

# Data-driven Approach to Rapid Prototyping Xhosa Speech Synthesis

*Justus C. Roux, Albert S. Visagie*

Centre for Language and Speech Technology, Stellenbosch University, South Africa  
jcr@sun.ac.za, avisagie@dsp.sun.ac.za

## Abstract

This paper presents work in progress towards building a Xhosa speech synthesizer. HTS is being used for this purpose due to certain desirable properties. As a minority language, linguistic resources for Xhosa are limited despite a variety of impressionistic phonetic studies, prompting a minimalist approach and a preference for data-driven methods. Xhosa is an agglutinative language, and is also held to be a tonal language, which therefore requires morphological analysis and tonal information in order to generate intelligible speech. By taking into account more recent findings on the nature of Xhosa prosody, it appears that a minimalist approach that excludes tone information is possible. We implement the system using HTS. Such a data-driven TTS system is a useful tool to test various syntactic and other features in text that influence Xhosa prosody.

## 1. Introduction

This paper reflects on ongoing work towards the development of a text-to-speech (TTS) system for an African tone language, Xhosa. No serious attempts have yet been made to develop a general TTS system for this language. A limited domain synthesis application has been built in the *African Speech Technology project* (<http://www.ast.sun.ac.za>). An attempt towards the development of such a system for a sister language, Zulu, has recently been made by Louw et al. [1]. A feature of this ‘general-purpose’ synthesizer, however, was that it did not attempt to model intonation.

Xhosa like many other minority languages, lacks linguistic resources that are required for TTS. For example, the language is held to be a tone language, however, impressionistic tonal descriptions are extremely diverse in nature as has been previously indicated by Roux [2], [3]. This leads us to seek a minimalistic approach. In this work, we join current debate in the field, cf. Roux [4], Downing [5], and Kuun et al. [6], and suggest that tone assignment on individual syllables may not be that necessary to construct a highly intelligible Xhosa TTS system. This position challenges entrenched assumptions about the tonal nature of the language. In Section 4, we discuss subjective tests with encouraging results.

The immediate research aims of a broader study in this field are

- to determine what linguistic features are salient for text-to-speech synthesis of Xhosa,
- to build a front-end capable of deriving the features from the text, and
- to create a test bed from which the tonal and/or accentual properties of the language could be assessed through further experimentation.

This paper will reflect on a particular approach followed to create an intelligible Xhosa TTS system. We chose to use HTS for its ability to automatically draw correlations between symbolic features derived from the text and the observed acoustics cf. [7]. This is ideal for this work, since the text-analysis front-end is the only language dependent part of the resulting synthesizer. We also hope that using HTS will let us gauge the importance of various features by judging their effect on the output, and so provide further insight into what is needed for Xhosa TTS.

## 2. Linguistic features of Xhosa

In this section some of the basic linguistic features of Xhosa are listed which need to be taken into account in the development of a TTS system for this language.

### 2.1. Tone and vowel duration

Xhosa is regarded as a **tone language** belonging to the Nguni group of Bantu languages. It is spoken in South Africa by approximately 7,5 million people, i.e. by nearly 16% of the total population.

The language is highly agglutinative which means words are formed by combining a wide range of morphemes with word stems either as prefixes, infixes or suffixes. Hence, the word for a preacher ‘umfundisi’ derives from a verbal stem /-fund-/ ‘teach’ with the following morphemes attached:

$$/u + m(u) + fund + is + i / (1)$$

Although tone is not indicated orthographically lexical tone is realized on each syllable in the final surface form, hence

$$[úmfúndi:sì] \text{ (preacher/ one who teaches)} \quad (2)$$

The low tone of the deleted /u/ is maintained on the preceding nasal /m/. As morphemes are added to this form the tonal pattern may change:

$$\begin{aligned} &/u + m(u) + fund + is + ana \\ &\text{(-ana denotes diminutive)} \\ &[úmfúndisà:nà] \text{ (small preacher)} \end{aligned} \quad (3)$$

Note the H(igh) tone shifts to the antepenultimate syllable, whilst syllable length on /i/ is likewise shifted to the penultimate syllable. This is an important point that will permeate further discussions.

Three tones are traditionally distinguished in Xhosa, i.e. H(igh) [ˈ], L(ow) [ˌ] and F(alling) [ˆ]. Apart from dialectal variations in tonal patterns, impressionistic descriptions are extremely inconsistent, as has been pointed out in some detail by Roux [3]. Claughton [8] for example, introduces the use of superscript x tonal markings to indicate “free variation” in tonal realization in Xhosa, whilst trying to establish particular

tonological rules. The point is that the empirical bases of impressionistic tonal descriptions in Xhosa are suspect; descriptions rarely stretch beyond observations of the production of a single “ideal” mother-tongue speaker of the language. Tonological descriptions more than often reflect the impressionistic interpretations of the researcher, generalizing on the performance of a single mother-tongue speaker; references or access to large speech databases from which conclusions are drawn are non-existent.

