TONE REALISATION IN A YORÙBÁ SPEECH RECOGNITION CORPUS

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

[1,2] Multilingual Speech Technologies, North-West University, Vanderbijlpark, South Africa.
[1] Human Language Technologies Research Group, Meraka Institute, CSIR, Pretoria, South Africa.
dvniekerk@csir.co.za, etienne.barnard@nwu.ac.za

ABSTRACT

We investigate the acoustic realisation of tone in short continuous utterances in Yorùbá. Fundamental frequency (F0) contours are extracted for automatically aligned syllables from a speech corpus of 33 speakers collected for speech recognition development. Extracted contours are processed and analysed statistically to describe acoustic properties in different tonal contexts. We demonstrate how features useful for tone recognition or synthesis can be successfully extracted from a corpus of this nature and confirm some previously described phenomena in this context.

Index Terms— Yorùbá, tone language, fundamental frequency

1. INTRODUCTION

Many African tone languages of which Yorùbá is a well known example from the Niger-Congo family, distinguish words based on two or three distinct level tones (register tones) realised on each syllable. In a three-tone system, as in the case of Yorùbá, these tones are labelled high (H), mid (M) and low (L) relying on changes in pitch between consecutive syllables. Such a system stands in contrast to contour tone systems (for example in Chinese languages) where tones are identified by changes in pitch within a syllable.

Given the significance of linguistic tone in the interpretation of semantic information (amongst others), it is important for the development of speech technology in these languages to understand the tone system in detail [1]. Developing systems such as text-to-speech (TTS) and automatic speech recognition (ASR) requires knowledge in two areas, namely (1) deriving surface tone assignments from text, i.e. tone assignments of syllables in target context after linguistic processes (e.g. sandhi) have been applied and (2) understanding the relationship between acoustic parameters (such as pitch) and these surface tones.

Yorùbá is a relatively well studied African tone language of which the linguistic details of the tone system have been thoroughly described. The result is that tones are marked explicitly on the orthography, making the automatic derivation of surface tone from text relatively simple (compared to other African tone languages [1]). Consequently, in this work we focus on the acoustic realisation of tones (specified in this fashion) in continuous utterances.

In the following section we will briefly present related work, and state the approach and contribution of the current study. In Section 3 we summarise relevant details of the Yorùbá tone system and describe the speech corpus and methodology used, followed by results in Section 4. We conclude in Section 5 with an outline for future work.

2. RELATED WORK AND CURRENT APPROACH

Recent work on the realisation of tone in Yorùbá for the development of a speech technologies has been described by Ọdẹ́jọbí et al. Their work involved the development of prosodic models (using Stem-ML [2] and a novel approach) for the synthesis of F0 contours suitable for use in a speech synthesiser [3][4][5]. The work of Ọdẹ́jọbí et al draws primarily on previous acoustic analyses by Connell and Ladd [6] and later by Laniran and Clements [7] which aimed to systematically investigate linguistic concepts such as *downstep* and *high tone raising* alongside effects such as *declination*. These studies relied on relatively small samples (3 to 4 speakers) based on carefully designed corpora in order to investigate these specific properties of tone realisation.

In the current work we attempt a general description of tone realisation based on statistical analysis of a (multi-speaker) speech recognition corpus, relying on automatic alignments and feature extraction methods that are expected to be applicable in eventual systems. In doing this, we investigate relevant methods for manipulation of F0 contours and highlight aspects of tone realisation such as speaker-specific variation and co-articulation in natural continuous utterances that need to be addressed for the development of speech systems.

3. EXPERIMENTAL SETUP

3.1. Language details

Literary or Standard Yorùbá has a fairly regular orthography with graphemes generally corresponding directly to underly-

ing phonemes with the inclusion of a few simple digraphs (such as gb and certain vowels followed by an n). The syllable structure is relatively simple, with all syllables being open or consisting of syllabic nasals with no consonant clusters; thus any of consonant-vowel (CV), vowel only (V) and syllabic nasal (N). A more detailed description of the relevant language details can be found in Section 2 of [3]. The Yorùbá tone system is based on 3 tonemes (H, M and L), with rising and falling tones considered to be phonetic variations of H and L in certain contexts respectively [6]. These tones are marked in the standard orthography using diacritics on vowels and nasals, with the acute accent (e.g. ń), grave accent (e.g ǹ) and unmarked letters representing H, L and M respectively (in the case of M-toned nasals the macron (e.g. n̄) is used).

These properties of the language were used in the automatic alignment and tone assignment process described in the following section.

3.2. Corpus alignment and F0 estimation

The speech corpus used in this study consisted of a subset of 33 speakers from an ASR corpus currently under development at the University of Nairobi, Nigeria and North-West University, South Africa. Each speaker recorded between 115 and 145 short utterances from the pool of selected sentences, amounting to about 5 minutes of audio per speaker. Audio is broadband, collected in Nigeria using a microphone attached to a laptop computer. In some cases significant amounts of background noise is present; data from one speaker was omitted because of the presence of power line noise which greatly affects F0 estimation.

For this analysis a set of basic hand-written rewrite rules were used for grapheme to phoneme conversion based on a description of the Standard Yorùbá orthography. In addition, a simple syllabification algorithm was implemented based on the description presented in the previous section. Syllable tones were obtained from the orthography (diacritics). Using this, we performed automatic phonemic alignment of the audio by forced alignment of Hidden Markov Models (HMMs) as described in [8] on each speaker's audio separately.

