SEAME: a Mandarin-English Code-switching Speech Corpus in South-East Asia

Dau-Cheng Lyu[1,4], *Tien-Ping Tan*[2], *Eng-Siong Chng*[1,4], *and Haizhou Li*[1,3,4]

[1] School of Computer Engineering, Nanyang Technological University, Singapore 639798
[2] School of Computer Sciences, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia
[3] Institute for Infocomm Research, 1 Fusionopolis Way, Singapore 138632
[4] Temasek Laboratories, Nanyang Technological University, Singapore 639798

dclyu@ntu.edu.sg, tienping@cs.usm.my, aseschng@ntu.edu.sg, hli@i2r.a-star.edu.sg

Abstract

In Singapore and Malaysia, people often speak a mixture of Mandarin and English within a single sentence. We call such sentences intra-sentential code-switch sentences. In this paper, we report on the development of a Mandarin-English code-switching spontaneous speech corpus: SEAME. The corpus is developed as part of a multilingual speech recognition project and will be used to examine how Mandarin-English code-switch speech occurs in the spoken language in South-East Asia. Additionally, it can provide insights into the development of large vocabulary continuous speech recognition (LVCSR) for code-switching speech. The corpus collected consists of intra-sentential code-switching utterances that are recorded under both interview and conversational settings. This paper describes the corpus design and the analysis of collected corpus.

1. Introduction

Code-switching, which is defined as the usage of more than one language, variety, or style by a speaker within an utterance or discourse, has been increasingly reported on speech technology and linguistic studies in recent research works [1]. For some bilingual and multilingual societies, for example in United States and Switzerland, we could often hear Spanish-English and French-Italian code-switching speech [2]. In Hong Kong, Cantonese-English code-switching speech, where English words are embedded into colloquial Cantonese, is a common speaking style among the young generation [3]. Similarly in Taiwan, Mandarin-Taiwanese codes-witching speech, has become widespread in recent years [4]. Code switching is very common as it enable people to maintain a sense of social belonging, and provide a convenient way to express themselves [5].

Singapore and Malaysia are multi-racial societies. According to the Singapore government statistics [6], Chinese (74%) forms the majority of the population, followed by Malays (13%) and Indians (9%) in Singapore. In Malaysia, Malay makes up of 50% of the population, followed by Chinese 24%, Indian 7%, and 19% others. Although the biggest ethnic group in Singapore is Chinese, the working language is English. As a result, a particular speaking style, Singapore vernacular English, which consists of words originating from English, Mandarin, Malay, and other Chinese dialects, is created [7]. Most Singaporean can also communicate using their respective native languages, e.g. Mandarin and Malay and Tamil. In this paper, we focus in code-switching speech between Mandarin and English, which is common in the Chinese community of Singapore and Malaysia. In Figure 1, we show a typical example of Mandarin-English code-switching speech.

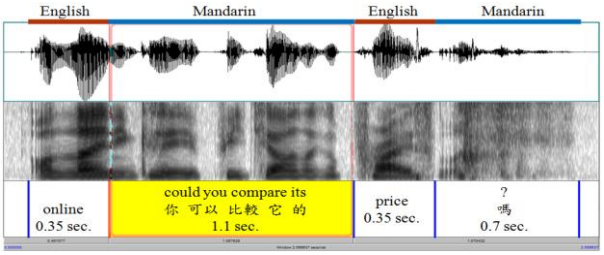

Figure 1. An example of Mandarin-English code-switching utterance, that has three language changes, namely English-Mandarin-English-Mandarin, in 2.6 seconds.

Several code-switching corpora have been reported in the literature, e.g. Cantonese-English, Mandarin-Taiwanese and Mandarin-English code-switching speech corpus [8-10]. Most of the studies are focused on the tasks of language boundary detection (LBD), language identification (LID) and automatic speech recognition (ASR) using bi-phone probabilities or delta-BIC and LSA-based GMMs [10-11]. Some corpora are created by manually interconnecting mono-lingual segments [10], others are read speech following given scripts.

A simple way to enable a monolingual LVCSR system to recognize a bilingual code-switching speech input is to augment the monolingual acoustic model and language model with those of the second language. During decoding, if the system is provided with prior information as to when code-switch occurs, the recognition accuracy of intra-sentential code-switching speech will improve. This motivates us to develop a corpus of spontaneous speech - we hope that the analysis of such a corpus will provide statistics such as lexical triggers, switching probability, and prosodic patterns that may be helpful to LVCSR implementation.

