National Speech Corpus

Technical Reference for the Creation of Speech Corpora for Technology

1. Introduction

The advancement of machine learning methods aided by the improvement in computing power has boosted the creation of new technological capabilities and industries. Though the field of data collection and annotation is not new, the new wave of technological advancements has certainly drawn attention to the demand for large amounts of high-quality data and the opportunities available in this industry.

The purpose of this technical reference is to provide an effective source of information for which local companies that are interested in pursuing the creation of large-scale speech datasets, or speech corpora, can leverage on.

1.1. What is a speech corpus?

Speech corpora consist of speech audio files, corresponding text transcriptions of these audio files, as well as a lexicon. Speech corpora, especially large-scale ones that are used for speech-enabled applications, often require intensive resources to build. This is why certain languages, varieties, and even accents tend to perform better in these applications than the others. Some prominent corpora used in the research community include the Wall Street Journal read speech corpus (Robinson, Fransen, Pye, Foote, and Renals, 1995) and Switchboard spontaneous speech corpus (Godfrey, Holliman, and McDaniel, 1992) which are often used for benchmarking automatic speech recognition systems and algorithms, while CMU ARCTIC is commonly referenced in text-to-speech research (Kominek and Black, 2003).

1.2. National Speech Corpus

A point of reference that will be made throughout this technical reference is the National Speech Corpus (NSC). NSC was initiated in November 2017 by the Infocomm Media Development Authority of Singapore (IMDA) to provide companies and research institutions that are interested in developing speech-enabled tech applications with large-scale Singapore English accented speech datasets that is open and publicly available for use under the Singapore Open Data License. The corpus was initiated as part of the IMDA Industry Transformation Map to spur on speech-enabled tech applications and research efforts in Singapore since the construction of corpora is often resource-intensive and cannot be easily undertaken by start-ups or companies that wish to take advantage of these emerging technologies but are outside of the industry.

As of December 2019, NSC has contributed more than 3000 hours of Singapore English speech data and their corresponding transcriptions and lexicons, of which over 2000 hours are read speech, and more than 1000 hours conversational speech. Around half of the read speech corpora are recorded with phonetically balanced scripts to provide a baseline of English sentences read in the Singapore accent, while the other half focuses on named entities in Singapore such as food names and place names so as to aid technology providers who wish to localise their applications to the Singapore context. NSC has helped local technology providers such as Sentient.io to give their algorithms and solutions an edge over what international

companies offer. For access to the corpus and latest updates, visit www.imda.gov.sg/nationalspeechcorpus.

1.3. Scope of this reference

It is assumed that readers already possess a good grasp of related topics and notable works in the field of speech technology. Readers should also have some linguistic knowledge especially in terms of phonetics and phonology, as these would be essential especially when designing scripts for read speech corpora and in the construction of lexicons. Recommended texts for further reading are listed in Section 8.

It is important to keep in mind that different speech applications place different requirements and constraints on the speech corpora to be used, such as in terms of the number of speakers to record, or the amount of data to record from each speaker. This technical reference covers the creation of corpora, from the design to the post-processing of recorded speech data, mainly with the needs of speech recognition applications in mind.

Though the chapters are ordered according to the workflow of corpora building, they should be read and considered in entirety, and not as separate steps. The planning and consideration of practical, logistical, and ethical issues is essential to the success of the corpus.

2. Designing read speech corpora

A read speech corpus refers simply to one where speakers are recorded reading a script or set list of words, phrases, or sentences. One of the main advantages of constructing read speech corpora is that these tend to be relatively easier to obtain and allow for more control over the content. On the other hand, one of the biggest trade-offs is the mismatch between the style of speech, where read speech corpora is ill-suited to train models that are meant to perform on more conversational or natural, spontaneous speech.

Of the 2000 hours of read speech in NSC, around half was recorded using a phonetically balanced script, and the other half a script that features words that are pertinent to the Singapore context. These will be elaborated in the following Sections 2.2 and 2.3. We will first look at the considerations for designing a phoneme inventory which often works in tandem with the designing of the scripts.