An important observation by Downing [5] regarding tone, stress and focus in phonological phrases, provides a new angle when she argues that High tone realisations in Xhosa shows “culminativity effects” that make the tone system resemble stress-accent systems. In stress-accent systems main stress tends to occur on syllables “...at the edge of a stem or word.” Likewise High tones in Xhosa are restricted to occur at word edges, i.e. they regularly appear on the antepenultimate, penultimate or final syllable. Compare examples (2) and (3) above where the High tone on the antepenultimate syllable /fũ/ in (2) shifted to the antepenultimate /di/ in (3) when more syllables were added. This perceived preference for a High tone to appear in an antepenultimate syllable corroborates results of an informal investigation by Roux [4] on the allocation of ‘prominence’ (expressed in terms of H and concomitant increase in amplitude) to successive syllables by mother-tongue speakers of Xhosa. Results obtained for Zulu nouns and adjectives in the experimental work of Kuun et al. [6] also suggest a positional bias for H tones in the penultimate or antepenultimate syllable of the sister language of Xhosa.

Another important phenomenon that contributes to the metrical structure of Xhosa is the predictable assignment of **length (duration)** to particular syllables in a phonological word and/or phrase. Vowel lengthening normally takes place in the penultimate syllable of a word (in isolation), a phrase (demarcated by a following colon, or particular conjunctive words) or sentence (demarcated by a following full stop).

Given the query above on the representativeness of existing tonal data for Xhosa, and taking the observations of Downing [5], Roux [4] and Kuun et al. [6] into account, we adopt a simple syllable counting approach as features for the prediction of tone and duration as mentioned in 3.2.1 and 4.1 below. The observed position of High tone placement on the antepenultimate syllable of a long word, indicating some form of prominence (accent), as well as the predictability of vowel duration, are two aspects under investigation with the aim to create acceptable intonation contours for Xhosa.

## 2.2. Orthography, morphemes and letter-to-sound rules

Xhosa employs a conjunctive orthography, which together with the agglutinative nature of the language, poses a challenge for the construction of a lexicon.

Hence, a single ‘word’ may actually represent a phrase or a sentence:

$$/u + za + ku + ba + fund + is + a/ > uzakubafundisa$$

“He/she will teach them.” (4)

The form above actually comprises concordial morphemes (/u/ and /ba/), morphemes indicating future tense (/za/, /ku/ and /a/), a verbal stem (/fund/), and a causative morpheme (/is/). In order to identify these morphemes (and other parts of speech such as nouns, verbs, adverbs) it is

necessary to invoke a **morphological analyzer** for Xhosa. This analyzer identifies parts of speech, which may be useful for experimentation with prosody prediction in HTS (see also 3.2.2 below).

Fortunately the orthographic representation of Xhosa is very phonetic in nature which simplifies the creation of **grapheme-to-phoneme rules** for the language. An original set of rewrite rules developed by Roux [9] was recently updated and improved by Louw [10], and forms the basis for transforming orthographic forms into appropriate canonical phonetic representations for synthesis.

## 2.3. Segmental phonetic issues

In the development of a TTS system for Xhosa a few idiosyncratic segmental features of the language need to be taken into account. One of the most characteristic features of the language is the presence of **click sounds**; three different click sounds are identified: a dental click (represented orthographically as ‘c’), an alveo-palatal click (represented orthographically as ‘q’), and an alveo-lateral click (represented orthographically as ‘x’). Each of these (unvoiced) click types have four further phonetic attributes, rendering a total of fifteen different click sounds as listed below as represented in the orthography:

	<u>Dental</u>	<u>Alveo-palatal</u>	<u>Alveo-lateral</u>
Unvoiced	c	q	x
Aspirated	ch	qh	xh
Voiced	gc	gq	gx
Nasalized	nc	nq	nx
Voiced Nasal	ngc	ngq	ngx

Due to the fact that many of these clicks are rare, and in view of the desire to minimize the size of the phoneme set of the synthesizer to the most succinct possible set, the unvoiced and aspirated varieties, as well as the nasalized and voiced nasalized were lumped together.

