Building synthetic voices for under-resourced languages: a comparison between audiobook and studio data

Febe de Wet, Nkosikhona Dlamini, Willem Van der Walt, Avashna Govender

Human Language Technology Research Group, Meraka Institute
Council for Scientific and Industrial Research, Pretoria, South Africa
{fdwet, ndlamini3, wvdwalt}@csir.co.za, avashna002@gmail.com

Abstract

Creating synthetic voices that are both natural and intelligible is a daunting challenge for well-resourced languages. The challenge is much bigger for languages in which the speech and text resources required for voice development are not available. In previous studies, audiobooks have been considered as an alternative source of speech data. The aim of the current study was to compare the quality of voices derived from audiobook data with voices based on data recorded by professional voice artist under studio conditions. Two sets of voices were evaluated: male voices built using a very small data set (around 3 hours, representing a severely resource constrained scenario) and female voices trained on almost 10 hours of speech data. The results of subjective listening tests indicate that, while the majority of the listeners preferred the voice artists' voices over the audiobook voices, the difference in naturalness was not perceived to be substantial. Results also showed that the artists' voices outperform the audiobook voices in terms of intelligibility, especially if a limited amount of training data is available. However, if more training data is used, the difference in intelligibility can be reduced substantially.

Index Terms: speech synthesis, text-to-speech, audiobooks, under-resourced languages

1. Introduction

Collecting and annotating the speech data required to develop a state-of-the-art text-to-speech (TTS) system requires a considerable amount of time and expert knowledge. This requirement constitutes a major challenge for under-resourced languages. Only a few languages have resources available like the audio and text that are available through, for example, the Gutenberg project[1]. However, most countries do provide audiobooks to print-disabled readers through special libraries [1, 2]. The books are mostly read by volunteers, but the speech could still be used as an alternative source of data to develop synthetic voices for TTS systems.

However the use of audiobook data does pose a number of limitations to voice development, especially if the voices are to be distributed. For instance, most publications are protected by copyright, which means that neither the text nor the audio my be used without the permission of the copyright owner(s). For the purposes of voice development this restriction could be overcome by using books whose copyrights have expired, but using "old" books and voices introduce other complicating factors to the voice building process, some of which will be addressed in this paper (see Section 3.1).

A synthetic voice cannot be built and distributed using a single speaker's voice without his/her consent. A possible solution to this restriction could be to combine the voices of different speakers to create an average voice model that does not contain biometric information of a specific person, e.g. [3, 4]. However, if basic voice development has not been implemented for a particular language, it is highly unlikely that voice adaptation would be available.

Previous studies have also shown that the speech data in audiobooks require special pre-processing to enable voice development, e.g. [5, 6, 7]. The appropriate resources and technologies may not always be available in under-resourced languages, even if appropriate audiobook material is. For example, to perform forced alignment between the audio and text versions of a book, the text data needs to be available in electronic format. For most under-resourced languages, electronic books are not readily available. If older books are used to avoid copyright issues, the associated text is rarely available in electronic format and printed versions are often difficult to get hold of [2].

The printed version of a book requires optical character recognition (OCR) is required to convert the text to electronic format. Although language-independent OCR systems are available, recognition performance can be enhanced significantly by adding language specific information to the process [8]. Again, this type of information may not be readily available in under-resourced languages, which could have a negative impact on the quality of the OCR process.

Although researchers in the field seem to agree on a number of standard practices, there is also some scholarly discussion around how synthetic voices should be evaluated [9, 10]. It is therefore not immediately apparent exactly how voices in under-resourced, unknown languages should be evaluated. While the concepts of preference testing and Mean-Opinion-Scores (MOS) are not difficult to transfer between languages, the design of Semantically Unpredictable Sentences (SUS) has only been specified for a number of languages [11, 12].

Once again, the resources required to construct an SUS test could pose a challenge. For example, the languages considered in this study, Afrikaans, is closely related to Dutch and the design proposed for Dutch SUS could therefore be followed to some extent [13]. However, the only text corpus of substantial size that is available for research was sourced from government documents and contains mostly government specific lexical items and hardly any monosyllabic words [14].

