10th ISCA Speech Synthesis Workshop
20-22 September 2019, Vienna, Austria

Subset Selection, Adaptation and Gemination for Amharic Text-to-Speech Synthesis

Elshadai Tesfaye Biru[1], *Yishak Tofik Mohammed*[1], *David Tofu*[1], *Erica Cooper*[2], *Julia Hirschberg*[1]

[1]Columbia University, [2]National Institute of Informatics, Tokyo

tesfaye.biru@columbia.edu, m.yishak@columbia.edu, david.tofu@columbia.edu, ecooper@nii.ac.jp, julia@cs.columbia.edu

Abstract

While large TTS corpora exist for commercial systems created for high-resource languages such as Mandarin, English, and Spanish, for many languages such as Amharic, which are spoken by millions of people, this is not the case. We are working with "found" data collected for other purposes (e.g. training ASR systems) or available on the web (e.g. news broadcasts, audiobooks) to produce TTS systems for low-resource languages which do not currently have expensive, commercial systems. This study describes TTS systems built for Amharic from "found" data and includes systems built from different acoustic-prosodic subsets of the data, systems built from combined high and lower quality data using adaptation, and systems which use prediction of Amharic *gemination* to improve naturalness as perceived by evaluators.

Index Terms: text-to-speech synthesis, Amharic, gemination

1. Introduction

In recent years, text-to-speech synthesis (TTS) has become widespread in the form of mainstream consumer products such as mobile virtual personal assistants (Siri, Google Assistant), in-home devices (Amazon Echo), and other applications such as speech-to-speech translation. However, collecting the type of data required to build a high-quality TTS voice is typically very costly (at least $1M for each new voice), and is thus only undertaken with a major commercial incentive. Typically, a professional voice talent reads dozens of hours of text with good coverage of the target domain in a soundproof room with a high-quality microphone and in as neutral and even a style as possible. They are typically instructed to maintain constant f0, energy, speaking rate, and articulation throughout.

However, even without the resources to collect such data, it is still possible to create a high-quality voice. With the advent of statistical parametric speech synthesis (SPSS) such as Hidden Markov Model (HMM) based synthesis and neural network based synthesis, it is possible to create voices without necessarily having to collect large amounts of high-quality, single-speaker, in-domain speech. Large amounts of available speech such as audiobooks and radio broadcast news present a promising source of data for building new voices. In this paper, we describe the creation of TTS systems for Amharic using such "found" data which achieves reasonable ratings of intelligibility and naturalness from native listeners.

In Section 2 we describe prior efforts at building TTS voices for Amharic and in training voices from found data. In Section 3 we describe the "found" corpora we used in building our Amharic TTS. In Section 4 we describe the process we used to build voices. In Section 5 we describe experiments using acoustic-prosodic subset selection from "found" Amharic data. In Section 6 we describe adaptation experiments which improved TTS naturalness by combining data from multiple Amharic corpora. In Section 7 we describe our approach to modeling gemination and the improvement our system obtains using this model. Finally, in Section 8 we discuss our results and our plans for future research.

2. Related Work

There have been some previous efforts to build TTS voices for the Amharic language. For instance, [1] built a unit selection voice for Amharic using their own recorded data, and identified the language-specific challenges of Amharic regarding epenthesis and syllabification. Building on that work, [2] further identified gemination as a major challenge in Amharic language processing, and designed an improved syllabification algorithm that incorporates both gemination and epenthesis. [3] built a pronunciation modeling pipeline for Amharic which conducts morphological analysis of Amharic text to disambiguate gemination and vowel epenthesis. Audiobooks have been a popular source of found data for building speech synthesis systems because of their relatively clean recording conditions and the fact that they typically contain large amounts of speech from a single speaker. Furthermore, they are often freely and widely available in a large variety of languages. The challenge of this type of data is its typically more expressive style than more conventional TTS corpora, as we have empirically measured in [4]. There have been a number of efforts to reduce the variability in audiobook data to make it more appropriate for use in TTS in different ways, such as by identifying clusters of the most neutral utterances [5] or by choosing utterances based on ASR confidence scores or human judgment [6]. A similar approach was also explored in a multilingual setting by [7] who built a corpus of 60 hours of speech from audiobooks in 14 languages, filtered by including only utterances with high automatic alignment confidence scores. AudioBibles are a good potential source for found speech due to the fact that they exist in a very large variety of languages. In fact, the recently-released CMU Wilderness Multilingual Speech Dataset [8] contains aligned speech and text from the

10.21437/SSW.2019-37

New Testament in over 700 languages, collected online.