The phenomenon of **tonal depression** is widely mentioned in literature. It implies that an H tone following a voiced consonant will be relatively lower in pitch than an H tone following a voiceless consonant. This phenomenon as well as other phenomena such as **segmental deletions** and **vowel devoicing** at word endings have not been treated in any special way as this will be derived from context information by HTS.

## 3. Implementation

This work used HTS for synthesis and Festival [11] for front-end processing. Following [12], we implement various standard Festival modules for Xhosa. The resulting Festival utterance structures are used to obtain features for HTS.

The only language resources available to us at the outset were the aforementioned manually developed letter-to-sound rules. Consistent with the plight of all minority languages, this scarcity of resources is a major constraint in building Xhosa TTS systems.

### 3.1. HTS back-end

HTS was chosen as a synthesizer for its desirable characteristics [7]. Specifically, HTS draws correlations

between acoustic features and symbolic input features derived from text, making it possible to use it as a black-box, more than other methods. It is reported to work well on small datasets [13].

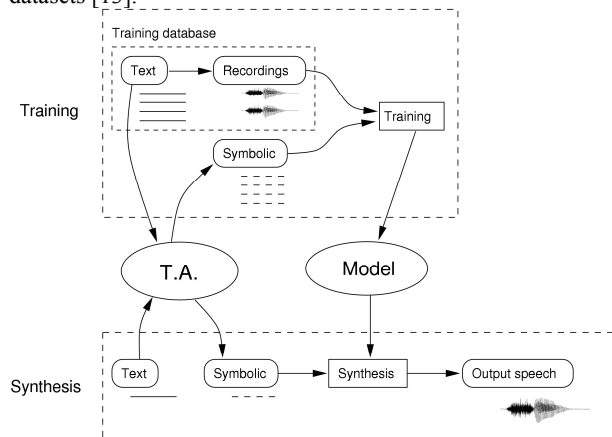


Figure 1: High-level overview of HTS system development. The text-analysis (T.A) component is constructed first. It is used both in training and during synthesis.

Figure 1 shows a high level summary view of how HTS works. The input to the system is a training database consisting of matching text and audio – one prompt per file.

The text component is converted to a phoneme sequence in the front-end (labeled Text-Analysis (T.A.)). Each phoneme label is augmented with symbolic features that describe its context. The contexts used in our system are described in Section 3.3.1.

The signal processing step extracts spectral information and average F0 and a voiced/unvoiced decision every 5ms. These features alone allow resynthesis of the audio.

The training process then finds the locations of the HMM-state-sized segments, 5 per phoneme, and clusters acoustically similar segments together using decision tree clustering.

An important characteristic of the clustering is that the duration of each HMM-state is performed separately for F0, duration and spectrum. Each of these factors are influenced by different contextual features, and as such it does not segment the training set unnecessarily.

HTS allows experimentation with various different features and tree clustering questions. To this end, it is an excellent tool to explore and test theories about the influence of various syntactical and morphological properties of the text on the synthesizer's ability to predict prosody, and allow guided incremental improvement of the text processing front-end.

The HMM synthesis framework trains Hidden Markov Models on a set of features extended from the normal speech recognition usage. The features contain Mel cepstrum parameters for modeling the spectrum, and F0 and duration distributions. It performs context clustering using decision trees separately for the spectrum, duration and F0 components, since different contextual information influences these properties of the surface realization.

## 3.2. Data

Text was collected in the form of two issues of a local tabloid, a novel used in Xhosa language teaching and several government documents explaining various services. The goal was to use edited texts such as these in order to get good quality sentences. It proved to be rather difficult to find appropriate material in electronic format.

Finally, we had 357 recordings, some containing entire paragraphs. There are 3339 words in the recordings. After cutting the recordings into 759 phrases, 43 minutes of speech remained.

The front-end of the synthesizer (described below) was used to obtain phoneme sequences for each utterance. Initial phonetic alignments were made using eHMM, bundled in the FestVox distribution [12]. eHMM produces alignments using forced alignment with a set of HMM models in the Festival voice's own phoneme set. The means and variances of the Gaussian components of the models are flat-started to the global mean and variance of the acoustic data, and then trained using embedded re-estimation. Roughly 10% of the alignments were checked manually, and all were found to be very accurate.

## 3.3. Front-end

Festival applies several modules during its text processing stages: tokenization, POS tagging, syntactic analysis, phrasing, orthographic to phonetic conversion, syllabification and post-lexical rules. The remaining modules' functions (F0, duration, loudness etc.) are performed in HTS.

The aforementioned letter-to-sound rules fit perfectly in Festival's module for rewrite rules, and so were easy to incorporate. The synthesizer training set contained only handful of loan words, and these constituted the lexicon. The lexicon had no stress or tone assignment.

