A target approximation intonation model for Yorùbá TTS

Daniel R. van Niekerk[1], *Etienne Barnard*[2]

[1]Centre for Text Technology, North-West University, Potchefstroom, South Africa
[2]Multilingual Speech Technologies, North-West University, Vanderbijlpark, South Africa
daniel.vanniekerk@nwu.ac.za, etienne.barnard@nwu.ac.za

Abstract

A complete intonation model based on quantitative target approximation is described for Yorùbá text-to-speech (TTS) synthesis. This model is evaluated analytically and perceptually and compared to a fundamental frequency (F0) model using the standard HTS implementation. Analytical results suggest that the proposed approach more efficiently models F0 contours given typical data constraints in under-resourced environments and perceptual results comparing the proposed model with HTS are encouraging.

Index Terms: speech synthesis, text-to-speech, intonation model, target approximation, tone language, Yorùbá, under-resourced.

1. Introduction

Increasingly powerful and efficient models and algorithms for speech and language processing have resulted in a suite of open source tools that have enabled the construction of successful corpus-based speech synthesis systems in under-resourced environments [1, 2]. In many cases, acoustic models for a basic speech synthesiser in a new language can be constructed automatically from a relatively small corpus of speech recordings and little language-specific development; typically less than one hour of audio, a phoneme set, small pronunciation dictionary or hand-written grapheme-to-phoneme rules and a simple syllabification algorithm based on the phonotactic constraints of the language will suffice.

Building basic speech synthesis systems of this nature for tone languages, however, presents a distinct set of challenges. In addition to the supplementary pronunciation information required to specify underlying lexical tones on the syllable level, complex linguistic processes such as tone sandhi and tone spreading often need to be modelled to arrive at a *surface tone* specification that may be used during acoustic modelling and speech synthesis. Furthermore, F0, the main acoustic correlate of tone, is known to have several additional linguistic and para-linguistic communicative functions [3] as well as physiologically motivated patterns such as *declination* [4]. This multiplexing of patterns in F0 poses a challenge to robust and accurate acoustic modelling, which in the case of non-tone languages largely affect naturalness (pragmatics), but in tone languages also threatens lexical intelligibility especially in under-resourced environments. These potential problems in TTS systems with limited tone specification have been documented and the impact on synthesised speech quality differs depending on the language [5, 6, 7].

In this work, we develop a suitable speech corpus and describe the development of a complete tone-aware TTS system for Yorùbá, including a the evaluation of a new intonation model. In the following section we briefly present some background on Yorùbá as well as previous work and motivate our approach. This is followed by descriptions of the corpus and system development. Finally, analytical and perceptual results are presented in an attempt to evaluate the utility of the intonation model presented, followed by a discussion and proposal for further work.

2. Background and approach

Yorùbá is a relatively well studied African tone language in the Niger-Congo family of which the linguistic details of the tone system have been thoroughly described. Three level (register) tones, labelled High (H), Mid (M) and Low (L) are associated with syllables and have a high functional load [8]. Tones are marked explicitly on the orthography (shallow marking), making automatic derivation of surface tone from text possible. In previous work we investigated pitch contours in different tone contexts and the distribution of syllable pitch height targets in complete utterances in a multi-speaker speech corpus [9, 10, 11], confirming the following core observations: **(1)** The nature of pitch contours in various tone contexts is most strongly affected by the preceding context due to inertia and to a lesser extent by following context in the form of dissimilative H tone raising and gradual rises and falls over sequences of H and L tones respectively. These observations confirm reports in [3, 12, 13]. **(2)** Both the height and gradient of a syllable pitch contour should be considered acoustic indicators of tone in Yorùbá and thus need to be modelled appropriately. We found that in certain tone contexts the gradient is the best indicator of H and L tones, while the most reliable indicator of M tones was height. This finding agrees with previous observations of *rising* and *falling* tones [12] and supports the idea that these patterns have become phonologised or have undergone "target alternation" [14]. **(3)** A pattern of gradual downward pitch movement emerges on average during the analysis of pitch height targets in utterances, with regression models based on a linear declination (in semitones) or syllable index as independent variable leading to more accurate prediction of height than models based solely on *downstep* [15].

