An Analysis of a Mandarin-English Code-switching Speech Corpus: SEAME

4 authors:

Dau-Cheng Lyu
Nanyang Technological University
23 PUBLICATIONS 326 CITATIONS
SEE PROFILE

Tien-Ping Tan
Universiti Sains Malaysia
54 PUBLICATIONS 311 CITATIONS
SEE PROFILE

Eng Siong Chng
Nanyang Technological University
285 PUBLICATIONS 4,013 CITATIONS
SEE PROFILE

Haizhou Li
National University of Singapore
817 PUBLICATIONS 11,614 CITATIONS
SEE PROFILE

An Analysis of a Mandarin-English Code-switching Speech Corpus: SEAME

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

SEAME (South East Asia Mandarin-English) is a 30 hours spontaneous Mandarin-English code-switching speech corpus recorded from Singapore and Malaysia speakers. In this paper, we report a series of analyses on the recording, processing time and voice activity rate (VAR) of the speech recording, transcription, validation and language boundaries labeling processes. In addition, the duration of the monolingual segment in the code-switching utterance and the analysis of the speakers' behavior in language switching during conversation are also described. The results of the analysis show that 80% and 72% monolingual segments of English and Mandarin in the code-switching utterance are shorter than one second. In over 80% of the cases, speakers directly switch language without any short pause and discourse particle between two adjacent different languages.

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 multiracial countries [1]. For example, the following mixed Mandarin/English utterance: "我 actually 很喜歡 travel" (English: 'I actually like to travel very much') is an example of intra-sentential English-Mandarin code-switching sentence which can be commonly observed from the Chinese population speakers of Singapore and Malaysia. The population of the former consists of Chinese (74%), Malays (13%), Indians (9%) and others (4%) while the latter is made up of Malay (50%), Chinese (24%), Indian (7%), and others (19%) [2-3]. In other bilingual and multilingual societies, such as in United States, Switzerland, Hong Kong and Taiwan, we could often hear code-switching speech of Spanish-English, French-Italian, Cantonese-English and Mandarin-Taiwanese, respectively [4-6]. Code-switching is a common speaking style in daily conversation as it enables people to maintain a sense of social belonging and provide a convenient way to express speakers' idea.

The research into code-switching speech processing is still at its infancy [7-8]. Several code-switching corpora have been built, e.g. Cantonese-English, English-Mandarin and Mandarin-Taiwanese code-switching speech corpora [6-9]. The transcriptions of such corpora have been designed based on real-world spontaneous code-switching speech extracted from the Internet or TV programs. According to [8-9], the purpose to develop code-switching speech corpora is mainly for language boundary detection (LBD), language identification (LID) and automatic speech recognition (ASR) tasks. The commonly used methods for LBD and LID are bi-phone probabilities, delta-BIC (delta-Bayesian information criterion) and LSA(latent semantic analysis)-based GMMs (Gaussian mixture model). For the LID task, it is challenging to achieve high accuracy for code-switching speech because the monolingual part of such utterance may only contain a single word. Previous studies [7, 11] have reported that the length of the monolingual segment in an intra-sentential code-switching speech is usually less than three seconds.

As part of an ongoing project, we are particularly interested in Mandarin/English code switching speech from the South-East Asia region. The corpus development project is therefore codenamed SEAME (South East Asia Mandarin English) [12]. As there is currently no existing corpus of this demography, our first task was to develop a 30 hours corpus from the Singapore and Malaysia population. In addition, as code switching occurs only naturally in conversation rather than in read speech, our corpus is developed from spontaneous speech in conversations and interview settings. In this paper, we report on the statistics such as lexical triggers, switching probability and prosodic patterns that are helpful to LVCSR implementation.

The paper is organized as follows: section 2 presents an overview of the SEAME corpus. Section 3 reports on the voice activity rate (VAR) in the recordings, as well as the transcription, validation and language boundaries labeling process. We define the VAR as the ratio between the actual code-switching speech and the total audio recordings. Section 4 presents an analysis on the length of the monolingual segment in code-switching speech and section 5 reports on the correlation between switching terms and short pause. The conclusions are given in section 6.

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

SEAME contains 97 speakers and approximately twenty-five thousands intra-sentential English-Mandarin code-switching utterances. The speakers took part in the recording are from Malaysia and Singapore.

