Rapid development of TTS corpora for four South African languages

Daniel van Niekerk[1], *Charl van Heerden*[1], *Marelie Davel*[1], *Neil Kleynhans*[1], *Oddur Kjartansson*[2], *Martin Jansche*[2], *Linne Ha*[2]

[1]Multilingual Speech Technologies, North-West University, South Africa
[2]Google Inc.

daniel.vanniekerk@nwu.ac.za

Abstract

This paper describes the development of text-to-speech corpora for four South African languages. The approach followed investigated the possibility of using low-cost methods including informal recording environments and untrained volunteer speakers. This objective and the additional future goal of expanding the corpus to increase coverage of South Africa's 11 official languages necessitated experimenting with multi-speaker and code-switched data. The process and relevant observations are detailed throughout. The latest version of the corpora are available for download under an open-source licence and will likely see further development and refinement in future.

Index Terms: text-to-speech corpus, under-resourced languages

1. Introduction

Recent work has been successful at developing text-to-speech (TTS) resources and systems for under-resourced languages by combining inexpensive methods such as crowd-sourcing and bootstrapping with powerful statistical acoustic modelling techniques [1]. The South African (SA) context with 11 official languages, most of them under-resourced but some closely related, provides an interesting scenario for further investigating rapid development of practical systems with similar efficient techniques. Another unique aspect of the SA context is that code-switched speech is important in any practical system due to proper names originating from different languages and indispensable loanwords. To provide useful TTS systems in this context, a broad goal is to find an efficient way to build a set of speech corpora that covers the necessary language spectrum.

Given the above-mentioned successes, conditions and overarching goal, the approach followed here is summarised as follows:

1. As a starting point, develop independent sets of speech recordings that may be used to build complete TTS systems in some of the 11 official languages where recordings are done by multiple speakers for each language subset.
2. Select the set of languages to maximise the coverage of the combined phone set required in SA, with the idea of being able to add languages at a later stage, possibly by sharing recordings from this initial set: *Afrikaans*, *isiXhosa*, *Sesotho* and *Setswana*. As expected, this set of four languages also represents the main language families.
3. Rely on relatively inexpensive methods for doing recordings, i.e. informal recording environments and untrained volunteer speakers.

This paper details the development process in the form of a case study. Section 2 discusses the script development, followed by the recording process (Section 3), pronunciation dictionaries and resources (Section 4) and quality control in Section 5. In Section 6, future work is proposed for both the currently developed resources and broader context.

2. Script development

A typical development strategy for an open-domain TTS script involves collecting a large representative reference text and selecting a subset (target text) to cover relevant phonetic and prosodic contexts, specifically: phonetic units (e.g. diphones), sentence types (e.g. questions and statements), and utterance lengths. When the number of target sentences is limited it is sensible to select short sentences primarily based on phonetic units [2]. This is under the assumption that prosodic patterns in clauses of longer sentences may be approximated from shorter sentences. Furthermore, short sentences should increase readability which may result in better speech quality (this is presumably important given untrained speakers).

Since one of the goals of the project was an open-source release of the corpora, it was necessary to collect text for the script from legally unencumbered sources. In recent years, a useful source of reference text in a growing number of languages is Wikipedia[1] [3]. Unfortunately, at this time, the only South African languages with a sufficient Wikipedia presence are Afrikaans and English.

2.1. Source text

In the context, four approaches were considered ranging from text selection to generation. (1) Simple text selection and preparation from Wikipedia or other freely available text corpora in South Africa [4]. This was only feasible for Afrikaans. Experience from past projects [5] showed that the largely government-domain texts that are available in all the South African languages [4] are difficult to prepare as source for TTS corpora (due to problems with readability). (2) Text generation by translation from Wikipedia. Using Wikipedia as a source for translating to the target language either from a closely related language or English has the potential advantage of generating less arcane sentences (compared to government-domain documents). Another advantage is that an expensive proofreading stage (necessary to ensure readability) may be combined with the translation stage. (3) Text generation by crowd-sourcing. This is a less direct approach which would require more time to set up and ensure that the process results in text that covers all contexts. (4) Semi-automatic text generation depending on application domain. This process is not relevant for an open-domain script, but can be used to generate sentences in narrow domains such as weather, navigation, and sport.

