A Quantitative Model of Yorùbá Speech Intonation Using Stem-ML

Ọdẹ́túnjí Àjàdí, Ọdẹ́jọbí

Room 109, Computer Buildings,
Computer Science & Engineering Dept.,
Ọbáfẹ́mi Awólọ́wọ̀ University,
Ilé-Ifẹ̀, Nigeria.

Abstract. We present a quantitative model of Standard Yorùbá (SY) intonation; it is designed to have parameters that are linguistically interpretable. The model is built and trained on speech data from a native speaker of SY. The resulting model reproduces the data well: its Root Mean Square prediction error (RMSE) is 14.00 Hz on a test set. We find that intonation is used to mark sentence and phrase boundaries: beginning syllables are systematically stronger, while ending syllables are systematically weaker than the medial syllables. The M tone is the strongest and the H tone is the weakest, though the differences are modest. We see comparable amounts of carry-over and anticipatory co-articulation. The resulting model for SY shows similar characteristics when compared to Mandarin and Cantonese intonation models.

Keywords: Intonation modelling, Speech synthesis, Quantitative model

(Received September 15, 2006 / Accepted May 27, 2007)

1 Introduction

In the past two decades, research efforts have been directed towards the development of ubiquitous Human-Computer Interface (HCI). Speech synthesis and speech recognition are two of the most important technologies deployed in this regard. The speech recognition technology allows a human speaker to give verbal commands to a computer system. Speech synthesis technology, on the other hand, allows a computer system to generate spoken responses to human requests or commands.

The combination of these two technologies into the computer user interface provides a number of advantages. First, they have the potential to improve the rate, ease, and accuracy of data entry and retrieval. Second, their application will reduce the physical restrictions on the use of the hands and eyes while operating a computer terminal. Third, and most important, their application will reduce, considerably, the amount of formal training required for operating computer systems. The third advantage is of importance to African countries in that a large percentage of the population do not have access to formal education.

This paper focuses on text-to-speech (TTS) synthesis technology. We describe a quantitative approach to modelling intonation in the context of a TTS system for the Standard Yorùbá language. The problem of intonation modelling in TTS systems is very important because a poor intonation will result in poor speech prosody. A poor speech prosody makes synthesised speech to sound mechanical and difficult to listen to. The task of producing an accurate intonation model for a language requires multidisciplinary approach which involves experts in diverse areas of study such as linguistics, phonetics, computer engineering and artificial intelligence.

Generally speaking, there are two classes of languages: tone and non-tone languages. In non-tone languages such as the English langauge, the stress pattern on a lexical item, i.e. word, determines its syntactic class and hence its meaning. For example, the English words *re*cord (noun) and re*cord* (verb) differ

in meaning because of the stress pattern on their constituent syllables. In the noun *re*cord the first syllable is stressed whereas in the verb re*cord* the second syllable is stressed. Other examples of this type of word in English include comment, read, commission. Many European languages are non-tone languages.

In tone languages such as SY, the meaning of a lexical item depends on the tones (not stress) associated with each syllable that constitutes the lexical item. For example, the SY mono-syllable words *kí* (to greet), *ki* (to be thick), and *kì* (to praise) differs in meaning because of the tones associated with them. The first word *kí* carries a high tone, the second word *ki* carries a mid tone and the third word *kì* carries a low tone. Many African, (e.g. Yorùbá, Igbo) and Asian (e.g. Mandarin, Cantonese) languages are tone languages.

It is important to note that the acoustic correlate of tone (as found in tone languages) and stress (as found in non-tone languages) are the fundamental frequency (f_0), duration or timing, and intensity or loudness. However, the fundamental frequency is putatively the acoustic correlate of intonation in both types of languages.

Accurate intonation modelling is particularly important in the synthesis of speech for tone languages. This is because the variation in pitch determines not only the intelligibility and naturalness of the speech but also the semantic implication of the utterance. The goal of this research is to build an intonation model for the SY language using the Soft Template Mark-up Language (Stem-ML) [5]. Stem-ML is particularly useful for this task because it can produce accurate models that are simple (in the sense of having very few adjustable parameters) and can be trained with relatively little data. In addition, the Stem-ML model is valuable in both speech synthesis application and linguistic interpretation of the resulting model. The Stem-ML model has been successfully applied to the modelling of tones and intonation in Mandarin [6, 7] and Cantonese [10].

1.1 A brief description of the Standard Yorùbá language

Standard Yorùbá (SY) is one of the three major languages spoken in Nigeria: Hausa and Igbo being the remaining two. SY is the native language of more than 30 million people in West Africa [1]. It is the native language of people in Lagos, Ọ̀yọ́, Ògùn, Oǹdó, Èkìtì, Ọ̀sun and part of Kwara and Kogi states of Nigeria. There are several dialets of the language but Standard Yorùbá (SY) is understood by all speakers. SY is used in language education, the mass media and everyday communication.

