An intonation model for TTS in Sepedi

Daniel R. van Niekerk & Etienne Barnard

Human Language Technologies Research Group, Meraka Institute, CSIR, Pretoria, South Africa

dvniekerk@csir.co.za, ebarnard@csir.co.za

Abstract

We present an initial investigation into the acoustic realisation of tone in continuous utterances in Sepedi (a language in the Southern Bantu family). An analytic model for the generation of appropriate pitch contours given an utterance with linguistic tone specification is presented and evaluated. By comparing the model output to speech data from a small tone-marked corpus we conclude that the initial implementation presented here is capable of generating pitch contours exhibiting some realistic properties and identify a number of aspects that require further attention. Lastly, we present some initial perceptual results when integrating the proposed model into a Hidden Markov Model-based speech synthesis system.

Index Terms: speech synthesis, tone languages, Sepedi

1. Introduction

Southern Bantu languages are tone languages in which word-level pitch variations generally convey both lexical and grammatical meaning. In contrast to tone languages like Chinese, they are agglutinative languages, i.e. several morphemes are joined together in a word. Although most Southern Bantu languages only have two level tones, namely high tone (H) and low tone (L), modelling of their prosody is complicated by the agglutinative morphology, the significant influence of grammar and the occurrence of tone sandhi within and across words. Given the role of word-level prosody in processes such as semantic interpretation and the production of natural speech, it is important that a detailed and systematic account of the prosody be given. Such an account is complicated by the fact that tonal information is not indicated in the orthography of many Bantu languages (including Sepedi, which is the focus of the current study).

We have recently presented an overview of intonation in the Southern Bantu languages [1], from which we concluded that a detailed understanding of the tone system of these languages is especially important for the creation of natural-sounding text-to-speech (TTS) systems. Our earlier work focused on two areas, namely (a) deriving tone assignments from text [2] and (b) understanding the relationship between physical parameters (such as pitch frequency) and the tone levels [3]. Here, we build on the findings of those investigations to develop an initial pitch model for TTS in Sepedi. We develop an algorithm that is used to generate fundamental-frequency contours for speech synthesis using Hidden Markov Models (HMMs).

Below, we briefly review a number of pertinent facts on tone in the Sotho-Tswana languages (of which Sepedi is a representative – Section 2), and summarise the approach to prosodic modelling employed in current state-of-the-art approaches to HMM-based TTS (Section 3). Section 4 presents the experimental methodology and corpus employed in our investigation. Our results are contained in Section 5, and Section 6 contains a discussion of our main conclusions and future work that is required to complete the current investigation.

2. Tone in the Sotho-Tswana languages

Most Southern Bantu languages are tone languages whose surface tones can be captured by two level tones, namely high (H) and low (L) [4]. The high tone is the active tone in Sotho-Tswana languages such as Sepedi, as it participates in tone spread and is subject to positional restrictions. As is the case for most Bantu languages, the Sotho-Tswana languages show an asymmetry in the tonal characteristics of its noun and verb system with nouns being more tonal than verbs: whereas nouns can contrast tone on every syllable, verbs only contrast tone on their stem-initial syllable.

By definition, the primary distinctive feature of a level tone is the value of the pitch frequency within the nucleus of a given syllable, with H generally having a higher pitch frequency than L. This general observation was confirmed in our earlier investigations [5], which focused on the temporal alignment of a single high tone within the verbal domain. (As is common practice, we measure the fundamental frequency (F0) as a physical indicator of the pitch frequency.)

The relationships between these pitch values in a complete utterance, as well as the details of the temporal trajectories of F0 within and between syllables, were investigated in [3]. In the current paper, we describe the creation of an analytic model that builds on that work, in order to supply appropriate pitch values to an HMM-based TTS system.

3. Prosodic models in HMM-based speech synthesis

In current HMM-based TTS systems, parameters required to generate prosodic features such as F0 contours and syllable durations are modelled statistically around phone-sized segments of speech. More specifically, F0 contours are modelled along with spectral information in an integrated fashion as part of the HMM framework [6] by employing multi-space probability distribution HMMs (MSD-HMMs). This supports models of parameters with varying dimensionality (enabling the modelling of continuous and discrete features) [7], while segment durations are modelled by employing Gaussian distributions describing the HMM state durations (and in some cases also modelling durations of complete phones [8]).