In this paper, we report on the development of a large Mandarin-English code-switching speech corpus, SEAME, collected from residents of Malaysia and Singapore. The corpus developed consists of Mandarin dominant utterances with intra-sentential code-switch between Mandarin and English. SEAME is a thirty hours word-level transcribed speech data with time-aligned language boundary markings.

This paper is organized as follows. In section 2, the approach used for collecting the spontaneous code-switching speech is presented. The transcription and verification issues are discussed in section 3. A series of analysis carried out on the acquired corpus is described in section 4, and finally we conclude in section 5.

2. Corpus Design

The design of SEAME is motivated by a number of considerations. First, the corpus will be used for LBD and LID studies to support the decoding of a multi-lingual LVCSR. A

word-level transcription with language boundary alignment is required. To account for the regional variations, we collected data from two countries: Singapore and Malaysia. As the corpus is developed for spontaneous code-switching speech research, our recordings consist of interviews and conversations without scripts. Furthermore, the recording setting is a quiet office room and the speakers are recorded using close talk microphones.

Table 1 present an overview of the overall statistics with regard to the corpus collected. The sampling rates and resolution for the corpus are 16K Hz and 16-bit per sample respectively. The orthographic transcription is used UTF-8 code.

CS speech	Singapore	Malaysia
number of speakers	77	20
number of utterances	13,810	11,313
number of hours	15	15
speaking rate.	181	157
number of turns	3.1	2.8

Table 1. The overall statistics of intra-sentential code-switching speech corpus collected from Singapore and Malaysia. Speaking rate is measured by number of words per minute.

2.1. Interview speech

The interviewer speech corpus setup is described below: there are two speakers in each interview setup, an interviewer who asks questions and an interviewee who answers. Only the interviewee's speech is recorded using a close talk microphone. In our experience, the design of the questions will greatly influence the amount of interviewee's code-switching response. For example, if the question itself contains two languages, the interviewee will probably use code-switching speech to respond. On the other hand, if the interviewer asks a question only in English, the interviewee answers will most likely to be in English. Hence, questions can be designed to trigger and motivate interviewee to respond with code-switching speech. The questions are also designed to be of interest to the speakers, for example, our topics include hobbies, movies, books, university life, working life, special topics and others. The following are two examples of questions.

- 你 参加 什麼 CCA
  (Which co-curricular activity do you participate in?)
- 谈谈 你 喜欢 的 fruits
  (talk about your favorite fruits)

2.2. Conversational speech

Like interview recording, there are two speakers in conversational speech recording. In this case, each speaker uses an individual close-talk microphone to minimize cross-talk effect. We first suggest some topics for each recording section, e.g. family, school life, sport and relationship. Each of our recording session is about one hour. As the recordings are spontaneous speech, the conversation is mainly informal and non-speech sounds often occur, such as laugh and cough. Furthermore, as most of the speakers are college students who usually speak English in the campus, less intra-sentential code-switching utterances were collected. On average, we can extract about ten minutes of intra-sentential code-switching speech from one hour of speech recording. The rest are either inter-sentential code-switching speech or mono English sentence. In general, the amount of intra-sentential code-switching utterances in an interview method is higher than that in conversational method.

3. Transcription and Verification

The ELAN annotation tool [12] is used to transcribe both interview and conversational code-switching speech. The transcription of each utterance includes language boundary labels and word transcription. Following [13], to describe the transcription of spontaneous speech, additional annotation principles were also developed to classify utterances into the following six categories.

Target speech: this category dictates that the utterance is intra-sentential code-switching speech, and it contains both Mandarin and English segments within one utterance.

Particle: In colloquial speech, additional words such as /ah/, /leh/, and /hoh/, are included in the conversation. These words usually loan word from southern dialects of Chinese are often used to indicate the attitude of speakers. The particles can also borrow from Malay for example '-lah' and '-pun'.

Other languages and dialects: the formers including Malay, Japanese and French are used by some speakers in the conversation. The latter such as Min-Nan and Hakka, which are major dialects of the Chinese language spoken predominantly in southern China, are also used to express speakers' emotion. For example, the word /pai-seh/ in Min-Nan means 'sorry' is a frequent word detected in the corpus. Hence, a tag is used to mark word as dialect or other languages when they occur.

Abbreviation and proper noun: eg. 'CCA', is the abbreviation for co-curricular activity and 'Choa Chu Kang', is the name of a road Singapore.