In this study synthetic voices built using audiobook data are compared with voices derived from professionally recorded voice artists' speech. The aim of the comparison was to investigate the feasibility of audiobooks as an alternative data source, given the challenges and limitations associated with using audiobook material. Despite numerous efforts, permission could

[1] `https://www.gutenberg.org/wiki/Main_Page`

[2] Moreover, if different editions of the same book were published, it is often difficult to find the edition that matches the audio.

not be obtained from publishers to use recently published books for this study. As a consequence, only audiobooks of which the copyrights have expired were used. Two sets of voices were evaluated: male voices built using a very small data set (around 3 hours, representing a severely resource constrained scenario) and female voices trained on almost 10 hours of speech data[3].

The comparison addressed two research question: firstly, how voices derived from the speech in audiobooks compare with voices derived from professional recordings in terms of naturalness and intelligibility and secondly how the amount of training data available influences voice quality.

2. Background

2.1. Grapheme based audio and text alignment

Audiobooks have been investigated as an alternative resource for building TTS systems. Audibooks are usually stored as .wav or .mp3 files. Older audiobooks are often still stored on cassette. These audio formats have accompanying text version of the books. In order to make the available audio and text suitable for training HMM-based speech synthesis models, pre-processing is necessary. The pre-processing step involves alignment of text and audio. Aligning audiobook speech and text data is a challenge because forced alignment procedures, like the Viterbi algorithm, requires short audio clips, yet audiobook files are speech data is usually longer, covering a whole chapter [6, 15].

According to [15] aligning audio and text is a speech recognition problem. A biased language model can be used to this end, the output of the recognition can be aligned with the transcript using dynamic programming. However, speech recognition relies on well tested, trained acoustic models, these models may not be available in poorly resourced languages. Poorly trained grapheme models were used in [15], in conjunction with a skip network. The purpose of the skip network was to allow Virtebi decoding to match audio segment to any point in the transcription, but only to sequence of words seen on a network. Sections of the text not seen in the audio are automatically skipped. Text and audio used in training the initial poor models was obtained manually. Aligned transcriptions obtained by these models are used to re-estimate new acoustic models aimed to harvest more aligned transcriptions from the training set. This method was reported to harvest about 55% accurately aligned transcriptions, with SER of 7.64% and WER of 0.5%.

An improved version of the above approach was proposed [6], the use of discriminative training of grapheme models was investigated. Five-state left-to-right, mono grapheme with eight mixture components per state, used in [15] were used, and the first step in building final acoustic models was to extend mono-grapheme to tri-graphemes. Tri-graphemes list was obtained from available text. The aim was to minimize SER, therefore the objective function criterion selected was maximizing the margin at which utterances are correctly classified. A 70% correctly aligned transcripts was reported.

These authors further proposed some improvements on the discriminative training of Grapheme models for alignment [7]. The improvements added a pre-processing step, where a GMM-based VAD detects sentence boundaries before alignment is done. 50 sentences were manually selected and aligned from audiobook data and used for training GMM-VAD. This AVD was evaluated, using metrics [16]. Mid-speech clipping (MSC) which is a metric that indicates the percentage of real speech classified (misclassified) as noise was 1.12% for the GMM-AVD. The other metric used, called (OVER), which is a measure of classifying noise as speech was 2.05%. These values show how well the GMM-AVD performed in detecting sentence boundaries and allowing an automatic alignment of using grapheme based automatic alignment.

The speech data that was used during this study is described in the next section. The voices and the tests that were conducted to evaluate them are introduced in Section 4 and the results are presented in Section 5. Concluding remarks are presented in Section 7.

3. Data

To build the voices that were evaluated in this study, data was sourced from the speech in *Audiobooks* (of which the copyrights have expired) as well as studio quality speech data recorded by professional voice *Artists*. The duration of the four data sets is summarised in Table 1.

Table 1: *Total duration of the four different data sets used for voice development*

	Male	**Female**
Audiobook	206	593
Artist	180	591

3.1. Audiobook data

Audiobooks were obtained from a local library for the blind. One of the books is *Loeloeraai* by C.J. Langenhoven, published in 1923. The book was read by a male speaker in October 1958. This recording was chosen because it matches a sub-set of the data produced by the male voice artist. The other books were all Afrikaans narratives read by the same female speaker between 1962 and 1980.