2.1. Phoneme inventory

Phonemes refer to the distinguishing or contrasting units of sounds in a language, and a phoneme inventory simply refers to the set of sounds. English has been described to possess around 44 phonemes, though the number varies by dialect and definition. For instance, there are around 20 vowel phonemes in British English and 15 to 19 for American English (Bizzocchi, 2017). For Singapore English, Deterding (2007) tabulated 22 consonants and 8 vowels, which the NSC has adapted upon to include the addition of two affricates and six diphthongs. Given that there are also non-English words found in the corpus, an additional 11 consonants and 1 vowel were included in the phoneme inventory.

The phoneme inventory is used as the basic units in acoustic modelling and places limits on the selection of the scripts used for the recording, as well as in the construction of the lexicon in the later part of the corpus-building (Section 6.2). It also ties in with the phonetic characteristics of the corpus, which will be covered in the next section.

2.2. Phonetic balance and richness

For robust study of a particular dialect (in the linguistic sense) or accent, the phonetic coverage of the corpus is often taken into consideration. This is done by calculating the phone (i.e. phoneme), diphone (adjacent pair of phones), and triphone (combination of three consecutive phones) statistics of the corpus. A corpus is phonetically balanced when the coverage of phones, diphones, and triphones that occur in the corpus aligns with the frequency of occurrence in natural usage. In contrast, a corpus is phonetically rich when all the phones, diphones, and triphones that occur in a language are uniformly distributed (i.e. roughly same amount of training data) in the corpus. Speech corpora that are meant for the purposes of speech recognition have the primary need to be phonetically balanced, while those that are designed for speech synthesis need to be phonetically rich.

The NSC includes a phonetically balanced portion to provide a baseline of English speech spoken in the Singapore accent. A minimum of 1000 hours of speech was targeted to be recorded from 1000 speakers (i.e. 1 hour of read speech from 1 speaker). To do so, all speakers had to read a set of 200 sentences that were designed to include the phonemes that have been documented to occur in Singapore English. They also had to read another set of 600 sentences that were randomly selected from a large pool of 72,000 sentences that were crawled from a range of online local news sources. Each sentence could be repeated a maximum of 8 times so as to obtain more speaker variation and more coverage of linguistic phenomena.

When designing prompts, it is recommended that the number of words in each sentence be kept between 10 to 20 words. Prompts that are too short risk not having enough linguistic phenomena being captured in the corpus, while prompts that are too long are difficult for speakers to read without too much disfluencies. It is also important to check and obtain the necessary permissions for published texts before using them as prompts.

2.3. Local words/ named entities

Names are often a challenge for speech-enabled applications, given their highly diverse nature and the general difficulty in capturing the range of names that are and will be used comprehensively. This means that names tend to be out-of-vocabulary, i.e. they do not appear in the training data. While researchers in named entity recognition are still looking for new and better ways to solve this problem, the ability of technology providers and their speech applications to localise and tailor to the local context will no doubt play an influential role in the usability, and consequently, the take-up rate of the applications.

To give an example, even though the working language in Singapore is English, a large proportion of “local words” such as food names, place names, people’s names are not in English but originate instead from the other languages that Singaporeans speak, like Chinese, Hokkien, Malay, and Tamil. Speech applications that are not trained with such data would only be able to recognize and synthesize these named entities with the nearest English variants that may not even be phonemically similar, affecting the usability of applications that rely on accurate recognition or synthesis of named entities, including hands-free calling and navigation systems.

To aid technology providers working to tailor their products for Singaporean consumers, the NSC has provided more than 1000 hours of read speech data recorded using prompts with local named entities. Lists of words belonging to the aforementioned categories, in addition to other categories such as brand names and abbreviations (which are used frequently in Singaporean

discourse and speech), were compiled and parsed into grammars to automatically generate scripts to be read as prompts for NSC. Speakers read a total of randomly selected 896 sentences. A common challenge with such corpora is that speakers often make errors when reading unfamiliar words or phrases (Hughes et al., 2010), though patterned errors, such as Malay and Indian speakers in NSC pronouncing /y/ in Chinese names as /ju/, could still prove useful for developers.

