Lang Resources & Evaluation (2015) 49:581–600
DOI 10.1007/s10579-015-9303-x

ORIGINAL PAPER

Mandarin—English code-switching speech corpus in South-East Asia: SEAME

Dau-Cheng Lyu[1] · Tien-Ping Tan[4] ·
Eng-Siong Chng[1,2] · Haizhou Li[1,2,3,5]

Published online: 20 May 2015

Abstract This paper introduces the South East Asia Mandarin–English corpus, a 63-h spontaneous Mandarin–English code-switching transcribed speech corpus suitable for LVCSR and language change detection/identification research. The corpus is recorded under unscripted interview and conversational settings from 157 Singaporean and Malaysian speakers who spoke a mixture of Mandarin and English within a single sentence. About 82 % of the transcribed utterances are intra-sentential code-switching speech and the corpus will be release by LDC in 2015. This paper presents an analysis of the code-switching statistics of the corpus, such as the duration of monolingual segments and the frequency of language turns in code-switch utterances. We also summarize the development effort, details such as the processing time for transcription, validation and language boundary labelling.

✉ Dau-Cheng Lyu
daucheng@gmail.com

Tien-Ping Tan
tienping@usm.my

Eng-Siong Chng
aseschng@ntu.edu.sg

Haizhou Li
hli@i2r.a-star.edu.sg

[1] Temasek Laboratories, Nanyang Technological University, Singapore 639798, Singapore

[2] School of Computer Engineering, Nanyang Technological University, Singapore 639798, Singapore

[3] Institute for Infocomm Research, 1 Fusionopolis Way, Singapore 138632, Singapore

[4] School of Computer Sciences, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia

[5] The University of New South Wales, Sydney, NSW 2052, Australia

Lastly, we present textual analyses of code-switch segments examining the word length of monolingual segments in code-switch utterances and the most common single word and two-word phrase of such segments.

Keywords Code-switching speech · Spontaneous spoken corpus development · Mandarin–English · Speech recognition · Language recognition

1 Introduction

Code-switching occurs when a speaker alternates between different languages in a conversation. It is a common and informal way of communication by multilingual speakers (Bullock and Toribio 2009). As more people become multilingual, code-switching speech becomes more common. For example, in the United States, more people of Spanish origin speak a mixture of Spanish and English. In Switzerland, English and Swiss-German code-switching can often be heard in daily conversation (Auer 1998). In Hong Kong, the young generation Hongkongers often embed English words into colloquial Cantonese due to the influence of its British colonial past (Chan 1992; Chan and Lee 2005). Similarly in Taiwan, Mandarin/Taiwanese code-switching has also recently become popular due to the aim of maintaining a sense of social belonging among speakers from the same culture (Su 2001; Lyu et al. 2006a).

Code switching is also a common speaking style among the multiracial society of Singapore and Malaysia (Kwan Terry 1992). In Malaysia, Malay makes up of about 52 % of the population, followed by Chinese 23 %, Indian 7 %, and the rest from other races (Demographic Transition in Malaysia, Demographic Statistics Division, Malaysia). In Singapore, the ethnic distribution is approximately: Chinese (74 %), Malays (13 %), Indians (9 %), and others (4 %) (Census of Population 2010 Statistical Release 2011; Population Trends 2012). Although different races do communicate using their respective native languages, English is widely used in daily life in both countries (Gopinathan 1998; Deterding et al. 2005). This study focuses on the Mandarin/English code-switching speech among the Chinese speakers in Singapore and Malaysia.

There are many reasons why code-switching speech occurs. According to (Myers-Scotton 1993; Malik 1994; Milroy and Muysken 1995), some possible reasons are as follows: lack of facility, lack of registral competence, semantic significance, to address different audience, to show identity with a group, to amplify and emphasize a point, mood of the speaker, habitual expressions, pragmatic reasons and to attract attention. Another reason is that skillful code switch operates like metaphor to enrich communication without any change in the situation, and hence speakers actively inject meanings into conversation by adding varieties (Su 2001).

Current literatures on code-switch research have the following foci: speaking style (Myers-Scotton 1989), grammar (MacSwan 2013), functions (Reyes 1994) and spoken language processing (Lyu and Lyu 2008; Chan et al. 2004; Shia et al. 2004). In (Myers-Scotton 1989), two types of speaking style of code switching speech were defined: inter-sentential and intra-sentential code-switching. The following shows example sentences of inter-sentential and intra-sentential Mandarin/English code-switching:

- *很容易罷了。* Put the pumpkin in, fried it then add the chicken stock.
  (The literal translation in English is:
  "*It's very easy*. Put the pumpkin in, fried it then add the chicken stock.)
- *我今天想去* Starbucks *買一杯* ice-coffee.
  (The literal translation in English is:
  "*I, today, am thinking of going to* Starbucks *to buy a cup of* ice-coffee")

In (MacSwan 2013), researchers examined the changes in the grammatical structure of a language due to code-switching. In (Reyes 1994), the motivations due to social and linguistic reasons were highlighted, and the authors found that socio-psychological factors as well as the linguistic factors contributed to the predominance of code-switching.

There have also been a number of published works of language identification (LID) and automatic speech recognition (ASR) for code-switching speech (Lyu and Lyu 2008; Chan et al. 2004; Shia et al. 2004). For the LID task, it is challenging to achieve high accuracy on code-switching utterances as the duration of the embedded monolingual segments can be very short (Lyu et al. 2013a). Similarly for the ASR task, the recognition performance remains poor as a code-switch ASR system must take into account the various languages acoustic and language modelling simultaneously. Furthermore, due to the difficulty of developing a spontaneous code-switching corpus, most of these works are based on read speech with pre-defined code-switching transcripts, or they are small scale corpora (Chan and Lee 2005; Lyu et al. 2006a, 2010; Wu et al. 2006; Li et al. 2012). In addition, these corpora are not accessible to the other researchers and hence have limited further study. In this paper, we introduce a large-scale Mandarin/English code-switching corpus (SEAME)—the corpus is named SEAME to reflect its characteristics, namely "South East Asia (SEA)" and "Mandarin/English (ME)" language foci. The corpus can be used for large vocabulary continuous speech recognition (LVCSR) research as well as LID on code-switch sentences.

