INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

# Building the Singapore English National Speech Corpus

*Jia Xin Koh*[1], *Aqilah Mislan*[2], *Kevin Khoo*[2], *Brian Ang*[2], *Wilson Ang*[2], *Charmaine Ng*[1], *Ying-Ying Tan*[1]

[1]School of Humanities, Nanyang Technological University, Singapore
[2]Digital Services Lab, Info-communications and Media Development Authority, Singapore

jkoh033@e.ntu.edu.sg, nuraqila001@e.ntu.edu.sg, kevin_khoo@imda.gov.sg, brian_ang@imda.gov.sg, wilson_ang@imda.gov.sg, char0091@e.ntu.edu.sg, yytan@ntu.edu.sg

## Abstract

The National Speech Corpus (NSC) is the first large-scale Singapore English corpus spearheaded by the Info-communications and Media Development Authority of Singapore. It aims to become an important source of open speech data for automatic speech recognition (ASR) research and speech-related applications. The first release of the corpus features more than 2000 hours of orthographically transcribed read speech data designed with the inclusion of locally relevant words. It is available for public and commercial use upon request at "www.imda.gov.sg/nationalspeechcorpus", under the Singapore Open Data License. An accompanying lexicon is currently in the works and will be published soon. In addition, another 1000 hours of conversational speech data will be made available in the near future under the second release of NSC. This paper reports on the development and collection process of the read speech and conversational speech corpora.

**Index Terms**: automatic speech recognition, Singapore English, read speech corpus, conversational speech corpus

## 1. Introduction

Though present speech recognition systems of English have been reported to have reached almost "human" levels of accuracy[1], this is true only for highly resourced varieties such as Standard American English, and progress for "non-native" accents is still lagging. A key reason for this is the lack of large-scale speech data available in the required accent to train ASR systems.

While Malay is the national language, English has been Singapore's official working language since its independence in 1965. Today, all Singaporeans undergo a formal education primarily in English, and 36.9% of families now speak English at home, making it the most frequently spoken home language [2]. A few spoken Singapore English corpora have been published so far, such as the NIE Corpus of Spoken Singapore English (NIECSSE), the Singapore component of the International Corpus of English (ICE-SIN), and the Grammar of Spoken Singapore English Corpus (GSSEC). An audio corpus for developing a Computer-Assisted Language Learning system was also produced [3]. However, these are smaller in scale. The NIECSSE consists of 3.5 hours of interviews and dictation of a phonetically designed passage [4]. The GSSEC has about 8 hours of conversations incorporated into the roughly 600,000-word ICE-SIN [5], [6]. The CALL resource appears to be the largest, to the best of our knowledge, with approximately 125 hours of speech collected from 83 university educated speakers [3]. In comparison, open large-scale English speech corpora such as LibriSpeech, Common Voice and TED-Lium 3 contain around 1000 hours[7], 803 hours[8], and 452 hours [9] respectively.

To help spur on speech technology research and adapt speech-related applications to Singapore, the Info-communications and Media Development Authority (IMDA) has commissioned the production of a large-scale speech corpus of Singapore English known as the National Speech Corpus (NSC), free for use under the Singapore Open Data License. Currently, NSC consists of three parts: 1) 1000 hours of read speech using phonetically balanced scripts (henceforth '*PB*'), 2) 1000 hours of read speech featuring words pertinent to the Singapore context (henceforth '*LW*'), and 3) 1000 hours of conversational, spontaneous speech (henceforth '*CS*'). NSC is thus, to the best of our knowledge, the largest open Singapore English speech corpus, and possibly one of the largest in the world as well. NSC is also a comprehensive speech resource in terms of the demographic spread, representing the major ethnic groups in Singapore and covering a range of ages (18 to 60+ year old) as well as educational levels. Furthermore, with read speech and conversational speech thoughtfully designed with considerations for sociolinguistic variables, NSC provides training data to suit a variety of applications and purposes.

This paper presents the development of the two read speech corpora (PB and LW) and the conversational speech corpora (CS), collected from Singapore English speakers. Section 2.1 to Section 2.3 documents the design and demographic statistics of the read speech corpora, and Section 2.4 explains how the data was processed. Section 3 continues with the report of the conversational corpus' design, collection procedures, demographic numbers and current status, before concluding in Section 4.

## 2. Read speech corpora

A total of 1,379 speakers were recruited, of which 1036 recorded PB, and 1032 recorded LW. PB and LW corpora were recorded simultaneously and speakers were allowed to participate in both of the recordings, of which 689 of them did. Recordings took around 3 months to complete.