3. Designing conversational speech corpora

Unlike a read speech corpus, a conversational or spontaneous speech corpus contains speech data that is produced freely and naturally. Past findings have revealed that speech recognition systems that are solely trained on read speech do not perform as well as systems trained on spontaneous speech when faced with spontaneous speech input, unless it can be ensured that the input is consistently fluent, that is, no fillers, false starts, repairs, or long pauses (Butzberger, Murveit, Shriberg, and Price, 1992). The match between the speaking styles of the training and test data is therefore important in developing and providing effective speech-enabled applications.

However, there are also some disadvantages to building conversational or spontaneous speech corpora. Unlike read speech, spontaneous speech means that there is less control over what occurs in the data, and exerting control on what speakers can or cannot say may affect spontaneity instead (Furui and Kawahara, 2007). Transcribing the data also requires significantly more human effort and labour, which is why it is expensive to collect and build spontaneous speech corpora.

The 1000 hours of conversational speech recorded for NSC were recorded in two modes: same-room, where speakers talked to their partners face-to-face, and telephone, where speakers were put into different rooms and talked to each other through the telephones provided. Each recording session was two hours long – assuming each speaker would contribute around an hour of data – with an additional buffer of 15 minutes to cater for a short break and long silences that sometimes occur during conversations. Speakers were encouraged to bring a family member or a friend whom they could speak at least two hours with as their partner for the recording as people tend to be more open to talking freely around someone they are familiar or comfortable with. However, not all speakers are able to bring a partner, and this is an important point take note of for future conversational speech corpora as well. These speakers were paired up with another speaker, i.e. a stranger, in NSC, and their relationship was recorded in the metadata.

Conventionally, conversational speech corpora may record speakers conversing freely. However, to minimize the occurrences of silences throughout the session, three methods of elicitation were used. The first was a spot-the-difference task using diapix materials from Baker and Hazan (2011) where speakers were asked to spot twelve differences in two similar pictures without looking at each other's pictures. This served as a warm-up to the recording environment while eliciting descriptive and directional phrases, and generally took between 10 to 20 minutes to complete. Next, speakers played a conversational card game where they had to take turns asking each other questions on the cards, which lasted around 45 minutes to the whole of the recording session, depending on how much the speakers could elaborate on the questions. Finally, for the speakers who still had some time before the end of the recording session, they were given a list of topic prompts as reference for discussion.

An analysis by Tan (2019) of the performance of the three elicitation methods on the quality of conversations found that while diapix tasks may be designed with target words in mind, there would usually be a speaker who takes the lead in suggesting answers and the other confirms. Likewise, speakers generally stuck close to the questions on the conversational card games, resulting in a smaller set of lexical items with repeated tokens as compared to free-talk prompts that were the easiest to execute but unpredictable in terms of what the speakers would say. In a way, conversational card games also allowed for equal contribution as speakers took turns to ask and answer questions, and thus may be a useful way of eliciting balanced and more controlled conversations.

4. Recruitment of speakers

4.1. Demographic considerations

While corpora that are built for the purposes of speech synthesis require a large amount of data from a few individuals, speech recognition corpora require speech data from a wide variety of speakers. An understanding of the targeted linguistic variety and the underlying sociolinguistic variables is therefore crucial to the representative quality of the corpus.

According to the targeted number of speakers, decide on the demographic distribution that needs to be fulfilled. For NSC, a few demographic variables were prioritised, specifically gender, age, and ethnicity. The distribution of male and female speakers was needed to be largely equal. Three target age groups were also used – 18 to 30 years old, targeted at around 50% of the corpus, 31 to 45 years old, at 30%, and over 46 years old at 20%. To avoid an over-representation of Chinese Singaporeans – which comprises 74.3% of the 4.0 million Singaporean population in 2017 – in the corpus, a target of 50% Chinese, 25% Malay and 25% Indians was implemented, with some allowance for speakers who do not fall into these categories.