The SY alphabet has 25 letters which is made up of 18 consonants (represented by the graphemes: **b, d, f, g, gb, h, j, k, l, m, n, p, r, s, ṣ, t, w, y**) and seven vowels represented by the graphemes: (**a, e, ẹ, i, o, ọ, u**) [29]. Note that the consonant **gb** is a diagraph, i.e. a consonant written in two letters. There are five nasalised vowels in the language (**/an/, /en/, /in/, /ọn/, /un/**) and two pure syllabic nasals (**/m/, /n/**). The consonant and vowel systems as well as a more detailed description of SY are presented in [35].

The SY syllable can be analysed using the ONSET-RHYME theory [31]. The ONSET occupies the first position in the syllable structure. If it is present, it is any of the SY consonants. The RHYME occupies the rest of the syllable and it is either a vowel, a nasalised vowel or a syllabic nasal. With this analysis, we can specify five syllable configurations for SY: CV, CVn, V, Vn, and N (where the N symbol represents the syllabic nasal configuration). Table 1 shows the distribution of the phonological structures of SY syllables. There are fewer syllable classes in SY when compare with Mandarin and Cantonese [4]. There are no closed syllables or consonant clusters in SY.

SY has three phonologically contrastive tones [8]: High (H), Mid (M) and Low (L). Figure 1 shows sample f_0 curves for the three tones on the syllable ẹ. Phonetically, the tone system is normally described with two additional allotones or tone variants: rising (R) and falling (F) [2]. A rising tone occurs when an L tone is followed by an H tone, while a falling tone occurs when an H tone is followed by an L tone. This situation normally occurs during assimilation, elision or deletion of phonological objects as a result of co-articulation during fluent speech. The Stem-ML model we propose here does not utilise the allotones: instead, tone shapes are generated by the processes of tonal co-articulation that is assumed to act everywhere in the continuous speech.

Table 1: Phonological structure of SY syllables.
The number of types of each unit is indicated in parentheses.

Tone syllables (690)				
Base syllables (230)				Tones(3)
Onset(18)	Rhyme(14)			
Consonant	Nucleus		Coda	
	Vocalic	Non-vocalic		
C	V(7)	N(2)	n(1)	H, M, L

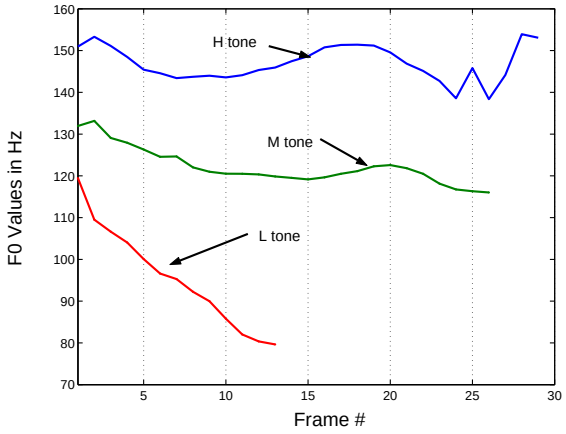

Figure 1: f_0 *curves for H, M & L tone on syllable ẹ, spoken in isolation.*

2 Approaches to Intonation Modelling

The time course of the fundamental frequency, i.e. f_0 curve, during the production of a syllable is central to the perception of tones. In spoken SY, tone is manifested in the f_0 curve over the voiced portion of a syllable. The RHYME is regarded as the voiced portion of the syllable while the ONSET is regarded as the unvoiced portion. The f_0 curves on syllables distinguish otherwise identical lexical items. The f_0 curves on each syllable that constitutes a multi-syllable utterance combine to define the f_0 contour over the utterance. In such utterances, co-articulation causes the f_0 curve of adjacent tones to influence each other resulting in the *tone Sandhi*[32] phenomenon. Other speech phenomena [3], e.g downstep, f_0 reset, etc., culminate on the f_0 curves of each syllable in the process of generating the f_0 contour of the utterance. The focus in intonation modelling is to capture these speech phenomena in a model with the aim of generating a corresponding f_0 contour from input data, e.g. from input text.