In this paper we use these observations to develop a tone-aware HMM-based TTS system for Yorùbá using HTS [2] and an intonation model based on the quantitative target approximation (qTA) approach [16] suitable for use in such a system.

3. Corpus

A single-speaker speech corpus was developed to support the building of a general purpose TTS system as follows: **(1)** Text was sourced from the Yorùbá language Wikipedia (database

dump dated 2013-05-25)[1] to select a subset for diphone coverage [17] and the result professionally edited. the Yorùbá translation of the Universal Declaration of Human Rights document dated 1998-11-12[2] (109 sentences and 2455 tokens) was added to serve as a pristine test set. **(2)** Sentences were randomised, recorded and post-processed in a professional studio environment over the course of two consecutive days. Phonemic alignment of utterances and transcriptions was done by forced alignment using HTK [18] with parameters and model initialisation as described in [19] and with a phone set defined for Standard Yorùbá and simple grapheme-to-phoneme and syllabification rules [20]. Pauses between words were detected automatically using re-alignment with `HVite` and the result was post-processed to remove inserted pauses shorter than 100 ms and insert breath-group breaks for durations longer than 300 ms. **(3)** Transcriptions for the recorded utterances were corrected using a semi-automatic process based on cepstral distance measurements [21] resulting in a total of 788 and 109 utterances in the "train" and "test" sets respectively, with a subset of the train set ("clean") where all tokens could be processed by the Yorùbá text-analysis components (having valid tone labels). The final corpus properties are summarised in Table 1.

Corpus	Utts.	Brthgs.	Words	Sylls.				Phones	Dur.
				H	M	L	N		
train	788	1704	13778	8650	9471	7753	408	45863	86
clean	654	1377	11171	7234	7564	6543	0	37248	70
test	109	276	2543	1609	1382	1753	0	8183	15

Table 1: *Corpus properties with syllable counts by tone (*`N` *indicates "None", mostly resulting from foreign words or names that were not processed by the Yorùbá text-analysis components). The number of phones and corpus duration (in minutes) exclude pauses.*

4. System

4.1. Pitch extraction

All F0 contours were extracted in the range **50 to 250 Hz** using Praat's autocorrelation method [22] and converted to semitones relative to 1 Hz at 1 ms intervals for the target approximation processing and log F0 at 5 ms for HTS.

For extracting the qTA parameters, *height*, *gradient* and *strength*, we implemented the analysis-by-synthesis process as described in [16] and applied this process to each breath-group detected in the corpus separately. As the process of extracting target parameters is quite sensitive to the placement of syllable boundaries – sometimes finding unexpected parameters given small misalignments – if the search space is left unconstrained, the search space was limited according to previous observations (Section 2) about tone contexts:

- **Height**:
$$x \in linspace(a = 67.73, b = 95.59, c = 100)$$
where $linspace$ generates a linearly spaced vector including endpoints a and b in semitones with size c (intervals $c - 1$), resulting in a resolution of about 0.28 st and corresponding to the limits used above (50 to 250 Hz).
- **Gradient**: Absolute limits on gradient were set at $-60 \leq g \leq 60$ st.s^{-1} with a resolution of 4 st.s^{-1}, corresponding approximately to the findings on maximum speed of pitch change in [23]. Ranges were limited more specifically for different tones:
$$g_H \in linspace(0, 60, 16)$$
$$g_L \in linspace(-60, 0, 16)$$
$$g_M \in \{0\}$$
- **Strength**: A single set of constraints was applied over all tone contexts. The minimum strength limits the potential distance between targets and the actual F0 contour and the maximum constraint allows for fully realising targets in short syllables [16]. The resolution is 5 s^{-1}:
$$\lambda \in linspace(40, 120, 17)$$

Given the extracted F0 and qTA parameters we discuss the considerations and implementation of pitch modelling using HTS and qTA in the follow subsections.