We have previously explored the use of acoustic and prosodic criteria for selecting suitable utterances for use in TTS voice training in English. In particular, we were able to create more natural-sounding and intelligible TTS voices from English radio broadcast news speech and speech data collected for building automatic speech recognition systems by selecting training utterances based on knowledge of the speech characteristics that are best suited to TTS data [9, 10, 11]. In particular, we found that a voice trained on a subset of the training data comprised of the utterances with the lowest mean f0 was preferred over a voice trained on all of the data, and that selecting utterances with lower levels of articulation (faster speaking rate and less variation in f0) improved ratings of both naturalness and intelligibility. In the first set of experiments in this area, we aim to examine which acoustic-prosodic subsets are useful for creating Amharic voices from found data.

We have also previously explored the use of model adaptation for improving voices trained on found data. In [10] and [12], we explored adapting found-data voice models towards the portions of the data that were acoustically and prosodically most similar to TTS data. In this paper, we extend this approach to train voice models on mixed high- and low-quality data, and adapt towards the high-quality portion, with the hypothesis that this approach will produce better-sounding voices than training on just a small amount of high-quality data alone. This approach has a precedent in [13], who trained an average voice model on data collected in an office environment and then adapted it to cleanly-recorded speech. They found that using both noisy and clean data together produced a voice with a slightly (but not statistically-significantly) higher mean opinion score than a voice trained on the clean data alone, and concluded that more data, even of a lesser quality, can be beneficial.

3. Corpora

Our TTS work on Amharic has been based on three main corpora. The first is the Amharic data collected for the IARPA BABEL project [14] which collected 25 corpora in low-resource languages to support the training and evaluation of spoken keyword search systems. Each corpus consists of a set of recorded and transcribed phone conversations as well as recorded scripted speech, with both male and female speakers. While the goal of BABEL was primarily speech recognition and spoken keyword search, we used some of this multi-speaker, read and conversational telephone data to build TTS voices for some of these languages including Amharic, Turkish and Telugu. The Amharic data (IARPA-babel307b-v1.0b full language pack) consists of about 40 hours of conversational speech from 300 different speakers, and about 10 hours of speech from 230 different speakers for the scripted portion. Due to the multiple speakers and speaking conditions these telephone recordings can be challenging to use for TTS.

The second corpus was prepared from a publicly available AudioBible, consisting of a single male speaker and about 55 hours of recorded speech. The audio was obtained from `http://amharicniv.com` and the corresponding text came from `http://www.bible.com`. Each original audio file is an entire chapter. Such single-speaker cleanly recorded data can be quite useful as "found" data.

The Amharic Read Speech (ARS) corpus [15] was originally collected by the University of Hamburg for the development of Automatic Speech Recognition systems. It consists of 20 hours of transcribed speech with 44 female speakers and 56 male speakers. Although it is read speech, the multiple speakers in this corpus, as in most corpora for ASR purposes make it quite different from typical single-speaker TTS corpora collected under better conditions.

4. Building TTS Voices

To prepare the AudioBible data for synthesis, each audio file was segmented into utterance-sized clips using word-level alignments obtained from Prosodylab-aligner [16] and the end-of-sentence punctuation marks present in the transcripts. The transcript was also segmented using the same punctuation marks to finally obtain utterance-sized audio-transcript pairs. These audio-transcript pairs were aligned at the phoneme level using Festival [17]. Due to the large number of out-of-vocabulary words, pronunciation and syllabification were inferred from the transcript itself (Amharic has a highly phonemic orthography).

We used the University of Edinburgh's deep-learning-based speech synthesis toolkit Merlin [18] to train voice models on each of the Amharic corpora. The sampling rate for all corpora was 16 kHz. We used the WORLD vocoder to extract F_0 in log-scale (lf0), mel-cepstral coefficients (mgc) and band aperiodicity (bap) as acoustic features. The synthesized voice consists of both an acoustic and a duration model, each consisting of 6 TANH layers of size 1024. Batch size was 64 for the duration model and 256 for the acoustic model. There was a fixed learning rate of 0.002 and number of training epochs was 25. We used the Merlin "build your own voice" recipe to generate the voice models, with a custom questions file that we created for our Amharic phoneset. Of the three corpora we initially trained Amharic voices on, only the AudioBible corpus produced a voice intelligible enough to evaluate. So our baseline voice for our remaining experiments was trained on the entire 55 hours of AudioBible data.