Colloquialism: for example, 'roomy' and 'sem' are the colloquial speaking style for the words of 'roommate' and 'semester' in the spoken language.

Non-speech signal: it refers to the paralinguistic phenomena such as laugh and cough during recordings.

For the purpose of LID task, each segment has a language tag, either Mandarin or English, but for segments belonging to particle, dialect and non-speech signal, we do not mark the language. Fig. 2 shows an example of a transcription output in a speech segment.

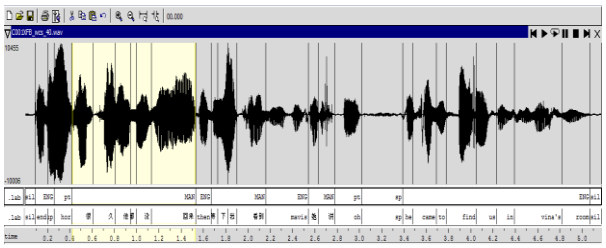

Figure 2. An example of a transcription output in a speech segment

After transcription, we further verify to ensure quality. The procedures of the verification consist of several stages. First, we correct all typos and ensure the boundary of each code-switching segment. After the boundary of the each segment is verified, we extract utterances from the long recording speech. Second, we add new words found in the recordings, such as proper noun, abbreviation and colloquialism, into the pronunciation dictionary to reduce out-of-vocabulary words. Third, each code-switching utterance is force-aligned to determine the language boundary tag. Fourth, we manually adjust language boundary to improve the alignment precision by using WaveSurfer tool [14]. The training data of the

acoustic mode for the forced alignment is the collected code-switching speech and the duration of English and Mandarin training data are 12 and 13 hours, respectively.

The output form of the transcription is described below. For English, we adopt the CMU Pronunciation Dictionary version 0.7. ARPABET is used and the phoneme set includes thirty-nine phonemes without lexical stress. For Chinese, we use simplified Chinese characters encoded with UTF-8 standard. Mandarin phonemes are labeled with Formosa Phonetic Alphabet which is designed for southern Mandarin accent which is similar with Singaporean Mandarin.

4. Code-switching Speech Analysis

Using the collected transcribed corpus, we analyzed the characteristics of Mandarin-English code-switching. As we recorded the corpus from two countries, we separately examine the statistics to gain insights into regional variation. The analysis includes general information of the speaker, such as language preference, characteristics of the intra-sentential code-switching speech.

4.1. Speaker Profile

The average age of the Singaporean speakers is 21.3 (from 18 to 23), while the average age of Malaysian speakers is 25.8 (from 21 to 34). About 90% and 46% of the Singaporean and Malaysian speakers are students respectively. We present in more detail the language preference of Singaporean speakers. For Singaporean speakers, the language preference of the speakers is reported in three language situations, at home, in campus or with friends to speak in Mandarin, English or both, as shown in Figure 3.

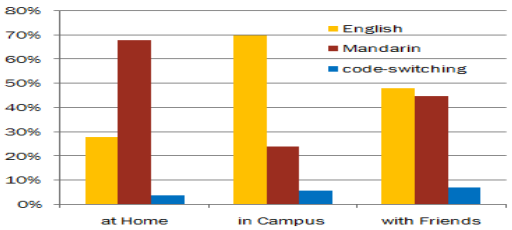

Figure 3. The statistics of language preference of the speakers at home, in campus and with friends.

From Figure 3, we can see that about 68% of the speakers whose main language at home is Mandarin, but 70% speakers whose main language in campus is English. This perhaps is because English is the language of instruction for most subjects in Singapore. As such, we find that the language preference, especially in campus or with friends, influences the extent of code-switching during the conversation. In our observation, the speakers who usually speak in Mandarin in campus could have more intra-sentential code-switching utterances, while others whose main language is English speak more inter-sentential code-switching utterances. In other words, the latter tends to speak the whole long session of mono-lingual speech, e.g. English, then change to another short session of mono-lingual speech, e.g. Mandarin.

4.2. Intra-sentential code-switching speech

We collected 120 hours of speech recordings in total, with 30 and 90 hours being recorded from Malaysia and Singapore, respectively. From this 120 hours speech data, we extracted and transcribed only about 30 hours of intra-sentential code-switching speech, leaving out mono-lingual or non-speech segments. The final amount of the code-switching speech extracted from the two countries is almost equal. However the duration of recordings captured in Singapore is almost three times longer than those from Malaysia. The reason is that both interview and conversational recordings are used for collecting Singaporean Mandarin-English code-switching speech, but only interview recording is used for collecting that in Malaysia. As mentioned previously, more intra-sentential code-switching speech occur in interview recordings.