[1]https://dumps.wikimedia.org/

2.2. Implementation

With possible applications in mind, the following structure was adopted for scripts in each language (ordered by size):

1. *Main*: The main in-language set of sentences designed for broad diphone coverage (described in more detail below).
2. *English*: All (593) sentences in "set A" of the CMU Arctic databases[2] [2] to cover accented English that may be useful in code-switching, proper nouns and loan words.
3. *Navigation*: Sentences giving directions and instructions typically used in satellite navigation devices.
4. *Questions*: A set of yes-no questions.
5. *Sport*: Sentences reporting on soccer and cricket events.
6. *Weather*: Sentences announcing temperature and weather conditions in various places.
7. *Numbers*: Short fragments consisting of numbers and digit clusters. For languages other than Afrikaans these were in English. Previous work found a preference for using English in every day use for larger numbers and telephone numbers [6] (a fact that was confirmed by the volunteer speakers).

The relatively small sets of domain specific sentences (navigation, sport, weather and questions) were manually created or randomly generated with simple grammars in Extended Backus-Naur form (EBNF) and included lists of relevant people and place names. The set of sentence fragments containing numbers and digits were selected to give minimal but complete coverage of diphone units.

2.2.1. Afrikaans

The main in-language sentences were selected from Wikipedia and open textbooks;[3] introductory paragraphs[4] for all articles on Wikipedia were extracted and partitioned into subsets by length. From each of these subsets, sentences were selected to approximate the target diphone distribution (distribution of the full set of Wikipedia sentences) using the Kullback-Leibler (KL) divergence [7]. Phonetisation was performed using grapheme-to-phoneme conversion (G2P) as described in Section 4, excluding words contained in a common English word list and capitalised tokens (proper nouns). A larger number of sentences were selected from the shorter subsets; this resulted in sentences of varying length and with reasonable diphone context. Finally, diphone distributions were checked against the Lwazi 2 TTS corpora for missing units.

2.2.2. isiXhosa, Sesotho and Setswana

An initial attempt was based on sentences translated from Wikipedia, however, this was discarded when recording started (see next section). The final scripts were developed using a simple greedy text selection process [8] from children's stories published online.[5]

[2] `http://www.festvox.org/cmu_arctic/`

[3] Available online at `http://www.siyavula.com/` under a Creative Commons licence.

[4] Introductory sentences may be more readable or familiar than sentences from specialised sub-sections.

[5] `http://www.africanstorybook.org` and `http://nalibali.org/` used with permission under a Creative Commons licence.

2.3. Observations and comments

During text selection for Afrikaans, it was found that the KL selection algorithm was sensitive to the long tail in the reference distribution (presumably Zipfian); the algorithm selected a large amount of unlikely sentences and preferred shorter sentences when presented with sentences of varying length. This necessitated the removal of capitalised tokens from the reference distribution and the stratified selection described above. During the final comparison of diphone distributions against the Afrikaans Lwazi 2 TTS corpus, it was found that a few important units associated with common pronouns such as *jy* and *jou*, amongst others, were completely missing from the script. This is likely due to the formal register used in Wikipedia. Sentences containing these units were inserted manually.

For isiXhosa, Sesotho and Setswana, the initial approach of translating sentences selected from introductory paragraphs of English Wikipedia pages from the "South African portal" was attempted. To do sentence selection based on approximate diphone distribution before human translation, statistical machine translation[6] was used to produce translations which were phonetised. However, it was discovered that the subsequent manual translations were unfamiliar to the volunteer speakers (Section 3.2). This necessitated the simplified approach described in the previous section.

For languages that are less associated with economic activity, such as most of the South African languages except English and possibly Afrikaans, it is difficult to find significant amounts of relevant and accessible text online. While fully crowd-sourced approaches were not attempted in this project due to time constraints, it may be more useful to think of script development rather in terms of text generation than selection. As such, more time should be budgeted for defining and implementing creative solutions for this process in the particular context or application. The failure to successfully use scripts translated from Wikipedia by language experts indicates that it is also important to assess and increase the familiarity of the language used with reference to the target speaker (reader). This suggests that involving eventual speakers in the script development process may be beneficial.

3. Speech recordings

Recording multiple speakers reduces the number of sentences individual speakers are expected to read and allows the building of an average voice model without the specific qualities of any particular speaker (an anonymous voice). With this approach an ideal process would involve recruiting a larger number of speakers and selecting those with similar voice quality, accent and delivery.