### 2.1. Recording procedure

Only speakers who are at least 18 years old were recruited so that they can legally agree to participate in the corpus collection. Other criteria for speakers include:

- Must be a Singaporean citizen raised in Singapore, or
- Residents who have lived in Singapore for at least 18 years, and
- Have literacy in English, and
- Have undergone a formal education in English in Singapore public schools for at least 6 years.

Recording studios were set up in various locations in Singapore. Some rooms were professional recording studios, while

http://dx.doi.org/10.21437/Interspeech.2019-1525

others were quiet rooms in co-working spaces. Each room was set up with three microphones: a close-talk headset microphone or standing microphone, a far-field boundary microphone, and a mobile phone. Audio recordings for the former two microphones were made in 48kHz and 16 bits. The mobile phones recorded in 16kHz and 16 bits. All recordings were saved as WAV files.

Each speaker was asked to read a total of 800 sentences for PB, and 896 for LW, in a manner that was most natural and comfortable to them. The sentences were shown individually on a computer screen, and speakers were taught to navigate the prompts and recording controls. Speakers were generally left alone to complete the recordings and given a break at the halfway mark, though recording assistants checked in periodically to ensure that there are no issues with the microphones and sound quality. Most speakers who requested to participate in both PB and LW recordings were agreeable to carrying out the recordings on two separate days so as to prevent any physical overexertion that may affect the voice quality of the data as well.

### 2.2. Demographics

The aim of NSC is to represent the local population as adequately as possible. In terms of gender distribution, a 5% deviation was allowed. Female and male speakers in PB made up 54.8% and 45.2% respectively, while LW had a slightly more balanced gender distribution with 49.6% of the speakers female, and 50.4% of the speakers male.

A major demographic consideration for NSC was the ethnic make-up. Out of 4.0 million Singapore citizens in 2017, 74.3% are Chinese, 13.4% are Malay, and 9.0% are Indian [10]. Adhering strictly to the actual ethnic distribution would limit the amount of data for minority groups and in turn limit the usefulness of the corpora. We thus targeted a higher intake of Malay and Indian speakers so as to ensure that there would be adequate representation of the minority groups in the data. Table 1 reports the targeted and actual distribution of speakers by ethnicity in the corpora.

Table 1: *Proportion of speakers distributed by ethnicity*

| **Ethnicity** | **Target(%)** | **PB Actual(%)** | **LW Actual(%)** |
|---|---|---|---|
| Chinese | 50.0 | 60.7 | 60.4 |
| Malay | 20.0-25.0 | 19.3 | 19.4 |
| Indian | 20.0-25.0 | 19.3 | 19.4 |
| Others | 0.0-5.0 | 0.7 | 0.9 |
| **Total** | 100 | 100 | 100 |

We categorized speakers into three broad age groups as well. Table 2 summarizes the distribution of speakers by their age groups.

Table 2: *Proportion of speakers distributed by age*

| **Age group** | **PB(%)** | **LW(%)** |
|---|---|---|
| 18-30 | 51.2 | 50.0 |
| 31-45 | 30.0 | 30.8 |
| >46 | 18.8 | 19.2 |
| **Total** | 100 | 100 |

A further consideration was to further recruit speakers from a range of educational levels, so as to obtain a more diverse range of linguistic abilities and socioeconomic backgrounds. Table 3 reports the distribution of education backgrounds among the speakers.

Table 3: *Proportion of speakers distributed by education*

| **Education level** | **PB(%)** | **LW(%)** |
|---|---|---|
| University or higher | 42.4 | 42.2 |
| Junior College/ Polytechnic | 41.3 | 42.0 |
| Secondary or below | 16.3 | 15.8 |
| **Total** | 100 | 100 |

### 2.3. Scripts

#### *2.3.1. Phonetically balanced corpus*

The entire script for PB features about 72,000 different sentences that were crawled from a variety of online news agencies in Singapore. These sentences were algorithmically selected to provide an adequate coverage of phones, diphones, and triphones. Another 200 sentences were crafted, adapting from Deterding's works [11],[12] so as to ensure that the documented phones in Singapore English are represented in the corpus. Excluding the 200 specially designed sentences that every speaker had to read, all the sentences were read no more than 8 times in total. This means that each sentence was read by a maximum of 8 speakers, thereby giving additional coverage of phonetic phenomena to the corpus.

