognition output. By applying character based error rates for Mandarin, the performance does not dependent on the applied word segmentation algorithm for Mandarin and thus performance can be compared across different segmentations, giving more flexibility for future investigations.

**Table 1**. *Statistics of the SEAME corpus*

| | Train set | Dev set | Eval set | Total |
|---|---|---|---|---|
| # Speakers | 139 | 8 | 10 | 157 |
| Duration(hours) | 58.4 | 2.1 | 2.3 | 62.8 |
| # Utterances | 48,040 | 1,943 | 2,162 | 52,245 |

## 3. BASELINE CODE-SWITCH SYSTEM

### 3.1. Bilingual Pronunciation Dictionary

The CMU English [9] and the Mandarin pronunciation dictionary [10] is merged into one bilingual pronunciation dictionary - the number of English and Mandarin entries in the lexicon is 135K and 130k respectively. Due to large differences between American English and Singaporean/Malayan English, we applied some rules for extending the CMU dictionary to adapt to the Singaporean/Malayan English. In [11], Chen et al. introduced 21 rules based on linguistic knowledge to adapt the Cambridge pronunciation dictionary for Singaporian accent. Due to rapid growth of the dictionary which leads to a higher confusion of words during decoding, we only applied three rules, in which a phone is deleted or switched with another phone. We experienced a deterioration in performance if the forced alignment for acoustic model training was done without these pronunciation variations. The following rules were used [11]:

- Syllable-final voiceless plosive omitted if preceded by another consonant: /p/, /t/, /k/ might be deleted
- Word-final /t/, /d/ omitted if preceded by another consonant: /t/, /d/ might be deleted
- Word-final metathesis from 'sp' to 'ps'

### 3.2. Bilingual Language Model

With the help of the SRI Language Modeling Toolkit [17], we built trigram language models (LMs) from the SEAME training transcriptions (*Training TRL*) containing the full 16k-vocabulary of the transcriptions. Those models were interpolated with two monolingual language models. Both monolingual language models were created from 350k English sentences from NIST (*EN-mono*) and 400k Mandarin sentences from the GALE project (*CH-mono*) which had been collected from online newspapers. The interpolation weights were tuned on the transcriptions of the SEAME development set by minimizing the perplexity of the model. Supplemental vocabulary was selected from *CH-mono* and *EN-mono* by selecting frequent words which are not in the transcriptions. In total, the vocabulary size is 30k words. The resulting model was used as our baseline language model for the decoding (*Baseline LM*) which has a perplexity of 489.4 and an out-of-vocabulary (OOV) rate of 1.21% on the SEAME development set.

### 3.3. Baseline Recognition Performance

Based on the SEAME corpus, we developed an initial baseline speech recognition system. The preprocessing consists of feature extraction applying a Hamming window of 16ms length with a window overlap of 10ms. A 143 dimensional feature vector was extracted by stacking 11 adjacent frames with 13 MFCC coefficients each. An LDA transformation reduced this to 42 dimensions. The acoustic model uses a fully-continuous 3-state left-to-right HMM. The emission probabilities are modeled by Gaussian Mixtures with diagonal covariances. The phone set contains all English and Mandarin phones with +noise+, +breath+ and +laugh+ to model continuous speech. Since particles are very often used in Singaporean and Malayan language, the phone +particle+ was added to the phone set. For context dependent acoustic modeling, we stopped the decision tree splitting process at 3,500 quintphones. We then applied merge&split training with a maximum of 64 mixtures per state and a global Semi-Tied Covariance (STC) matrix [13] to all the acoustic models followed by three iterations of Viterbi training. Our baseline speech recognition system is a 2-pass system which consists of two different acoustic models. The first acoustic model AM1 is speaker-independent. The second AM2 is trained

by applying Speaker Adaptive Training (SAT) with Feature Space Adaptation (FSA). In addition, we performed boosted Maximum Mutual Information Estimation (bMMIE) [14] to improve performance. The column "Baseline" in Table 2 shows the results of this baseline system on the SEAME development set.

## 4. CODE-SWITCH ACOUSTIC MODELS

Due to the fact that we have Mandarin and English data spoken by the same speaker we expect the Mandarin and English phones to share some characteristics which hopefully may lead to an improved acoustic model. In the following section we describe our approaches to phone merging, results are given in Table 2.