To highlight the uniqueness of the SEAME corpus, Table 1 lists other known Chinese (Mandarin/Cantonese/Taiwanese) /English code-switching speech corpora (Shen et al. 2011; Chan et al. 2009; Lyu et al. 2006b) for comparison. The following information presented in Table 1 include speaking style, total number of utterances, number of distinct utterances, number of speakers, number of hours, vocabulary size, languages and speaker ages. Some entries are labelled as N/A as no information is publicly available. One unique feature of the SEAME corpus is that the collected speech was unscripted dialog as compared to corpora that were read speech of code-switch sentences, i.e., the speakers were prompted to read code-switch transcripts extracted from newspaper, magazine, internet blogs or radio stations. The reason why scripted read speech was recorded as opposed to having recording of spontaneous code-switching speech data is attributed to the fact that significantly more cost and time would be required to develop such a corpus. The read-speech corpus however may not truly reflect the true nature of code-switching

Table 1 A summary of reported Chinese (Mandarin/Cantonese/Taiwanese)/English code-switching corpora

Ref. of corpus	Speaking style	Number of utterances (distinct)	Number of speakers	Number of hours	Vocabulary size	Code-switch language	Speaker ages
Shen et al. (2011)	Read	6650 (1321)	77	12.1	2000	Mandarin English	20–22
Chan et al. (2009)	Read	11,740 (4343)	74	11.76	NA	Cantonese English	Avg. 22
Lyu et al. (2006b)	Read	3000 (300)	12	3.1	1500	Taiwanese Mandarin	18–24
SEAME	Dialog	52,145 (49,001)	157	63	15,700	Mandarin English	18–34

speech characteristics and hence the SEAME corpus offers an opportunity to study code-switch speech under more natural settings in a larger scale than previously possible. SEAME corpus involved 157 speakers, consisting of a total of 63 h of transcribed conversational speech recordings, with a vocabulary size of over fifteen-thousand words.

To the best of our knowledge, the SEAME corpus is the first large-scale spontaneous Mandarin/English code-switching speech corpus. It will be published by Linguistic Data Consortium (LDC) for research on code-switch speech recognition and related topics. SEAME offers a good coverage of speakers and speaking conditions for robust acoustic modeling, such as Hidden Markov Models (HMMs), subspace Gaussian mixture model and Deep Neural Network (DNN) acoustic models (Young 1996; Rabiner 1989; Povey et al. 2011; Hinton et al. 2012). Our preliminary experience on this corpus reports a word error rate (WER) of 36.6 % for a code-switch LVCSR in (Vu et al. 2012) which suggests that code-switch LVCSR remains a challenging research problem.

In addition, the SEAME corpus may also be attractive to linguists who study and analyze code-switching conversational speech. For example, (Li 1998) surveyed syntactic and morphological patterns in code-switch speech, (Sankoff and Poplack 1981) studied specific grammatical rules and specific syntactic boundaries for where code-switching might occur, (Myers-scotton and Myers 1993) investigated the structure of code-switching and analyzed the interaction between speakers. The SEAME corpus hence provides a new source for such research.

The organization of this paper as follows: Sect. 2 presents an overview of the SEAME corpus and speakers' profile. Section 3 reports the various processes taken to develop the corpus. Section 4 describes the corpus characteristics such as: length of collected utterances, duration of monolingual speech segment in intra-sentential code-switching speech, frequency of language switch and language turns. Section 5 presents the conclusions.

2 The SEAME corpus

The section presents an overview of the SEAME corpus and then highlights the speakers' profile.

2.1 Corpus overview

The SEAME corpus is collected at Nanyang Technological University (NTU, Singapore) and Universiti Sains Malaysia (USM) from 2009 to 2010. The corpus is collected mainly from students and staff of the Universities. There are a total of 157 distinct speakers in 174 h of recordings. From these recordings, 63 h were transcribed. Table 2 lists the following statistics of the corpus: number of speakers and gender, the number of utterances, the number of hours and number of vocabulary collected from the Singapore and Malaysia site. The total number of distinct words is over fifteen thousands, the numbers of English and Mandarin words are 8415 and 6923 respectively.

Table 2 Overall statistics of the SEAME corpus

SEAME ALL	Singapore		Malaysia Interview	Total
	Conversation	Interview		
Number of distinct speakers (F: Female, M: Male)	61 (F: 56%, M: 44%)	67 (F: 54 %, M: 46 %)	29 (F: 48 %, M: 52 %)	157 (F: 53 %, M: 47 %)
Number of utterances	13,112	19,586	19,447	52,145
Number of hours transcribed	11.5	22.7	28.8	63
Number of distinct vocabularies (English, Mandarin)	7160 (Eng: 4543, Man: 2617)	9864 (Eng: 5606, Man: 4258)	9376 (Eng: 4469, Man: 4907)	15,338 (Eng: 8415, Man: 6923)

Table 3 Statistics of the SEAME corpus by type: code-switch utterances, Mandarin-only utterances, or English-only utterances

Type of utterances	No of hours and ratio	Singapore		Malaysia Interview	Total
		Conversation	Interview		
Code-switch utterances	Number of hours	9.9	18.7	23.1	51.7
	(ratio)	(86%)	(82%)	(80%)	(82%)
Mandarin-only utterances	Number of hours	0.3	1.7	5.4	7.4
	(ratio)	(3%)	(8%)	(19%)	(12%)
English-only utterances	Number of hours	1.3	2.3	0.3	3.9
	(ratio)	(11%)	(10%)	(1%)	(6%)
Overall	Number of hours	11.5	22.7	28.8	63

The breakdown by language type, i.e., by classifying each sentence as either intra-sentential code-switch, English-only or Mandarin-only utterances is listed in Table 3. As the focus of the corpus was to examine code-switching speech, most of the transcribed utterances were of this type. Specifically, 82 % of the sentences are intra-sentential code-switching speech followed by Mandarin-only (12 %) and English-only (6 %) sentences. We will continue to transcribe the rest of the corpus and release it to LDC in subsequent updates in future.