3.1. Implementation

3.1.1. Speaker selection

A call for volunteers (age range 19-30 years) in the four languages was launched on social media, and offered a financial reward. Volunteers were asked to send audio clips of themselves reading sentences from the declaration of human rights in English and the primary language to be recorded. Through this process, Afrikaans and isiXhosa were recorded in Hermanus in the Western Cape province during November to January 2015-2016 and Sesotho and Setswana in the first two weeks of February 2016 at the campus of the North-West University (NWU),

[6] `https://translate.google.com/`

Potchefstroom, North West province.

Between 5 and 9 speakers were selected per language based on reading fluency. Each section of the stratified script was divided into non-overlapping parts, with the exception of sentences containing the rarest diphone units which were read by all speakers.

3.1.2. Recordings

The audio was recorded using a Neumann KM-184 microphone, an USB A/D converter on a fanless Acer Chromebook laptop running Chrome OS. A web-based recording tool, ChitChat, was used to manage the recordings. ChitChat presents each user with the sentences which have been assigned to them. ChitChat records the audio in 48kHz, detecting silences and excessive ambient noise for quality purposes. The audio is uploaded to a server for storage and later quality checks.

3.2. Observations and comments

One aspect that may be important (in small TTS corpora) but was difficult to control for during the rapid development process is speaker accent variation. In the case of Afrikaans, several phonetic variations associated with accents from different regions were recognised during recording which were not fully appreciated during the screening process. This was also noted to some extent for the other languages: for isiXhosa the difference between "urban" and more pure "rural" dialects were noticed, while for Sesotho and Setswana which are closely related languages, it was difficult to find speakers in Potchefstroom that were not influenced in their pronunciation by language contact. Similarly, English second-language accents also varied widely.

Since the phone set and G2P components for each language needs to be developed before script preparation and given that it is possible to perform automatic phonetic alignment with as few as 20 utterances [9, 10], it may be worthwhile attempting to develop a tool that can automatically flag potentially significant divergences in pronunciation for manual inspection once a prototypical speaker has been identified.

Assessing and obtaining fluent volunteer speakers (readers) was challenging. The generational cohort of speakers recruited during this project rarely continued with their first language as a formal subject through secondary school. Speakers that did have this background were clearly more fluent.

Due to challenges with these aspects of speaker selection, no attempts were made during this project to find speakers with similar voice profiles.

4. Pronunciation dictionaries

Pronunciation dictionaries developed for and released with the corpora are divided into "regular", "irregular", and "pronunciation addenda". The first two categories contain "standard" pronunciations for each entry: regular pronunciations are considered to follow the spelling conventions of the specific language and are therefore useful for G2P training, while irregular pronunciations (foreign words, loanwords, names) do not. As there is a high occurrence of English words, these are modeled separately (Section 4.2) and not included with non-English foreign words in the irregular in-language dictionaries. The pronunciation addenda contain pronunciations that are considered to be speaker-specific, and were manually produced based on acoustic (phonetic) realisations in specific utterances in the corpus.

4.1. In-language dictionaries

In-language pronunciation dictionaries were bootstrapped using the *NCHLT resources* [11, 12].[7]

4.1.1. Afrikaans

The initial Afrikaans dictionary was created using G2P rules extracted from NCHLT, with syllable stress manually added (only primary stress is indicated). Words that are potentially irregular were identified using automatic language identification (word lists and joint-sequence model-based [12]). Seed pronunciations for these words were created by applying language-specific G2P and mapping back to the Afrikaans phone set. Rule-based syllabification was implemented based solely on phonotactic constraints adapted from the approach in [13]. Possible errors were flagged throughout the process, using the verification tools described in Section 4.3, and errors found were manually corrected.

4.1.2. isiXhosa, Sesotho and Setswana

Dictionaries for isiXhosa, Sesotho and Setswana were created automatically by applying G2P rules derived from the NCHLT dictionaries and performing phonotactic rule-based syllabification [14]. As the orthographies of these languages are more phonemic than Afrikaans, less intervention was required. In the case of Sesotho, differences in orthographic conventions originating in South Africa and Lesotho caused some inaccuracies in the standard G2P rules. In most such cases the G2P rules could be manually updated to correctly map the orthographic variants to the correct pronunciation. However, some ambiguous cases remain and were investigated in the corpus during the quality control process (Section 4.3).

Although isiXhosa, Sesotho and Setswana are considered tone languages, no tonal information is included in the dictionaries. Underlying tone can be marked on the lexical level, however, in sentence context the realised surface tone is a function of underlying tone and other linguistic structures (morphological, syntactic, etc.) [15]. A tonal feature was therefore not included, and requires further study.

