
Description of the SEAME: Mandarin-English Code-switching in South-East Asia Corpus

---

Created on: August, 2014

Updated on: July, 2015

CONTENTS

1. Introduction

2. Speaking style and recording environment

3. Transcription

4. Description of corpus

5. Speaker information

6. Publications

---

1. Introduction

SEAME (South-East Asia Mandarin-English) is a spontaneous Mandarin-English code-switching speech corpus recorded from Singapore and Malaysia speakers. The SEAME corpus can be applied for language boundary detection, language identification, and LVCSR of English/Mandarin code-switch speech.

SEAME corpus was recorded in Singapore and Malaysia by Nanyang Technological University (NTU) and Universities Sains Malaysia (USM) respectively. The recordings have two different speaking styles: conversational and interview conditions. For the conversational speech, each speaker's speech is recorded separately. The topics discussed in the corpus ranges from hobbies, friends, and daily activities. In the interview speech, only the single interviewee speech is recorded. All recordings are performed by close-talk microphone in quiet room. The speakers are age between 19 and 33, almost balanced in gender (49.7% of female and 50.3 % of male). The total number of distinct speakers is 156 (36.8% are Malaysian while the rest are Singaporean). There are total 192 hours of wav audio records. Initially, all conversational speech were recorded in with 16K Hz sampling rates and 16-bit resolution. About 80 hours of interview speech were recorded with the same format with conversational speech, the rest, about 20 hours of interview speech were recorded with 22050 Hz sample rates, 16-bit resolution. The 20 hours of 22050 Hz sampling interview speech then were converted into 16K Hz sample rates. We then converted the audio from original wav format into flac format, which is recommended for lossless audio compression.

This release contains audio recordings of duration from 20 minutes to 120 minutes. Selected segments of these audio recordings have been transcribed. The transcription file of each audio file is

stored in tab-separated text file format.

---

2. Speaking style and recording environments:

All of the recordings are recorded using a close-talk microphone with either 16K Hz or 22050 Hz sampling rates (see the files docs/22050Hz-list.txt for detail file name) and 16-bit resolution in a quiet environment.

a) Speaking style

The corpus is developed for spontaneous code-switching speech research, and hence the recordings are un-scripted interviews and conversations.

- Interview speech: In this setup, there are two speakers - an interviewer who asks questions and an interviewee who answers. Only the interviewee's speech is recorded. Each of the recording session is about one hour and the topics discussed include: from hobbies, friends, and daily activities.
- Conversational speech: There are two speakers conversing freely to each other. The topics for each recording section are family, school life, sports, etc. Each recording session is approximately one hour.

b) Recording channel

The corpus was recorded by using several microphones and recording devices. Each device setup is labelled as X, Y, P, Q and Z. The specification of these channels as follows:

- X: Sennheiser Microphone + PC sound card (recordings in NTU)
- Y: Sennheiser Microphone + notebook sound card (recordings in NTU)
- P, Q: two channels of MOTU (recordings in NTU)
- Z: USM recording devices

c) Recording session

Each volunteer speaker is allowed to record more than one recording session.

---

3. Transcription

The corpus transcription process was organized into two phases:

a) In "Phase I", selected audio segments, mainly code-switching segments, were transcribed.

b) In "Phase II", the majority of all audio segments in the corpus, including monolingual

and code-switching segments, were transcribed. In addition, the language label for each utterance is updated. The language labels are: English (EN), Mandarin (ZH) or Code-Switch (CS). The detail information of transcription files are explained in following paragraphs.

In each transcript, the annotations were categories in different types:

- Speech: Will be transcribed to either Mandarin/English.
- Discourse particle: We use both English and Mandarin words to represent discourse particle. When the discourse occurs in Mandarin segment, we use words such as [啊], [喔], etc. to represent. When the discourse occurs in English segment, we use words such as [ah], [oh], etc. to represent.
- Hesitation and filled pause: Filled pause refers to sounds made by people when they are hesitating or thinking. These sounds are indicated by a pair of brackets ( ), such as (er), (erm), etc.
- Other languages: This group includes languages and dialects other than English and Mandarin. E.g, we find Japanese, Korean and Indian words. E.g, these words were spoken to describe food and places. In addition, the Chinese dialects such as Cantonese and Hokkien, can also be found in the recordings. These foreign language words are indicated by a pair of # within the word, e.g, #nasi-lemak# (a Malay word food dish):
  - #ah-mah#: It is Hokkien dialect and it means “old woman”.
  - #ayumi-hamasaki#: a Japanese singer-songwriter.
- Non-speech: Non-speech sounds are tagged by (xxx), where xxx indicates the class of non-speech sound. Currently we tag non-speech sounds such as breathing, coughing, laughing and other paralinguistic phenomena noises are labelled. The labels (ppb) means paralinguistic phenomena breathing, (ppc) means paralinguistic phenomena coughing, (ppl) means paralinguistic phenomena laughing and (ppo) means paralinguistic phenomena others.