For frontend processing in Amharic, we used the Festival [17] toolkit. The main resources one needs to provide to Festival to create a linguistic frontend for a new language are a pronunciation lexicon and a phoneset definition. For the lexicon, we started with the Amharic lexicon from BABEL [14]. Since there were many OOV words that were present in the Bible and ARS text that were not present in the BABEL lexicon, we had to create pronunciations for these words and add them to the lexicon. We did this using the CMU Sphinx G2P tool [19]. We trained a g2p (grapheme-to-phoneme) model on the existing lexicon, and then used the model to generate phoneme sequences for our OOV words. The phoneset definition required a list of phonemes used in the lexicon, as well as indication of which ones are vowels, which we selected hand. For the words unrecognized by our g2p model, we manually generated pronunciations.

5. Acoustic-Prosodic Subset Evaluation

We prepared a number of test voices designed to determine whether selecting data from the AudioBible based on acoustic-prosodic-defined subsets would improve over the baseline voice trained on the entire corpus. Our initial test voices were four-hour subsets chosen using one of the following: f0, energy (computed using Praat [20]), speaking rate (in syllables per second), and level of articulation (computed as mean energy divided by speaking rate, so that high levels of articulation are characterized by loud and slow speech). For each feature, we sorted utterances by feature value, and then selected three subsets as follows: we initialized three empty subsets (subset "low", "mid", and "high") and then (1) added the next lowest scoring utterance to the "low" subset until subset was four hours long, (2) added the next highest scoring utterance to the "high" subset until subset was four hours long, and (3) added the "median-scoring" utterance to the "mid" subset, and then "expanded" the subset in both directions, alternating the direction one utterance at a time, until our subset was four hours long.

We trained a Merlin voice on each of the three subsets for all the above mentioned features, keeping all other factors the same among the voice models.

Due to the difficulty of finding Amharic-speaking raters on crowdsourcing websites such as Amazon Mechanical Turk, we conducted an initial automatic evaluation for intelligibility by obtaining Word Error Rate (WER) for our synthesized sentences using a speech recognizer trained by IBM for Amharic for the BABEL project. This ASR system for Amharic is described in [21]. WER results are shown in Table 1, with the best five voices shown in bold.

Voice	IBM WER	Voice	IBM WER
Baseline	45.00%		
High Mean Energy	**20.0%**	High Mean F0	50.7%
Med Mean Energy	31.4%	Med Mean F0	50.7%
Low Mean Energy	64.3%	**Low Mean F0**	**26.4%**
High Stdv Energy	61.4%	High Stdv F0	38.6%
Med Stdv Energy	**30.7%**	Med Stdv F0	40.0%
Low Stdv Energy	**27.1%**	**Low Stdv F0**	**27.1%**
Slow Speaking Rate	62.9%	High Articulation	52.1%
Med Speaking Rate	41.4%	Med Articulation	40.0%
Fast Speaking Rate	45.7%	Low Articulation	41.4%

Table 1: *ASR word error rates for voices trained on 4-hour subsets of Amharic AudioBible data.*

We found that high mean energy was the best selector for creating an intelligible voice from our AudioBible data. This is consistent with our findings in [11], where we also found that high mean energy was one of the most useful selectors for read telephone speech in US English. Furthermore, we observed higher values for mean energy when we compared TTS data to other genres in [4], indicating that this may in fact be a salient feature of good TTS data. The other features that were good selectors were ones that we might expect to be in line with a consistent, neutral style of TTS data: lower ranges for variation in energy and f0, as well as lower mean f0.

6. Adaptation Experiments

While we created a number of voices from different portions of the BABEL corpus, based on gender, on conversational vs. scripted speech, and on location where these were recorded, none produced an intelligible voice – even when we hand-selected the cleanest audio. In the low-resource language setting, there is often only a very small amount of high-quality data available. Then, the question arises whether it is better to use this data by itself, or to combine it with a larger amount of lower-quality data. To test the latter possibility, we first tried to adapt 10 hours of male BABEL speech (first of scripted speech and then of conversational) each to 10 and then to 20 minutes of AudioBible speech using Merlin's Speaker Adaptation recipe. We used the "fine-tune" adaptation method (described in [22]) implemented by back-propagating the adaptation data through the model to re-tune all the weights. However, none of these experiments proved successful. Nevertheless, voices trained on any amount of AudioBible data alone continued to produce better voices than either of the BABEL adapted voices.