Considerations were also made with regards to educational level of the speakers so that there would be representation especially from speakers of lower educational backgrounds, but this had some difficulty to achieve due to the increasingly educated population of Singapore.

4.2. Recruitment process

Different corpora with different purposes naturally require different types of participants. For read speech, it would be more important to recruit or prioritise speakers who are able to read sentences fluently without too much practice, and without too many mistakes. It may thus prove worthwhile to conduct interviews to assess the literacy and fluency levels of the speakers prior to confirming their participation in the recordings. One simple way to do so is through phone interviews, where a set of test sentences are sent to the speaker shortly before or during the phone interview, and the speaker is asked to read out the sentences on the spot.

After determining the targeted demographics of the corpus, list out the criteria that would be used when assessing the suitability of the participant.

In recruiting the speakers for NSC, some criteria were used in the selection. Speakers had to firstly reach at least 18 years of age so that they are legally (according to the law in Singapore) able to consent to participating in the recordings. To ensure that the speakers grew up in contact with the prototypical local accent, the following criteria were also used:

- Must be a Singapore citizen raised in Singapore, or

- Residents who have lived in Singapore for at least 18 years, and
- Have undergone a formal education in English in Singapore **public** schools for at least 6 years

4.3. Participant data collection

For researchers and developers working in the field of speech technology, participant data is an important and essential source of information that allows them to understand the speech recordings with greater depth so as to derive useful insights and develop better applications. As such, the collection of participant data should be conducted in a purposeful manner, considering what data is likely to be necessary and useful for the target users of the corpus, and the various resources that are needed to store and publish the information in a manner that still protects the anonymity of the participants. The following is a non-exhaustive list of participant information that may be useful in the context of speech corpora as they are sociolinguistic variables that can explain reasons for language variation that possibly occur in the corpus:

- Gender
- Age
- Educational attainment
- Socioeconomic status
- Linguistic repertoire and proficiency
- First language/s (languages first acquired when born)
- Region in which the speaker was brought up in

Table 1 lists out some types of information that are sensitive, i.e. high likelihood of identifying the corpus speakers, and solutions for presenting the information during publishing such that anonymity is still preserved. Keep in mind that information that are not sensitive on its own (e.g. gender, ethnicity) may be combined with other information such that it becomes possible to deduce the identity of the speakers.

Table 1. Types of sensitive information and how to anonymise it

Sensitive information	Solution
Name	Give each speaker a unique speaker ID instead
Exact address	Replace with state, city, or town. The larger the grouping, the less identifiable it will be
Contact information	Omit from publishing and protect in a secure and confidential manner
Age	Replace with age groups
Income	Replace with income brackets and expect that participants may not be willing to provide such information

For conversational speech recordings, it would also be useful to make a note of the partner's speaker ID and the relationship that the speakers have with each other to facilitate further data selection and analysis. Keeping field notes of notable occurrences during each recording is also a good practice to maintain.

5. Recording setup

5.1. Recording equipment

For the read speech corpora in NSC, three microphones were used: a close-talk or standing microphone, a boundary microphone, and a mobile phone. The first two microphones were connected to a laptop through an audio interface or audio card while the mobile phone was placed near the speaker. The audio interface used was one from the Focusrite Scarlett series. An in-house recording software displayed the prompts on the laptop and speakers were taught how to navigate and record themselves reading the prompts.

The microphone set-up also differed according to the mode of recording that the speakers participated in for the conversational speech portion of NSC. For same-room recording, a close-talk microphone was given to each speaker, and a boundary microphone was placed on the table to record both speakers. For telephone recording, standing microphones were placed in each room in front of the speaker in addition to the telephone set that was connected internally through VoIP using an Interactive Voice Response system. Similar to the previous set-up, all microphones apart from the telephone were connected to a laptop through an audio interface. Adobe Audition – a digital audio workstation – was used to initiate and end the recordings.