The approach to intonation modelling in the context of speech synthesis, can be grouped into two classes: (i) rule-based and (ii) machine-learning [9]. In rule-based intonation modelling [23, 24, 25, 27, 26, 28], a set of rules is designed to transform underlying tones into f_0 curves. The approach often employs rules that convert discrete phonological symbols into phonetic symbols. A final interpolation step is then applied to produce the f_0 contour of the target utterance. If the resulting set of rules is simple, this approach has the advantage that the resulting intonation model can be understood and can be easily adapted to different speakers and situations. However, generating an accurate set of rules is a very difficult task which involves the creation and analysis of a large body of language resource and their interpretation using a multi-disciplinary perspective.

In a machine learning approach, techniques such as Artificial Neural Networks[14, 16, 15, 17], Hidden Markov Models [18, 19, 20, 21] or Decision Trees [22] are used to predict the intonation contour of an utterance. Usually, the predicted f_0 contour is based on information that is extracted from the input data, e.g. annotated texts and labeled speech files. A strength of this approach is its simplicity of design and implementation. However, while this technique can be successful if sufficient data is available, the resulting models are typically opaque, and often cannot usefully be applied outside the specific conditions under which they were trained.

The Stem-ML [5][7], model attempts to integrate the ideas of data-driven and rule-rule driven approaches in intonation modelling. The underlying aim is to combine the strengths of the rule-driven and data-driven approach into one model.

2.1 Overview of the Stem-ML Model

Stem-ML allows the introduction of linguistically motivated rules where appropriate, but it contains a set of learnable parameters that control the phonetic realisation of f_0 contour. These parameters are automatically learnt by fitting the Stem-ML model to a corpus of data. This is achieved by setting soft intonational targets which need not be precisely realised. It chooses the f_0 curve that minimizes the sum of violations of two types of constraints. First, a global constraint that f_0 (as a proxy for the vocal fold tension) should be smooth and continuous everywhere. Second, a local constraint on each syllable that f_0 curve should follow the syllable's template.

Stem-ML brings several ideas into intonation modelling. First, in its design, it is assumed that people plan their utterances several syllables in advance. In addition it is assumed that people produce speech that is optimized to meet specific communication needs. Based on these assumptions, a physically reasonable model for the dynamics of the muscles that control pitch is derived. A linguistically reasonable concept of a strength that is associated with each lexical item is introduced.

Each syllable has a strength parameter, which controls the relative importance of each constraint. We interpret this parameter as the prosodic strength of the syllable. In our model, the Stem-ML templates represent lexical tones. The actual realisation of the f_0 curve on a syllable depends on the template, the neighbouring tones and their prosodic strengths.

Despite the introduction of linguistically motivated concept, the Stem-ML model is not based on any lin-

guistically motivated theory [5]. This is because it allows a description of any physiologically realizable prosody in terms of linguistic concepts, without imposing a restrictive theory on the data. For example, Stem-ML allows, but does not require, descriptions involving phrase curves. The tags can contain adjustable parameters that can be used to explain phonological phenomena observed in the target language. It can also be valuable linguistically, as many of the model parameters can be directly used to answer linguistic questions.

3 Design of the Stem-ML model for SY

For intonation modelling using the Stem-ML, a fundamental assumption is that the surface forms of the f_0 curves of underlying tones vary extensively and that these variations are related to articulatory constraints [6]. In our Stem-ML model, an intonation contour is calculated from a set of tags. Some of the tags set global parameters (such as f_0 range) which are properties of the speaker. Other tags represent intonational events (such as f_0 reset). In order to adapt the Stem-ML model to the modelling of SY intonation a number of assumptions specific to SY is required. The assumptions in our model, beyond those that are generic to Stem-ML (see [5]), are that:

- Each syllable carries a soft intonation target with one of the three shapes, chosen by the lexical tone (i.e. High, Mid, Low). Each target is a line segment.
- The prosodic strength of a syllable both affects the precision with which a tone is realised and scales the f_0 range of that tone's template. One can expect that a linguistically stronger syllable will have both a larger f_0 range and also be articulated more carefully. We include an adjustable parameter in the model (*atype*) to account for such a correlation.
- Minimal syntactic information is required to model the intonation of SY. Our model includes five syntactic classes related to the position of syllables. These are: phrase initial, phrase final, sentence initial, sentence final, and medial (i.e. all linguistic elements that are not initial or final).
- Syllables in one- and two-syllable words have the same phonetic realisations, and we assume that there are no intrinsic differences between the first and second syllables in a two-syllable word.
- Segmental effects that depend on the phoneme sequence are relatively small.
- Intonation contour can be generated without the explicit use of phonological rules.