Using these mechanisms to model static and dynamic aspects of the pitch (F0, δF0 and δ^2F0)from training data based on criteria such as maximum likelihood (ML) or minimum generation error (MGE) [8] in conjunction with a parameter generation algorithm which accounts for dynamic features [9], one can generate relatively detailed and accurate contours associated with individual phones.

In practical systems such as that described in [10] (where an English synthesiser is constructed), "full-context" phone models are constructed by considering phonetic, syllabic, phrase- and sentence-level context as well as including various linguistic and prosodic features such as part-of-speech, stress and ToBI labels [11] and performing tree-based clustering in order to tie models where appropriate. This yields an integrated overall F0 intonation model which suitably models English prosody.

Although the mentioned techniques are very powerful and usually result in highly accurate models given appropriate training data, there are a number of things to consider when trying to develop systems with appropriate intonation for tone-languages in resource-scarce environments:

- Designing and developing speech corpora to ensure synthesised speech with linguistically correct prosody by relying on modelling based on phone models in context is non-trivial and requires expertise and effort which is not often found in resource-scarce environments.
- By developing an analytic model based on well understood linguistic phenomena, the potential is there to easily adapt such a model to different dialects/languages (e.g. other Sotho-Tswana languages in our case) without having to rely on additional data collection efforts at each stage.

4. Methods and corpus

Firstly we present an algorithm for generating an F0 contour suitable for creating a voiced excitation signal for synthesis with the HTS vocoder, followed by a description of our test corpus and methods of evaluation.

4.1. Generating an F0 contour from linguistic specification

The main conclusions reached in [3] regarding the link between tone levels and F0 in the context of complete sentences can be summarized as follows:

- The *changes* in mean F0 between syllables (and especially in the syllable nucleus) are the strongest indicator of tone, and
- The influence of tone or F0 of surrounding syllables and the segmental make-up of syllables on the perceived tone level could not be clearly established,

We therefore start with a process of defining relative mean F0 values per syllable in a specific utterance given the initial F0 value ($f0_1$), final F0 value ($f0_n$) and sequence of tone levels (t_1, t_2, ... where $t_i \in \{H, L\}$) corresponding to each syllable, by defining a table mapping tone transitions ($\Delta t(\delta t)$ where $\delta t \in \{HH, HL, LL, LH, FF\}$) to relative changes in mean F0 between syllables. The symbol FF represents a special case in Sepedi based on the observation that the penultimate syllable in a sentence always exhibits a marked fall in F0 (usually also with an increase in syllable duration) [1]. This is used to determine the absolute value of F0 for each syllable i given the initial and final F0 values as follows:

$$\begin{aligned} \delta t_i &= t_{i-1}t_i \quad \text{where } i \neq n-1 \\ \delta t_i &= FF \quad \text{where } i = n-1 \\ f0_i &= \begin{cases} f0_1 & \text{where } i = 1 \\ f0_{i-1} + \Delta t(\delta t_i)\frac{f0_n - f0_1}{\sum_{i=2}^{n} \Delta t(\delta t_i)} & \text{where } 1 < i < n \\ f0_n & \text{where } i = n \end{cases} \end{aligned}$$

Given these mean F0 levels for each syllable, a smooth F0 contour is constructed by creating a step-like sample sequence given the syllable (and segment) durations, low-pass filtering this sequence in order to obtain a relatively smooth contour and deleting samples for segments that are defined as unvoiced (an example is shown in Figure 1).