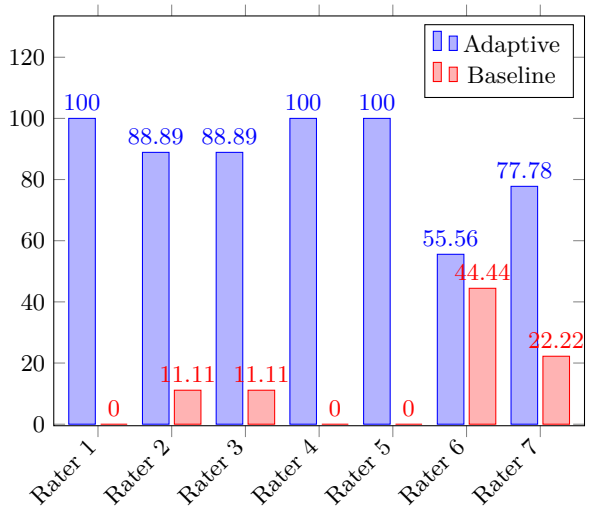
In another experiment, we trained a voice on the multi-speaker ARS corpus, but found that this did not produce a high quality voice, even when training only on a single speaker. However, when we next tried adapting the ARS lower quality speech using a small quantity of the higher-quality AudioBible data, we created a voice superior in naturalness to both the ARS-only trained voices and the AudioBible-only trained voices.

The adapted voice was generated from an Average Voice Model (AVM) which was trained on the entire ARS corpus combined with 10 minutes of AudioBible data; this voice was then adapted to the same 10 minutes of AudioBible data which was used for the AVM.

We evaluated this voice using local crowdsourcing on a simple custom web interface we created specifically for the evaluation of our Amharic TTS voices since it was difficult to find raters using standard crowdsourcing systems. There were seven participants in this evaluation task and all were native speakers of Amharic. Each participant listened to one pair of sentences at a time – a nonsense sentence synthesized from the combined adapted voice and the same sentence synthesized from the baseline voice trained on 10 minutes of AudioBible data alone. We asked each participant to select the one from each pair that sounded to them more natural. Nine pairs of sentences were presented to each rater. Results for each rater are shown in Table 2.

Rater	Adapt Voice	Baseline
Rater 1	100.00%	0%
Rater 2	88.89%	0.11%
Rater 3	88.89%	0.11%
Rater 4	100.00%	0%
Rater 5	100.00%	0%
Rater 6	55.56%	44.44%
Rater 7	77.78%	22.22%

Table 2: *Naturalness Evaluations and Preference Comparisons per Rater for the Adapted Voice.*



As we see from Table 2, the evaluators showed an **87.3%** preference overall for the adapted voice. We performed a z-test to calculate a p-value of 3.19×10^{-9}.

7. Gemination Experiments

In Amharic, *gemination* is one of the most distinctive characteristics of the cadence of speech. It carries a very heavy semantic and syntactic functional weight [3]. Gemination in Amharic can be either lexical or morphological. In its lexical sense, it cannot be predicted. A typical example is **ገና**, which may be read as /gəna/, to mean 'still/yet', or /gənna/, to mean 'Christmas'.

Because these are infrequent, it is usually not difficult for Amharic speakers to distinguish between them from context. From a speech synthesis perspective, however, it becomes difficult to produce these distinctions because Amharic's orthography fails to represent geminates. The morphological form of gemination, on the contrary, **is** possible to predict since it can be identified from the orthography of the language. As an example, consider the root verb made of the consonant sequence sbr ('break') and two words derived from it – **ስበረው** and **ይሰበራሉ.** The first is /sɨlbərəw/, 'break (masc.sing.) it', the second is /yɨssəbbərallu/, 'they are broken'. This distinction can be inferred from the pattern of stem vowels — that /s/ and /b/ are not geminated in the first word and that both are geminated in the second, and that the /r/ is geminated in neither word [3].

7.1. Experiments

Initially, we used the same set of labels created for generating a baseline voice for the AudioBible corpus to reconstruct words including the duration information of each phoneme in the word. These labels were obtained using Festival [17] and pronunciations were obtained primarily from the BABEL lexicon. Many OOV pronunciations had to be created manually. EHMM [23] was used for alignment. We aimed to use this information to identify the duration ranges for normal and geminate forms of each Amharic phoneme. We could identify whether a phoneme was geminated or not based on its duration, and label it as such at the frontend. To do this, we needed to use the already generated duration information per-phoneme and match each phoneme to its respective word in the utterance. The algorithm we used to reverse engineer words from the phonemes in the baseline labels failed to provide actual words and boundaries or pauses between words despite multiple iterations through the data with various editions of the algorithm. Secondly, we experimented with setting a general threshold of duration (for all phonemes) that would distinguish between geminated and normal forms. We experimented with various threshold ranges and narrowed these down to 0.15 seconds as optimal. We further specified a threshold for each phoneme, creating a dictionary of phonemes mapping to their average duration thresholds between geminate and singleton forms. We then added an extra feature to the labels for the AudioBible corpus that indicated whether the phoneme duration was beyond its declared threshold or not: the first indicated that it identified a geminate for the word and the second meaning that it did not. This second approach turned out to be considerably more useful in improving the naturalness of voices synthesized.