a) Transcription file format in “Phase I”

In “Phase I”, 52,145 selected segments were transcribed. The selected segments were mainly segments containing code-switching utterances. The transcriptions were formatted in tab-separated text files, using UTF-8 form. Fig 3.1 shows an example of the format. Each line represents an utterance. There are four columns of information in each line.

i) The first column is the audio file name without extension;
ii) The second column is the start time of the utterance in millisecond;
iii) The third column is the end time of utterance in millisecond and
iv) The last column is the transcription of the utterance.

```
04NC07FBX_0101  567160  569520  来看 见 我们 来 meet us 啦
04NC07FBX_0101  572640  576320  因为 他的 眼镜 like super big 他的 那个 frame super big
04NC07FBX_0101  599440  602480  spectacles 他的 那个 框框 那个 width 很大
04NC07FBX_0101  630240  633320  他 有 short sightedness 他 戴 contacts 的 leh
04NC07FBX_0101  633519  639399  then hor on top of that 他 戴 那种 大大 个 的 spectacles 啦 没有 degree 的
04NC07FBX_0101  643399  645399  他 只是 戴 玩玩 而已 for the day
04NC07FBX_0101  806239  808919  哎 有时 你 讲话 太 大声 会 affect 到 我的 leh
04NC07FBX_0101  809399  811079  我也 是 会 affect 到 你 leh
04NC07FBX_0101  816319  818239  你的 是不是 比较 sensitive 啊
```

Figure-3.1 – Phase I transcription in text formatted file

Most of the transcribed segments are code-switching segments. There is a small selection of monolingual segments: 12% and 6% of the transcribed segments are Mandarin and English monolingual utterances respectively.

b) Transcription file format in “Phase II”

In this phase, most of the speech segments, including monolingual and code-switch segments were transcribed. In total, 110,145 segments were transcribed. The distribution of the segments by language are 28,655 (English) and 24,438 (Mandarin) and 57,052 (Code-Switch). The language label for each segments were also added in this phase. The transcription files were still formatted in tab-separated text files, using UTF-8 form. In addition, erroneous labels and transcriptions were updated. Fig 3.2 shows an example of the file format. By adding language label for each utterance, now each transcription line represents one utterance and contains 5 columns:

i) The first column is the audio file name without extension;
ii) The second column is the start time of the utterance in millisecond;
iii) The third column is the end time of the utterance in millisecond;
iv) The fourth column is the language label, it is either EN (for English) or ZH (for Mandarin) or CS (for code-switch) and
v) The last column is the transcription of the utterance.

```
04NC07FBX_0101  577004  579697  EN      then in the end right what
04NC07FBX_0101  580033  596610  CS      then in the end right (呃) 我 下面 讲 说 [诶] jason
你 应该 那个 戴 dorous 的 眼镜 [啦] then i <unk> specticles for him to 戴 戴 看 [哇] okay [咧
] then 我 讲 改次 你跟 那个 眼镜 店 老板 讲说 (呃) 我 要 小号 小号 一点 的
04NC07FBX_0101  597367  602480  CS      他 的 他 的 很大 spectacles 他 的 这个 框 框 这个
width 很大
04NC07FBX_0101  603118  604577  EN      (oops)
04NC07FBX_0101  607949  614317  EN      his very big [eh] than than mine right on him right
is like just nice [lo] his really is like so big
04NC07FBX_0101  615102  615579  EN      yup
04NC07FBX_0101  618076  624816  EN      (ppb) (ppl) why are we talking about spectacles
(ppb)
04NC07FBX_0101  626660  627501  EN      (erm)
04NC07FBX_0101  628238  630240  ZH      我 有 一个 朋友 她 叫 婉玲 [咯]
04NC07FBX_0101  630240  633320  CS      她 有 short-sightedness 她 戴 contacts [诶]
```

Figure-3.2 – Phase II transcription in text formatted file

---

4. Description of corpus

a) Directory structure

The directory structure of the SEAME corpus is shown in figure 5.1. The directory has 5-levels:

- Data: SEAME\data\{speech-type}\{data-type}\{transcription-type}\{file}
  - {speech-type} could be *conversation* or *interviewee*
  - {data-type} could be either *transcript* or *audio*
  - {transcription-type} could be either *phaseI* or *phaseII* or just empty in *flac* folder
  - {file} depend on which {data-type} is *transcript* or *audio*, {file} could be plain text file (.txt) or audio file (.flac) file.
- Documents: SEAME\docs\{document-file}

