Spontaneous Speech Elicitation for Large Speech Corpus in Multilingual Singapore

Ying-Ying Tan

School of Humanities, Nanyang Technological University Singapore
yytan@ntu.edu.sg

Abstract

In November 2017, the Singapore government tasked the Info-communications and Media Development Authority of Singapore (IMDA), the technological and media arm of the Singapore government, to build the National Speech Corpus (NSC). The objective of building the NSC is to have a corpus of standard Singapore English that can serve to improve the accuracy of speech recognition engines to handle locally-accented English, and to encourage the creation of innovative speech-enabled applications for various industry sectors. The driving motivation behind this is the apparent inability of existing commercial off-the-shelf speech recognition systems to understand the Singaporean English accent with a reasonable level of accuracy. The IMDA has, since mid-2018, started the process of building the speech corpus, and the completed corpus will have over 3000 hours of voice recordings of Singaporean speakers from different groups, in both read and spontaneous speech, making it the largest, most comprehensive corpus of any language in the country thus far. The author of this paper was tasked to be involved in the building of the NSC, and has been the only academic on the team since its inception. And as the only linguist on the team of engineers and government officials, I was asked to check on the design for data elicitation. Bearing in mind that the main purpose of the NSC is not for academic research, but for the development of speech technologies in Singapore, one had to work within the guidelines given by the authorities. However, this does not mean that this corpus cannot be used for academic research. To make it work for academic research however, a few questions need to be answered in the process, namely,

(1) What are the current gaps in the phonetic research on Singapore English, and how can the NSC help address these gaps?
(2) What are the best methods to elicit speech data that can serve the needs of *both* speech technologies and academic researchers?

This paper covers both the above questions. Specifically, this paper details the process of spontaneous data elicitation, and aims to describe and compare three methods of spontaneous speech elicitation for a large speech corpus such as the NSC. This paper will look specifically at one phonetic feature, namely the postvocalic-*r*, as a basis of comparison across the three methods.

1. Introduction

In November 2017, the Singapore government tasked the Info-communications and Media Development Authority of Singapore (IMDA), the technological and media arm of the Singapore government, to build the National Speech Corpus (NSC). The objective of building the NSC is to have a corpus of standard Singapore English that can serve to improve the accuracy of speech recognition engines to handle locally-accented English, and to encourage the creation of innovative speech-enabled applications for various industry sectors [1]. The driving motivation behind this is the apparent inability of existing commercial off-the-shelf speech recognition systems to understand the Singaporean English accent with a reasonable level of accuracy. The IMDA has, since mid-2018, started the process of building the speech corpus, and the completed corpus will have over 3000 hours of voice recordings of Singaporean speakers from different groups, in both read and spontaneous speech, making it the largest, most comprehensive corpus of any language in the country thus far.

The author of this paper was tasked to be involved in the building of the NSC, and has been the only academic on the team since its inception. And as the only linguist on the team of engineers and government officials, I was asked to check on the design for data elicitation. Bearing in mind that the main purpose of the NSC is not for academic research, but for the development of speech technologies in Singapore, one had to work within the guidelines given by the authorities. However, this does not mean that this corpus cannot be used for academic research. To make it work for academic research however, a few questions need to be answered in the process, namely,

(3) What are the current gaps in the phonetic research on Singapore English, and how can the NSC help address these gaps?
(4) What are the best methods to elicit speech data that can serve the needs of *both* speech technologies and academic researchers?

This paper covers both the above questions. Specifically, this paper details the process of spontaneous data elicitation, and aims to describe and compare three methods of spontaneous speech elicitation for a large speech corpus such as the NSC. For the purpose of comparison across the three methods, this paper will look only at one phonetic feature, namely the postvocalic-*r*.

2. Challenges in phonetic research on Singapore English

Linguists consider Singapore English to be the "standard" form of English found in Singapore, whose syntax and lexicon are not distinctly different from other "standard" British, American or Australian varieties [2]. While Singapore English phonology has been well researched (see [2] for more details), much of the current research on Singapore English sounds are based on small and skewed sampling. As most academic research is done in universities, researchers tend to use

university students as their participant pool, and few succeed in getting access to participants outside university grounds. What this means then is that past research in this area can be said to be representative only of a small segment of the Singaporean population, i.e. university-educated Singaporeans in their early twenties. Little, if nothing, is known about the speech patterns of Singaporeans beyond this group. What are the speech patterns of Singaporeans of different age groups, educational profiles, and socio-economic status?