2.2 Speakers' profile

The speakers who took part in this data collection effort are typically undergraduate students or staff working in the two universities. The Singaporean speakers are of ages 18–23, and 95 % of them are NTU's students. The Malaysian speakers are of ages 21–34, and 46 % of them are USM students. The gender is almost balanced. The Singapore speakers are asked to state their preferred language of communication in two locations, home and campus. The language choices are: English, Mandarin or code-switch (Mandarin/English). Figure 1 shows the survey results: about 69 % of the speakers use

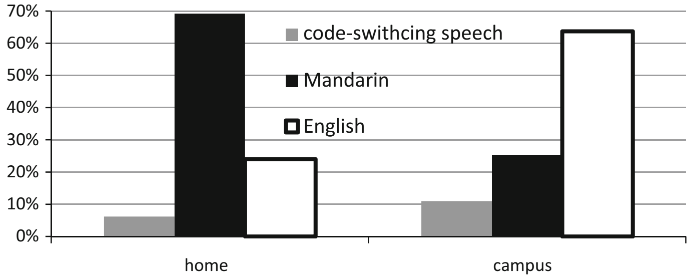

Fig. 1 The Singaporean speakers' language type preference at home and in campus

Mandarin at home, but 64 % speakers use English in campus. This is because English is the medium of instruction, while Mandarin is the speakers' mother tongue. Thus, speakers typically speak Mandarin with their family members but English to their classmates. From our observations, we find that language preference influences the extent of code-switching during the conversation. Speakers who prefer to speak Mandarin in campus produce more intra-sentential code-switching utterances, i.e., typically embedding pockets of English words in Mandarin dominated utterances. While speakers who prefer to speak English at home produce more inter-sentential code-switching utterances, i.e., these speakers speak whole utterances in English, followed by complete short segments of Mandarin.

3 Corpus development

This section describes the various processes taken to develop the corpus: speech recording, manual word-level transcription, process of validation and information for corpus release. The time and effort spent for the corpus development are also summarized.

3.1 Speech recording

All the recordings were recorded in a quiet room. The speech signal were recorded using close talk microphone and sampled at 16 kHz with 16-bit wav format. We recorded over 100 h of raw speech from 128 speakers in Singapore, and over 50 h of raw speech from 29 speakers in Malaysia. To have spontaneous code-switching speech, we have two recording setup: either conversation or interviews between two speakers. In Singapore, both setups were recorded. In Malaysia, only the interview speech setup was recorded. The details of these two setups are described below.

3.1.1 Conversational speech recording setup

To record conversational speech, two speakers were asked to sit facing each other in a quiet room and allowed to speak freely. Their conversations were recorded using individual close-talk microphones. The topics of discussion include: family, school life, vacation, sports and friendship. Each recording session was about 1 h. As the recordings were spontaneous speech, the conversation was informal and non-speech sounds such as fillers, laughter and cough occur often. In such recording session, each speaker spoke approximately half the time. Hence the amount of available useful audio for transcription is low. On average, we extracted only about 10 min of code-switching speech from 1 h of speech recording per speaker as some of monolingual sentences and in-complete sentences; most of silence, laughter and non-speech sounds were not used.

3.1.2 Interview speech recording setting

In the interview recording setup, the two speakers, the interviewer and interviewee, were placed in a quiet room. In this setting, only the interviewee's speech was recorded using a close-talk microphone. The interviewer's questions were designed to be non-sensitive with discussion topics such as: hobbies, movies, books, university life, part time job, etc.

From our experience, we observe that the design of the questions can have a big influence on the amount of interviewee's code-switching response. For example, if the questions include code-switching, then the interviewee will probably follow suit in their response. The followings are two examples of questions asked by the interviewer:

Example 1 *你有去过哪些* countries? (Literal translation: *You have gone to which* countries?)
Example 2 Talk about your national service *的經驗* (Literal translation: Talk about your national service *experience*)

Hence, there is more code-switching speech from interview speech recording as compared to conversational speech recording. In addition, as the discussion is mainly dominated by the interviewee, the amount of code-switch utterances from 1 h speech recording is quite high—almost 30 min of code-switching speech could be extracted from 1 h of speech recording.

3.2 Word-level transcription

The ELAN (http://www.lat-mpi.eu/tools/elan/) and Praat (Ćhttp://www.fon.hum.uva.nl/praat/) annotation tools were used for transcription. We extracted utterances that are self-contained semantically or separated by an obvious pause. Both code-switch utterances and some mono-lingual (English/Mandarin) utterances were selected and transcribed.

Table 4 provides the details of the number of recording hours as well as providing a summary of the time spent to develop the SEAME corpus—the time spent on speech recording, transcription, validation and meta-inflation boundary labelling effort. We observed that the yield of useful extracted speech data from conversational settings is significantly lower than interview settings. One reason why there is more code-switching speech under the interview settings is that the

Table 4 The statistics of the SEAME corpus and transcription effort

	Singapore Site		Malaysia Site
	Conversation	Interview	Interview
1. Raw speech recording	68 h	54 h	52 h
2. Extracted speech data (monolingual + code-switch utterances)	11.5 h (17 %)	22.7 h (42 %)	28.8 h (55 %)
3. Transcription time	34 RT	32 RT	32 RT
4. Validation time	7 RT	6 RT	6 RT
5. Meta-information boundary labelling time	20 RT	20 RT	20 RT

interviewer's role in the conversation is minimized. This is as opposed to the conversational settings in which about 50 % of each recording channel is silence as speakers take turn to speak. In addition, for the conversation speech recordings, overlapping speech (two speakers simultaneously speak) and incomplete sentences are not transcribed. The transcribed recordings for the conversational speech setting is approximately 17 % of raw speech as opposed to over 42 % for interview speech of Singaporean data and 55 % for interview speech of Malaysian utterances.