#### *2.3.2. Local words corpus*

An ongoing challenge for speech recognition is in recognizing named entities, and more so when facing unforeseen foreign words that are from diverse language origins. The LW corpus is thus aimed at minimizing the out-of-vocabulary gap for Singapore-based systems. Scripts for this part of NSC were generated by parsing word lists into a set of grammars. The word lists were compiled based on unique local entities, and range from place names and addresses, local food names, brands, names of prominent people in Singapore, local abbreviations, and so on.

### 2.4. Read speech post-processing procedure

The 48kHz and 16 bits audio recordings were downsampled to 16kHz and 16 bits for online publication. All 3 channels of recordings are published, with each speaker's individual utterances saved into a single ZIP file identified by their individual speaker ID. Around 1113 hours of speech data have been collected for PB, and 1057 hours for LW. These add up to an approximate total of 2170 hours of edited speech data.

#### *2.4.1. Transcription*

Both PB and LW were transcribed orthographically. Transcribers logged on to an in-house online transcription platform, and were provided with an audio segment, as well as the corresponding prompt that the speaker had read. Their task was to check if the speaker had read according to the prompt, and to transcribe any deviations or other acoustic events based on a set of transcription instructions. Accuracy of the transcriptions was determined on the word level, and refinements like making use of the lexicon to sieve out spelling errors are currently ongoing.

### *2.4.2. Speaker information*

Information about the speakers were collected and compiled into a spreadsheet. Information released have been anonymized and speakers are represented by their own unique speaker ID. The speaker ID is made up of two parts - a four-numeral PB script number followed by a four-numeral LW script number. Speakers who did not read either part will have corresponding 0000 in the speaker ID.

The following types of information are published:

- Speaker's gender
- Speaker's ethnicity
- Recording equipment

### *2.4.3. Lexicon*

The lexicon, or pronunciation dictionary of PB consists of around 30,000 unique words, while the lexicon of LW has about 22,200 unique words, derived from the orthographic transcriptions. Prior to generating the phonetic transcriptions of each word, a Singapore English phoneme inventory was first constructed in both ARPAbet and X-SAMPA by Singaporean linguists on the team. The phoneme inventory consists of 24 consonants, 8 vowels, and 5 diphthongs, for transcribing English words. An additional 11 consonants and 1 vowel were included to account for non-English phonemes reflected in LW. Symbols in X-SAMPA that did not have an equivalent in ARPAbet were approximated with available ARPAbet symbols. The X-SAMPA transcriptions in the lexicon can thus be said to be narrower than ARPAbet's. An example of the lexicon is provided in Table 4.

Table 4: *Example of the lexicon*

| Word | ARPAbet | SAMPA |
|---|---|---|
| **algae**<br>/ɛlge/ | EH L G EY | E l g e |
| **Bandung**<br>/banduŋ/ | B AA N D UW NG | b a n d u N |
| **Blangah**<br>/blaŋa/ | B L AA NG AA | b l a N a |
| **Blangah**<br>/blaŋga/ | B L AA NG G AA | b l a N g a |
| **think**<br>/θiŋk/ | TH IY NG K | T i N k |
| **think**<br>/tiŋk/ | T IY NG K | t i N k |
| **Yunnan**<br>/ynnan/ | IY N N AA N | y n n a n |
| **Yunnan**<br>/junnan/ | Y UW N AA N | j u n n a n |

As the PB lexicon is mostly made up of English words, we made use of mapping rules [3] to approximate phonetic transcriptions from British English to Singapore English. For the LW lexicon, a grapheme-to-phoneme model was used to generate a base transcription before both lexicons were merged and reviewed manually by the linguists. Illegitimate words or mispronunciations that had been included in the orthographic transcriptions were also extracted out from this version.

Constructing a lexicon for LW is in comparison a more complicated task. Apart from not having an officially standardized pronunciation, many entries in LW are loanwords that originate from a diverse set of languages that are present in Singapore, such as Malay, Hokkien, Cantonese, Tamil, and Hindi [13]. Though theoretically we could look towards finding the canonical pronunciations of these words through their corresponding language references, pinpointing accurately which loanword originate from which language is not a trivial task given the linguistic changes these languages have also undergone over time. Ethnic variation may also have exerted influence on the pronunciation especially when the word originates from a language that a speaker is not a native of. This could mean that a local word such as *blangah* /blaŋa/ in the place name *Telok Blangah* "cooking pot bay" [14, pp. 379] be mispronounced by a speaker who is unfamiliar with the pronunciation rules of Malay as /blaŋga/.