7.2. Evaluation and Results

We created voices with the added gemination feature on 4.5 hours of our AudioBible corpus and ran a crowdsourced evaluation for naturalness against voices generated **without** the added "gemination" feature. Naturalness was evaluated by nine native Amharic speakers through our custom website. A total of 13 sentences were generated from a 4.5-hour baseline mode and the same sentences were generated using a 4.5-hour model that incorporated gemination information. Thirteen pairs were thus created and were evaluated by the nine raters. Table 3 and the bar chart below present the results of this evaluation.

Rater	Geminated Voice	Baseline
Rater 1	61.53%	38.47%
Rater 2	100.00%	0.00%
Rater 3	84.62%	15.38%
Rater 4	92.31%	7.69%
Rater 5	100.00%	0.00%
Rater 6	38.47%	61.53%
Rater 7	76.93%	23.07%
Rater 8	69.23%	30.77%
Rater 9	61.54%	38.46%

Table 3: *Naturalness Evaluations and Preference Comparisons for Geminated Voices*

The evaluation results shown in Table 3 show that our native raters found a **77.7%** preference for voices using gemination as an added feature over voices that did not. Upon performing a z-test, we arrived at a p-value of p-value of 2.7×10^{-10}. So we conclude that, adding the gemination feature to the rest of the corpus' labels enabled us to obtain significantly more natural sounding sequence of voices, as was predicted by Anberbir et al.[3].

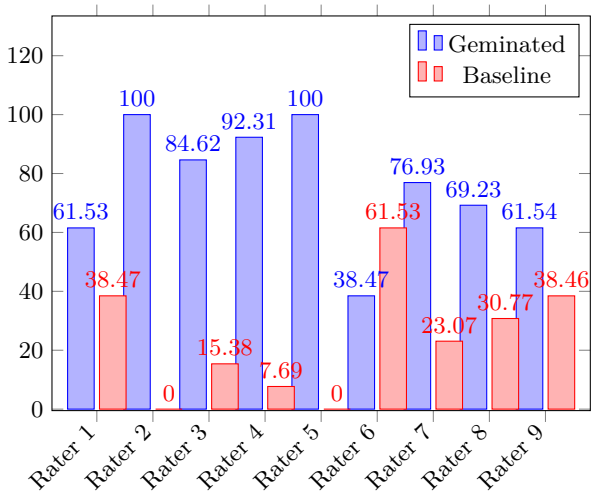


Figure 1: *Naturalness Evaluations and Preference Comparisons for Geminated Voices*

8. Discussion, Conclusions, and Future Work

We have demonstrated a number of ways to build and improve TTS voices for Amharic using various types of found data. We were able to improve intelligibility using AudioBible data by selecting training data subsets based on high mean energy, middle and low standard deviation of energy, and low values for mean and standard deviation of f0. We also found that the combination of a very small amount of high-quality AudioBible data along with a larger amount of lower-quality ASR data can produce a better voice than either data source on its own. Furthermore, we found that incorporating gemination information improves naturalness substantially.

Since the data selection approach has proven to be promising, it would be interesting to extend it by combining multiple features, or by selecting utterances using a more automatic or clustering-based approach. For combination and adaptation using multiple types of data, we could explore different amounts and types of data, to determine the conditions under which adaptation is most beneficial. Furthermore, the question remains whether these methods can apply to more recent end-to-end style TTS such as [24, 25]; although these types of models typically require very large amounts of high-quality data, approaches such as adaptation may be applicable. Gemination gave a marked improvement; however the thresholds we picked for acoustically determining gemination are largely corpus-dependent. So, more generalizeable methods for determining the thresholds should be explored in future work as well.