When deciding your recording set-up, keep in the mind of the following points. Far-field microphones are omni-directional, which means all speakers can be recorded at the same time. However, these microphones also record ambient and environmental noises, as well as noises unintentionally generated by speakers when they fidget or knock on the tables. This will increase the difficulty of speaker separation (for multi-party data) and transcription, and possibly reduce the recording quality.

On the other hand, close-talk and standing microphones provide the highest audio quality which can allow for the use of automatic transcriptions prior to manual transcription. A pitfall is that they are prone to aspiration and spurts when they are positioned too closely to the speakers. Some speakers may also find close-talk microphones uncomfortable, resulting in more fidgeting and thus affecting the recording quality. However, close-talk microphones are more discrete as compared to standing microphones placed directly in front of the speakers.

Given that different microphones have different strengths and weaknesses, it is worth considering setting up multiple microphones in your recording environment as the weaknesses of certain microphones may be compensated by the strengths of another. For instance, data recorded by close-talk microphones could be used to train the automatic transcription of far-field microphone recordings, given that the audio times are in sync. A multi-microphone set-up will also allow for different kinds of audio data to be captured.

5.2. Recording environment

The NSC was recorded mainly in quiet office spaces, though some speakers recorded in professional studios which provided more soundproofing from external noise.

Some points to take note of when testing the acoustics of the room:

- Background noise such as from air-conditioning should be limited (noise may be added to clean audio through post-processing if noisy audio is the intended purpose of recording)

- Be aware of the noises generated from outside the room at different times of the day or week, e.g. traffic noises, building plumbing, opening/closing of doors from other rooms, etc.
- Echo or reverberation levels of the room will add extra noises in the recordings. Check beforehand using software that can help you to determine and measure the RT60 levels, or perform a clapping test to obtain a rough estimate

The location of the recordings should also be held in clean and professional spaces as far as possible. This ties in with the ethics of the recordings and ensures the safety of both recording staff and speakers. For conversational speech, comfortable environments aid in relaxing and opening up speakers to talk more freely, and so it may be worthwhile to decorate the room to look like a resting or leisure spot. Obvious microphones and recording equipment set-up may become a constant reminder to the speakers that they are being recorded and may lead to some people altering their natural way of speaking, and hence this is also a point to note when choosing microphones and setting up the recording environment.

Ensure also that the materials placed in the recording room serve a purpose. Having extra materials such as prompts or posters that are unintended for the particular group of speakers could result in wasted effort or incongruency with the other sets of data recorded. You may also need to weigh the pros and cons of the speakers carrying personal electronic devices such as smartphones or tablets during the recording – whether these devices will help to elicit more natural conversations or distract the speakers away from the conversation instead.

5.3. Recording process

5.3.1. Prior to the recording

It is always good practice to reconfirm the recording session with the participants before the day of the recording. For conversational speech formats that require speakers to attend in pairs or groups, expect and prepare for late attendees or no-shows in advance.

Prior to the start of the recording, ensure that all recording equipment are in working condition. Conduct a detailed briefing to let the speakers understand what they can expect to happen throughout the session, as well as what is expected of the speakers. For instance, a non-exhaustive list of information that would be good to cover:

- Purpose of the recording and what the speech data would be used for
- Purpose of collecting participant information
- Duration of the session and availability of breaks
- Any tasks that they would need to carry out
- Expectations or requirements regarding speech styles, languages, topics to talk about or avoid talking about, etc.

Written consent should be given before the start of the session, and speakers should be informed that they have the right to leave the recording session at any point in time. Ensure that documents filled in and signed by the speakers are placed securely in a designated position where only authorised personnel are able to access. As these documents tend to ask for identifying and sensitive information, it is crucial to ensure that both hardcopy and softcopy versions are stored safely and confidentially. A leak of information could bring about legal repercussions.