## 3. Conversational speech corpus design

Though conversational speech corpora are more expensive and difficult to obtain than read speech, it is often preferred as training data for ASR systems designed with the purpose of recognizing human-to-human interactions. The dissimilarity between the style and speech patterns of read speech and how people speak and communicate with each other in real life often limits the performance of ASR that serves these specific types of applications. In this part of NSC, we collected around 1000 hours of conversational speech produced in Singapore English, split into two modes of recording - one in a face-to-face (FTF) setting, and the other over the telephone in two separate rooms. Each mode recorded at least 250 pairs of speakers.

### 3.1. Recording procedure

Recruitment criteria were the same as PB and LW, though less focus was placed on literacy than on the ability of the speakers to converse fluently in Singapore English.

Speakers were recommended to bring a partner, preferably a friend or family member whom they could speak at least 2 hours with. All speakers (including referred partners) were only allowed to participate once, regardless of whether they recorded their session face-to-face or on the telephone. They were allowed, however, to have participated in PB and/or LW. Some speakers were also requested to bring a partner who are of a different ethnicity. Speakers who were unable to refer a partner were paired with other solo speakers. These pairs of speakers were therefore strangers prior to meeting for the first time at the recording venue.

Recording studios were set up in quiet rooms based in two different co-working offices. Each FTF room was set up with a close-talk headset microphone and a far-field boundary microphone. Each telephone room was set up with a standing microphone and a corded telephone set. The telephones were connected internally through VoIP using an Interactive Voice Response system. All microphones recorded in 48kHz and 16 bits, before down-sampling to 16kHz. The telephones recorded in 8kHz and 8 bits.

Each recording session was approximately 2 hours 15 minutes long. Speakers were generally left alone in the rooms so that they could talk comfortably and naturally with each other, though recording assistants periodically checked in to see how the recordings were going, and to adjust the microphones if necessary.

As it may be quite difficult to elicit natural conversations from speakers, especially when they first enter an unfamiliar setting and know that they are being observed, the following materials were used in both FTF and telephone recordings to get the speakers accustomed to the environment and facilitate the conversations:

- Spot-the-difference diapix
- Conversation card games
- Free-talk prompts

In the first task, speakers were asked to collaborate and find 12 differences between two similar pictures without looking at each other's pictures. The pictures were DiapixUK picture materials[15]. The main purpose of having the task was to help the speakers mitigate some of the initial awkwardness when recording, while eliciting descriptive and directional phrases that may be useful for some ASR applications. The pictures were switched out for new ones periodically so as to introduce some diversity in content. Speakers generally took around 10 to 20 minutes to complete the task.

In the second task, two different sets of conversation card games were procured to act as conversational prompters. The FTF sessions used *smol tok* - which consisted of a basic deck and a booster deck that was localized to the Singapore context. The telephone sessions used *Hypothetically Fun*. Depending on how conversational the speakers were, speakers could finish their deck of cards in as short as 45 minutes, or all the way until the recording session ends. Speakers who finished using the cards but still had time left were then given a list of suggested topics and questions to talk about.

### 3.2. Demographics

The same demographic targets were applied in the recruitment for CS speakers. Table 5 reports the gender distribution of each mode of recording, as well as the overall distribution in CS.

Table 5: *Proportion of speakers distributed by gender*

| Gender | FTF(%) | Tel.(%) | Overall(%) |
|---|---|---|---|
| Female | 52.5 | 54.9 | 53.7 |
| Male | 47.5 | 45.1 | 46.3 |
| **Total** | 100 | 100 | 100 |

Likewise, for ethnic distribution, Table 6 reports the ethnic representation of the speakers in each mode as well as in the overall distribution. Table 7 reports the age distribution, while Table 8 shows distribution of education level among the speakers.

Table 6: *Proportion of speakers distributed by ethnicity*

| Ethnicity | FTF(%) | Tel.(%) | Overall(%) |
|---|---|---|---|
| Chinese | 58.8 | 58.9 | 58.8 |
| Malay | 20.0 | 20.7 | 20.4 |
| Indian | 20.6 | 18.9 | 19.8 |
| Others | 0.6 | 1.6 | 1.1 |
| **Total** | 100 | 100 | 100 |

Table 7: *Proportion of speakers distributed by age*

| Age group | FTF(%) | Tel.(%) | Overall(%) |
|---|---|---|---|
| 18-30 | 49.0 | 46.9 | 48.0 |
| 31-45 | 29.4 | 32.0 | 30.7 |
| >46 | 21.6 | 21.1 | 21.3 |
| **Total** | 100 | 100 | 100 |