The phoneme-set was determined by the output of the letter-to-sound rules – a total of 82 phonemes. Of these, many were deemed to be very close to each other, and were merged, yielding a final phoneme set of 63 symbols. The variety of consonants mentioned above explains the need for such a relatively large set.

We used the punctuation decision tree in Festival for phrasing.

The current system does not perform any post-lexical changes on the utterances. As it seems very context dependent, and open to speaker specific interpretation, we relied on the data available to HTS.

### 3.3.1. Symbolic features & questions for HTS

At the time of writing, the system outputs these features into the HTS label files:

- Phonetic context, two segments preceding and two following.
- Word position in the sentence.
- Syllable counts from the end of the utterance, and end of the phrase. The observation that the phrase-penultimate syllable is always lengthened to indicate the end of a phrase motivates this.
- Syllable position in the word, both from the start of the word, as well as from the end of the word. For example:

“Okubaluleke” yields these segments and syllable positions: O: 1-6, k: 2-5, u: 2-5, b: 3-4, a: 3-4, l: 4-3, u: 4-3, l: 5-2, E: 5-2, k: 6-1 and E: 6-1.

Although Xhosa is generally held to be a tone language, recent studies [4,5,6] showed that the location of high tones is dependent on position within words and is regularly tied to the antepenultimate or penultimate syllable. The syllable position feature is a minimalist attempt to exploit this regularity in light of the absence of linguistic resources, and recent opinion in the field of Xhosa intonation study.

The question set includes the usual (in HTS) questions about various phoneme properties, such as phonemes types, voicing, place of articulation etc., adapted for Xhosa.

### 3.3.2. Role of morphological analysis

The next step in improving the synthesizer is to perform morphological analysis on the words.

As shown in examples (1-3) above, the tonological structure of the language is influenced by the specific prefixes and suffixes used to compose the word, whether or not each prefix or suffix carries its own high or low tone.

Morphological analysis will enable experimentation with prefix and suffix types as features for predicting prosody in HTS.

An analyzer for Zulu has been developed by Bosch and Pretorius [14], and work towards adapting it for Xhosa is currently underway. The first prototype, used in this work, contains a lexicon of all the morphological roots in the training set.

It is still possible however to interpret isolated words as containing various root morphemes or even different parts-of-speech. Some form of disambiguation given the sentence context of the word remains to be done.

Some classes of words, such as conjunctions, are not composed morphologically, or can be enumerated easily and therefore form small closed sets. Work is underway to produce a lexicon of these words that provides their parts-of-speech and supposed tone-markings. The system will consult this lexicon before attempting morphological parsing.

## 4. Experiments

The synthesizer was evaluated in a very small intelligibility test. Eight stimuli from two versions of the synthesizer, and eight obtained by resynthesizing the extracted spectral and F0 features were played to three mother-tongue, and two second language speakers.

The two versions of the synthesizer differed only in that one excluded the features indicated syllable position in words.

The mother-tongue speakers could understand all the stimuli nearly perfectly. Each of the three mother-tongue speakers indicated that they had trouble with at least one or two of the stimuli. Each of them had difficulty with different prompts. In each of these cases, the listeners were still able to give a very nearly correct “phonetic” transcription.

In each such case we feel that the segmental realization of the prompt was good, and that confusion was caused by bad prosody.

Both second language listeners understood the resynthesized prompts perfectly. However, they only understood slightly more than half of the synthesized

prompts. This shows that the mother-tongue speakers’ results are not quite as encouraging as it might seem.

### 4.1. Syllable position and accent or stress

Subjective comparisons between the same synthesized utterances before and after adding only the word-level syllable counts indicates a significant positive effect of syllable position in words on the rendition of rhythm and intonation. This is obtained without including any explicit accent or stress markings. Several comparative examples, including natural speech may be found at [http://www.sun.ac.za/su\\_clast/tts.html](http://www.sun.ac.za/su_clast/tts.html).

That this one feature made such a significant difference to the prosody seems to support the stress-accent side of recent debate about Xhosa tonology.

Syllabic prominence was generally predicted well for longer words. Short words such as pronouns were usually de-emphasized compared to the naturally pronounced versions. The classical tone markings and better parts-of-speech tagging are being explored as a means of providing information for predicting better prosody for these shorter words.

### 4.2. Clicks

As mentioned before, we lumped together aspirated and unvoiced click sounds. One listener felt that the unvoiced version was produced in a word that contains the aspirated version in one example. The dental click sound is dominated by examples of the unvoiced version. Experimentation is still needed to test the perception of clicks as produced by the synthesizer given information at various granularities.