It is encouraged to provide some bottled water for speakers, especially in long recordings of more than 30 minutes to ensure that speakers are not only physically comfortable when speaking but also that the audio quality does not get affected.

5.3.2. During and after the recording

A dilemma in corpus building concerns whether there should be a recording assistant present in the room during recording. A common reason not to have a recording assistant is that having a third party, effectively a stranger, may lead some people to alter their natural ways of speaking, or even avoid personal topics altogether, regardless whether it is deliberate or not. On the other hand, having a recording assistant means that common recording audio issues (such as when speakers adjust the microphones too close to themselves) can be easily and quickly rectified. A recording assistant who is trained as an interviewer may also help to elicit conversations, effectively minimising silences that naturally occur during spontaneous conversations. It is thus up to the corpus builder to decide what best fits the goals and purposes of the speech corpus.

Regardless of the type of speech data being recorded, it is best that speakers are given ample breaks throughout to rest their voices and buffer the length of the recording session accordingly. An acceptable duration of a recording session is around 1 hour, after which physical fatigue is likely to set in and in turn affect the audio quality.

After the end of the recording session, ensure that the recordings have been properly saved in the correct format and destination, and prepare the prompts and recording room for the next session.

6. Data processing

The work of processing recording data is perhaps the most laborious part of creating a speech corpus. As with all other steps in the building of corpora, decisions relating to how the data would be processed should be discussed in the initial development stage. A major part of data processing deals with transcribing the data, and though there are tools available such as speech recognition software to derive a base transcript first, delivering high quality transcripts still requires a great deal of human effort. In the following subsections, we will focus on some best practices for ensuring quality transcriptions as well as the building of the lexicon, otherwise known as the pronunciation dictionary. Audio processing will not be covered in this reference, though it may still be a necessary step to take depending on the audio quality and/or needs of the corpus.

6.1. Transcribing the data

Prior to transcribing the data, transcription rules should be discussed and agreed upon. How many levels of transcriptions (orthographic, phonetic, prosodic, boundary markings and time alignments, etc.) and how broad or narrow the transcription needs to be would depend on the

intended purposes and goals of the corpus. Commonly used transcription software includes Praat (Boersma and Weenink, 2020) and ELAN (Version 5.9; 2020). Transcribers should be trained and be familiar with the transcription rules so as to ensure accuracy and consistency in their transcriptions. A proper system in place to keep track of who has transcribed what would help when the need to rectify transcriptions arises. Each transcription work should be checked by at least another transcriber, and conflicting or varying arguments should be discussed to decide on the most appropriate transcription. This is known as multi-annotator agreement. Some may choose to have this process done in teams of threes or fives, with the majority argument taken as the checked transcription. Finally, transcribers should be native speakers of the language variety that the corpus is built for. This is especially important for corpora that require a high degree of local knowledge.

For read speech, most of the work required has to do with checking. Though speakers generally read according to the prompts, reducing the laboriousness of transcribing, disfluencies and misreadings or mispronunciations may still occur occasionally. These would need to be accounted for in the transcriptions.

Conversational speech, on the other hand, requires more human effort. Automatic transcriptions may be done first to generate a base transcription that can help transcribers speed up their work. Though often taken for granted, a major reason to why humans are able to disambiguate words easily from a string of sounds is due to the contextual and world knowledge that one possesses. As such, transcribers should be responsible for the entirety of a recording, rather than working in snippets of speech or on one speaker's recording but not the partner's. Transcription rules for conversational speech would also need to include the standardization of spelling of words that do not have a conventionalised form in general dictionaries. These could range from slang words or local lexical items, where transcribers are likely to vary widely in their spelling preferences.