We have collected audio recordings from 77 speakers of which 90% of these speakers are college students. For these 77 speakers, 68% of them speaks Mandarin at home but 70% speakers speaks mainly English in campus. On the other hand, there are 20 speakers from Malaysia and 62% the speakers whose main language at home is Mandarin and others are Hokkien (21%) and Cantonese (17%). Besides, all the speakers in Malaysia also speak Malay fluently. The other details of SEAME are described in Table 1. The sampling

rates and precision for both corpora are 16K Hz and 16-bit per sample. The orthographic transcription used is UTF-8 code.

Code-switch (CS) speech corpus	Singapore (NTU)	Malaysia (USM)
Number of speakers	77	20
Age group (Mean)	21.3	25.8
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. The speaking rate is defined as the number of words per minute.

This corpus is designed not only for ASR task but also for LID task where the latter is to automatically identify when and what languages are spoken in an utterance. Therefore, we transcribed the lexical items as well as the time stamps of language changes. In order to efficiently label the language boundary, three steps are carried out, namely manual word transcription, automatic forced alignment, and manual language boundary labeling. We have used the following categories for the language boundary labeling:

- *Target languages (ENG and MAN)*: The target languages here refer to either English or Mandarin.
- *Discourse particle and Non-speech signal (DP)*: In colloquial speech, words with the following pronunciation such as /ah/, /leh/, and /hoh/, can often be found in conversational speech. These are usually loan words from Hokkien, and are used to indicate the attitude of speakers. Some of the particles are from the Malay language such as /lah/ and /pun/. For example, like in Malay they can function as imperative marker [13]. For non-speech signal, it refers to the paralinguistic phenomena such as laugh and cough during recordings
- *Other languages (OL)*: some speakers who could also speak Min-Nan and Hakka, which are major dialects of the Chinese language spoken predominantly in southern China, used their familiar dialect words to express their emotion. For example, the word /pai-seh/ in Min-Nan means 'sorry' is a frequent word detected in the corpus. Besides Japanese and Malay have also been detected in the corpus, e.g. /sakura/ and / roti-prata /.
- *Proper noun (PN)*: The proper nouns include examples such as road names /Choa-Chu-Kang/ or names of speakers such as /kew-jin/. As these names can be pronounced either in English or Mandarin, it is not straightforward to identify these segments language ID.
- *Short pause (SP) (Silent segments)*: Only if the duration of the silent segments are longer than 0.15 seconds are these segments labeled as short pause. Otherwise, these segments are merged into the next language segment.

3. Statistics of Data Processing Effort

3.1. Recording

We have recorded two types of speech recording: interview and conversational speech. For the interview settings, we only recorded the interviewee's speech using close talk microphones. The recordings were carried out in Singapore and Malaysia at Nanyang Technological University (NTU) and Universiti Sains Malaysia (USM) respectively. To ensure the same recording environment, the same headset and sound card were used. We recorded 72-hour of speech from Singapore, and 27 hours of speech from Malaysia. The gender is almost balanced and 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).

3.2. Transcription

After the recording, we transcribe the speech by using [14] transcription tool, and extract intra-sentential Mandarin-English code-switching utterances if the utterance contains both Mandarin and English segments and the utterance is self-contained semantically or separated by an obvious pause. In NTU, we transcribed 7.5 hours of code-switching speech from 20.6 hours of interview speech recording - hence the voice activity rate of code-switching utterance is about 36%. On the other hand, the yield of code-switching speech from conversational speech is significantly lower with only about 15%, due to the fact that the interview is more formal than conversational type, therefore, non-speech sounds, such as laugh, cough, which often occur in the conversational speech is absent. This significantly changes the yield of code-switching speech between the two settings.

The evaluation of time spent on transcribing spontaneous speech is an important issue. We monitor the effort in transcribing spontaneous code-switching speech, and report a transcription rate of thirty to forty times real time (xRT). In other words, to transcribe one hour of code-switching speech, we will require approximately thirty to forth hours.