We collected a total of twenty-five thousand intra-sentential code-switching utterances. Each utterance contains about 12 words on average. Each utterance is roughly a complete sentence. Our analysis showed that speakers do not always speak in full sentences during spontaneous speech. Hence, we consider utterances to be extracted segments of speech that is self-contained semantically or separated by an obvious pause. Figure 4 shows the duration (in seconds) of the extracted utterances.

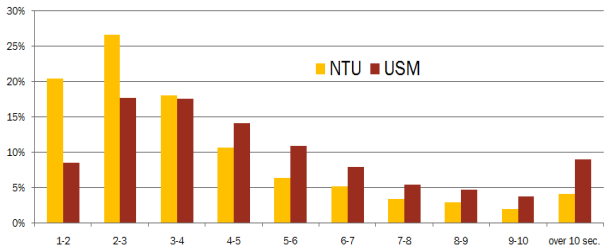

Figure 4. The duration distribution (in seconds) of the intra-sentential code-switching speech from NTU (Nanyang Technological University) and USM (Universiti Sains Malaysia).

Considering code-switch, we are particularly interested in language switching frequency in conversations. From the SEAME corpus, we find that on average, the number of turns of language switch for each code-switching utterance is 2.8 for Malaysian and 3.1 for Singaporean speakers. Take the following utterance as an example. This example has 3 language turn.

Example A:

你们 那些 *guys*, 每次 唱 的 时候, *sing so much louder*

It is straightforward to count the number of English word in each turn, but it is not as easy to do so for Chinese text. For consistency, we first segment a Chinese phrase/segment into lexical words with a forward maximal-length matching algorithm as shown in Example A. In this way, a Chinese word could have one to six characters. We find that about 56% of the words are two-character word, and 32% are one character word. In this example, the second Mandarin segment contains four Chinese words with two, one, one and two characters each, respectively. We report the duration statistics of the English and Mandarin turns in terms of the number of words in figure 5 and 6.

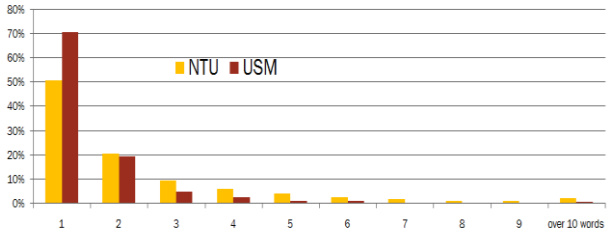

Figure 5. The "number of words" distribution of the English segments.

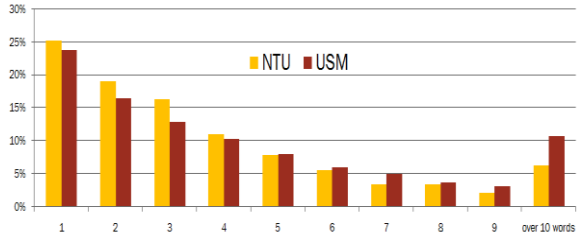

Figure 6. The "number of words" distribution of the Mandarin segments.

Based on above two figures, we can characterize the way of speaking in intra-sentential code-switching speech in Singapore and Malaysia. In general, people tend to switch to English just for one word. This accounts for 50% and 70% of the total sentences from Singapore and Malaysia respectively. This observation of speaking style in code-switching utterance coincides with what are reported in Hong Kong and Taiwan [8-9]. Example A is one of the typical examples of this kind. While we notice that the distribution of word numbers in Mandarin turns seems more even than that in English, one word turns still dominate, representing about 25% of all cases in both countries.

According to the distributions above, we analyze further the language turns to find out the type of lexical words speakers frequently used. We choose one and two word turns in English and Mandarin in the intra-sentential code-switching utterance and report the top ten most frequent words/phrases in Tables 2 and 3 for Singapore and Malaysia speech sources respectively.

	Mandarin		English	
	One word	Two words	One word	Two words
top 1	的	我 就	then	I think
top 2	我	的 时候	for	as in
top 3	你	我 觉得	right	but then
top 4	啊	他 就	but	after that
top 5	那个	的 吗	I	and then
top 6	了	的 那个	take	that's why
top 7	吗	的 人	hall	next semester
top 8	啦	你 可以	so	that means
top 9	就	我 要	one	but actually
top 10	他们	我也 是	the	you know

Table 2. The top ten most frequent words and two-word phrases in English and in Mandarin in Singapore.