### 4.1. Knowledge-based phone merging

Our knowledge-based approach uses the IPA [15] to identify phones common to both languages. In total, there are 21 symbols in the IPA tables which occur in both Mandarin and English. Hence, we merged their corresponding models and reduced the phone set to 60. All the phones in English and Mandarin which were merged in our bilingual acoustic model are vowel ( ɒ i oʊ eɪ ə u) and consonant (n l h f w k ɹ j b p ŋ d g m s t). In comparison to the baseline system, the improvement of MER from the SAT system is very small. For pure English and Mandarin sentences, we consistently got slight improvements, but an increased MER for code switch sentences.We assume that this is due to the fact that phone merging results in a higher confusion between words of different languages during decoding. Hence, code switches are harder to detect. Discriminative training (DT) is a convenient approach to compensate this effect. As shown in Table 2, after applying DT on top of the SAT system we achieved 37.1% MER on the development set.

### 4.2. Data-driven phone merging

While in the knowledge-based approach we only merged those phones which are represented by the same IPA symbol, the data-driven phone merging approach exclusively combines phones based on their phonetic similarity. First, we applied the Bhattacharyya distance [16] to compute the distance between all phone models in our set. Second, we iteratively merged the two clusters with minimal distance until there is only one cluster left. For phone merging, we only used clusters consisting of two phones. We observed three different categories of similar phone models, within English and Mandarin and across languages. Therefore, we trained three different speaker adaptive training systems by apply the following merging methods: 1) merge only phone across languages, 2) merge phones across languages and phones within English and 3) merge all of them. The best performing system merged phones across languages plus phones within English. On top of this system we applied discriminative training to improve the accuracy. The final MER is 37.2% i.e. the best system applying data driven phone merging could not outperform the straight-forward IPA-based phone merging. Table 2 summarizes all the results of our phone merging approaches.

**Table 2**. *System Performance [MER] on the SEAME dev set*

| System | Baseline | IPA-based | Data-driven |
|---|---|---|---|
| Speaker Adaptive (SA) | 39.7 | 39.6 | 39.6 |
| SAT + bMMIE | 37.3 | 37.1 | 37.2 |

## 5. SMT-BASED TEXT GENERATION FOR CS LANGUAGE MODELS

The SEAME corpus provides only little text data to reliably compute n-grams. Therefore, we applied different SMT-based methods to generate artificial code-switch texts. For our SMT-based text generation, we analyzed characteristics of code-switching from the SEAME training transcriptions. Based on this knowledge, we extracted monolingual English segments from the SEAME training transcriptions, translated them to Mandarin and searched the translations in a large monolingual Mandarin text. If the translations were found, we replaced them with their English counterparts. We build texts based on a large monolingual English text analogously. For the translation, we used the Moses Statistical Machine Translation Toolkit [18]. We started with a simple search-and-replace approach (*S&R*). To improve the probability distribution of *S&R*, we limited the replacements to segments which occur at least twice in the training text (*MinThres2*). Further we used context information: Found segments are only replaced if the word preceding the segment is a trigger word (*TriggerWords*) or a trigger part-of-speech tag (*TriggerPOS*) to limit the replacements of the segments. A trigger is a token found in the training text before a code-switch. Additionally, we investigated a probability improvement approach which sets a maximum number of replacements per segment, based on the segment frequency in the SEAME training text (*FreqAlign*). Finally, we analyzed a combination of the last approach with the part-of-speech trigger approach (*FreqAlign+TriggerPOS*). Their perplexities, out-of-vocabulary (OOV) rates and n-gram coverages on the SEAME development set transcriptions are illustrated in Table 3. Due to lower perplexities and higher n-gram coverages than the baseline language model, we used the *FreqAlign* and *FreqAlign+TriggerPOS* LMs to decode the development set. The *FreqAlign* LM shows most improvement with 36.9% MER.

## 6. INTEGRATE LID INFORMATION INTO DECODING

The proposed LID framework for code-switching speech includes feature extraction, voice activity detection, GMM-based classification and a post processing procedure. The LID framework outputs the language identity along with a

**Table 3**. *Quality of Language Models based on Artificial CS Texts (Vocabulary size: 30k, OOV rate: 1.21%)*

| | Baseline LM | S&R LM | MinThres2 | TriggerWords | TriggerPOS | FreqAlign | FreqAlign+TriggerPOS |
|---|---|---|---|---|---|---|---|
| Perplexity | 489.4 | 507.6 | 500.4 | 503.1 | 492.1 | 483.9 | 485.3 |
| 1-gram coverage (%) | 98.87 | 98.87 | 98.87 | 98.87 | 98.87 | 98.87 | 98.87 |
| 2-gram coverage (%) | 77.89 | 78.01 | 78.90 | 78.28 | 79.80 | 79.77 | 79.40 |
| 3-gram coverage (%) | 29.43 | 25.90 | 27.18 | 25.93 | 28.98 | 29.87 | 29.89 |

confidence score on a frame-by-frame basis. We used the same features as in the ASR. An HMM-based voice activity detector is used to separate speech and non-speech segments in each utterance. The speech segments are then evaluated by two GMM acoustic-based LID classifiers to produce two log likelihood scores for each speech frame. The post processing eliminates too rapid language changes as the language identity classification is done at frame level. In the post processing module, we decide the language identity of the $i$-th frame by averaging the log likelihood scores generated from the Mandarin GMM and English GMM from the $(i - w)$-th frame to the $(i + w)$-th frame, where $w$ is the length of the window. We used Hamming windows to emphasize the weighting of a current frame over the log likelihood scores. The frame error rate for voice activity detection and language identification on the development set is 5.88% and 70.64%, respectively. The LID suffers from the fact that the language segments are very short and the changes between languages are very quick and smooth.