The audio versions of all the books were in analogue format. The analogue recordings were digitised and noise reduction was performed on the digitized versions using the `denoise` function of `SpeexDSP`[4].

The text version of some of the books is publicly available in electronic format [5]. The other books were scanned and OCR was performed using `ABBY FineReader`[6] with the Afrikaans language model activated. The resulting text contained almost no errors and very little post-editing was required.

3.2. Voice artists data

The *Loeloeraai* text was included in the Afrikaans text that was read by the voice artists. The recording made by the male voice artist was used to build the *artist* version of the corresponding audiobook voice[7].

The professional female voice was built with all the data that was available at the time of writing, comprising almost 10

[3] Ideally, all the voices in the study should have been either male or female. However, the experimental design was determined by the data that the authors had access to.

[4] `https://github.com/xiph/speexdsp`

[5] `https://wikisource.org/wiki/Main_Page`

[6] `https://www.abbyy.com/en-apac/finereader/`

[7] Unlike the volunteer at the library for the blind, the voice artist did not read the entire *Loeloeraai* book - hence the slight difference in duration.

hours of speech data. The majority of the text data that was used during the recordings was selected to ensure phonetically balanced data with adequate diphone coverage in Afrikaans. In addition, examples of fiction and non-fiction were included from text on which the copyrights have expired.

4. Voice evaluation

Four voices were built using the data described in Section 3 and the `Speect` TTS system [17]. The male and female voices were subsequently compared using objective measures as well as subjective listening tests.

4.1. Objective evaluation

The voices were compared in terms of two objective measures, the mel-cepstral distance (MCD) and the root-mean-square-error (RMSE) of logF_0. MCD is used to measure the accuracy of the spectral envelope and calculated as the average Euclidean distance between the mel-cepstral coefficients of two samples [18]. The RMSE logF_0 quantifies the accuracy of the F_0 contour generated by the model. Since the F_0 is only observed in voiced regions, the RMSE of logF_0 was only calculated for the voiced parts of the speech signals.

A subset of 99 sentences from the *Loeloeraai* text was used to calculate the objective measures for the two male voices. An overlapping set of 44 sentences was identified for the two female voices. The MCD distance and RMSE of logF_0 between the natural speech and the synthesised speech for each of these sentences were calculated. Dynamic Time Warping (DTW) was applied to ensure that the temporal differences between the two samples did not influence the distance measures.

4.2. Subjective evaluation

Subjective evaluations were conducted using formal perceptual listening tests administered via a web interface. The two female voices were evaluated first, followed by an evaluation of the male voices about two weeks later. The listening tests were performed by 20 adult listeners native to South Africa who all speak or understand Afrikaans. The majority of the listeners are not familiar with speech technology or synthetic voices.

A paired comparison test was used to evaluate user preference and a Mean-Opinion-Score (MOS) test to evaluate naturalness. Intelligibility was evaluated by means of a transcription test. Fifty test sentences were synthesized, of which 40 were not included in the training data. The remaining 10 sentences were taken from the training data as examples of natural speech (required to compile the MOS test). These 10 sentences were the same for the male and female voices.

4.2.1. Paired comparison test

Participants listened to 20 voice samples. Each question comprised of two samples, one *Audiobook* sample and one *Artist* sample. All samples were ordered randomly. Listeners were required to select either ‘Sample A’ or ‘Sample B’.

4.2.2. MOS test

In the MOS test, participants listened to 20 voice samples with 10 samples synthesised using each system. Participants were asked to listen to both the natural and synthesised samples and to rate each sample based on its naturalness. The rating was performed in terms of a 5-point scale, defined as follows: 1 - Completely Unnatural, 2 - Unnatural, 3 - Slightly natural, 4 - Natural and 5 - Completely Natural.

4.2.3. Transcription test

Participants were required to listen to each speech sample and transcribe the audio. Twenty semantically unpredictable sentences (SUS) were transcribed per listener, of which 10 random samples corresponded to each of the two voices (Audiobook or Artist). Although the two evaluations were conducted with a two week interval in between, many participants' browsers “remembered” the first set of transcriptions. Two different sets of SUS were therefore used to evaluate the female and the male voices.