	Mandarin		English	
	One word	Two words	One word	Two words
top 1	的	这样 咯	so	first year
top 2	啊	的 那个	OK	and then
top 3	咯	的 东西	for	online shopping
top 4	是	的 时候	but	based on
top 5	我	的 吗	I	I think
top 6	你	的 咯	customer	second year
top 7	了	的 啊	Malaysia	let say
top 8	这样	这样 子	actually	make sure
top 9	的话	的 啦	in	credit card
top 10	然后	来 的	project	of course

Table 3. The top ten most frequent words and two-word phrases in English and in Mandarin in Malaysia.

In Table 2, the two most frequent two-word phrases in English are 'I think' and 'as in', and these phrases usually occur at the beginning of the utterance. On the other hand, for Mandarin, the two-word phrases, "我就" (I just) and "的时候" (of the time) usually occur in the middle of the utterance. Besides, the particles, "啊", "啦" and "吗", are usually uttered after an English word, and this type of speaking style has become a particular characteristics in Singapore [7].

4.3. Annotation scheme

In the annotation scheme, each speaker is given a speaker ID. The ID starts with either 'F' or 'M' for speaker gender and followed by a three-digit integer then end with an alphabet country code, 'A' or 'B', for the nationality, Singaporean or Malaysian, of the speakers. For example, the speaker ID, F012A, refers to the speaker 012, who is a female speaker from Singapore. Besides, the other information of the speaker, such as age and basic language preference are also included in the documents.

5. Conclusion

We have collected 30 hours of spontaneous Mandarin-English intra-sentential code-switching speech from Singapore and Malaysia. We find that Mandarin is a dominant language in the code-switch utterances and the duration of English segment is very short with 50% and 70% in English only having one word in Singapore and Malaysia, respectively. Detecting the change of language, short-segment LID becomes important. In addition, we gain insights from the lexical statistics, that are especially useful when we build code-switch bilingual language model.

6. References

[1] Barbara E. Bullock and Almeida J. Toribio, The Cambridge Handbook of Linguistic Code-switching, Cambridge University Press, 2009.
[2] P. Auer, Code-Switching in Conversation: Language, Interaction and Identity, London: Routledge, 1998.
[3] David C.S. Li, Cantonese-English code-switching research in Hong Kong: a Y2K review, World Englishes, 19-3, 2000.
[4] C.-M. Chen, "Two types of code-switching in Taiwan," paper presented in Sociolinguistics Symposium15, Newcastle, 2004
[5] H. Y. Su "Code-switching between Mandarin and Taiwanese in Three Telephone Conversation: The Negotiation of Interpersonal Relationships among Bilingual Speakers in Taiwan," In Proc. of the Symposium about Language and Society, 2001
[6] Population Trends 2009, http://www.singstat.gov.sg/pubn/popn/population2009.pdf
[7] D. Deterding, Singapore English, Edinburgh: Edinburgh University Press, pp.90-91, 2007
[8] Joyce Y. C. Chan, P. C. Ching and T. Lee, "Development of a Cantonese-English Code-mixing Speech Corpus," In Proc. of Eurospeech, 2005
[9] D.-C. Lyu, R.-Y. Lyu, Y.-C. Chiang and C.-N. Hsu, "Speech Recognition on Code-switching Among the Chinese Dialects," In Proc. of ICASSP, 2006.
[10] C.-H. Wu, Y.-H. Chiu, C.-J. Shia, and C.-Y. Lin "Automatic Segmentation and Identification of Mixed-Language Speech Using Delta-BIC and LSA-Based GMMs," IEEE Transactions On Audio, Speech, And Language Processing, 14-1, Jan. 2006
[11] Y.C. Chan, P.C. Ching, Tan Lee and Houwei Cao "Automatic speech recognition of Cantonese-English Code-Mixing utterances," In Proc. of ICSLP, 2006.
[12] ELAN, http://www.lat-mpi.eu/tools/elan/
[13] S.-C. Tseng "Spoken Corpora And Analysis Of Natural Speech," Taiwan Journal of Linguistics, Vol. 6.2, 1-26, 2008
[14] http://www.speech.kth.se/wavesurfer/.