4.2. English dictionary

English code switching is prevalent in all the corpora. In addition to the in-language dictionaries, an English dictionary was therefore created per corpus, using South African English (SAE) pronunciation. An in-house Google SAE dictionary was used to generate a seed dictionary: these pronunciations were substantially changed during manual verification. A set of conventions was developed to improve consistency and specific attention paid to vowels (including the KIT split, using rules from [16]). Syllables were included with primary stress markers.

The protocol followed tends towards phonemic rather than phonetic pronunciations, e.g. where vowels are often reduced in utterances (e.g. in function words), pronunciations expected in "careful" speech were retained. Because speakers' English accents vary significantly this dictionary will not generally be an accurate phonetic description of pronunciations in the recordings, but may be used to make language or speaker dependent English models.

[7] Available online: `https://sites.google.com/site/nchltspeechcorpus`

Table 1: *Afrikaans corpus*

Afrikaans				
	Utterances		Duration (mins.)	
Speaker	In-lang.	English	In-lang.	English
01	297	65	16.00	2.81
02	248	60	18.11	3.51
03	254	66	15.66	3.05
04	246	64	14.91	3.16
05	250	67	15.38	3.16
06	258	65	15.33	3.18
07	247	68	16.33	3.51
08	248	64	15.91	3.16
09	250	64	14.91	3.06
Total	**2298**	**583**	**142.66**	**28.71**

Table 2: *IsiXhosa corpus*

isiXhosa				
	Utterances		Duration (mins.)	
Speaker	In-lang.	English	In-lang.	English
01a	147	–	14.25	–
02a	143	–	12.45	–
01	–	161	–	8.28
02	281	126	18.73	5.66
03	268	147	18.95	6.66
04	252	–	16.78	–
05	308	143	19.48	6.76
06	289	144	19.65	6.25
Total	**1688**	**721**	**120.30**	**33.63**

Table 3: *Sesotho corpus*

Sesotho				
	Utterances		Duration (mins.)	
Speaker	In-lang.	English	In-lang.	English
01	279	145	15.39	6.39
02	274	—	14.76	—
03	282	144	16.01	6.82
04	283	154	18.31	7.13
05	274	139	16.27	6.96
06	—	143	—	6.66
Total	**1392**	**725**	**80.77**	**33.98**

Table 4: *Setswana corpus*

Setswana				
	Utterances		Duration (mins.)	
Speaker	In-lang.	English	In-lang.	English
01	302	139	16.74	6.41
02	304	143	16.64	6.21
03	303	144	16.33	6.00
04	312	153	18.23	6.50
05	297	145	15.77	6.14
Total	**1518**	**724**	**83.73**	**31.28**

4.3. Pronunciation verification

Semi-automatic verification of transcriptions and dictionaries entailed manually reviewing words flagged by automatic techniques as potential inaccuracies [17, 18]:

1. Pronunciations of capitalised words flagged using the mel-cepstral distance [18] were reviewed.
2. Pronunciations of standard words flagged using the PDP score [17] were reviewed for Afrikaans, the only of the target languages where standard words do not have a regular spelling system.
3. Some tokens with possibly consistent orthographic or dialectic variation were examined.

In addition, tokens added to the pronunciation addenda (ref. Section 4) were manually reviewed. In order to produce the dictionaries quickly and effectively, only certain categories of words were reviewed; not all entries in the dictionaries.

5. Quality control

After recording, the following tasks were performed towards readying corpora for use in TTS voice building:

- Removal of recordings containing buffer-underrun audio artefacts (detected after recording).
- Manual text normalisation based on the audio and mark-up of embedded foreign words.
- Transcriptions were reviewed by comparing expected durations of words with automatic phone alignments and gross transcription and pronunciation deviations fixed.
- Pronunciation verification (Section 4.3) occured during different stages of the process.

Post-processing tasks that were not performed but may be useful before voice building, include: gain normalisation over utterances, de-reverberation filtering, trimming of start and end silences and manual marking of intonation phrase breaks (breath-groups).

6. Conclusion and future work

The main technical and operational difficulties experienced during the development process related to script development and volunteer speaker recruitment. While these two components have typically been treated separately in a straightforward manner in similar projects [2], the shortage of freely available online text and indeed fluent volunteer speakers (readers) necessitated an integrated development approach and additional processes such as speaker evaluation.