5. Results

5.1. Objective evaluation

The objective distance measures calculated for the *Audiobook* and *Artist* voices are shown in Table 2.

Table 2: *Average MCD [db] and RMSE logF_0 [cent] values for the Audiobook and Artist voices*

	Male		**Female**	
	MCD	RMSE	MCD	RMSE
Audiobook	1.7	206	2.1	223
Artist	1.9	185	2.1	151

The results for the MCD indicate that the male voice trained on the audiobook data is slightly closer to the natural voice than the male voice based on the TTS specific data, but the difference is insignificant. The MCDs for the two female voices are exactly the same. These results indicate that objectively the spectral features extracted from the audiobooks data are comparable with the accuracy of the features obtained for the commercial voices.

In terms of the RMSE of logF_0, both the *Artist* voices are closer to the natural voice than the corresponding *Audiobook* voices. This result is expected, given the quality of the audiobook recordings as well as the difference in prosodic delivery.

5.2. Subjective evaluation

5.2.1. Paired comparison test

More than 75% of the participants showed a clear preference for the synthetic voices derived from the professional voice artists' data. Given the comparatively poor quality of the audiobook data and the fact that it was *found* data rather than carefully designed data, this result is not completely unexpected. Many participants remarked that the audiobook voices sounded old-fashioned, but could not explain exactly which properties of the voices made them sound that way.

5.2.2. MOS test

The average MOS scores for the two sets of voices are shown in Table 3. The fact that the average values for the natural voices are not 5 indicates that not all listeners assigned the maximum value to the examples of natural speech that were included in the test. Moreover, the average values for the natural examples of the *Audiobook* data are lower than the corresponding values for the *Artist* data. This trend may be linked to the participants' comment that the *Audiobook* voices sound old-fashioned.

Table 3: *Average MOS scores for the synthetic (S) and natural (N) voices*

	Male		**Female**	
	S	N	S	N
Audiobook	2.9	4.2	2.9	4.1
Artist	3.3	4.9	3.3	4.6

Given the fact that the two sets of voices were trained on substantially different amounts of training data, it is rather unexpected that the MOS scores for the two *Audiobook* voices and the two *Artist* voices are the same. This observation seems to suggest that the additional data that was available to train the acoustic models for the female *Audiobook* voice could not compensate for the inferior acoustic quality of the data.

5.2.3. Transcription test

The average word error rate (WER) for the transcription test was calculated using the formula provided in the Blizzard 2007 challenge guidelines [9]. Spelling mistakes and typographical errors were corrected before the WER was calculated. The results in Table 4 show that the intelligibility of the *Audiobook* voices are substantially worse than the *Artist* voices.

Table 4: *WER (%) values for the transcription of the SUS sentences*

	Male	**Female**
Audiobook	83	44
Artist	17	21

According to the results in Table 4 the WER for the male *Artist* voice is lower than the WER measured for the female voice, despite the fact that the voice was trained on much less data. This observation could be explained - at least to some extent - that the male voices were evaluated after the female voices and that the listeners had gained some experience in transcribing SUS, despite the fact that two different sets of sentences were used for the evaluations.

While the difference in intelligibility between the male and female *Artist* voices are fairly small, the additional data seems to make a substantial difference to the intelligibility of the *Audiobook* voices: the male voice, trained on the smaller data set, has almost double the WER of the female voice. With a WER value of 83% the voice could very well be classified as unintelligible.

6. Discussion

In the absence of vast text and speech resources, the aim of this study was to investigate the feasibility of using audiobooks as an alternative source of data to develop synthetic voices in under-resourced languages. To achieve this aim, two voices built from audiobook material were compared with similar voices derived from professionally recorded voice artists' speech.

Due to copyright restrictions the audiobooks that were used to build the synthetic voices were old and in analogue format. The analogue to digital conversion and denoising operation did not result in acoustic data of the same quality as the studio data. The results of the subjective listening test seem to indicate that less data of superior quality yields a better quality voice than much more data of inferior quality. The authors are still trying to obtain permission to use more recently recorded audiobooks and hope to conduct a comparative study with a new audiobook in the near future.

Another property of especially the female audiobook data that is not very desirable is the fact that the recordings were made over a period of almost twenty years. The change in the reader's voice is audible in some of the recordings. In contrast, the female voice artist's speech was recorded within a single week.