Singapore's diverse population also yet presents another challenge for speech research. We currently know very little about how Singaporeans from different ethnic groups speak. Most studies on Singapore English have been mainly focused on the speech of ethnically Chinese Singaporeans, who constitute an overwhelming majority of Singapore's population at 76.2% [3], and little has been done to understand the speech patterns of speakers belonging to the other two major ethnic groups in Singapore. Approximately 15% of Singapore's population are Malays and 7.4% are ethnically Indian [3]; all of whom do speak a different language apart from English. Research over the past two decades on Singapore English has shown that there are some specific segmental and prosodic patterns that are unique to the three major ethnic groups in Singapore [4], [5]. In the few descriptive studies focusing on segmental features, researchers tend to show how there are "prototypical" sounds that are characteristic of each ethnic group. In the area of prosody, which tends to present more conclusive research showing ethnic differentiation in Singapore English, there is evidence to show that sounds from the speakers' 'first' or 'native' language are assumed to have made an imprint on the speakers' English. Beyond these snippets of features however, there has been no large-scale research showing how Singaporeans from different ethnic backgrounds speak, and this is particularly important given Singapore's multicultural and multi-ethnic make-up.

With the exception of the research done on the NIE Corpus of Spoken Singapore English (NIECSSE) [6], almost none of the above-mentioned research on Singapore English has been based on spontaneous speech. Yet we know that speech produced during an interaction differs from read speech, especially in terms of segmental phonetic features and in prosody [7], [8]. Furthermore, read speech and spontaneous speech are perceptually distinguishable even if they contain the same speech material [9]. However, spontaneous speech, especially the ones in conversation, are not easily obtainable as they are more difficult to process and analyse as compared to read speech. Conversational speech corpora are also more expensive and difficult to obtain than read speech. And more importantly, especially for phonetic research that demands for some control of phonological environments, spontaneous speech takes the control away from the researcher. What then are the best ways of eliciting useable data for phonetic research in spontaneous speech?

Given the opportunity to build a large corpus like the NSC on Singapore English, how should one build the corpus such that it can also facilitate speech research in Singapore and cover the issues as highlighted above? The next few sections discuss the building of this large spontaneous speech corpus of Singapore English.

3. The National Speech Corpus

There are a few existing speech corpora for English in Singapore, but none with the scale and scope of the NSC. The NIE Corpus of Spoken Singapore English (NIECSSE) consists of some 20 odd interviews between Singaporean teacher trainees and their British professor and dictation of a phonetically designed passage [10]. There is also the Singapore component of the International Corpus of English (ICE-SIN), and the Grammar of Spoken Singapore English Corpus (GSSEC). The GSSEC has about 8 hours of conversations incorporated into the roughly 600,000-word ICE-SIN [11], and the conversations were collected under natural conditions, making the noisy data unsuitable for phonetic or acoustic analysis. A Computer-Assisted Language Learning system was produced [12], and while it appears to be the largest in comparison, the 125 hours of speech collected from 83 university educated speakers were all read, and not spontaneous speech. As can be seen, the existing speech corpora for Singapore English seem sorely inadequate in terms of addressing the gaps in phonetic research in Singapore English, and there is also a lack of resource for spontaneous speech.

3.1. Describing the corpus

Currently, the NSC consists of three parts: 1) 1000 hours of read speech with randomised sentences drawn from periodicals and phonetically balanced scripts, 2) 1000 hours of read speech featuring local words and items, many of which are from the other languages in Singapore, and 3) 1000 hours of conversational, spontaneous speech. See [13] for a full description of the NSC. This paper focuses only on the elicitation of the 1000 hours of conversational spontaneous speech. Section 3.2 is an adaptation of [13] in terms of the description of the data elicitation procedure and participant demographic information.