  There are two documents in the *docs* folder:
  - SEAMEV4.0.doc is this document. It contains the description of the SEAME corpus, recording setup, transcription details, and so on.
  - Speaker-info.xls contains the following information of each speaker: speaker identification, age, gender and nationality. In the xls file, there are three sheets, namely: Conversation, Interview-NTU, Interview-USM. In these sheets, the details of the speaker's information of conversation recordings recorded in NTU and interview recordings recorded in NTU and USM is given. The conversation recordings were only made in NTU.
- REAME: The introduction plain text file which introduces directory structure of whole SEAME folder.

b) File name convention

Within the conversation and interview folders, the data has been divided into two groups, transcript and audio. We named both transcript and audio files according to 1) Conversation group number id, 2) Recording location, 3) Speaking style, 4) Speaker id nunber, 5) Gender, 6) Nationality of speakers, 7) Microphone channel, 8) The sequence of recording of that speaker and 9) The order of actual file in the recording. We use two examples to describe the name of the recording file to show different conditions.

- The filename 08NC16FBQ_0101 means:
  - 08: conversation group number id. The 2 digit id ranges from 01 to 46 as there is a total of 46 conversations in the corpus.
  - N: recording location (N: NTU, U: USM)

- C: conversational speaking style where (C: conversational, I: interview)
- 16: speaker identity (from 01 to 61)
- F: gender, female (F: Female, M: Male)
- B: speaker's nationality, Singaporean (A: Malaysian, B: Singaporean)
- Y: channel identity (NTU: P, Q, X,Y. For USM: Z)
- 01: the first record session of speaker 16
- 01: the first part of recording 01 of speaker 16

o The filename UI06MAZ_0105 has the following interpretation:

- U: recording location USM (N: NTU, U: USM)
- I: interview speaking style, where (C: conversational, I: interview)
- 06: speaker identity (NTU: from 01 to 67, USM; from 01 to 29)
- M: gender, male
- A: speaker's nationality, Malaysian (A: Malaysian, B: Singaporean)
- Z: channel identity (NTU: P, Q, X,Y. For USM: Z)
- 01: the first record of speaker 06, from USM
- 05: the fifth parts of speaker 06's first record session

Figure 4.1 shows the screen capture of folder structure of SEAME corpus.

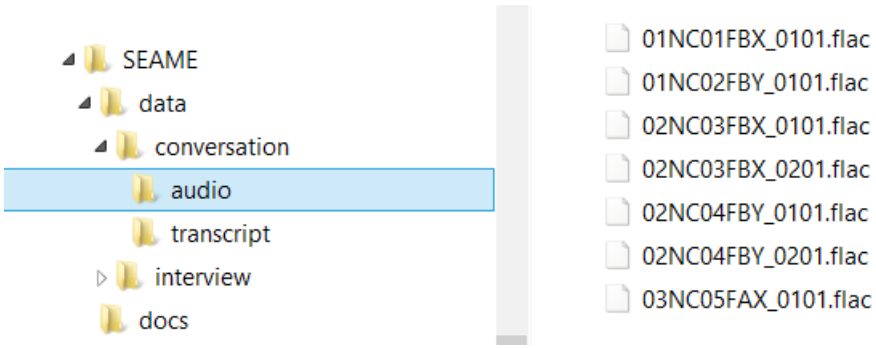

Figure 4.1 - The folder structure of SEAME corpus.

---

5. Speaker information

The speaker's information can be found in speaker-info.xls in the same directory with this file.

---

6. Publications

The following papers related to SEAME corpus have been published:

[1] Dau-Cheng Lyu, Tien-Ping Tan, Eng-Siong Chng, and Haizhou Li : **Mandarin–English code-switching speech corpus in South-East Asia: SEAME**. Language Resources and Evaluation 2015, 49: 581-600

[2] Dau-Cheng Lyu, Tien Ping Tan, Engsiong Chng, Haizhou Li: **SEAME: a Mandarin-English code-switching speech corpus in south-east asia.** INTERSPEECH 2010: 1986-1989

[3] Ngoc Thang Vu, Dau-Cheng Lyu, Jochen Weiner, Dominic Telaar, Tim Schlippe, Fabian Blaicher, Engsiong Chng, Tanja Schultz, Haizhou Li: **A first speech recognition system for Mandarin-English code-switch conversational speech.** ICASSP 2012: 4889-4892

[4] Dau-Cheng Lyu, Eng-siong Chng, Haizhou Li: **Language diarization for code-switch conversational speech with pronunciation dictionary adaptation**. ChinaSIP 2013

[5] Dau-Cheng Lyu, Tien-Ping Tan, Eng-Siong Chng, and Haizhou Li: **An Analysis of a Mandarin-English Code-switching Speech Corpus: SEAME**. OCOCOSDA 2010.