Table 8: *Proportion of speakers distributed by education*

| Education level | FTF(%) | Tel.(%) | Overall(%) |
|---|---|---|---|
| University or higher | 49.0 | 39.8 | 44.4 |
| Junior College/ Polytechnic | 31.9 | 38.8 | 35.3 |
| Secondary or below | 19.1 | 21.5 | 20.3 |
| **Total** | 100 | 100 | 100 |

### 3.3. Post-processing procedure

Similar to the former two corpora, CS is being transcribed on the word level manually and a first round of transcribing has been completed so far. A second round of transcribing is underway to bring the transcripts to a satisfactory level and to resolve various issues that the current transcript conventions do not yet account for. For instance, spelling of local expressions will be standardized, and conventions will also be added to deal with mispronunciations and incomplete words. Handling code-switching occurrences, that is the use of more than one language in a conversation, is however a trickier task for both the transcripts and the lexicon. This would require further inquiry into the data and research literature in order to derive a feasible solution.

## 4. Conclusion

The first release of NSC is currently the first large-scale read speech corpora of Singapore English, contributing to more than 2000 hours of data with accompanying orthographic transcripts for public use. The accompanying lexicon will also be made available in 2019, and may be used for text-to-speech applications as well. In addition, another 1000 hours of conversational speech along with the accompanying improved transcripts and lexicon is set to be released in the near future. Possible future work include training baseline ASR systems and analysing the phonetic variations in the speech of Singaporeans. It is hoped that the NSC can help speech technology developers and researchers to build and improve speech-related applications in Singapore. The corpus and future updates are accessible via "www.imda.gov.sg/nationalspeechcorpus".

## 5. Acknowledgements

We would like to thank the organizations who had provided invaluable input that contributed to the corpus, and the speakers for participating in the corpus.

## 6. References

[1] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving Human Parity in Conversational Speech Recognition," Microsoft Re-

search, Tech. Rep. MSR-TR-2016-71, 2016. [Online]. Available: https://arxiv.org/abs/1610.05256

[2] "General Household Survey 2015." [Online]. Available: https://www.singstat.gov.sg/-/media/files/publications/ghs/ghs2015/ghs2015.pdf

[3] W. Chen, Y.-Y. Tan, E. Siong Chng, and H. Li, "The development of a Singapore English call resource," in *Oriental COCOSDA*, Nepal, 2010.

[4] D. Deterding and E. L. Low. The NIE Corpus of Spoken Singapore English (NIECSSE). [Online]. Available: http://videoweb.nie.edu.sg/phonetic/niecsse/saal-quarterly.htm

[5] L. Lim, "The Grammar of Spoken Singapore English Corpus: Ground Rules & Conventions," 2009. [Online]. Available: https://english.hku.hk/staff/lisa_lim/GSSEC-GroundRules-2009.doc

[6] E. Y. M. Lai, L. Tan, V. Wong, L. T. T. Loke, and F. Bond, "The OPT-ional Phenomenon in Singapore English: A Corpus-based Approach Using Time Annotated Corpora," *Procedia - Social and Behavioral Sciences*, vol. 95, pp. 431–441, 2013.

[7] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Australia: IEEE, 2015, pp. 5206–5210.

[8] Common Voice by Mozilla. [Online]. Available: https://mzl.la/voice

[9] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Estève, "TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation," in *SPECOM*, Germany, 2018.

[10] "Singapore in Figures 2018." [Online]. Available: https://www.singstat.gov.sg/-/media/files/publications/reference/sif2018.pdf

[11] D. Deterding, "The North Wind versus a Wolf: Short texts for the description and measurement of English pronunciation," *Journal of the International Phonetic Association*, vol. 36, no. 2, p. 187, 2006.

[12] ——, "Phonetics and Phonology," in *Singapore English*. Edinburgh University Press, 2007, pp. 12–39.

[13] A. Pakir, "Lexical Variations in "Singapore English": Linguistic description and Language Education," in *Corpus Analysis and Variation in Linguistics*, Y. Kawaguchi, M. Minegishi, and J. Durand, Eds. John Benjamins Publishing Co., 2009, pp. 83–102.

[14] V. R. Savage and B. S. A. Yeoh, *Singapore Street Names: A Study of Toponymics*. Marshall Cavendish Editions, 2013.

[15] R. Baker and V. Hazan, "DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs," *Behavior Research Methods*, vol. 43, no. 3, pp. 761–770, 2011.