4.2. Pitch modelling and synthesis using HTS

We used the HTS toolkit version 2.2 [24] in combination with the HTS engine version 1.05.[3] A comprehensive overview of the theory of HMM-based acoustic modelling can be found in [25]. The standard training procedure implemented in the demonstration available from the HTS website,[4] was used with the following exceptions/details: **(1)** F0 extraction did not use the default tool in the HTS demo script, but *Praat*. **(2)** We implemented mixed excitation synthesis as described in [26]. **(3)** Relying on the HTS engine for synthesis, our acoustic models were single mixture per HMM state. **(4)** Models considering global variance [27] and parameter sharing using decision trees based on the minimum description length (MDL) criterion [28] were used.

Given this setup we experimented with four sets of features including tone context to different degrees; **No tone** (`none`), the "standard" HTS phone and positional features with the exclusion of "guess-part-of-speech" (`gpos`), **Immediate context** (`pt,tt,nt`), the standard features including the current, previous and following syllable tone identities, **Preceding context** (`ppt,pt,tt`) including the pre-previous, previous and current syllable tone identities and **Extended context** (`ppt,pt,tt,nt`) including all the features described above. For each of the tone features, questions were added to allow state clustering of tones in all possible contexts.

To investigate the effect of different feature sets analytically, we performed 10-fold cross-validation experiments on the clean test set, calculating the root mean squared error (RMSE) and Pearson correlation between synthesised and actual F0 contours. The experiment was repeated for 3 different random segmentations of the data by utterance (A, B and C) and to obtain meaningful RMSE and correlation values we found it necessary to smooth the synthesised contours using cubic smoothing splines. Measures were calculated only for non-zero (voiced) sections in both the reference and target contours: Table 2. A significant improvement over the baseline is seen when tone labels are included and the system with **extended context** features performed marginally better and most consistently. This and previous observations led us to adopt this system for further comparison.

	RMSE				Correlation			
Tone context	A	B	C	**Mean**	A	B	C	**Mean**
`none`	3.44	3.42	3.42	**3.43**	0.49	0.50	0.50	**0.50**
`pt,tt,nt`	3.21	3.15	3.18	**3.18**	0.57	0.59	0.58	**0.58**
`ppt,pt,tt`	3.20	3.16	3.18	**3.18**	0.57	0.58	0.58	**0.58**
`ppt,pt,tt,nt`	3.18	3.17	3.17	**3.17**	0.58	0.58	0.57	**0.58**

Table 2: *Root mean squared errors and correlations for the HTS cross-validation experiment. A, B and C refer to independent experiment iterations using different randomisations.*

[1] `http://dumps.wikimedia.org`

[2] `http://www.ohchr.org/EN/UDHR/Pages/Language.aspx?LangID=yor`

[3] `http://hts-engine.sourceforge.net/`

[4] `http://hts.sp.nitech.ac.jp/`

4.3. Pitch modelling and synthesis using qTA

To synthesise appropriate contours using qTA, we need to predict *height*, *gradient* and *strength* for each syllable. As mentioned in Section 2, both the *height* and *gradient* parameters are crucial to tone realisation. Although the importance of the *strength* parameter in this regard is unknown, we assume that it will affect the overall naturalness of synthesised speech depending on semantic context.

An appropriate regression model needs to consider the intra- and inter-syllable relationships between these parameters to ensure the continuity of the resulting contour. In the absence of sufficient knowledge to construct such a model, we assume a strong specification of *height* and *gradient*; using a simple set of regression models to predict these parameters independently in different tone contexts and considering the consequences on resulting contours. Figure 1a illustrates a potential problem; sharp acceleration especially at syllable boundaries which may be perceptually troublesome.