The effort to transcribe the collected corpus is 34 times real time (RT). In other words, to generate 1 h of transcribed speech, approximately 34 manpower-hours is required. There are two levels of information provided in the transcription, namely, one is the verbatim transcription at word level, and the other provides meta-information at the segment level. The word level information supports the LVCSR task and the meta-information can be used to support LID research. Figure 2 shows the screen shot of the transcription tool capturing these two levels of information.

The meta-information labelling consists of the followings:

(a) Language ID (*ENG/MAN)*: The target languages are either English or Mandarin.

(b) Other: Segments containing words from other languages will have this segment tagged as "Other". E.g, in Fig. 2, the word "angbao" (literal translation: red

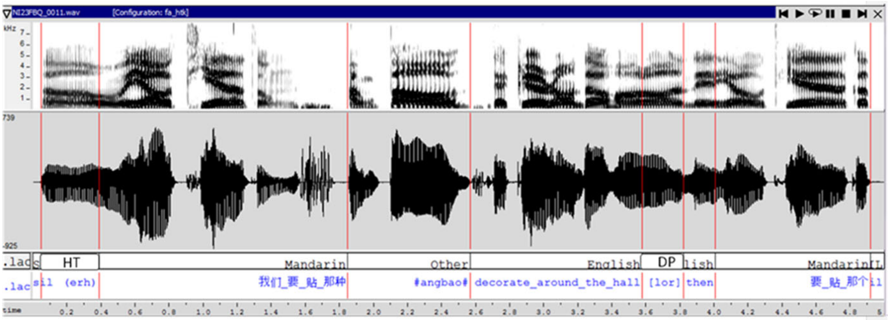

Fig. 2 An intra-sentential Mandarin–English code-switching utterance that shows the waveform, spectrogram, meta-information of segment and word-level transcription

envelop) is from the “Hokkien” dialect. Hokkien is a major dialect spoken predominantly in southern China and Singapore.

(c) Discourse particle and non-speech signal (DP): In colloquial speech, discourse particles with pronunciations such as [ah], [leh], and [hoh] are often found in conversational speech in Singapore and Malaysia. Some particles are originated from Malay, e.g. -lah [lah] and -pun [pun]. For the Malay language, particles can be appended to a based word and can function as imperative marker. the non-speech signal tag refers to the paralinguistic phenomena such as laugh and cough during recordings.

(d) Hesitation (HT): Sounds such as (erh), (mm) and (erm) are commonly found in our conversational speech recording and are tagged as hesitation. These sounds are often used to signal continued attention, agreement, and various emotional reactions to the other speaker’s speech and are also known as backchannel.

(e) Short pause (SIL) (Silent segments): Silent segments longer than 0.15 s are labeled as short pause. Otherwise, these segments are merged into the next language segment.

3.3 Verification process

To improve the accuracy of the transcription, a validation process consisting of several stages was carried out. Firstly, all typos and boundary label of each code-switching segment were checked. Secondly, new words found in the recordings, such as proper noun, abbreviation, colloquialism and other language’s words, were included into the pronunciation dictionary to support LVCSR research. The details of pronunciation dictionary generation are described in (Lyu et al. 2013b). Thirdly, the Chinese transcriptions were processed by the Stanford Word Segmenter (Tseng et al. 2005) to generate lexical words. We found that about 56 % of the words are two-character word, and 32 % are one character word.

An example of a Mandarin/English code-switching speech is shown in Fig. 2. The transcription of the speech is “(erh) 我们_要_贴_那种 #angbao# decorate_around_the_hall [lor] then 要_贴_那个” . For this sentence, under the meta-information tier, the following segments are labelled as: SIL, hesitation (HT), Mandarin, Other, English, discourse particle(DP), English, Mandarin and SIL. Ignoring silence, hesitation, other and discourse particles, the language turns for this segment will be Mandarin/English/Mandarin.

3.4 Corpus release

The SEAME corpus will be released though LDC (https://www.ldc.upenn.edu/) in 2015. It includes audio files, transcriptions and speaker information. The audio files are the whole sessions of interview and conversational speech recordings. Word level transcription with time stamp is provided. Information about the speaker such as speaker ID, gender, age and nationality is also included. An example of the transcription is shown in Table 5. The first column is speech recording ID, the

Table 5 Examples of the transcription

Speech recording ID	Start time	End time	Word-level transcription
13NC26MBQ	290295	293585	er 明天 只是 去 只是 flicks 而已 啦 没有 什麼 东西 的
13NC26MBQ	294595	298785	哎呀 没有 机会 玩 的 啦 最多 我们 下去 那边 show face
13NC26MBQ	302615	305315	ya 啦 politics 可以 这种 是 politics 来 的 right
13NC26MBQ	308665	310875	right 老实 讲 老实 讲 erm
13NC26MBQ	311085	313715	那时 wen-kai 出来 他 sub 出来
13NC26MBQ	320055	322005	as in 他 try to 给 他 玩 but
13NC26MBQ	322165	324235	讲 而已 lo but 基本上
13NC26MBQ	324335	327595	第一 场 的 时候 我们 when 我们 lead two 二 零 的 时候

second and third columns are start and end time of one utterance with time unit in millisecond. The last column is the word-level transcription.

4 Code-switching speech analysis

This section reports on the analysis of the SEAME corpus in term of duration of segments from monolingual speech and intra-sentential code-switching utterances. The comparison of meta-information boundary labelling using manual and automatic methods and the analysis of language turns are also described.

4.1 Duration of utterances

We examine the duration of monolingual utterances as well as the duration of the speech segment within intra-sentential code-switching utterances.