3.1 Data

We collected a text corpus of 100 sentences from SY newspaper and language education textbook domains [33]. Thirty of the sentences were selected after an analysis of the corpus. Two criteria were used for the selection of each sentence: (i) the sentence must be a statement sentence, (ii) the sentence must contain words in common, everyday use of SY, (iii) the sentence must not contain words written in foreign autography. In the selected sentences, the minimum number of syllables per sentence is 4 with a mean of 6.7 and maximum of 15 syllables. There are 198 syllables in all. The H and L tone syllables accounting for 40% each while the M tone syllables accounts for the remaining 20%. Nineteen out of the sentences collected are one-phrase sentences while the rest eleven are two-phrase sentences.

An adult male native speaker of SY read the sentences at an average rate of 5.6 syllables per second. Phrase boundaries, separated by commas in the text, are marked by a pauses during the recoding of the speech. We recorded three[1] productions each of the 30 selected sentences. The first production was to ascertain that the sentence could be read smoothly, with either no pause or a pause in the correct location, as appropriate. These decisions were made without access to f_0 information. The second production was used if acceptable, otherwise the third. If the third production is unacceptable, the sentence is dropped.

Five of the sentences were rejected based on the above criteria. Hence we used 25 of the 30 sentences for our model development and analysis. The sentences are recorded and annotated using the *Praat* [34] speech processing software. The annotated files for a one-phrase sentence (*"Bàbá àgbẹ̀ ti ta kòkó."*) and a two-phrase sentence ("Ópẹ́ kí ó tó dé, kò tètè lọ".) are shown in Figure 2 and 3 respectively.

In Figure 2, there are five panels. The topmost panel depicts the f_0 contour of the sentence. The panel below the topmost panel depicts the waveform of the speech sound. The next two panels depicts the syllable and word annotations respectively while the last panel depicts the sentence annotation. In Figure 3 the additional panel between the word and sentence panel is the phrase panel which depicts the annotations for the phrases in a sentence.

For certain types of syllables, we found that boundaries between syllables were hard to determine. This

[1] Some sentences were produced more times while the recording level was being adjusted.

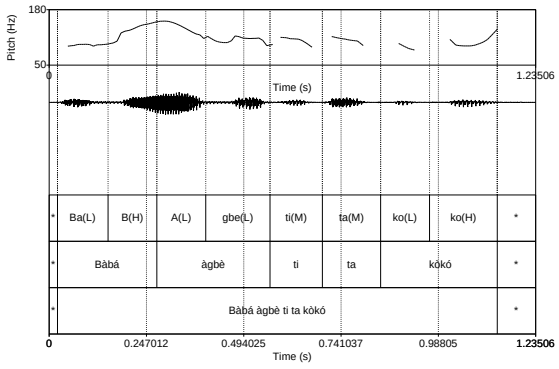

Figure 2: Annotate file for the one-phrase SY sentence *"Bàbá àgbẹ̀ ti ta kòkó."*

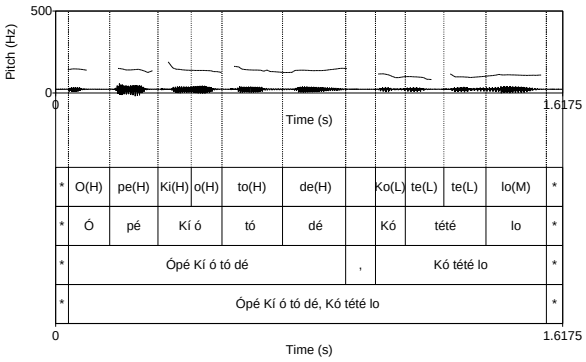

Figure 3: Annotate file for the SY sentence *"Ópẹ́ kí ó tó dé, kò tètè lọ́"*.

occurred between V, V pairs or between V, CV pairs where C is a semi-vowel such as /y/ or /w/. In this situation we employed listening tests in addition to speech spectrograph and waveform characteristics for syllable boundary detection. Where the boundaries were in doubt, we found the earliest reasonable position and the latest reasonable position, then placed the boundary half-way in between. For voiced plosives, e.g. /p/, /t/ and /k/, we placed the syllable boundary in the centre of the closure. We note that the voiced plosives show strong segmental effects on the f_0 curves of the syllable in which they occur.

4 Tag description

The Stem-ML tags used in our model are briefly described in this section. The data obtained for some of the parameters represented by the tags are documented in Table 2.

Table 2: Parameter used in our model

Ser. No.	Parameter	Values
1.	smooth	0.062 ± 0.006
2.	base	$105 \pm 0.5\ Hz$
3.	atype	0.41 ± 0.05

4.0.1 smooth tag

The smooth tag specifies the parameter that controls the rate at which the f_0 curve of the speech produced by the speaker changes when considering aweak tone. In our model, the smooth value is 0.062 ± 0.006. This value corresponds to a time of 106 ± 8 milliseconds. This is roughly double the value obtained for a Mandarin speaker [5].