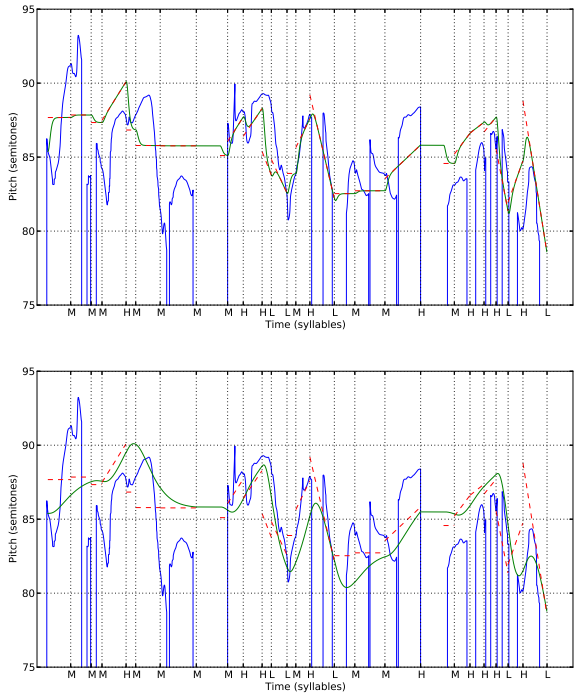

Figure 1: *Synthesised contours (green) from predicted height and gradient targets (red) compared to the original unseen F0 contour (blue) in the training corpus. The first figure (a) illustrates the result for high values of the strength parameter, while the second (b) shows the result using the strength limiting synthesis algorithm for a low value of minimum strength* ($10\ s^{-1}$).

Initial attempts at reducing such artefacts using regression models for *strength* conditional on the *height* and *gradient* parameters were unsatisfactory. This leads to a heuristic solution aimed directly at the observed problem; the requirement is the limiting of sharp acceleration in the opposite direction of height targets without infringing on the velocity of pitch movements required for the implementation of steep targets and natural effects of inertia at syllable boundaries. This was implemented by limiting the *strength* in an iterative synthesis process (Eq. 2 in [16]) until either no acceleration in the opposite direction of the height target is present or the minimum strength is reached. For this implementation we used cubic splines to estimate acceleration and limited the strength in steps of $5\ \mathrm{s}^{-1}$. An example resulting from this synthesis algorithm (Figure 1b) suggests a potential criticism; flat targets (M tones) will generally be implemented with a very low strength due to the deceleration required to approach such targets. While *height* seems to be the best indicator of M tones, the strength requirements have not been determined. In some works it has been suggested that the M tone is essentially "targetless" (an earlier work by Akinlabi is cited in [13]), while in work on the target approximation model the concept of a "targetless" syllable is disputed, but it is suggested that low strength may be a distinct feature of certain tones, with specific reference to the neutral tone in Mandarin and M tone in Yorùbá [14]. Considering this we adopt the current synthesis algorithm and extend the heuristic approach to completely determine the *strength* parameter by predicting a maximum strength for each syllable before letting the constraint on acceleration systematically determine the eventual value. This procedure is at least partially supported by the observation that the *strength* values found during the analysis-by-synthesis process seem to be more variable depending on constraints on the *height* and *gradient* parameters than vice versa [16], and inspection of the *strength* values in our corpus where a majority of values are found to be at either the minimum or maximum values of the search range (i.e. 5772 and 9325 syllables from 21341 at a minimum and maximum respectively). The determination of *strength* is thus reduced to two meta-parameters: minimum and maximum. Varying the maximum parameter beyond $100\ \mathrm{s}^{-1}$ had little effect on contours, however the minimum parameter broadly determined the smoothness of contours. This parameter was found using cross-validation experiments on the training set, selecting the lowest value before over-smoothing resulted in a significant increase in RMSE and decrease in correlation. Direct optimisation using RMSE and correlation was not possible as this did not lead to smooth contours.