3.2. Procedures for conversational data elicitation

1000 hours of conversational speech in Singapore English were collected, 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 around 250 pairs of speakers.

Speakers were recommended to bring a partner, preferably a friend or family member with whom they could speak for at least 2 hours. 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.

As mentioned in Section 2, the gaps in phonetic research in Singapore include not having large sample size, and having enough representation of speakers of different ethnic groups, age groups, and of different educational background. Effort was therefore made to ensure that there was a good distribution of speakers along these lines. In addition, speakers were also requested to provide information about their linguistic background, and language repertoire. Tables 1, 2, 3, and 4 are the demographic breakdown of the participants.

Table 1: *Proportion of speakers distributed by gender*

Gender	FTF(%)	Tel.(%)	Overall(%)
Female	52.5	54.9	53.7
Male	47.5	45.1	46.3

Table 2: *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

Table 3: *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

Table 4: *Proportion of speakers distributed by education*

Education	FTF(%)	Tel.(%)	Overall (%)
University or higher	49.0	39.8	44.4
Jr.College/ Polytechnic	31.9	38.8	35.3
Secondary or below	19.1	21.5	20.3

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 speaker was recorded on two channels, the first allowed for the collection of data from each individual speaker, therefore making it easy to work with the data without overlap and interruptions from the conversational partner. The second channel recorded the entire conversation of both speakers. Each recording session was approximately 2 hours 15 minutes long. Speakers were compensated for their time.

3.3. Spontaneous speech elicitation tasks

Bearing in mind that the NSC is a government project whose main objective is to elicit voice samples that can serve speech recognition technologies, the main consideration was to be able to have speakers converse for at least two hours in "standard" Singapore English naturally. Three tasks were presented to the participants to elicit natural conversations from them. They are:

1) Spot-the-difference diapix
2) Conversation card games
3) Free-talk prompts

3.3.1. Diapix

In the diapix task, speakers were asked to collaborate and pick out 12 differences between two similar pictures without looking at each other's pictures. The pictures were adopted from DiapixUK picture materials [9]. This task was useful for eliciting descriptive and directional phrases. It also allowed for some control over the lexical content of the interaction. The pictures were switched out for new ones periodically so as to introduce some diversity in content. Speakers on average took around 10 to 20 minutes to complete the task.

3.3.2. Conversation card games

In the second task, two different sets of conversation card games were procured to act as conversational prompters. The FTF sessions used *smol tok*, a card game set with prompters localized to the Singapore context [14]. Depending on how conversational the speakers were, speakers could finish their deck of cards in as short as 45 minutes.

3.3.3. Free-talk prompts

This third task is the simplest and most conventional method of eliciting spontaneous speech. In this task, the speaker pairs were instructed to converse spontaneously about a particular topic, and this topic can be chosen from a set of prompts which ranged from vacation spots to favourite food. *Spontal* is an example of such a corpus [15]. As expected, this approach allows for spontaneous and natural speech, but researchers have no control over the lexical content of conversations.

4. Comparing the methods

This section compares the three methods of spontaneous speech elicitation, and for the purpose of this paper, a small sample of five pairs of speakers will be used for this comparison. To eliminate possible effects of ethnicity, educational level, age, and language background, the ten speakers whose data is described here are somewhat similar. All ten speakers are aged between 18 to 23. They are all Chinese Singaporeans, bilingual in both English and Mandarin-Chinese, with English as their dominant first language. All ten speakers are current students at local universities. The speaker pairs are also friends with each other.

The comparison will first provide a general description, and then focusing on one phonetic feature, namely, looking for the presence of the postvocalic-*r*, to highlight the suitability of each method for phonetic research. The postvocalic-*r* has been chosen since it has been shown to be a new phonetic feature that is used increasingly by young Singapore English speakers [16].

4.1. General characteristics of the three tasks

This section outlines some observations of the three tasks across all six speaker pairs.