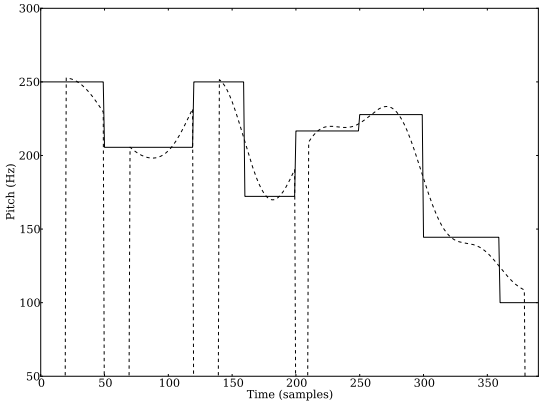

Figure 1: *An example contour generated with the input parameters:* $f0_1 = 250Hz$, $f0_n = 100Hz$, $t = [L, L, H, L, H, H, L, L]$, $\Delta t() = (LL \mapsto -1.6, LH \mapsto +1.6, HL \mapsto -2.8, HH \mapsto +0.4, FF \mapsto -3.0)$*, where relative changes are in Hz/Hz. Thus in this example there are 8 syllables with various lengths and unvoiced segments deleted.*

As can be seen, global effects such as the overall declination in pitch is accounted for by choosing appropriate values for the relative changes between syllables.

4.2. Speech corpus

For our current investigation, we use a single speaker speech corpus that was recorded for the development of a unit-selection TTS system. The speaker is a 30-year old male employing a standard dialect of Sepedi and in accordance with the requirements for TTS development with a limited corpus, the speaker was requested to speak naturally, but with a relatively flat intonation.

Of these utterances, 15 were selected for analysis (based on factors such as the absence of loan words and proper nouns, and limitations on the mood of the verb to limit the influence of dialectal variations). All syllables were subsequently labelled for tone by three labellers independently of each other, relying on perception and analysis of the F0 contour using the *Praat* software package [12] (refer to [3] for a discussion on the labelling process). These labelled utterances were used for comparison as described in Section 4.3, while the remaining utterances were employed in the construction of an HMM-based TTS system (described in Section 4.4). All utterances were automatically aligned as described in [13], resulting in the properties described in Table 1.

Set	Utterances	Duration	Sylls.	Phones
Complete	322	26 mins.	6223	13726
TTS	307	25 mins.	5959	13145
Tone-marked	15	1 min.	264	581

Table 1: Corpus properties.

4.3. F0 contour comparison

We evaluate the output of our model by calculating the mean square error (MSE) per syllable between the pitch contours extracted from the subset of the corpus described above using *Praat* and the generated pitch contours. Values are compared with baseline references including linearly declining and flat contours and contours generated with the HMM-based system described in the following section. To test our model with appropriate parameters, the tone-marked utterances are divided into three sets of five utterances each. The comparison then consists of three separate experiments where two sets of utterances are used to estimate the parameters and the third is used for evaluation (three-fold cross validation). Model parameters (including the relative pitch changes and start and end frequencies, where we modelled starting frequencies for utterances starting with L and H tone separately) are estimated using a sequential least squares optimisation algorithm [14] implemented in the *SciPy* software package [15] and segment and syllable durations are obtained from the phonetically aligned speech samples. Although the values of parameters mentioned are estimated from the data, we keep the signs constant (consistent with the example; Figure 1), thereby fixing the direction of mean F0 change for each step defined (e.g. FF having a fixed negative sign will always represent a fall in mean F0). In the case of the baseline contours we simply used the average start and end frequencies and average frequency from the training utterances for the linearly declining and flat contours respectively. Statistics from the same training utterances are used here as these contours represent competing baseline models. During comparison, slight differences between start and end times of contours within a syllable are handled by only comparing contours where they overlap, i.e. at points in time where both contours have defined values (voiced regions).

4.4. TTS system

Using the above-mentioned corpus, an HMM-based synthesis system was constructed. Training of HMM models was done via the standard demonstration script available as part of the HMM-based Speech Synthesis System (HTS) [16] with the addition of incorporating global variance as described in [17].

For the model tying decision tree, questions relating to phone and word contexts were constructed, while further questions were generated based on phonetic categories defined in the phone set (e.g. categories such as plosives, nasals and vowels and voicing etc.).