4.0.2 base

The base tag specifies the parameter that represents the speaker's base frequency. To reduce the number of parameters in our model, we calculated the base outside of Stem-ML. The base is estimated as the 25^{th} percentile of the f_0 values in each speech file. The computed value used in our model is $105 \pm 0.5Hz$.

4.0.3 atype

The atype tag specifies the parameter which describes how much the f_0 range of a template expands as the strength changes. In our model, the parameter has value in the range 0.41 ± 0.05 which indicates a fairly weak (but significant at $P < 0.01$) effect. For example, a 40% change in strength (e.g. changing from a sentence-initial to a medial syllable) will make the template of the stronger syllable to have an f_0 range just 14% larger than a comparable medial syllable. The value we obtained is about half of the 0.87 ± 0.7 value from [6].

4.0.4 ctrshift and wscale

These two tags specifies parameters that describe the scope of the template relative to the syllable boundaries. In our model, the length of the target is $85.0 \pm 2.0\%$ of the length of a syllable, similar to the $88.0 \pm 1.0\%$ for Mandarin. The target is nearly centered in the syllable, $1.0 \pm 1.0\%$ of its width (i.e. about 2 ms) before the syllable centre. This implies a nearly equal balance between anticipatory and carry-over co-articulation.

4.0.5 Intrinsic Strengths of Tones

The tone tags specifies the local f_0 curve corresponding to a syllable [5]. Each tone tag is defined by a number of parameters. The most important in our modelling is

Table 3: Parameter values for Tones

Tone	Relation to base	*atype*
H tone	8% above base to 112% above base	0.72 ± 0.06
M tone	28% above base to 6% above base	0.59 ± 0.06
L tone	15% below base to 68% below base	0.148 ± 0.04

the strength. The *strength* tag controls the interaction of tone with their neighbours [5]. The strength of each syllable is given by:

$$S_i = A[\tau(i)] \cdot C[P(i)], \quad (1)$$

where $\tau(i)$ returns the tone of the i^{th} syllable, $A[\tau]$ is the intrinsic strength of a syllable of tone τ, $P(i)$ returns the position in the sentence (e.g. sentence final, medial,...), and $C[P]$ is the strength factor for position P. For the intrinsic strength of tones, we obtain $A[H] = 1.8 \pm 0.12$, $A[M] = 2.5 \pm 0.15$, and $A[L] = 2.2 \pm 0.08$. From this values it can be deduced that the H tone is the weakest and the M tone is the strongest. This implies that there is a tendency for the shape of a H tone to be influenced by its environment than for the other tones. However, although the differences are significant ($P < 0.01$), they are not dramatic.

4.0.6 H tone

For the H tone f_0 model, the template slopes up from 8% above *base* to 112% above *base*. The target shape is both high and rising. The tone's *atype*=0.72 ± 0.06. This implies that both the height and the shape of the f_0 curve are important, but errors in the tone's average f_0 value are more important than errors in the tone's shape.

4.0.7 M tone

For the M tone, the template slopes down from 28% above *base* to 6% above *base*. This is a mid-level and weakly falling tone. The *atype*$= 0.59 \pm 0.06$ parameter indicates that both the f_0 curve height and slope are important.

4.0.8 L tone

For the L tone, the template slopes down from 15% below *base* to 68% below *base*. The *atype* for the L tone is 0.148 ± 0.04, indicating that the shape of the f_0 curve is more important than the average value. It might be best to describe this tone as *falling* with a tendency toward low.

4.0.9 Phrase Boundaries

The strength factors for syllables in phrase-initial and phrase-final positions (but not at the beginning or the end of a sentence) are $C[\text{SI}] = 1.09 \pm 0.03$ and $C[\text{SF}] = 0.73 \pm 0.06$ respectively. Again, with $P < 0.01$, phrase-initial syllables are stronger than medial syllables (though just slightly), and phrase-final syllables are also weaker than medial syllables.

4.0.10 Sentence Boundaries

The strength factors for syllables in sentence-initial and sentence-final positions (Equation 1) are $C[\text{SI}] = 1.42 \pm 0.06$ and $C[\text{SF}] = 0.70 \pm 0.03$ respectively. Thus, with $P < 0.01$, sentence-initial syllables are stronger (i.e. they are articulated more precisely and have wider f_0 swings) than medial syllables, and sentence-final syllables are weaker than medial syllables.

The results obtained for sentence- and phrase-boundaries are similar to those reported by Kochanski and Shih [5] for Mandarin and by Lee *et al.* [10] for Cantonese. We speculate that differences in strength may be cues that help the listener to find phrase and sentence boundaries.