4.1.1. Diapix

The shortest time taken to complete the diapix task is eight minutes, and the longest is 20 minutes. Most speaker pairs take around ten to twelve minutes on average. The strategies employed by the speakers in the diapix task are fairly consistent, making it possible therefore for researchers to design the diapix to elicit target words with a fair amount of control. The strategy typically involves a speaker in each pair to initiate the starting point, and this is usually at top left or top right of the picture. Speakers usually describe colors and numbers first. Speakers also tend of describe landmarks and use landmarks as points of confirmation. In general, speakers who do more turn-taking and feedback tend to complete the task faster, regardless of the difficulty level of the diapix.

However, this is not usually the case as it is more the norm for one speaker to lead, while the other confirms.

4.1.2. Conversation card games

The card game task yielded the longest recording time. All six pairs of speakers took an average of 90 minutes to complete one stack of cards. This task generally worked well in eliciting conversations. The framing of some of the questions in the card games created casual responses, and some questions also resulted in personal, intimate conversations. The flow of conversations generally stuck closely to the card game questions. While this means that researchers have no control over the lexical content of the conversations, the card game is useful for eliciting long stretches of spontaneous speech in a relaxed setting. One other advantage of the card game is that there is usually equal contribution between the two speakers, as speakers take turn to ask the question on the card, and responding in turn. One can think of this as a guided "interview" with a fun element between the speaker pair. What this also means is that this set-up gives us spontaneous speech data with relatively less overlap or interruptions between the speakers. Unlike the diapix task, the card game task can be used for eliciting speech beyond two speakers at each time.

4.1.3. Free talk prompts

Depending on the topic, the time taken for this task can be as short as 5 minutes, or as long as 30 minutes. Conversations were generally awkward and stilted, though occasionally, the speaker pair may get excited over a particular point, for example, a shopping trip, and ended up spending a little more time talking about it. This task, while the easiest to execute, is also the most unpredictable. It depends entirely on the speakers' interests, and willingness to talk about a topic. As can be expected, the content of the free talk is not within the control of the researcher, making it rather difficult if one's aim was to look for specific target words. The contribution from both speakers also tend to be uneven.

4.2. Locating a specific phonetic feature

In order to ascertain the suitability of each method in the analysis of specific phonetic features, I looked at the number of words in the data that had the potential for the occurrence of postvocalic-*r*. This includes words with *r* in the spelling in word final positions and syllable-final positions in polysyllabic words. For ease of reference, I will refer to them as *r*-words. Table 5 provides the percentages of the occurrences of *r*-words words in relation to the total number of words spoken per speaker. Table 6 provides the breakdown of *r*-words per task. Speakers are numbered, and letters next to the speakers indicate their pairing.

Table 5: *Total word count and postvocalic-r words per speaker*

Speaker	Word total	Total *r*-word	%	Speaker	Word total	Total *r*-word	%
1-A	14862	560	3.8	1-B	26191	1115	4.3
2-A	13415	581	4.3	2-B	17369	856	4.9
3-A	24214	1000	4.1	3-B	20406	1087	5.3
4-A	18894	847	4.5	4-B	14263	577	4.1
5-A	15353	580	3.8	5-B	22239	956	4.3

As can be seen from Table 5, there is a fairly consistent yield of a specific phonetic feature in relation to the total number of words spoken per speaker per recording session.

Table 6: *r-words in the three tasks*

	Diapix			Card game			Free talk		
Speaker	*r*-word	Total token	%	*r*-word	Total tokens	%	*r*-word	Total tokens	%
1-A	29	47	61.7	98	428	22.9	44	85	51.8
1-B	35	113	31.0	103	766	13.5	69	236	29.2
2-A	31	99	31.3	92	373	24.7	27	109	24.8
2-B	31	124	25.0	110	572	19.2	40	160	25.0
3-A	29	86	33.7	112	749	15.0	57	165	34.6
3-B	15	50	30.0	131	765	17.1	71	272	26.1
4-A	33	99	33.3	103	478	21.6	88	270	32.6
4-B	25	90	27.8	89	316	28.2	55	171	32.2
5-A	27	97	27.8	100	345	29.0	42	138	30.4
5-B	30	120	25.0	102	689	14.8	38	147	25.9
Avg	29	93	32.7	104	548	20.6	53	175	31.3

The percentages in Table 6 give an indication of how often the *r*-words are repeated in each task. The card game, due to the length of recording time, yields the largest number of *r*-word tokens. However, it also produces a relatively small number of unique words. In other words, on average, each *r*-word is repeated 5 times, whereas for the diapix task and free talk, each *r*-word is only repeated 3 times. What this means then is that the card game is useful to the extent of getting a smaller set of lexical items, but with repeated tokens.