That said, HTS models click sounds well. In initial subjective tests, listeners generally had no trouble distinguishing between the renditions of types of clicks.

## 5. Future work

We plan to experiment with various ideas of placing accent or predicting tone in the near future. Morphological analysis forms an integral part. The current system only used the morphological parsing results to determine (still ambiguous) parts-of-speech. In the near future we will incorporate information about the boundary between prefixes and the root morpheme first, and then add morpheme types, such as those indicating tense, negatives and diminutive forms.

Explicitly marking syllable prominence, especially for short the words in the current training and development set prompts, should form an interesting experiment to determine the validity of the stress-accent point of view.

Once subjective listening tests indicates acceptable performance, we want to construct a Blizzard style test [15], incorporating preference tests between a small number of systems and intelligibility tests.

The tests should incorporate synthesis of minimal pairs currently considered to be distinguished by tone. There are very few, and they tend to have different parts-of-speech.

## 6. Conclusions

The Xhosa and Zulu languages’ agglutinating nature and tone structure are generally held to be the greatest hurdles to

building TTS systems. We feel that the minimalist approach taken here indicates that good synthesis is already possible with simpler features. The modern data-driven approach relieves one from much of the theoretical effort.

This work is to be used in embedded applications for two projects building translation and educational reference systems at Stellenbosch.

## 7. Acknowledgements

Thanks to Nick Campbell, Kamiya Tosirou and the team at ATR for an enriching and stimulating visit to Keihanna, Japan.

Special thanks to Prof. Sonja Bosch and her team at the University of South Africa (UNISA) for the use of the morphological analysis tools for Zulu and Xhosa.

We express our appreciation towards the National Research Foundation for the sponsorship of this project (GUN 2074784) in terms of the Japan-South African Intergovernmental Science and Technology Cooperation programme. All opinions expressed here are that of the authors.

## 8. References

- [1] Louw, J.A., Davel, M. and Barnard, E. "A general purpose isiZulu TTS system." *South African Journal of African Languages*, 25(2): 92-100, 2005.
- [2] Roux, J.C. "On the perception and production of tone in Xhosa." *South African Journal of African Languages*, 15(4): 196-204, 1995.
- [3] Roux, J.C. "On the perception and description of tone in the Sotho and Nguni languages." *Proc. of the 3<sup>rd</sup> Int. Symposium on Cross Linguistic Studies of Tonal Phenomena*. Tokyo University of Foreign Studies, Tokyo. Ed. S Kaji, pp 155- 176, 2003.
- [4] Roux, J.C. "Xhosa: A tone-or pitch-accent language?" *South African Journal of African Languages*, Supplement 36, 33-50, 1998.
- [5] Downing, L.J. "Stress, Tone and Focus in Chichewa and Xhosa." *Stress and Tone: The African Experience*. Ed. R-J. Anyanwu. Ruediger Koeppel Verlag, Cologne, 2003.
- [6] Kuun, C., Zimu, V., Barnard, E. and Davel, M. "Statistical investigations into isiZulu intonation." *Proc. of the 16<sup>th</sup> Annual Symposium of the Pattern Recognition Association of South Africa*, 111-115, 2005.
- [7] <http://hts.sp.nitech.ac.jp/>
- [8] Claughton, J.S. *The Tonology of Xhosa*. Unpublished Doctoral Thesis, Rhodes University, South Africa, 291pp.
- [9] Roux, J.C. "Grapheme to phoneme conversions in Xhosa." *South African Journal of African Languages*, 9(2): 74-78, 1989.
- [10] Louw, P. "A new definition of Xhosa grapheme-to-phoneme rules for automatic transcription." *South African Journal of African Languages*, 25(2):71-91, 2005.
- [11] <http://festvox.org/>
- [12] Dijkstra, J., Pols, L.C.W., van Son, R.J.J.H., "Frisian TTS, an Example of Bootstrapping TTS for Minority Languages", in *5<sup>th</sup> ISCA Speech Synthesis Workshop*, 2004.
- [13] Maia, R. da S., Zen, H., Tokuda, K., Kitamura, T., Resende, F.G.V. Jr., "Towards the development of a Brazilian Portuguese text-to-tpeech system based on HMM", in *Eurospeech, Geneva*, 2003, pp. 2465-2468.
- [14] Bosch, S.E., Pretorius, L., "Finite-state computational morphology: An analyzer prototype for Zulu", *Machine Translation*, 18:191-212, 2003.
- [15] <http://festvox.org/blizzard/>