4.1.1 Duration of all utterances

In the first release of SEAME, we only selectively transcribed segments of speech if it is self-contained semantically or separated by an obvious pause. We shall call these transcribed segments as utterances even though the speakers sometimes do not express themselves using complete sentences. The total number of transcribed utterances is over 52 thousands and can be broken down into three classes of utterances based on their language type. The three classes are: intra-sentential code-switch utterances, Mandarin-only monolingual or English-only monolingual utterances. Specifically, there are 37 thousands intra-sentential code-switching utterances, about 10 thousands Mandarin-only utterances, and 5 thousands English-only utterances. The numbers of monolingual utterances are far less than code-switching utterances as some monolingual segments were not transcribed in this round of corpus development—the focus was on code-switching speech.

The average duration of the three classes of sentences are shown in Table 6. We have the following conclusions when analysing the table: First, the average duration of intra-sentential code-switching utterances is longer than monolingual Mandarin-

Table 6 The average duration of utterances from Singapore and Malaysia on interview and conversational speech recording

Average duration (sec.) (words) (WPS: word per sec.)	Singapore							Malaysia	
	Conversation				Interview			Interview	
	Seconds	Words	WPS	Seconds	Words	WPS	Seconds	Words	WPS
Code-switch utterance	3.2	11	0.29	4.3	14	0.31	5.3	16	0.33
Mandarin-only utterance	2.1	7	0.32	3.1	9	0.36	3.9	7	0.56
English-only utterance	2.6	8	0.33	3.1	8	0.39	4.0	11	0.36
Average	2.9	10	0.29	3.9	12	0.31	4.9	15	0.33

WPS word per second

only and English-only utterances in all recording settings. Second, the average number of words in intra-sentential code-switching utterances spoken by Singaporean speakers is 11 and 14 while Malaysian speakers is 16. It appears that Singaporean speakers prefer short utterances as compared to Malaysian speakers. In addition, the speaking rate of the Singaporean speakers is slightly faster than Malaysian speakers. Not surprisingly, we also note that the speaking rate in the conversational setting is faster than interview setting.

4.1.2 Duration of intra-sentential code-switching utterance

Figure 3 shows the duration (in seconds) of the intra-sentential code-switching utterances. The average duration of interview speech in Singapore and Malaysia is 4.3 and 5.3 s, respectively. The average duration of conversational speech is shorter than interview speech, and it is only 3.2 s. We find that the duration of intra-sentential code-switching utterances is very short—almost 75 % of our collected utterances' duration length is less than 5 s. We also notice that the average duration of an utterance produced by Singaporean speakers is shorter than those by Malaysian speakers. Most of the utterances produced by Malaysian speakers are 3–6 s long. On the other hand, utterances by Singaporean speakers are 1–3 s long.

We observe that conversational utterances are shorter because the two speakers change speaking turns more frequently in conversational setting than in interview setting. In interview speech recording, there is usually little of no interruption from the interviewer when the interviewee speaks. Hence, the interviewee's speech is typically longer than that of conversational speech. As for the longer duration of interview speech for Malaysian speakers than Singaporean speakers, we observe a slower speaking pace of Malaysian speakers and the usage of more words to express themselves.

4.1.3 Duration of monolingual speech segment

This subsection provides an analysis of the duration of the monolingual segment in code-switch utterances. The meta information of the corpus contains the language

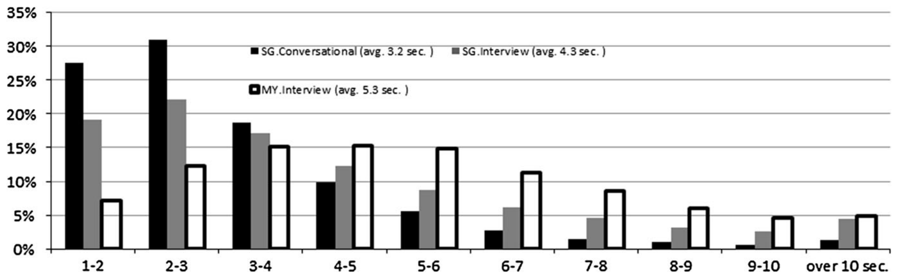

Fig. 3 The duration distribution (in seconds) of the intra-sentential code-switching speech from Singapore (SG) and Malaysia (MY) on interview and conversational speech recording

identity of the speech segment time-aligned to the transcription. The time stamp of the language identity is found by using an ASR to perform force-alignment on the word transcription and then these hypothesis are manually verified using the WaveSurfer (http://www.speech.kth.se/wavesurfer/) tool. The language meta-information include tags such as English, Mandarin and others (silence, hesitation, discourse particles and other languages). As the precision of the alignment is important, this section analyze the differences between the results of forced alignment and those by manual label correction. I.e, we compare the errors between the force alignment results to the manual label results and record their differences in time-units (seconds) as well as by their percentages. For example, given a force alignment result of a segment of one second, and a manual label correction which resulted in its duration being changed by 0.1 s, e.g. to either 0.9 or 1.1 s, the absolute change recorded is 0.1 s and the relative change is recorded as 10 %. Table 7 summarizes our findings, we see that the difference between manual and automatic language boundary process is around 10 %. It is clear that manual label correction of timing information must be carried out to ensure good language boundary tagging.

The duration of the monolingual segment affects LID performance. According to previous studies, the accuracy of the LID system is proportional to the duration of the utterance (Lyu et al. 2013b; Zissman 1996; Li et al. 2013; Santhosh Kumar et al. 2010). For example, in (Lyu et al. 2013b), the equal error rates of LID on 0.1–0.5, 0.5–1, 1–3, and 3–9 s. monolingual speech segments of intra-sentential code-switching speech are 17.3, 13.1, 8.1 and 3.8 % respectively. In (Ćhttp://www.

Table 7 The change of language boundary duration by human annotator over forced alignment suggestions

Languages	NTU				USM	
	Conversation		Interview		Interview	
	Seconds	%	Seconds	%	Seconds	%
Mandarin	0.07	10	0.08	10	0.09	10
English	0.08	10	0.08	11	0.06	12

speech.kth.se/wavesurfer/), the equal error rates of language identification task for 3 and 30 s monolingual speech are about 15 and 8 % respectively.