Our multi-stream approach operates on the acoustic level, where we apply the linear interpolation approach to combine the acoustic models score, and the LID scores. The decoding process determines the current language through an LID decision tree, choses the appropriate LID score, and then produces a linear combination of the acoustic model score and the LID score with the weight 0.9 and 0.1 respectively. The decoding then proceeds as usual, using the combination score instead of just the AM score. Since the LID performance is not accurate enough, the best MER with adding LID information is 36.6% (0.3% absolute improvement).

## 7. CONCLUSION AND FUTURE WORK

In this paper we have presented our first steps toward an LVCSR system for spontaneous conversational code-switching speech. State-of-the-art techniques such as speaker adaptive and discriminative training enhanced our first baseline system. For acoustic modeling, we applied two phone merging approaches based on IPA and Bhattacharyya distance in combination with discriminative training to improve accuracy. On language model level, we investigated statistical machine translation-based text generation approaches for enhancing code-switching language models and improved the MER by 0.2% absolute. Furthermore, we integrated the provided information from a language identification system (LID) into the decoding process by using a multi-stream approach, which gave 0.3% absolute improvement. Our best 2-pass system achieves a MER of 36.6% on the SEAME development set.

## 8. REFERENCES


[1] D. Lyu, T. Tan, E. Chng and H. Li, "An Analysis of a Mandarin-English Code-switching Speech Corpus: SEAME", Interspeech, Japan, 2010.

[2] P. Auer, Code-Switching in Conversation: Language, Interaction and Identity, London: Routledge, 1998.

[3] C. Yeh, C. Huang, L. Sun and L. Lee, "An integrated Framework for Transcribing Mandarin-English Code-mixed Lectures with Improved Acoustic and Language Modeling", ISCSLP, Taiwan, 2010.

[4] T. Tsai, C.Y. Chiang, H. Yu, L. Lo, Y.R. Wang and S.H. Chen "A Study on Hakka and Mixed Hakka-Mandarin Speech Recognition ", ISCSLP, Taiwan, 2010.

[5] S. Yu, S. Zhang, B. Xu. " Chinese-English Bilingual Phone Modeling for Cross-Language Speech Recognition',' ICASSP, Canada, 2004.

[6] H. Cao, T. Lee and P.C. Ching," Development of the Cantonese-English code-mixing speech corpora", in Computer Processing of Asian Spoken Languages, Shuichi Itahashi and Chiu-yu Tseng et al., eds., (Japan: Consideration Books, March 2010), pp. 204-207.

[7] K. Bhuvanagiri and S. Kopparapu, " An Approach to Mixed Language Automatic Speech Recognition ", Oriental COCOSDA, Nepal, 2010.

[8] D.C. Lyu, R.Y. Lyu, Y. Chiang and C.N. Hsu, " Recognition on Code-Switching Among the Chinese Dialects", ICASSP, France, 2006.

[9] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[10] R. Hsiao, M. Fuhs, Y. Tam, Q. Jin and T. Schultz, "The CMU-InterACT 2008 Mandarin Transcription System ", Interspeech, Australia, 2008.

[11] W. Chen , Y. Tan , E. Chng , H. Li. "The development of a Singapore English call resource", Oriental COCOSDA, Nepal, 2010.

[12] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries and M. Westphal, "The Karlsruhe Verbmobil Speech Recognition Engine", ICASSP, Germany, 1997.

[13] M. Gales, "Semi-tied covariance matrices for hidden Markov models", IEEE Transactions Speech and Audio Processing, vol. 7, pp. 272-281, 1999.

[14] D. Povey, "Discriminative training for large vocabulary speech recognition", Ph.D. dissertation, Cambridge University Engineering Dept, 2003.

[15] Handbook, IPA: Handbook of the International Phonetic Association, 1999.

[16] B. Mak, and E. Barnard, "Phone Clustering Using the Bhattacharyya Distance", In Proceedings of ISCSLP, Philadelphia, PA, USA, pp. 2005-2008, 1996.

[17] A. Stolcke, SRILM an Extensible Language Modeling Toolkit, ISCSLP, 2002.

[18] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation", Annual Meeting of ACL, demonstration session, Czech Republic, 2007.