The overall model uses 22 parameters: eleven to specify the templates, seven to specify the strengths, and four global, speaker-specific parameters. We kept the model simple to allow it to be trained on a small data set as this is a preliminary study.

5 Model implementation

There are three sets of input files into the Stem-ML program. The first set of files are text files **(*.f0)** which contain the f_0 data for the speech sound corresponding to the syllables in a sentence. This file was generated using the *Extract Pitch Tier* module of *Praat*. The second set of files are the label files **(*.lab)** which contains each syllable, associated tone symbol and the duration range. This data was generated from the *Praat* **.TextGrid** files for each of the speech file.

The third file, tag parameter generation output file, *(Output.out)* contains the initial values for the parameters described in Section 4 for each of the sentences in our speech database. The content of this file is generated using a program coded in C.

In order to develop our Stem-ML model, we used 18 sentences (10 one-phrase and 8 two-phrase) for our training set. The test set comprises 7 sentences (4 one-phrase and 3 two-phrase). The model was fitted to the corpus under the assumption that the f_0 data had independent Gaussian errors, using a *Bayesian Markov Chain Monte Carlo* algorithm that produced samples from the posterior distribution of the parameters. Once

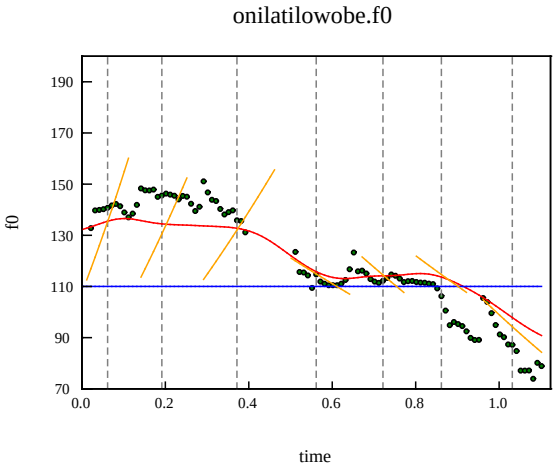

Figure 4: Model fit and raw data for SY sentence *"Óní láti lọ wobẹ̀"*.

the algorithm had converged to a stationary distribution, we collected the last 6000 samples. From that, we computed the average values of all the parameters and their uncertainties.

The Stem-ML implementation is quantised in $10ms$ increments, which raises the possibility of spurious local minima. To check for this, we started ten more *Monte Carlo* runs with different fixed values of the *centershift* parameter. We chose 10 random samples of parameters from the second thousand iterations from each run and computed the Root Mean Square Error (RMSE). The resulting set of samples traces out a minimum of RMSE against a *centershift* that is consistent with the error reported in the following subsection.

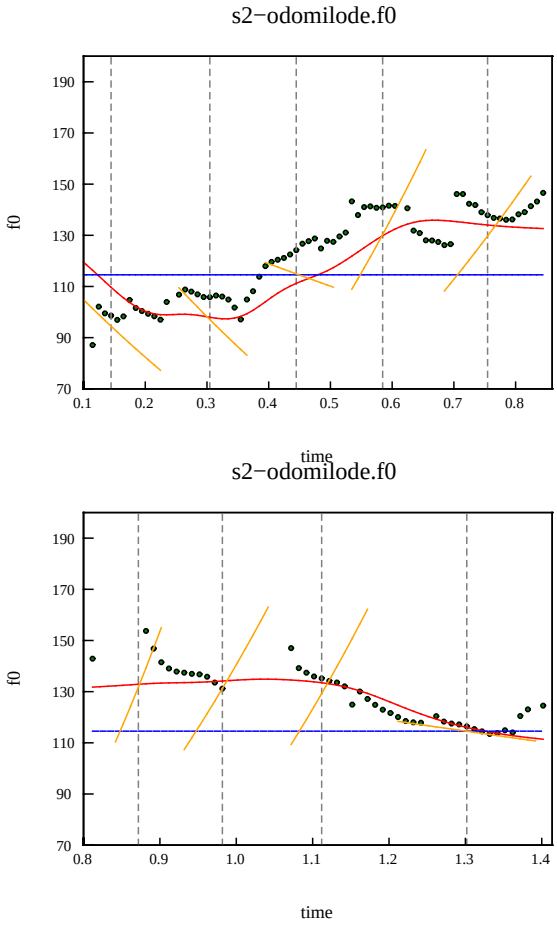

Figure 5: Stem-ML prediction of f_0 and raw data for the two phrases of SY sentence *"Ọ̀dọ̀mi lódé, Kó tó lọ."*. This data is in the test set; the model prediction is based on parameters derived from the training set, using syllable boundaries for this specific utterance.