To model syllable targets in utterance context, two mechanisms were tested, **(1)** a simple tree implementation subdividing estimation contexts using pre-ordered categorical features with a back-off mechanism motivated by previous results [11], employing a "minimum samples" (`minsamples`) meta-parameter to control model complexity and **(2)** independent modelling of the parameters using standard regression trees [29] as implemented in *Scikit-learn* [30] using a combination of categorical and numerical features, using minimum samples per leaf as meta-parameter. The categorical features employed included target tone (`tt`), utterance final syllable (`uf`), previous tone (`pt`), pre-previous tone (`ppt`) and following tone (`nt`), with utterance contexts categorical for (1) and numerical for (2): breath-group length (`bl`) based on 5-syllable chunks into 3 states (`bl53`), i.e. states for 1 to 5, 6 to 10 and >10 syllables, syllable index in breath-group (`si`) similarly using 5-syllable chunks and having up to 6 states (`si56`) and breath-group index (`bgi`) allowing for up to 5 states (`bgi5`).

For all estimates using mechanism (1) we performed 3-fold cross-validation to determine the `minsamples` meta-parameter by minimising the RMSE of the *height* parameter, the same value was used for *height* and *gradient* estimates. For mechanism (2) we optimised the trees independently using 3-fold cross-validation also using the MSE criterion. With mechanism (1) we evaluated estimates based on the mean (`mean`) and a joint estimate of *height* and *gradient* (`linR2`) using multiple linear regression given the previous syllable values. Results for the 10-fold cross-validation experiment are marginally in favour of the tree-based model (Table 3).

Model	Syllable features	RMSE	Corr.
Mean (eval. height)	bl53,si56,bgi5,tt,uf,pt,ppt,nt	**3.02**	**0.52**
LinR2 (eval. height)	bl53,si56,bgi5,tt,uf,pt,ppt,nt	**3.07**	**0.51**
Tree (indep.)	bl,si,bgi,tt,uf,pt,ppt,nt	**2.96**	**0.55**

Table 3: *Mean root mean squared errors and correlations for the qTA cross-validation experiments (three iterations).*

5. Results

5.1. Analytical

To compare the relative efficiency of the proposed models, we measured the RMSE and correlation on the unseen test set for portions of the training set used during estimation. This was done by executing 5 iterations of the following experiment: **(1)** Given the "clean" set, create 6 subsets that contain 100%, 75%, 50%, 25%, 10% and 5% of the original number of utterances by iteratively randomly discarding utterances. **(2)** Estimate models using HTS and qTA models for each subset. **(3)** Synthesise and compare utterances against the original speech samples in the "test" set (Figure 2).

While the interpretation of these results is not straightforward, owing to unknown level of perceptual significance, we make the following observations: **(1)** A lower RMSE is achieved using target approximation models, however, the HTS models are relatively successful at modelling appropriate pitch movement; the simpler qTA models are least successful at modelling pitch movement from the smaller subsets. **(2)** The HTS results are more sensitive to the data and significant variation is present throughout the tested range. The test set contained mostly long utterances and the training set a mixture of lengths, which may explain why some selections at 50% have a lower RMSE and better correlation than using the full corpus.**(3)** The independent trees modelling *height* and *gradient* were most consistent with better performance on both measures over the entire range tested.

Despite the approximate nature of this analysis, we suggest that the consistency and margin achieved by the tree-based qTA model indicates a more accurate and efficient intonation model given modest to very limited speech corpora such as tested here.

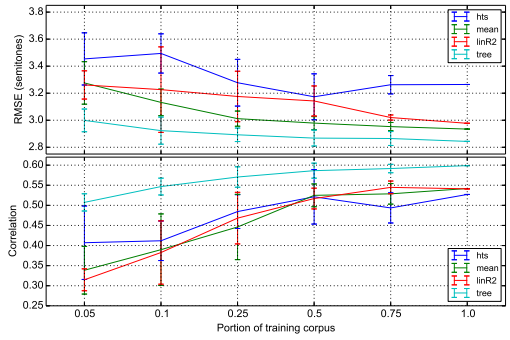

Figure 2: *RMSEs and correlations on the held-out test set for models estimated from portions of the clean training set. Plots show the mean of 5 iterations using different randomly selected subsets and error bars show the standard deviation.*