The current version of the corpora (Tables 1 to 4) will be made available online under an open-source licence. Speakers marked with "a" indicate recordings done with original translated text script where fluency was a concern (Section 2). Future work on the current set of data will involve TTS voice building experiments which may result in further refinement of the existing pronunciation and textual resources (especially since code-switching pronunciations have not been fully investigated at this point). It should be possible to use these corpora as a starting point for experimenting with a shared (multilingual) phone set for the SA context with the intention of leveraging this to efficiently expand coverage to the other SA languages.

7. References

[1] A. Gutkin, L. Ha, M. Jansche, O. Kjartansson, K. Pipatsrisawat, and R. Sproat, "Building Statistical Parametric Multi-speaker Synthesis for Bangladeshi Bangla," *Procedia Computer Science*, vol. 81 (SLTU-2016), pp. 194–200, 2016.

[2] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proceedings of the 5th International Speech Communication Association Speech Synthesis Workshop (SSW)*, Pittsburgh, PA, USA, 2004, pp. 223–224.

[3] S. Pammi, M. Charfuelan, and M. Schröder, "Multilingual Voice Creation Toolkit for the MARY TTS Platform," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta, May 2010, pp. 3750–3756.

[4] R. Eiselen and M. J. Puttkammer, "Developing Text Resources for Ten South African Languages," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, May 2014, pp. 3698–3703.

[5] K. Calteaux, F. de Wet, C. Moors, D. R. van Niekerk, B. McAlister, A. Sharma Grover, T. Reid, M. Davel, E. Barnard, and C. van Heerden, "Lwazi II Final Report: Increasing the impact of speech technologies in South Africa," Council for Scientific and Industrial Research, Pretoria, South Africa, Tech. Rep. 12045, Feb. 2013.

[6] A. Sharma Grover, O. Stewart, and D. Lubensky, "Designing interactive voice response (IVR) interfaces: localisation for low literacy users," in *Proceedings of Computers and Advanced Technology in Education (CATE)*, St. Thomas, US Virgin Islands, Nov. 2009, pp. 328–335.

[7] E. Gouvêa and M. H. Davel, "Kullback-Leibler Divergence-Based ASR training data selection," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Florence, Italy, Aug. 2011, pp. 2297–2300.

[8] J. P. Van Santen and A. L. Buchsbaum, "Methods for optimal text selection," in *Proceedings of EUROSPEECH*, Rhodes, Greece, September 1997, pp. 553–556.

[9] S. Brognaux and T. Drugman, "HMM-based Speech Segmentation: Improvements of Fully Automatic Approaches," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 1, pp. 5–15, 2016.

[10] D. R. van Niekerk and E. Barnard, "Phonetic alignment for speech synthesis in under-resourced languages," in *Proceedings of the Tenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Brighton, UK, September 2009, pp. 880–883.

[11] E. Barnard, M. H. Davel, C. J. van Heerden, F. de Wet, and J. A. C. Badenhorst, "The NCHLT speech corpus of the South African languages," in *Proceedings of the Workshop on Spoken Language Technologies for Under-resourced languages (SLTU)*, St. Petersburg, Russia, May 2014, pp. 194–200.

[12] M. H. Davel, W. D. Basson, C. van Heerden, and E. Barnard, "NCHLT Dictionaries: Project Report," North-West University, Tech. Rep., May 2013. [Online]. Available: https://sites.google.com/site/nchltspeechcorpus/home

[13] T. A. Hall, "English syllabification as the interaction of markedness constraints," *Studia Linguistica*, vol. 60, no. 1, pp. 1–33, 2006.

[14] C. Halpert, "Overlap-driven consequences of nasal place assimilation," *Consonant Clusters and Structural Complexity*, pp. 345–368, Oct. 2012.

[15] M. Raborife, "Tone labelling algorithm for Sesotho," M.Sc. Thesis, University of the Witwatersrand, 2011. [Online]. Available: http://wiredspace.wits.ac.za/handle/10539/11248

[16] L. Loots, M. Davel, E. Barnard, and T. Niesler, "Comparing manually-developed and data-driven rules for P2P learning," in *Proceedings of the Twentieth Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, Nov. 2009, pp. 35–40.

[17] M. H. Davel, C. J. van Heerden, and E. Barnard, "Validating smartphone-collected speech corpora," in *Proceedings of the Workshop on Spoken Language Technologies for Under-resourced languages (SLTU)*, Cape Town, South Africa, May 2012, pp. 68–75.

[18] D. R. van Niekerk, "Experiments in rapid development of accurate phonetic alignments for TTS in Afrikaans," in *Proceedings of the Twenty-Second Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, Vanderbijlpark, South Africa, Nov. 2011, pp. 144–149.