The following is a non-exhaustive list of phenomena that you may wish to consider in deciding your transcription rules:

- Pauses
    - Boundary-marking e.g. end of sentence/ commas
    - Mid-sentence hesitations
- Numbers
- Symbols e.g. $ spelt full form as *dollar*
- Titles e.g. mister, doctor, professor
- Acronyms e.g. SAFRA
- Initialisms e.g. IMDA
- Multi-word nouns
- Discourse particles e.g. *lah*/*la*/*luh*
- Capitalisations
- Fillers e.g. *uh*/*er*
- Other languages
- Unclear words/ segments
- Incomplete words/ false starts/ repairs
- Mispronunciations

- Identifying/ sensitive information
- Paralinguistic phenomena e.g. laughs, coughs, breaths
- Non-speech acoustic events

6.2. Lexicon

A lexicon, or a pronunciation dictionary, provides phonetic transcriptions to **all** the words found in the corpus. This is typically organized in a two-column text file with the orthographic form on the left and the phonetic transcription on the right. Though the International Phonetic Alphabet provides a comprehensive alphabet to transcribe phonetic sounds, the signs are often not easily machine-readable. As such, lexicons for speech applications have used other ASCII phonetic scripts such as ARPAbet or SAMPA. ARPAbet was developed and used in many American English systems and in the CMU Pronouncing Dictionary ("The CMU Pronouncing Dictionary", n.d.). On the other hand, SAMPA was developed first for European languages, with 24 languages – including Cantonese and Thai – being covered at this point of writing ("SAMPA computer readable phonetic alphabet", n.d.). A variant of SAMPA known as X-SAMPA was also developed to cover all of the symbols used in the International Phonetic Alphabet.

Similar to orthographic transcriptions, it is possible to generate phonetic transcriptions automatically such as through grapheme-to-phoneme (g2p) converters or dictionaries as a base for transcribers to work on. As mentioned earlier in Section 2.1, the phoneme inventory is crucial to the construction of the lexicon as it effectively places constraints on the automatic and manual transcriptions that will be obtained. Increasingly, pretrained g2p converters have been developed using state-of-the-art machine learning techniques and thus may help in generating base pronunciations more effectively, even for multilingual corpora.

In the case of NSC, a major challenge to building the lexicon has to do with the loanwords that typically occur in Singaporean speech, as mentioned earlier in Section 2.3. Given that there is no official standardized pronunciation of these non-English named entities, ethnic variation especially from speakers who do not speak the language that a particular word may originate from have resulted in multiple pronunciations for many of the loanwords covered in the corpus (Koh et al., 2019). These differing pronunciations have to be taken into consideration, even as they deviate from the canonical, especially if it has become part of the speakers' variety. It is therefore highly encouraged for corpora that cater to multilingual societies to have a diverse team of transcribers who are native speakers of the local variety so that there are credible sources to base the transcriptions on.

6.3. Checking and evaluation

As errors are guaranteed to occur, it is important to put in place a system for quality checking of the transcripts and lexicons. Spell-checking and formatting scripts may help to reduce the workload of correcting minor errors but are likely miss out on more significant errors such as where the transcriptions are spelt correctly but ultimately do not match what is being said. These errors are often only picked up on during manual checking, and hence we reiterate here the importance of keeping track of each transcriber's work such that consistent errors may be corrected by batches without causing unforeseen errors or delays in the delivery.

Finally, some corpora, especially those created for automatic speech recognition, may require an additional step of evaluation in the form of a baseline automatic speech recognition system so as to validate the quality of the data. This first requires portioning the corpus into train, test, and if required, development sets as well. An acoustic model and language model would also need to be built and trained from the corpus on available toolkits such as Kaldi (Povey et al., 2011), before testing the corpus on the automatic speech recognition system to derive benchmarking or performance results which is often tabulated in terms of word error rate. Some further readings have been provided in Section 8 for readers who are interested to know more about acoustic and language modelling.

7. Conclusion

In this reference, several considerations underlying the design, execution, and post-processing of speech data were discussed. Though the reference is written with speech corpora for speech recognition applications in mind, some general pointers regarding the recording and post-processing of speech data are still applicable to the building of speech corpora tailored to other aims. Given that building a speech corpus is a non-trivial task, in terms of both financial and labour-wise, effort should be taken to engage local stakeholders as much as possible to ensure that the corpus represents the targeted linguistic variety or accent well.