The distributions of the segment duration are shown in Fig. 4a–c. The figures show that the duration of the monolingual segments in a code switching utterance is very short—about 67 and 72 % of all segments made by Singaporean and Malaysian respectively are less than one second. In summary, the averages duration of monolingual speech segment of English and Mandarin from Singaporean data are 0.73 and 0.78 s for conversational speech, and 0.82 and 0.72 s for interview speech. The average monolingual speech segments of English and Mandarin from Malaysia data are 0.53 and 0.85 s respectively. These short segments contain only one to two words.

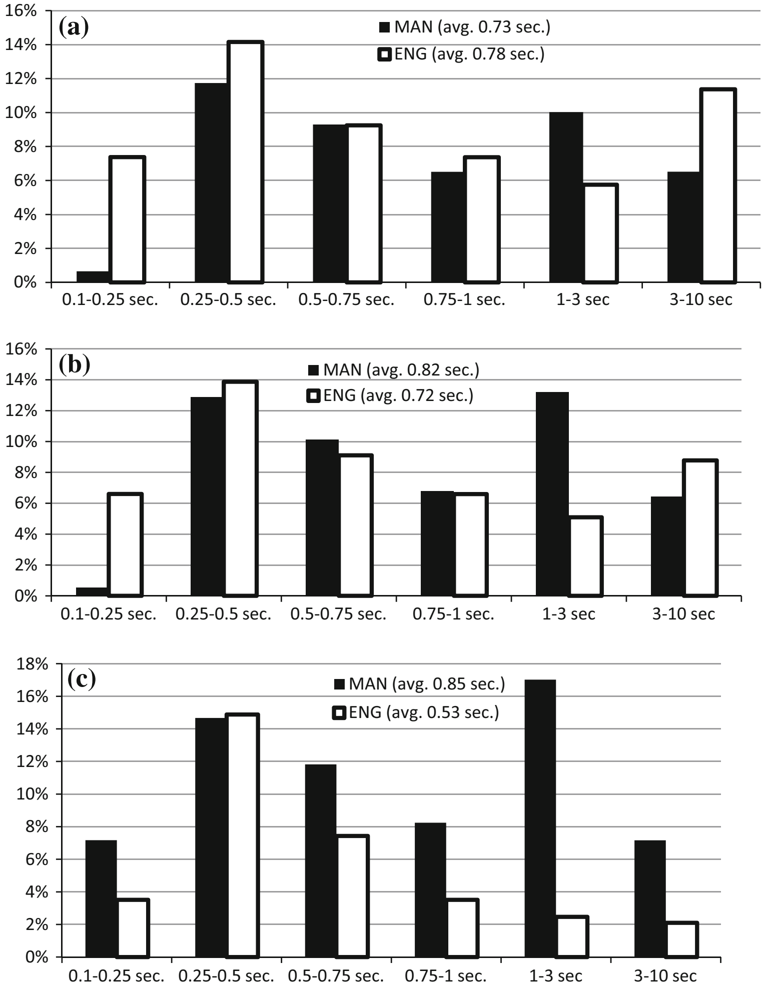

Fig. 4 **a** The distribution of the duration of monolingual segments for conversational speech in Singapore. **b** The distribution of the duration of monolingual segments for interview speech in Singapore. **c** The distribution of the duration of monolingual segments in Malaysia interview speech data

The ratios of Mandarin to English in Fig. 4b and c reveal a distinct difference in language preference among the speakers in Singapore and Malaysia, even though the speakers' profile such as age, gender and educational level are similar. The results show that Malaysian speakers prefer to use Mandarin compared to English since most of the average Mandarin segments are longer than the average English segments. The ratios of Mandarin segments and English segments in Malaysian speakers are about 66 and 34%, respectively, while the ratios of Mandarin and English segments in Singaporean speakers are balanced.

4.2 Language turns

This section focuses on the language turns and trigger words prior to language turns in intra-sentential code-switching utterance. The former is used to analyze language switch tendency and the latter is to find out the type of words that may trigger language switch. The details are described in the following subsections.

4.2.1 Language turns in speech

This section analyzes the frequency of language turns in intra-sentential code-switching utterance. The example in Fig. 2 has a Mandarin–English–Mandarin switching sequence, i.e. it has 2 language turns. Our analysis of the SEAME corpus shows that the average number of language switches is 1.9 for conversational speech, and 2.1 for interview speech for the Singaporean speakers. For the Malaysian speakers under interview speech, this value is 2.4. In other words, the average number of language turns in intra-sentential code-switching utterance for Singaporean is lower than Malaysian speakers. This may be due to the fact that Malaysian speakers' preference or higher competence in using Mandarin in their speech (Fig. 4c), and they only switch to English whenever necessary, therefore producing more languages switches and shorter duration of monolingual language segments as compared to the Singaporean speakers.

4.2.2 Language turns in text

In the previous section, we analyze the language turns in intra-sentential code-switching speech. In this subsection, we analyze the textual information of language turns. The average number of words in the intra-sentential code-switching utterances of the 3 types of data is: 11 (Sg conversational), 14 (Sg interview) and 16 words (Malaysia interview).

Our analysis shows that it is common to switch to another language with a single word before switching back to the original language in an intra-sentential sentence. This accounts for 50 % (Singapore) and 70 % (Malaysia) of all intra-sentential code-switching utterances.

To determine the type of lexical words speakers frequently switch to, we consolidate segments that consist of only one or two words in the intra-sentential code-switching utterance and list the top ten most frequent words/phrases in Tables 8 and 9 for Singaporean and Malaysian speech sources respectively.