3.3. Verification and Language Boundary Labeling

After the first round of transcription, a verification process is carried out. The procedures of the verification process consist of several stages: Firstly, all typos are corrected and the boundary labeling of each code-switching segment is checked. We then extracted utterances from the long recording speech. Secondly, we added new words found in the recordings, such as proper noun, abbreviation and colloquialism, into the pronunciation dictionary to reduce out-of-vocabulary words. Thirdly, each code-switching utterance is force-aligned to determine the language boundary tag. Fourth, based on the results of forced alignment, we manually adjust language boundary to improve the alignment precision by using WaveSurfer tool [15], an example is shown in figure 1. The training data of the acoustic mode for the forced alignment is the collected code-switching speech.

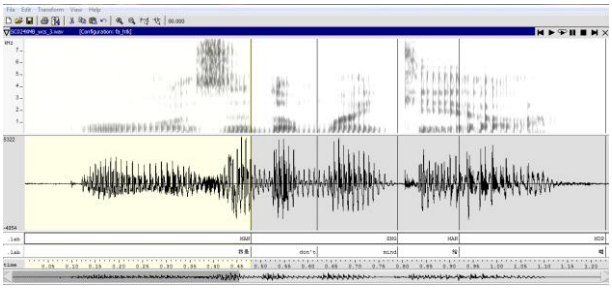

Figure 1. An example of code-switching speech with language boundary labeling (upper) and content (lower) using WaveSurfer tool. The segment with pink color is Mandarin segment, and the length of the segment is about 0.5 second.

We report that the effort on verification and language boundary labeling is about 36 and 35 xRT for NTU and USM data. The details of the time effort in speech recording,

transcription, validation and language boundary labeling are summarized in Table 2.

	NTU		USM
	Interview	Conversation	Interview
Recording	20.6 hours	49.6 hours	27 hours
CS Data	7.5 hours	7.5 hours	15 hours
VAR	36%	15%	56%
Transcription	35 xRT	36 xRT	35 xRT
Validation	7 xRT	9 xRT	7 xRT
LBL	24 xRT	24 xRT	20 xRT

Table 2. The statistics of transcription effort for major steps in developing SEAME corpus. LBL means language boundary labeling.(CS: code-switching)

4. Statistics of Language Durations

In this session, we report the statistics of monolingual segment duration in intra-sentential code-switching speech. As the corpus is collected in two locations, which are NTU, Singapore and USM, Malaysia, our analysis will focus on comparing the different characteristics of Mandarin-English code switching between these two sites.

The ratios of target languages and other categories in NTU data are: Mandarin (42.5%), English (38.3%), short pause (13.3%), discourse particles (5.4%), other language (0.2%) and proper noun (0.3%). The ratios of the categories in USM case are: Mandarin (48.5%), English (20%), short pause (25.5%), discourse particles (5.5%), and other languages (0.5%).

4.1. Durations of Target Languages

The distributions of the duration of the monolingual segments in the utterances are plotted in figure 2 and 3 respectively. The figures show that the duration of the monolingual segments is very short. In fact, about 80% of English and 72% of Mandarin segments are shorter than one second in NTU data, while 90% of English and 63% of Mandarin segments are shorter than one second in USM data.

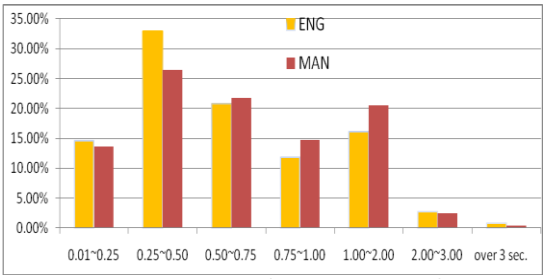

Figure 2. The distribution of the duration of monolingual segments for NTU

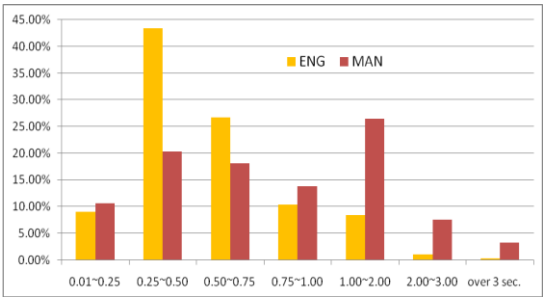

Figure 3. The distribution of the duration of monolingual segments for USM.

