Interspeech 2018
2-6 September 2018, Hyderabad

Mandarin-English Code-switching Speech Recognition

Haihua Xu[1], *Van Tung Pham*[1,2], *Zin Tun Kyaw*[2], *Zhi Hao Lim*[1], *Eng Siong Chng*[1,2], *Haizhou Li*[3]

[1]Temasek Laboratories, Nanyang Technological University, Singapore
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[3]Department of Electrical and Computer Engineering, National University of Singapore, Singapore

haihuaxu@ntu.edu.sg

Abstract

This work presents the development of a Mandarin-English code-switching speech recognition system. We demonstrate three key novelties in our system. First, we increase our lexicon coverage to 360K words, where phone sets of different languages are maintained separately. Secondly, we used over 1000 hours of training data combining both mono-lingual and code-switch corpus to develop the acoustic model. Finally, for language modelling, we applied context-aware text normalization and word-class language model. When testing on our internal code-switch close talk microphone recording, the system achieves recognition performance that can support real applications.

Index Terms: code-switch speech recognition, phone sets, context-aware language model, TDNN.

1. Introduction

Most existing state-of-the-art speech recognition systems were focused on monolingual speech recognition, i.e, they deal with only one language in an utterance at one time. Hence, such systems cannot recognize code-switching (CS) speech. CS speech refers to the case when an utterance contains more than one language. The development of automatic speech recognition (ASR) for Code-switching (CS) speech is important since such phenomenon widely occurs in multilingual societies such as Singapore, Hong Kong. Although CS-ASR has been investigated for over ten years [1, 2, 3], prior system's performances have remained poor, primarily due to the lack of training data.

In this work, we present our effort to develop a functional prototype CS system that can recognize live Mandarin-English code-switching speech. Several technical novelties are introduced to deal with the limited training data problem.

2. System development

Our system is illustrated in Figure 1. The system is a desktop application where users can record their CS speech and automatically obtain the transcribed text live. The system also supports monolingual English or Mandarin speech. Additionally, our system supports translation from English or Mandarin to other languages such as Malay, Vietnamese, etc using the Google API [1]. The system's GUI and samples of generated transcriptions are shown in Figure 2. When testing on our internal close talk microphone recording, the system achieves recognition performance that can support real applications.

The novelties of our code-switching speech recognition system are summarized as following:

Table 1: *Data distribution for Code-switching acoustic modelling. CTS stands for Conversational Telephony Speech (CTS), BN is for Broadcast News (BN) correspondingly.*

Corpus	Category	Length (hours)	LDC data
LDC98S69 [6]	Mandarin CTS	∼190	LDC data
HUB4 [7]	Mandarin BN	∼30	LDC data
SEAME [1]	Code-switching conversation	∼100	LDC data
FM938	English conversation	∼250	SG-CT (Singapore English close-talk)

- We increase our lexicon coverage to 360K words. Different from others researchers, we maintain two separate phone sets (Mandarin and English) in our lexicon.
- We build the acoustic models using the state-of-the-art Time-delay Neural Network (TDNN) with over 1000 hours of acoustic data, which is significantly more than what have used in previous research.
- To generate text data to build the code-switching LM, we applied language context-aware text normalization. Additionally, word-class language model have been utilized to deal with the data sparsity problem.

2.1. Lexicon construction

We have built a lexicon consisting of 360K words, of which approximately 240K Mandarin words and the rest are English words. It is derived from four sources: LDC Mandarin dictionary, CMU dictionary [4], TEDLIUM dictionary [5], and our own transcribed of about 3000 Singapore name entities. Those entities include Singapore road/street, local dish/food, company and restaurant names, institute, as well as Singapore famous person names. These word pronunciations are transcribed by our own linguists.

In our lexicon, we have maintained two separate phone sets for the two languages, which consists of 209 Mandarin tonal phones (initial and final) and 42 English phones. We have not employed cross-lingual phone set merging as this would undermine the discriminative capability of the context-dependent phones from either language.

2.2. Acoustic modelling

We have developed over 1000 hours of training data for acoustic modelling. The data consists of both monolingual and code-switch utterances, as shown in Table 1.

The acoustic model was built using the recent Kaldi toolkit

[1] https://cloud.google.com/translate/docs/

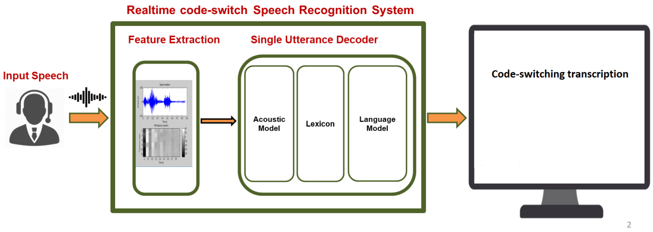
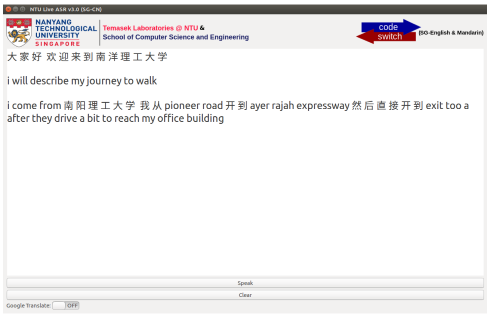