In addition, we examine the text to list the top 30 frequent code-switching points with one word in both languages for Singaporean and Malaysian data in Tables 10 and 11. Table 10 shows that “then” and “like” are the two most frequent words in the change points of code-switching speech for Singaporean speakers. For Malaysian speakers, “then” and “so” are two most frequent words of language turns for either Mandarin to English or English to Mandarin language switching. These words are conjunctions. This suggests that speakers like to switch languages before or after conjunction words. In Table 11, the most frequent Chinese word in language change point is “的” and followed by “我” (I), “你” (you) and “他” (he/she) where the word “的” in part of speech could be classified into many types such as a preposition, an adverb, an auxiliary verb, a suffix or a particle depending on the context.

In this section, we have analyzed the SEAME corpus in two aspects. The first aspect is the duration of all utterances and the duration of monolingual speech segment in intra-sentential code-switching speech. The second is about the occurrence of language turn. The findings in this analysis provide important clues to research in LID and LVCSR development of Mandarin/English code-switching speech. Specifically, the duration of monolingual speech segments in intra-sentential code-switching speech highly influence the performance of LID (Lyu et al. 2013b; Zissman 1996; Li et al. 2013). The analysis of language turn provides knowledge of how to develop language model for code-switch LVCSR system (Lyu et al. 2013a).

Table 8 The top ten most frequent words and two-word phrases in English and in Mandarin in Singaporean data

	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

The Mandarin words such as “那个”, “时候”, “觉得”, etc consisting of two syllables are considered as a single Chinese word in our segmentation

Table 9 The top ten most frequent words and two-word phrases in English and in Mandarin in Malaysian data

	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 10 The top 30 frequent language switch from Mandarin to English and English to Mandarin in Singaporean data

	MAN–ENG	ENG–MAN		MAN–ENG	ENG–MAN
Top 1	的 then	then 我	Top 16	东西 then	think 我
Top 2	了 then	then 你	Top 17	的 but	like 很
Top 3	住 hall	then 他	Top 18	很 interesting	then 然后
Top 4	在 hall	then 我们	Top 19	我 I	so 你
Top 5	是 like	then 就	Top 20	的 so	so 我们
Top 6	就 okay	then 他们	Top 21	因为 I	think 是
Top 7	就 like	but 我	Top 22	还 okay	that 我们
Top 8	的 right	so 我	Top 23	那个 bus	N.T.U. 的
Top 9	因为 like	then 那个	Top 24	的 like	then 他们就
Top 10	会 like	that 我	Top 25	有 like	interesting 的
Top 11	要 take	hall 的	Top 26	我们 like	like 我
Top 12	那种 like	actually 我	Top 27	什么 then	but 你
Top 13	去 like	like 你	Top 28	在 N.T.U.	then 去
Top 14	很 sad	but 他	Top 29	是 for	then 很
Top 15	我 join	then 有	Top 30	所以 like	then 我的

5 Conclusion

In this paper, we have presented the development and analyses of the SEAME Mandarin/English spontaneous code-switching speech corpus collected in Singapore and Malaysia. The uniqueness of this corpus is due to it being a true un-

Table 11 The top 30 frequent language switch from Mandarin to English and English to Mandarin in Malaysian data

	MAN–ENG	ENG–MAN		MAN–ENG	ENG–MAN
Top 1	在 Malaysia	then 我	Top 16	的 area	so 他们
Top 2	的 then	then 你	Top 17	就 like	that 我
Top 3	在 U.S.M.	then 他	Top 18	在 Penang	for 那个
Top 4	的 so	then 过后	Top 19	是 under	free 的
Top 5	了 then	so 我	Top 20	这个 course	okay 我
Top 6	很 stress	then 我们	Top 21	东西 then	so 他
Top 7	的 course	so 我们	Top 22	那个 server	but 他
Top 8	做 research	then 就	Top 23	一个 project	year 的时候
Top 9	了 so	so 你	Top 24	就 okay	management 的
Top 10	就 okay	Malaysia 的	Top 25	去 try	research 的
Top 11	在 K.L.	then 他们	Top 26	是 more	for 我
Top 12	去 K.L.	then 那个	Top 27	是 okay	that 你
Top 13	讲 okay	basic 的	Top 28	的 project	so 如果
Top 14	还 okay	but 我	Top 29	我们 Malaysia	but 你
Top 15	的 product	U.S.M. 的	Top 30	是 U.S.M.	okay 你

scripted spontaneous recording of speech, it is a relatively large corpus, has good speaker diversity, abundant vocabulary, as well as recording in 2 different countries. This corpus can be acquired through LDC for research into LID and Mandarin-English multilingual LVCSR for spontaneous speech.

In addition, we have carried out a detail analysis of the corpus in terms of duration of code-switching speech, length of monolingual speech segments, statistics of language turns in intra-sentential speech, as well as listing the top 30 frequent words of language turns in both Mandarin to English and English to Mandarin. Our analysis uncovers interesting differences between the Singaporean and Malaysian speakers.

References

Auer, P. (1998). *Code-switching in conversation: Language, interaction and identity*. London: Routledge.

Bullock, B. E., & Toribio, A. J. (2009). *The Cambridge handbook of linguistic code-switch*. Cambridge: Cambridge University Press.

Census of Population 2010 Statistical Release 1. (2011). Demographic Characteristics, Education, Language and Religion, Department of Statistics, Ministry of Trade & Industry, Republic of Singapore. January 2011.

Chan, H. S. (1992). Code-mixing in Hong Kong Cantonese–English bilinguals: Constraints and processes. M.A. Thesis, The Chinese University of Hong Kong, Hong Kong.

Chan, J. Y. C., Ching P. C., & Lee, T. (2005). Development of a Cantonese–English code-mixing speech corpus. In *Proceedings of Eurospeech*.

Chan, J. Y. C., Cao, H., Ching, P. C., & Lee, T. (2009). Automatic Recognition of Cantonese-English Code-Mixing Speech. *Computational Linguistics and Chinese Language Processing, 14*(3), 281–304.

Chan, J. Y. C., Ching, P. C., Lee, T., & Meng, H. M. (2004) Detection of language boundary in code-switching utterances by bi-phone probabilities. In *Proceedings of the international symposium chinese spoken language processing.*

Deterding, D., Brown, A., & Low, E. L. (2005). *English in Singapore: phonetic research on a corpus, Singapore*. New York: McGraw-Hill Education.