9. Acknowledgements

This work was supported by the National Science Foundation under Grants IIS 1548092 and 1717680. Many thanks to Alan Black for providing us with some scripts and assistance in using Festival in languages other than the default US English. We also thank Dr. Solomon Abate at Addis Ababa University for allowing us use of the ARS Corpus and Andrew Rosenberg for helping us evaluate the intelligibility of our Amharic TTS systems.

10. References

[1] S. H. Mariam, S. P. Kishore, A. W. Black, R. Kumar, and R. Sangal, "Unit selection voice for Amharic using Festvox," *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[2] N. Hailu and S. Hailemariam, "Modeling improved syllabification algorithm for Amharic," *Proceedings of the International Conference on Management of Emergent Digital EcoSystems, ACM*, 2012.

[3] T. Anberbir, M. Gasser, T. Takara, and K. D. Yoon, "Grapheme-to-phoneme conversion for Amharic text-to-speech system," *Proceedings of Conference on Human Language Technology for Development*, 2011.

[4] E. Cooper, E. Li, and J. Hirschberg, "Characteristics of text-to-speech and other corpora," *Speech Prosody*, 2018.

[5] A. Chalamandaris, P. Tsiakoulis, S. Karabetsos, and S. Raptis, "Using audio books for training a text-to-speech system," *LREC*, 2014.

[6] N. Braunschweiler and S. Buchholz, "Automatic sentence selection from speech corpora including diverse speech for improved HMM-TTS synthesis quality," *INTERSPEECH*, 2011.

[7] O. Watts, A. Stan, R. Clark, Y. Mamiya, M. Giurgiu, J. Yamagishi, and S. King, "Unsupervised and lightly-supervised learning for rapid construction of TTS systems in multiple languages from 'found' data: evaluation and analysis," *8th ISCA Speech Synthesis Workshop*, 2013.

[8] A. W. Black, "The CMU Wilderness multilingual speech dataset," https://github.com/festvox/datasets-CMU_Wilderness, 2018.

[9] E. Cooper, Y. Levitan, and J. Hirschberg, "Data selection for naturalness in HMM-based speech synthesis," *Speech Prosody*, 2016.

[10] E. Cooper, A. Chang, Y. Levitan, and J. Hirschberg, "Data selection and adaptation for naturalness in HMM-based speech synthesis," *INTERSPEECH*, 2016.

[11] E. Cooper, X. Wang, A. Chang, Y. Levitan, and J. Hirschberg, "Utterance selection for optimizing intelligibility of TTS vocies trained on ASR data," *INTERSPEECH*, 2017.

[12] E. Cooper and J. Hirschberg, "Adaptation and frontend features to improve naturalness in found-data synthesis," *Speech Prosody*, 2018.

[13] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, Y. Guan, R. Hu, K. Oura, Y.-J. Wu, K. Tokuda, R. Karhila, and M. Kurimo, "Thousands of voices for HMM-based speech synthesis - analysis and application of TTS systems built on various ASR corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, 2010.

[14] M. Harper, "IARPA solicitation IARPA-BAA-11-02," 2011.

[15] S. T. Abate, W. Menzel, and B. Tafila, "An Amharic speech corpus for large vocabulary continuous speech recognition," *INTERSPEECH*, 2005.

[16] K. Gorman, J. Howell, and M. Wagner, "Prosodylab-Aligner: A tool for forced alignment of laboratory speech," *Canadian Acoustics*, vol. 39, no. 3, pp. 192–193, 2011.

[17] A. W. Black, P. Taylor, and R. Caley, "The Festival Speech Synthesis System," http://www.festvox.org/festival/.

[18] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," *9th ISCA Speech Synthesis Workshop*, 2016.

[19] CMU Sphinx Sequence-to-Sequence G2P Toolkit. [Online]. Available: https://github.com/cmusphinx/g2p-seq2seq

[20] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9–10, pp. 341–345, 2001.

[21] J. Cui, B. Kingsbury, B. Ramabhadran, G. Saon, T. Sercu, K. Audhkhasi, A. Sethy, M. Nussbaum-Thom, and A. Rosenberg, "Knowledge distillation across ensembles of multilingual models for low-resource languages," *ICASSP*, 2017.

[22] B. Bollepalli, M. Airaksinen, and P. Alku, "Lombard speech synthesis using long short-term memory recurrent neural networks," *ICASSP*, 2017.

[23] K. Prahallad, A. W. Black, and R. Mosur, "Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis," *ICASSP*, 2006.

[24] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," *INTERSPEECH*, 2017.

[25] W. Ping, K. Peng, and J. Chen, "ClariNet: Parallel wave generation in end-to-end text-to-speech," *ICLR*, 2019.