4.2. Durations of Other Categories

Table 3 below shows the ratio of different segments. The duration length of the other categories, such as discourse particle (DP), other language (OL), proper noun (PN) and short pause (SP) are illustrated in figure 4 (NTU) and Figure 5 (USM). We can see that, the data collected from NTU, 93% discourse particles, 84 short pause, 78% proper noun and 74% other languages are under only 0.5 seconds. The most case of PN of USM's data are belonging to OL, therefore, the other categories in USM are only three, and they are: DP, OL and SP because we map proper names to their respective language based on the pronunciation used by the speaker. The tendency of the length duration of the other categories in USM is similar with that in NUT. About 90% discourse particles, 80% short pause and 52% other languages are under 0.5 seconds.

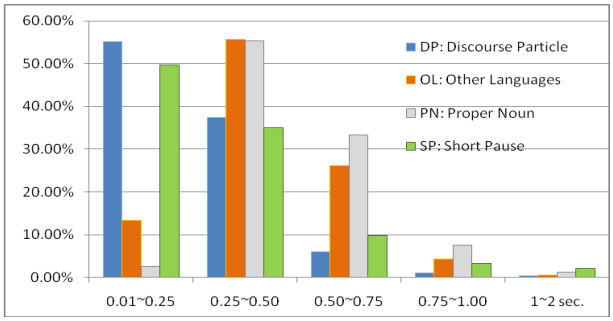

Figure 4. The duration distribution (NTU) of discourse particle (DP), other language (OL), proper noun (PN) and short pause (SP) in intra-sentential code-switching speech with several length levels.

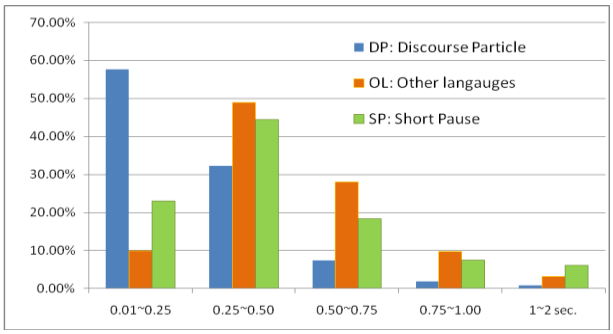

Figure 5. The duration distribution (USM) of DP, OL and SP in intra-sentential code-switching speech with several length levels.

4.3. Automatic vs Manual Language Boundary Labeling

As discussed previously, we used the results from forced alignment as references to human annotator to manually refine the language boundaries. In this section, we will analyze the changes made by human annotators over that of the initial force alignment suggestions for 6 categories. In order to clearly present the differences between the results of forced alignment and those of the manual label correction, we use the two values of the average absolutely changed duration of each category before and after the manual label correction to represent. One is calculated in second, and the other is measured by percentage. For example, if the duration of one Mandarin segment is one second in the forced alignment result, after the manual label correction, the duration of the Mandarin segment is changed to 0.9 second or 1.1 second. Then the average absolutely changed duration measured in second is 0.1

absolute second and calculated in percentage is 10%. Table 3 summarizes our findings.

Categories	NTU		USM	
	sec.	%	sec.	%
ENG	0.106	13.29	0.161	25.82
MAN	0.114	13.57	0.265	22.68
DP	0.139	49.74	0.240	79.85
OL	0.101	22.91	0.218	33.01
PN	0.098	19.42	NA	NA
SP	0.094	17.25	0.153	23.29

Table 3. The change of language boundary duration by human annotator over forced alignment suggestions.

In this table, we can see that the part of ENG and MAN is similar which are about 0.11second (11 frames) difference in NTU case. On the other hand, the length difference in DP part is about 0.14 second, but the length differences in percentage are 50% and almost 80%. The reason is that the most casts (over 50%) of DP length are under 0.25 second (Figure 4 and Figure 5) and this result in DP length variation between forced alignment and manual labeling to be very high. This result suggests that forced alignment performance need to be improved for the DP category.

5. Statistics of Language Turns

In this section, we further analyze the structure of the intra-sentential code-switching speech in its language switching characteristics. First, we investigate the language switching frequency in the corpus. There is an example of the code-switching sentence extracted from the corpus shown below:

他 会 做 一些 adjustment 好像 noise reduction 之類 的 吧
(He will do some adjustment, like the kind of noise reduction.)