Figure 1: *Schematic diagram of our Mandarin-English code-switching speech recognition system.*


NTU Live ASR v3.0 (SG-CN)
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
Temasek Laboratories @ NTU &
School of Computer Science and Engineering
code switch
(SG-English & Mandarin)
Speak
Clear
Google Translate: OFF

大 家 好 欢 迎 来 到 南 洋 理 工 大 学

i will describe my journey to walk

i come from 南 阳 理 工 大 学 我 从 pioneer road 开 到 ayer rajah expressway 然 后 直 接 开 到 exit too a after they drive a bit to reach my office building

Figure 2: *Examples of output transcriptions. Our system is able to transcribe both monolingual and code-switching speech.*

[8]. In this work, we have used the Time Delay Neural Network (TDNN)[9] for acoustic models. Compared with conventional feed-forward DNN [3], TDNN can capture even longer temporal context. Additionally, it is simpler to train. We have also applied the Lattice-free Maximum Mutual Information (LF-MMI) [9] criterion to train the TDNN and this yields sharper acoustic models in terms of discriminative capability.

2.3. Language modeling

To build the code-switching language model, we need to collect and normalize large amount of text data. To improve the word coverage of code-switching speech, we apply language aware context based text normalization. For instance, given sentences, such as “ 那家酒店的顾客好评率在4.2分。” and “he gave 4.2 points for his score”, the numeral 4.2 needs to be converted correctly according to the language of the utterance. By doing so, we will improve the coverage for that language. This language context aware paradigm is applied to the other entities such as temperature, time, currency amount, percentage, etc.

We have collected and performed language aware context normalization on approximate 100M word text corpus. The raw text corpora were crawled from our local newspaper websites such as Zaobao [2], Straits Times [3], etc. Then we used the SRILM toolkits [10] to build the 4-grams language model.

Estimating the n-gram language model for code-switching is challenging due to data sparsity problem. To deal with such problem, we proposed to use the word-class n-gram language modeling. However, different from the conventional word-class n-gram method [11], only infrequent words are clustered while frequent words are treated as singleton classes themselves. This helps to alleviate the data sparsity issue for infrequent n-gram, while preserving the discriminative capacity for frequent n-grams.

3. Conclusions

We have demonstrated the development of our Mandarin-English code-switching speech recognition system. We first introduced lexicon construction, where words from various dictionaries are combined to form a single lexicon of 360K entries. We then described acoustic modelling, where we used over 1000 hours from various corpora to train a state-of-the-art acoustic model, i.e. the TDNN. We finally demonstrated the language model training, where a language context aware normalization and word-class language model techniques are applied. Our system is able to generate real-time transcriptions with recognition performance satisfied real applications.

4. References

[1] D. Lyu, T. P. Tan, E. Chng, and H. Li, “SEAME: a mandarin-english code-switching speech corpus in south-east asia,” in *Proc. of Interspeech*, 2010, pp. 1986–1989.

[2] N. T. Vu, D. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E. Chng, T. Schultz, and H. Li, “A first speech recognition system for mandarin-english code-switch conversational speech,” in *Proc. of ICASSP*, 2012, pp. 4889–4892.

[3] C. Yeh and L. Lee, “Transcribing code-switched bilingual lectures using deep neural networks with unit merging in acoustic modeling,” in *Proc. of ICASSP*, 2014, pp. 220–224.

[4] G. Chen, H. Xu, M. Wu, D. Povey, and S. Khudanpur, “Pronunciation and silence probability modeling for ASR,” in *Proc. of Interspeech*. ISCA, 2015, pp. 533–537.

[5] A. Rousseau, P. Deléglise, and Y. Estève, “Ted-lium: an automatic speech recognition dedicated corpus,” in *In Proc. of LREC*, 2012, pp. 125–129.

[6] LDC, “Hub5 mandarin telephone speech corpus,” in *https://catalog.ldc.upenn.edu/LDC98S69*, 1998.

[7] ——, “1997 english broadcast news speech (hub4),” in *https://catalog.ldc.upenn.edu/ldc98s71*, 1997.

[8] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer, “The kaldi speech recognition toolkit,” in *Proc. of ASRU workshop*, 2011.

[9] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, “Purely sequence-trained neural networks for ASR based on lattice-free MMI,” in *Proc. of Interspeech*, 2016, pp. 2751–2755.

[10] A. Stolcke, “Srilm – an extensible language modeling toolkit,” in *Proc. of ICSLP*, 2002, pp. 901–904.

[11] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, “Class-based n-gram models of natural language,” *Comput. Linguist.*, vol. 18, no. 4, pp. 467–479.

[2] http://www.zaobao.com.sg/

[3] http://www.straitstimes.com/