6 Evaluation of model fit

The result of applying the Stem-ML model to SY prosody modelling on the sample sentences: *"Óní láti lọ wobẹ̀"* and *"Ọ̀dọ̀mi lódé, Kó tó lọ."* are shown in Figures 4 and 5 respectively. These figures show fits of our model to the f_0 contours of typical one-phrase (Figure 4) and two-phrase (Figure 5) sentences.

In these figures, the black dots mark measured f_0, the grey curve is the predicted f_0, the grey lines show the Stem-ML templates, and the vertical dashed lines show the syllable centers. Our model fits the test set with a RMSE (Root Mean Square Error) of 14 Hz; the fit to the training set has a RMSE of 12 Hz. Much of this error can probably be accounted for by segmental effects [11, 12] due to changes in the vowel and the syllable onset consonant.

Generally speaking, the worst-fitting syllable are those with the largest and fastest f_0 excursions. These are conditions where Stem-ML's approximations between templates and the realised pitch curve may be furthest from the actual perceptual metric, but segmental effects or un-modelled differences in the strength of syllables may also play a role. The results of the fits are generally similar to other Stem-ML based intonation models.

7 Conclusions

We have presented a Stem-ML intonation model for the Standard Yorùbá (SY) language. The model attains a fitting accuracy of 14 Hz on the testing set, using only 21 adjustable parameters and a very small training corpus. The model shows that the H tone is high and rising while the M tone is mid-level and weakly falling. The L tone is falling with tendency toward low. We also found that the M tone has the highest strength followed by the L tone and that the H tone is the weakest. Some of the properties of SY seem similar to Mandarin and Cantonese, for instance the strength of syllables at the

beginnings of sentences and phrases is enhanced, and syllables at the end of sentences and phrases are especially weak.

On-going work is directed towards expanding the speech corpus and building a more robust prosody model for SY based on an improved Stem-ML model. Since the domain of application of our prosody model is text-to-speech synthesis, we hope to implement a TTS system based on the obtained results.

8 Acknowledgements

I thank the Association of Commonwealth Universities in the United Kingdom for providing the funds for the research reported in this paper. I also thank Dr. Greg Kochanski and Dr. John Coleman of the Oxford University Phonetics Laboratory for their assistance during the research reported in this paper. I also that the Oxford University Phonetics Laboratory for allowing me to use their research facility.

References

[1] D. H. Crozier and R. M. Blench, "An Index of Nigerian Languages", Summer Institute of Linguistics, Dallas, USA, 2^{nd} Ed., 1976,

[2] B. Connell and D. R. Ladd, "Aspect of pitch realisation in Yorùbá", Phonology, 7:1–29, 1990.

[3] Y. O. Laniran and G. N. Clements, "Downstep and high rising: interacting factors in Yorùbá tone production", J. of Phonetics, 331(2):203–250,2003.

[4] H-M. Wang, T-H. Ho, R-C. Yang, J-L. Shen, B-O. Bai, J-C. Hong,W-P. Chen, T-L. Yu and L-S. Lee,Complete recognition of continuous Mandarin speech for Chinese languages with Very large vocabulary using limited training data, IEEE Trans. on Speech & Audio Processing, 5 2:195–200, 1997.

[5] G. Kochanski and C. Shih, "Prosody modelling with soft templates", Speech Comm, 39:311–352, 2003.

[13] G. Kochanski and C. Shih, "Stem-ML: language-independent prosody description", In ICSLP-2000, vol.3, 239-242.

[6] G. Kochanski, C. Shih , and H. Jing, "Quantitative measurement of prosody strength in Mandarin," Speech Comm., 41:625–645, 2003.

[7] G. Kochanski, C. Shih, and H. Jing, "Hierarchical Structure and Word Strength prediction of Mandarin," Int. J. of Speech Tech., 6:33–43, 2003.

[8] O. A. Ọdẹ́jọbí, A. J. Beaumont and S. H. S. Wong, "Experiments on stylisation of standard Yorùbá language tones," Tech. Report:KEG/2004/003, Aston University, Birmingham, U.K., 2004.

[9] A. Sakurai, K. Hirose and N. Minematsu, "Data-driven generation of f_0 contours using a superpositional model,"Speech Comm., 40:535–549, 2003.

[10] T. Lee, G. Kochanski, C. Shih and Y. Li, "Modeling Tones in Continuous Cantonese Speech," in Proc. of Int. Conf. on Spoken Language Processing, Denver, Colorado, Sept., 2002.

[11] K. E. A. Silverman, *The Structure and Processing of F_0 contours.*, Ph.D. Thesis, Cambridge University, 1987.