5. Results

5.1. Contour comparison

In Table 2, statistics of the RMSEs calculated on all of the test sets for each of the models (as described in Section 4.3) are presented. In the first row the values are calculated over the entire voiced portion of each syllable and in the second row we consider only the nucleus of each syllable as these are considered to be more relevant in the perception of tone.

It is evident that the speech used here employs a rather flat intonation: the RMSE values for the flat and linearly declining contours are closely comparable. The tone-based model fares slightly better than these baselines, but is slightly further from the reference contours than the HMM-generated contours trained from the remaining utterances of the same speaker (which, of course, have many more adjustable parameters).

tone-based		linear		flat		HTS	
μ	σ	μ	σ	μ	σ	μ	σ
5.98	2.82	6.43	2.89	6.31	3.17	5.23	2.45
5.69	3.15	6.19	3.23	6.22	3.65	5.11	3.02

Table 2: Mean RMSE values (Hz) calculated over complete syllables and syllable nuclei respectively.

Although the size of the tone-marked corpus does not lend itself to a comprehensive statistical analysis of the comparison results, we have identified a number of characteristics consistently exhibited in the natural F0 contours not accounted for in the tone-based model which compromises the precision of this model. Figures 2 to 4 illustrate some of these factors, including:

- F0 contours at the start of an utterance often climb to a higher value relatively late despite being perceived as H (Figure 2).
- New phrases often result in a jump in the F0 contour to a higher value after which the pattern of relative changes continue (Figure 3, around time sample 900).
- Some sequences of tones are not modelled well by only considering local changes (e.g. a string of H tones realising as a slight decline in F0, Figure 4).

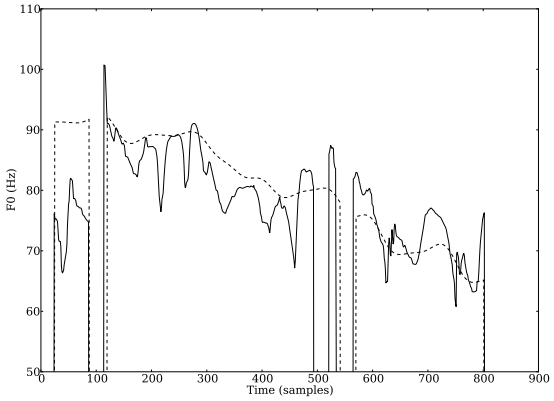

Figure 2: *Example comparison where a large error is seen on the first syllable.*

As one can see there is also a fair amount of intra-syllable detail present in the natural contours that is not modelled, although it is unclear how much of this detail is perceptually important.

5.2. Perceptual evaluation

To get an initial indication of the perceptual significance of the F0 contours generated here, a small perceptual evaluation was performed where 2 Sepedi speakers rated synthesized samples generated with the system described in Section 4.4. Each of the utterances in the tone-marked set was synthesised with excitation signals derived from the standard HMM-based models, the tone-based model and the linearly declining contours discussed above. Listeners were asked to rate each sample using integers ranging from 1 (poor) to 5 (excellent) on the overall quality of the sample.

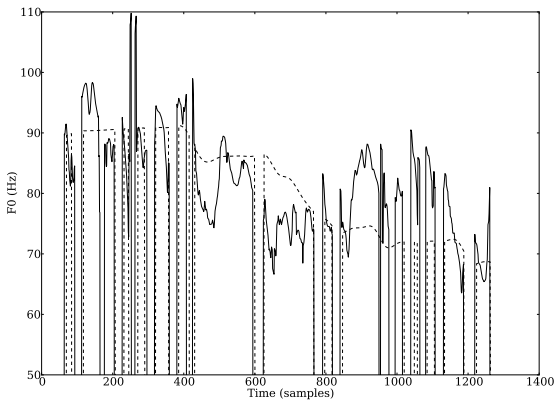

Figure 3: *Example comparison where the error in the latter part of the utterance is high because of the occurrence of a phrase break.*