5. Future Directions

As a point of summary, the three tasks explored here come with their pros and cons, and researchers working on a large-scale corpus such as the NSC can choose the tasks they would like to work with given their needs. The following outlines the key features of each task.

Pros	Cons
Diapix task • Short, consistent recording time per picture (averaging 10 minutes) • Control of lexical items can be done by manipulating pictures • Predictable speaker strategies • Good yield of tokens (averaging 3 tokens per word)	• Restricted to dyad • Unequal contribution from each speaker
Conversation card game • Capable of having long recording time (average 90 minutes) • Able to elicit natural, relaxed speech • Control of lexical items not easy, but speakers tend to	• Due to the relaxed nature of the game, codeswitching can happen • Yields large number of tokens, but high occurrences of repetition

stick closely to cards, and manipulation can be achieved there. • Equal contribution between speakers • Less overlap and interruption between speakers • Able to accommodate multi-speaker setting	
Free talk • Easy to execute as no planning is required • If conversation goes well, can yield large number of tokens with an average of 3 repetitions per word	• Recording time varies according to speakers' inclination • Can be awkward • No control over lexical items • No control over codeswitching.

The NSC is planning for future phases of spontaneous speech elicitation, and new ways of eliciting spontaneous speech are currently being explored.

6. Acknowledgements

The author acknowledges the Info-Comm Media Development Authority of Singapore (IMDA) for the use of the National Speech Corpus.

7. References

[1] I. Tham, "Artificial intelligence library with voice samples in Singaporean English to launch next year," *The Straits Times*, 2017. [Online]. Available: https://www.straitstimes.com/singapore/artificial-intelligence-library-with-voice-samples-in-singaporean-english-to-launch-next?xtor=CS3-18

[2] Cavallaro, F., B.C. Ng, and Y.Y. Tan. (2019, in press) "Singapore English". In Kingsley Bolton and Andy Kirkpatrick (eds.), *Handbook of Asian Englishes*, Blackwell-Wiley.

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

[4] L. Lim, "Ethnic group differences aligned? Intonation patterns of Chinese, Indian and Malay Singaporean English," *The English language in Singapore: Research on pronunciation*, pp. 10-21, 2000.

[5] Y.Y. Tan, "The acoustic and perceptual properties of stress in the ethnic subvarieties of Singapore English". Unpublished Doctoral Dissertation. Department of English Language and Literature, National University of Singapore, 2002.

[6] D. Deterding, A. Brown and E.L. Low (eds). *English in Singapore: Phonetic Research on a Corpus.* Singapore: McGraw Hill.

[7] R. Baker, and V. Hazan, "LUCID: A corpus of spontaneous and read clear speech in British English," in *Proceedings of the DiSS- LPSS Joint Workshop 2010*, Tokyo, Japan.

[8] G.P.M. Lann, and D. R. Van Bergem, "The contribution of pitch contour, phoneme durations and spectral features to the character of spontaneous and read aloud speech". In *Proceedings of Eurospeech*, Berlin, Germany, pp. 569–572, 1993.

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

[10] 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

[11] 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

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

[13] J. Koh, A. Mislan, K. Khoo, B. Ang, W. Ang, C. Ng, and Y.Y. Tan, "Building the Singapore English National Speech Corpus" in *Proceedings of Interspeech* 2019.

[14] https://www.starknicked.com/

[15] J. Edlund, J. Beskow, K. Elenius, K. Hellmer, S. Stronbergsson, and D. House, "Spontal: A Swedish spontaneous dialogue corpus of audio, video and motion capture," In *Proceedings of the International Conference on Language Resources and Evaluation,* Valletta, Malta, 2010.

[16] Tan, Y.Y. "To r or not to r: social correlates of /ɹ/ in Singapore English" In *International Journal of the Sociology of Language*. 218: 1-24, 2012.