[12] K. Dusterhoff, Synthesizing Fundamental Frequency Using Models Automatically Trained from Data, Ph.D. Thesis, University of Edinburgh, 2000. *http://www.cstr.inf.ed.ac.uk/downloads/publications/2000/Dusterhoff/_2000_a.pdf*

[13] A. Botinis, B. Granström and B. Möbius, "Development and paradigms in intonation research", Speech Comm., Vol. 33: 263–296, 2001.

[14] M. Vainio, Artificial neural network based prosody models for Finnish text-to-speech synthesis, Ph.D. thesis, Department of Phonetics, University of Helsinki, 2001, Helsinki

[15] J. W. A. Fackrell, H. Vereecken, J. P. Martens, and B. V. Coile, Multilingual prosody modelling using cascades of regression trees and neural networks, EuroSpeech '99, 1999, Visited: Sep 2004,*http://chardonnay.elis.rug.ac.be/papers/1999_0001.pdf*

[16] T-L. Burrows,Trainable Speech Synthesis, Speech Processing with linear and neural network models, 1996, Cambridge, Mar

[17] P. Taylor, Using neural networks to locate pitch accents,Proceedings of EuroSpeech '95, 1995, 1345–1348, Madrid, Sep

[18] A. Ljolje and F. Fallside, Synthesis of natural sounding pitch contour in isolated utterances using Hidden Markov Models, IEEESpeech, 1986, ASSP-34, 1074–1080.

[19] R. E. Donovan and P. C. Woodland, Improvements in an HMM-based speech synthesiser, Proc. of EuroSpeech Conference, 1995, 573–576, Madrid

[20] M. Plumpe, A. Acero, H. Hon and X. Huang, HMM-based smoothing for concatenative speech synthesis,URL:*http://research.microsoft.com/srg/papers/1998-plumpe-icslp.pdf*,1998,Visited:Aug 2004

[21] K. Tokuda and T. Yoshimura and T. Masuko and T. Kobayashi and T. Kitamura, Speech parameter generation algorithms for HMM-based speech synthesis,Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000, 3, 1315–1318, Istanbul, Jun

[22] R. E. Donovan, Topics in decision tree based speech synthesis, Computer Speech and Language, 2003, 17, 43–67

[23] N-C. Chan and C. Chan, Prosodic rules for connected Mandarin synthesis, Journal of Information Science & Engineering, 1992, Vol. 8, No. 2,261–281

[24] S. Raptis and G. V. Carayannis, Fuzzy logic for rule-based formant speech synthesis, EuroSpeech '97, 1997, 1599–1602

[25] J. S. Coleman, "Synthesis-by-rule" without segments or rewrite-rules, Talking Machines: Theories, Models and Designs, Elsevier, 1992, G. Bailly, C. Benoit and T. R. Sawallis, 43–60, Amsterdam

[27] S. D. Young and F. Fallside, Synthesis-by-rule of prosodic features in word concatenation systems, Intl. Journal of Man-Machine Studies, 1980, 12, 241–258

[26] J. t'Hart and A. Cohen, Intonation by rule: a perceptual quest, J. of. Phonetics, 1972, 1, 309–327

[28] D. R. Ladd, A model of intonation phonology for use in speech synthesis by rule, European Conference on Speech Technology (ESCA),1987,21–24

[29] L. O. Adéwọlé, The categorical status and the function of the Yorùbá auxiliary verb with some structural analysis in GPSG, Ph. D thesis, University of Edinburgh, 1988, Edinburgh

[30] K. Owólabí, Ìjìnlẹ̀ Ìtúpalẹ̀ èdè Yorùbá:Fònẹ̀tîkì àti Fonọ́lọ́jì, Onibonoje Press & Book Industries (Nig.) Ltd., 1998, 1, Ìbàdàn,1^{th}

[31] J. Goldsmith, Autosegmental and metrical phonology, Blackwell, 1990, Oxford

[32] C. Shih and R. Sproat, Issues in text-to-speech conversion for Mandarin, Computational Linguistics and Chinese Language Processing,1996, 1, 1, 37–86

[33] O. A. Ọdẹ́jọbí, A Computational model of prosody for Yorùbá text-to-speech synthesis, Ph.D thesis, Aston University, 2005, United Kingdom

[34] P. Boersma and D. Weenink,*Praat*, doing phonetics by computer, http://www.fon.hum.uva.nl/praat/,Mar,2004,Visited: Mar 2004

[35] O. A. Ọdẹ́jọbí, A. J. Beaumont and S. H. S. WongIntonation contour realisation for Standard Yorùb'a text-to-speech synthesis: A fuzzy computational approach, Computer Speech and Language, 2006, 20, 563-588