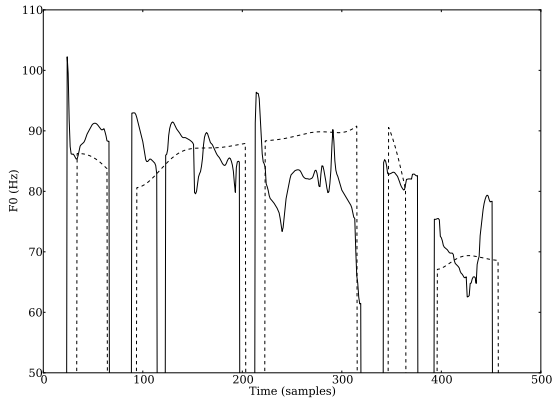

Figure 4: *Example comparison where the occurrence of a string of five H tones from around sample 100 to sample 300 result in an inaccurate tone-based contour.*

Table 3 summarises the scores assigned to each of the systems by each respondent. The scores obtained on this small test set do not allow us to properly investigate the perceptual implications of our model. However, the relatively small differences seen here suggest that some fine differences measurable during the F0 contour comparisons might not be very significant perceptually.

system	listener 1	listener 2	overall
linear	2.27	2.87	2.57
tone-based	2.60	2.53	2.57
HTS	2.47	2.87	2.67

Table 3: Mean opinion scores.

6. Conclusions and future work

Our initial implementation of an analytic intonation model for Sepedi has demonstrated the ability to successfully model aspects of the F0 contour of natural speech given a linguistic tone description (Section 5.1). However, a number of shortcomings requiring further attention have also been identified:

- The implementation of a tone reset mechanism is necessary where a phrase boundary causes the F0 contour to rise, this might be even more important when speech with more naturally varying intonation is modelled.
- Intra-syllable variation in the F0 contour such as the effect of the segmental make-up of a syllable needs to be investigated, both in terms of perceptual relevance and contour accuracy.
- The addition of further contextual information needs to be considered especially for some cases such as repeated H tones.
- Exceptional variation of the F0 contour at the start and end of utterances or phrases need to be investigated.

Future work should focus on investigating and incorporating the above points into the model, further work into using information available via HMM-based modelling of the F0 contour as well as a more extensive evaluation of the intonation model using a larger corpus of more natural speech data.

7. References

[1] S. Zerbian and E. Barnard, "Phonetics of intonation in South African Bantu languages," *Southern African Linguistics and Applied Language Studies*, vol. 26, no. 2, pp. 235–254, 2008.

[2] S. Zerbian and E. Barnard, "Word-level prosody in Sotho-Tswana," in *Proceedings of Speech Prosody 2010*, 2010.

[3] E. Barnard and S. Zerbian, "From tone to pitch in Sepedi," in *Proceedings of the 2nd International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU'10)*, 2010, pp. 29–34.

[4] C.W. Kisseberth and D. Odden, "Tone," in *The Bantu Languages*, D. Nurse and G. Philippson, Eds. Routledge, London, New York, 2003.

[5] S. Zerbian and E. Barnard, "Realizations of a single high tone in Northern Sotho," *Southern African Linguistics and Applied Language Studies*, vol. 27, no. 4, pp. 357–379, 2009.

[6] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-Based speech synthesis," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[7] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden markov models based on multi-space probability distribution for pitch pattern modeling," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing- Proceedings*, 1999, vol. 1, pp. 229–232.

[8] Keiichiro Oura, Yi-Jian Wu, and Keiichi Tokuda, "Overview of NIT HMM-based speech synthesis system for Blizzard Challenge 2009," University of Edinburgh, Sept. 2009.

[9] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings*, 2000, vol. 3.

[10] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *IEEE Speech Synthesis Workshop*, 2002.

[11] M. E Beckman and J. Hirschberg, "The ToBI annotation conventions," 1993.

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

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

[14] D. Kraft, "A software package for sequential quadratic programming," *Forschungsbericht- Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*, 1988.

[15] Eric Jones, Travis Oliphant, Pearu Peterson, et al., "SciPy: Open source scientific tools for Python," 2001–2010.

[16] 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*, 2006.

[17] 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.