8. Recommended readings

Hardcastle, W. J., Laver, J., & Gibbon, F. E. (2010). The handbook of phonetic sciences. Chichester, U.K: Wiley-Blackwell.

Gibbon, D., Moore, R., & Winski, R. (Eds.). (1998). *Spoken language system and corpus design*. Berlin, Boston: De Gruyter Mouton.

Jurafsky, D., & Martin, J. (2009). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Upper Saddle River, N.J: Pearson Prentice Hall.

O'Keeffe, A., & McCarthy, M. (2010). *The Routledge handbook of corpus linguistics (1st ed.)*. London; Routledge.

Schultz, T., & Kirchhoff, K. (2006). *Multilingual speech processing*. Elsevier Academic Press.

Singapore Open Data License (v. 1.0) https://data.gov.sg/open-data-licence

Wynne, M. (Ed.). (2005). *Developing linguistic corpora: a guide to good practice*. Oxford: Oxbow Books. Available online from http://ota.ox.ac.uk/documents/creating/dlc/ [Accessed 2020-05-23].

9. References

Baker, R., & Hazan, V. (2011). DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods*, *43*(3), 761–770.

Bizzocchi, A. L. (2017). How many phonemes does the english language have. *International Journal on Studies in English Language and Literature (IJSELL)*, *5*(10), 36–46.

Boersma, P., & Weenink, D. (2020). Praat: doing phonetics by computer [Computer program]. Version 6.1.15, retrieved 20 May 2020 from http://www.praat.org/

Butzberger, J., Murveit, H., Shriberg, E., & Price, P. (1992). Spontaneous speech effects in large vocabulary speech recognition applications. In *Proceedings of the workshop on Speech and Natural Language (HLT '91)*. (pp. 339–343.). Association for Computational Linguistics, USA.

Deterding, D. (2007). *Singapore English*. Edinburgh University Press.

ELAN (Version 5.9) [Computer software]. (2020). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

Furui S. & Kawahara T. (2008) Transcription and distillation of spontaneous speech. In Benesty J., Sondhi M.M., Huang Y.A. (Eds.), *Springer Handbook of Speech Processing. Springer Handbooks*. (pp. 627-652). Springer, Berlin, Heidelberg

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: telephone speech corpus for research and development. *Proceedings of ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, *1*, 517-520.

Hughes, T., Nakajima, K., Ha, L., Vasu, A., Moreno, P. J., & LeBeau, M. (2010). Building transcribed speech corpora quickly and cheaply for many languages. *INTERSPEECH 2010*, 1914-1917.

Koh, J. X., Mislan, A., Khoo, K., Ang, B., Ang, W., Ng, C., & Tan, Y.-Y. (2019). Building the Singapore English National Speech Corpus. *INTERSPEECH 2019*, 321–325.

Kominek, J., & Black, A. W. (2003). *CMU Arctic databases for speech synthesis*. (Report no. CMU-LTI-03-177). Retrieved from https://www.lti.cs.cmu.edu/sites/default/files/CMU-LTI-03-177-T.pdf

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., & Vesel, K. (2011). The Kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Robinson, T., Fransen, J., Pye, D., Foote, F., & Renals, S. (1995). WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition. *Proceedings of ICASSP-95: 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing*, *1*, 81-84.

SAMPA computer readable phonetic alphabet. (n.d.). Retrieved from https://www.phon.ucl.ac.uk/home/sampa/

Tan, Y-Y. (2019). Spontaneous speech elicitation for large speech corpus in multilingual Singapore. *Proceedings of the LPSS.* Academia Sinica, Taipei Taiwan.

The CMU Pronouncing Dictionary. (n.d.) Retrieved from http://www.speech.cs.cmu.edu/cgi-bin/cmudict