Note: The top sentence is the original code-switching sentence, the bottom sentence is the translated version in English.

The above sentence has a Mandarin-English-Mandarin-English and Mandarin code-switching sequence. We consider this utterance to have 4 language turns. In the SEAME corpus, we find that on average the number turns of language switch for each code-switching utterance is 2.8 for Malaysian and 3.1 for Singaporean speakers.

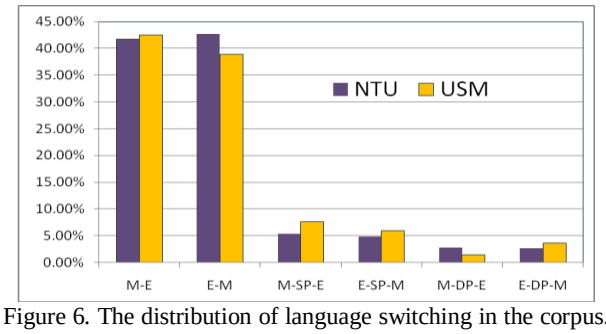

Figure 6. The distribution of language switching in the corpus. M-E = Mandarin-English, E-M = English-Mandarin, DP = discourse particle, SP = short pause.

Second, we investigate the behavior of language switching when speakers talk to each other in a spontaneous speaking style. Our analysis on the behavior of language switching from English/Mandarin, Mandarin/English, with or without a short pause (SP) or a discourse particle (DP) is shown in figure 6. Although we observe that some speakers insert a discourse particle or a shot pause prior to language turn, our analysis show that in most cases (80%), speakers switch languages smoothly. That is, there is no discourse particle or a shot pause, either for Mandarin/English or English/Mandarin turns. The results show that we cannot exploit the detection of SP or DP for language turns.

6. Conclusion

In this paper, we introduced a Mandarin-English code-switching speech corpus in South-East Asia: SEAME and analyzed the corpus in many aspects, such as time processing in developing the corpus, durations in Mandarin and English and behavior of the language turns. Our analysis on the developed corpus shows that 80% and 72% the length of English and Mandarin in code-switching speech is shorter than one second, and 96% monolingual segments which length are shorter than two seconds. Additionally, 80% of language turns occur smoothly without short pause or presence of discourse particle.

7. References

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

[2] Population Trends 2009, http://www.singstat.gov.sg/pubn/popn/population2009.pdf

[3] C. I. Agency, "CIA - The World Factbook - Malaysia," 2010, https://www.cia.gov/library/publications/the-world-factbook/geos/my.html.

[4] David C.S. Li, Cantonese-English code-switching research in Hong Kong: a Y2K review, World Englishes, 19-3, 2000.

[5] C.-M. Chen, "Two types of code-switching in Taiwan," paper presented in Sociolinguistics Symposium15, Newcastle, 2004

[6] Joyce Y. C. Chan, P. C. Ching and T. Lee, "Development of a Cantonese-English Code-mixing Speech Corpus," In Proc. of Eurospeech, 2005

[7] D.-C. Lyu and R.-Y. Lyu, "Language Identification on Code-Switching Utterances Using Multiple Cues," In Proceedings of Interspeech, Brisbane, Australia, Sep. 2008

[8] 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.

[9] C.-H. Wu, Y.-H. Chiu, C.-J. Shia, and Chun-Yu 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

[10] 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.

[11] Hou wei Cao,P.C.Ching and Tan Lee, "Effects of Language Mixing for Automatic Recognition of Cantonese-English Code-Mixing Utterances," In Proceedings of Interspeech, Brighton, U.K., Sep. 2009

[12] Dau-Cheng Lyu, Tien-Ping Tan, Eng-Siong Chng and Haizhou Li "SEAME: a Mandarin-English Code-switching Speech Corpus in South-East Asia," In Proc. of Interspeech 2010

[13] B. Ranaivo-Malacon, "Computational Analysis of Affixed Words in Malay Language," presented at International Symposium on Malay/Indonesian Linguistics, Penang, 2004.

[14] ELAN, http://www.lat-mpi.eu/tools/elan/

[15] Wavesurfe http://www.speech.kth.se/wavesurfer/.