5.2. Perceptual

We attempted to determine the perceptual appropriateness by means of a simple preference test comparing the HTS synthesised samples with samples when replacing F0 before vocoding with contours generated by the tree-based qTA model. These contours did not contain voicing information, relying on the mixed-excitation bandpass strength models. For the perceptual experiment we selected 30 sample pairs from the test set by dividing the utterances into three sets based on length; "short", "medium" and "long" having from 1 to 10, 11 to 20 and more than 20 words and randomly selecting ten utterances from each set. The preference test was conducted via a simple HTML form with respondents being able to listen to samples without limitation and simply having to select either of the samples or "no preference". The task was performed by 7 respondents (Table 4). According to McNemar's test statistic using the chi-squared distribution with 1 degree of freedom and Yates' continuity correction, the 95% confidence level is given by $\frac{(|b-c|-0.5)^2}{b+c} \geq 3.841$. These results show that respondents significantly favoured the qTA samples in the case of short utterances and no significant difference was found for medium to longer utterances, however there is a weak preference for the HTS model in longer utterances. We suspect the result for longer utterances is due to the HTS model reproducing other – prosodic – aspects of F0 more clearly.

Utterance length (words)	HTS	qTA	No preference	Total	χ^2
$n \leq 10$	3	51	16	70	41.782
$10 < n \leq 20$	17	16	37	70	0.008
$20 < n$	22	13	35	70	2.064
All	42	80	88	210	11.527

Table 4: *Perceptual preference.*

6. Conclusion

In this work we have developed and evaluated an intonation model for Yorùbá based on quantitative target approximation (qTA). Analytical and perceptual results showed respectively that such a model is a promising option for speech synthesis development in under-resourced environments and that the models and synthesis algorithm developed here are applicable in HMM-based synthesisers.

As well as being a more accurate and efficient model, a working intonation model based on simple interpretable parameters provides developers of TTS systems the opportunity for intervention when implementing higher level prosodic effects using smaller amounts of additional data or based on theory (e.g. in the PENTA framework [3]) and for robust modelling of tone in the face of noisy or non-ideal data. The synthesis algorithm described allowing relatively independent modelling of *height* and *gradient* targets without resulting in disturbing synthesis artefacts makes this implementation an ideal vehicle for further theoretical work on dynamic models to predict *height* targets [10]. Immediate future work should investigate synthesis of expressive prosody within this framework and more detailed perceptual tests in the form of "Blizzard-style" naturalness and intelligibility tests [31, 32] to compare this work with other efforts to develop tone-aware synthesisers for African tone languages such as Ibibio [7].

7. Acknowledgements

The authors would like to thank Oluwapelumi Giwa and Olamma Iheanetu for crucial assistance with the development of the speech corpus and perceptual evaluation.

8. References

[1] M. Davel and E. Barnard, "Pronunciation prediction with Default&Refine," *Computer Speech and Language*, vol. 22, pp. 374–393, 2008.

[2] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *The 6th International Workshop on Speech Synthesis*, Bonn, Germany, August 2006, pp. 294–299.

[3] Y. Xu, "Speech melody as articulatorily implemented communicative functions," *Speech Communication*, vol. 46, pp. 220–251, 2005.

[4] A. Cohen, Collier, R., and t'Hart, J, "Declination: Construct or Intrinsic Feature of Speech Pitch?" *Phonetica*, vol. 39, no. 4-5, pp. 254–273, 1982.

[5] J. A. Louw, M. Davel, and E. Barnard, "A general-purpose IsiZulu speech synthesizer," *South African journal of African languages*, vol. 2, pp. 1–9, 2006.

[6] M. Ekpenyong, E.-A. Urua, and D. Gibbon, "Towards an unrestricted domain TTS system for African tone languages," *International Journal of Speech Technology*, vol. 11, no. 2, pp. 87–96, 2008.

[7] M. Ekpenyong, E.-A. Urua, O. Watts, S. King, and J. Yamagishi, "Statistical parametric speech synthesis for Ibibio," *Speech Communication*, vol. 56, pp. 243–251, 2014.

[8] K. Courtenay, "Yoruba: a terraced-level language with three tonemes," *Studies in African Linguistics*, vol. 2, no. 3, pp. 239–255, 1971.