7. Conclusion

The results of this study indicate that, even if a relatively small amount of data is used, TTS voices derived from professional, studio quality data are more natural and intelligible than similar voices built using audiobook data. While using more audiobook data improves the quality of an audiobook voice, the "bigger" voice is still outperformed (in terms of intelligibility) by an artist voice trained on much less data. If audiobooks are to be used as a source of data for voice development, the books should therefore be selected based on the quality of the data rather than the quantity that is available.

8. Acknowledgements

This work was supported by the Department of Arts and Culture of the South African government. The authors would like to express their sincere gratitude to the staff at the Pioneer Printers for their friendly and enthusiastic support.

9. References

[1] H. Kawamura, *International Directory of Libriaries for the blind*. IFLA Plublications, 1990.

[2] M. Nomura and M. Yamada, Eds., *International Directory of Libriaries for the blind.* DE GRUYTER SAUR, 2000.

[3] J. Yamagishi, "Average-voice-based speech synthesis," Ph.D. dissertation, Tokyo Institute of Technology, 2006.

[4] J. Yamagishi, O. Watts, S. King, and B. Usabaev, "Roles of the average voice in speaker-adaptive HMM-based speech synthesis," in *INTERSPEECH*, September 2010, pp. 418–421.

[5] N. Braunschweiler, M. J. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings." in *Proccedings of Interspeech*, 2010, pp. 2222–2225.

[6] A. Stan, P. Bell, J. Yamagishi, and S. King, "Lightly supervised discriminative training of grapheme models for improved sentence-level alignment of speech and text data." in *Proceedings of Interspeech*, 2013, pp. 1525–1529.

[7] Y. Mamiya, J. Yamagishi, O. Watts, R. A. Clark, S. King, and A. Stan, "Lightly supervised GMM VAD to use audiobook for speech synthesiser," in *Proceedings of IEEE ICASSP 2013*. IEEE, 2013, pp. 7987–7991.

[8] J. Hocking and M. J. Puttkammer, "Optical Character Recognition for South African languages," in *Proceedings of the 27^{th} Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, Stellenbosch, South Africa, December 2016.

[9] R. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard challenge 2007 listening test results," *Proceedings of the Blizzard Workshop, BLZ3-2007 (in Proc. SSW6)*, 2007.

[10] M. Wester, C. Valentini-Botinhao, and G. E. Henter, “Are we using enough listeners? no!-an empirically-supported critique of interspeech 2014 tts evaluations.” in *INTERSPEECH*, 2015, pp. 3476–3480.

[11] C. Benoît, M. Grice, and V. Hazan, “The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences,” *Speech Communication*, vol. 18, no. 4, pp. 381–392, 1996.

[12] M. Sulír, J. Staš, and J. Juhár, “Design of phonetically balanced sus test for evaluation of slovak tts systems,” in *ELMAR (ELMAR), 2014 56th International Symposium*. IEEE, 2014, pp. 1–4.

[13] W. Heeringa, F. De Wet, and G. B. Van Huyssteen, “Afrikaans and Dutch as closely-related languages: A comparison to West Germanic languages and Dutch dialects,” *Stellenbosch Papers in Linguistics Plus*, vol. 47, pp. 1–18, 2015.

[14] R. Eiselen and M. J. Puttkammer, “Developing text resources for ten south african languages.” in *LREC*, 2014, pp. 3698–3703.

[15] A. Stan, P. Bell, and S. King, “A grapheme-based method for automatic alignment of speech and text data,” in *Proceedings of Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 286–290.

[16] J. Richiardi and A. Drygajlo, “Evaluation of speech quality measures for the purpose of speaker verification,” in *Proceedings of Odyssey*. Citeseer, 2008, p. 5.

[17] J. A. Louw, G. I. Schlünz, W. Van der Walt, F. De Wet, and L. Pretorius, “The Speect text-to-speech system entry for the Blizzard Challenge 2013,” in *Blizzard Challenge Workshop 2013*, Barcelona, Spain, September 2013.

[18] R. F. Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” in *Communications, Computers and Signal Processing, 1993., IEEE Pacific Rim Conference on*, vol. 1. IEEE, 1993, pp. 125–128.