Gopinathan, S. (1998). Language policy changes 1997: Politics and pedagogy. In S. Gopinathan, A. Pakir, H. W. Kam, & V. Saravanan (Eds.), *Language, society and education in Singapore* (2nd ed.). Singapore: Times Academic Press.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, *29*, 82–97.

Kwan Terry, A. (1992). Code-switching and code-mixing: The case of a child learning English and Chinese simultaneously. *Journal of Multilingual and Multicultural Development, 13*(3), 243–259.

Li, W. (1998). The ‘Why’ and ‘How’ questions in the analysis of conversational codeswitching. In P. Auer (Ed.), *Code-switching in conversation: Language, interaction, and identity*. London: Routledge.

Li, H., Ma, B., & Lee, K.A. (2013). Spoken language recognition: From fundamentals to practice. In *Proceedings of the IEEE.*

Li, Y., Yu, Y., & Fung, P. (2012). A Mandarin–English code-switching corpus. In *Proceedings of the eight international conference on language resources and evaluation (LREC'12).*

Lyu, D.-C., Chng, E.-S., & Li, H. (2013a). Language diarization for code-switch conversational speech. In *Proceedings of ICASSP.*

Lyu, D.-C., Chng, E.-S., & Li, H. (2013b). Language diarization for conversational code-switch speech with pronunciation dictionary adaptation. In *Proceedings of ChinaSIP.*

Lyu, D.-C., & Lyu, R.-Y. (2008). Language identification on code-switching utterances using multiple cues. In *Proceedings of the international speech communication association.*

Lyu, D.-C., Lyu, R.-Y., Chiang, Y.-C., & Hsu, C.-N. (2006a). Speech recognition on code-switching among the Chinese dialects. In *Proceeding of ICASSP*.

Lyu, D.-C., Lyu, R.-Y., Chiang, Y.-C., & Hsu, C.-N. (2006b). Language identification by using syllable-based duration classification on code-switching speech. ISCSLP, volume 4274 of Lecture Notes in Computer Science (pp. 475–484). New York: Springer.

Lyu, D.-C., Zhu, C.-L., Lyu, R.-Y. & Ko, M.-T. (2010). Language identification in code-switching speech usingword-based lexical model. In *Proceedings of the 7th international symposium on chinese spoken language processing (ISCSLP ‘10), Tainan, Taiwan, December 2010* (pp. 460–464).

MacSwan, J. (2013). Code-switching and grammatical theory. In T. Bhatia & W. Ritchie (Eds.), *Handbook of Multilingualism* (p. 2013). Cambridge: Blackwell.

Malik, L. (1994). *Sociolinguistics: A study of code-switching*. New Delhi: Anmol.

Milroy, L., & Muysken, P. (1995). *One speaker, two languages. Cross-disciplinary perspectives on code-switching*. Cambridge: Cambridge University Press.

Myers-Scotton, C. (1989). Codeswitching with English: Types of switching, types of communities. *World Englishes*, *8*(3), 333–346.

Myers-Scotton, C. (1993). *Social motivations for code-switching: Evidence from Africa*. Oxford: Clarendon Press.

Myers-scotton, C., & Myers, C. (1993). *Duelling languages: Grammatical structure in codeswitching*. Oxford: Clarendon Press.

Povey, D., Burget, L., Agarwal, M., Akyazi, P., Kaie, F., Ghoshal, A., et al. (2011). The subspace Gaussian mixture model—A structured model for speech recognition. *Journal of Computer Speech and Language, 25*(2), 404–439.

Population Trends. (2012). http://www.singstat.gov.sg/Publications/publications_and_papers/population_and_population_structure/population_trend.html.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–287.

Reyes, I. (1994). Functions of code switching in schoolchildren's conversations. *Bilingual Research Journal, 28*(1), 77–98.

Sankoff, D., & Poplack, S. (1981). A formal grammar for code-switching. Papers in Linguistics.

Santhosh Kumar, C. P., Li, H., Tong, R., Matějka, P., Burget, L., & Černocky, J. (2010). Tuning phone decoders for language identification. In *Proceedings of ICASSP.*

Shen, H.-P., Wu, C.-H., Yang, Y.-T. & Hsu, C.-S. (2011). CECOS: A Chinese–English code-switching speech database. In *Proceedings of the international conference on speech database and assessments (Oriental COCOSDA '11)*, *Hsinchu, Taiwan, October 2011* (pp. 120–123).

Shia, C. J., Chiu, Y. H., Hsieh, J. H., & Wu, C. H. (2004) Language boundary detection and identification of mixedlanguage speech based on map estimation. In *Proceedings of the IEEE international conference on acoustics, speech, and signal.*

Su, H. Y. (2001). Code-switching between Mandarin and Taiwanese in three telephone conversation: The negotiation of interpersonal relationships among Bilingual speakers in Taiwan. In *Proceedings of the symposium about language and society.*

Tseng, H., Chang, P., Andrew, G., Jurafsky, D., & Manning, C. (2005). A conditional random field word segmenter. In *Fourth SIGHAN workshop on Chinese language processing.*

Vu, N. T., Lyu, D. C., Weiner, J., et al. (2012). A first speech recognition system for Mandarin–English code-switch conversational speech. In *Proceedings of the 37th IEEE international conference on acoustics, speech and signal processing (ICASSP '12), Kyoto, Japan, March 2012* (pp. 4889–4892).

Wu, C.-H., Chiu, Y.-H., Shia, C.-J., & Lin, C.-Y. (2006). 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), 266–275.

Young, S. (1996). A review of large-vocabulary continuous-speech. *IEEE Signal Processing Magazine, 13*(5), 45–57.

Zissman, M. A. (1996). Comparison of four approaches to automatic LID of telephone speech. *IEEE Transactions on Acoustics, Speech and Signal Processing, 4*(1), 31–44.