[9] D. R. van Niekerk and E. Barnard, "Tone realisation in a Yorùbá speech recognition corpus," in *The Third International Workshop on Spoken Language Technologies for Under-resourced Languages*, Cape Town, South Africa, May 2012, pp. 54–59.

[10] D. R. van Niekerk and E. Barnard, "Predicting utterance pitch targets in Yorùbá for tone realisation in speech synthesis," *Speech Communication*, vol. 56, pp. 229–242, 2014.

[11] D. R. van Niekerk, "Tone realisation for speech synthesis of Yorùbá," Ph.D. dissertation, North-West University, South Africa, forthcoming.

[12] B. Connell and D. R. Ladd, "Aspects of pitch realisation in Yoruba," *Phonology*, vol. 7, no. 1, pp. 1–29, 1990.

[13] A. Akinlabi and M. Liberman, "Tonal complexes and tonal alignment," in *Proceedings of the North East Linguistic Society*, vol. 31, Georgetown University, 2001, pp. 1–20.

[14] Y. Xu, "Tone in Connected Discourse," in *Encyclopedia of Language & Linguistics (Second Edition)*, K. Brown, Ed. Oxford: Elsevier, 2006, pp. 742–751.

[15] Y. O. Laniran and G. N. Clements, "Downstep and high raising: interacting factors in Yoruba tone production," *Journal of Phonetics*, vol. 31, no. 2, pp. 203–250, 2003.

[16] S. Prom-On, Y. Xu, and B. Thipakorn, "Modeling tone and intonation in Mandarin and English as a process of target approximation," *The Journal of the Acoustical Society of America*, vol. 125, pp. 405–424, 2009.

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

[18] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Veltchev, and P. Woodland, *The HTK Book (for HTK Version 3.3)*. Cambridge University Engineering Department, 2005. [Online]. Available: http://htk.eng.cam.ac.uk/

[19] D. R. van Niekerk and E. Barnard, "Phonetic alignment for speech synthesis in under-resourced languages," in *Proceedings of INTERSPEECH*, Brighton, UK, September 2009, pp. 880–883.

[20] O. A. Ọdéjọbí, A. J. Beaumont, and S. H. S. Wong, "Intonation contour realisation for Standard Yorùbá text-to-speech synthesis: A fuzzy computational approach," *Computer Speech & Language*, vol. 20, no. 4, pp. 563–588, 2006.

[21] D. R. van Niekerk, "Experiments in rapid development of accurate phonetic alignments for TTS in Afrikaans," in *Proceedings of PRASA*, Vanderbijlpark, South Africa, 2011, pp. 144–149.

[22] P. Boersma, *Praat, a system for doing phonetics by computer*. Amsterdam: Glott International, 2001.

[23] Y. Xu and X. Sun, "How fast can we really change pitch? Maximum speed of pitch change revisited," in *The Sixth International Conference on Spoken Language Processing*, Beijing, China, 2000, pp. 666–669.

[24] H. Zen, K. Oura, T. Nose, J. Yamagishi, S. Sako, T. Toda, T. Masuko, A. W. Black, and K. Tokuda, "Recent development of the HMM-based speech synthesis system (HTS)," in *Proceedings of th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Sapporo, Japan, 2009, pp. 121–130.

[25] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[26] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Proceedings of EUROSPEECH*, Aalborg, Denmark, 2001, pp. 2263–2266.

[27] T. Toda and K. Tokuda, "A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.

[28] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modelling for speech recognition," *Journal of the Acoustic Society of Japan*, vol. 21, no. 2, pp. 79–86, 2000.

[29] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1984.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python ," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[31] S. King and V. Karaiskos, "The Blizzard Challenge 2011," in *Blizzard Challenge 2011*, Turin, Italy, 2011. [Online]. Available: http://www.festvox.org/blizzard/blizzard2011.html

[32] S. King, "The Blizzard Challenge 2012," in *Blizzard Challenge 2012*, Portland, Oregon, USA, 2012. [Online]. Available: http://www.festvox.org/blizzard/blizzard2012.html