Lastly, in an attempt to discard samples where automatic alignment might have failed (due to for example mismatches between transcriptions and audio), we only retained utterances where all syllables have durations of more than 30ms. This might have the additional side-effect of discarding utterances which tend to be relatively fast, this was deemed an acceptable compromise for the current study.

The resulting usable corpus amounted to 33 speakers, each having between 82 and 127 single-phrase utterances. Utterance lengths ranged from 2 words (4 syllables) to 10 words (28 syllables) with an average length of 5 words (10 syllables). The total number of syllables amounted to 34570 (H: 12777, M: 10743, L: 11050).

To extract F0 contours, we used *Praat* [9], specifically the autocorrelation method, and applied a small amount of smoothing to reduce measurement/estimation noise (in this study we are not considering the finer movements in F0 due to the segmental make-up of syllables, i.e. microprosody). Pitch ranges were determined for each speaker manually, by plotting histograms of F0 samples extracted using the range 60 to 600 Hz and subsequently resetting and re-extracting contours for a narrower, more suitable, range. For each utterance contour extracted, we use cubic spline interpolation to obtain non-zero F0 values for unvoiced regions (see Figure 1).

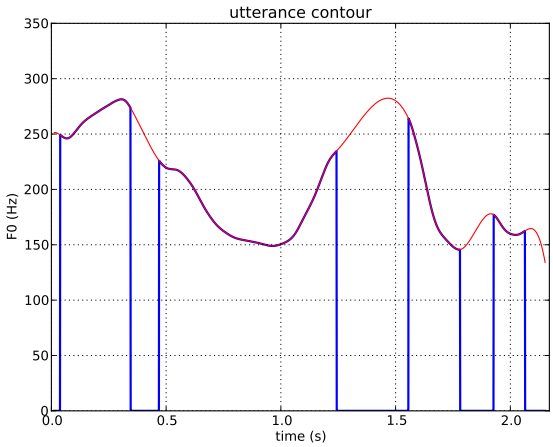

Fig. 1. Example of spline interpolation for utterance contours, the originally estimated contour is in *blue* with the interpolated contour in *red*.

For an indication of the reliability of this process we randomly selected a short sample; one utterance from each speaker, manually determining and counting the number of gross errors in alignment and F0 extraction. Syllables were inspected in Praat (using spectrograms and F0), counting gross errors when a significant part of the syllable is misrepresented (approximately 50% or more). This is reported in Table 1.

Alignment error rate	5%
F0 error rate	8%
Number of syllables	355

Table 1. Gross error rates observed in a small subset of the corpus.

In the following section we describe the details of further investigations along with results.

4. EXPERIMENTAL RESULTS

We start our investigation to see if phenomena similar to those that were observed for Sepedi (a Southern Bantu language

with a 2-tone system) hold for a 3-tone system [1]. Thus, whether it is the general trend that H $>$ M $>$ L with regard to mean absolute pitch values of syllables and whether mean pitch change is a good indicator of tone.

Calculating the mean pitch for syllables based on interpolated contours (as in the red curve in Figure 1) results in 31 of the 33 speakers indeed showing this trend, with typical distributions as shown in Figure 2 (using speaker `08` as an example). Most speakers seem to have very similar distributions where H tones show a higher degree of variance (in some cases more skewed towards the higher parts of the speaker's pitch range) and L tones generally concentrated in the lower part of a speaker's pitch range. In the case of most speakers the L and M means are relatively close while H is somewhat higher.

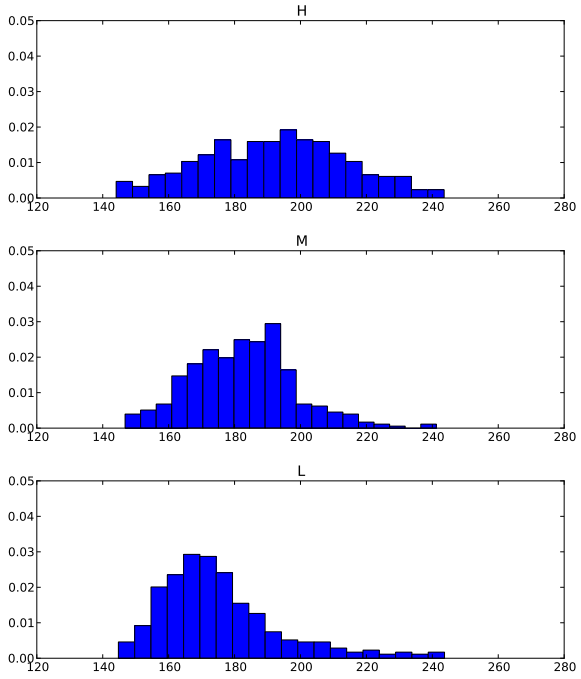

Fig. 2. Example of mean F0 distributions for syllables of different tones by a female speaker (x is the mean pitch in Hertz and y the fraction of samples).

In [1], overlap between mean absolute pitch values was to some degree explained by the general trend of declining pitch within an utterance and the difference in mean F0 between syllables was thus found to be a good indicator of tone. The general trend of declining pitch (referred to as *downtrend*) is also found in our corpus: we performed a least-squares linear fit for each utterance with the result that 2949 of the 3435 utterances have a declining trend with a mean gradient of about -12 Hz/s over the complete corpus. Obtaining a robust utterance-wide estimate for expected pitch level due to *downtrend* is however reliant on understanding phenomena such as *downstep* and *pitch reset*, which should be investigated in future work (see Section 5).

In Figure 3 we thus plot the distributions of change in mean F0 for different transitions. Histograms obtained for single speakers are generally similar in nature to the corpus-wide distribution. Interesting features of these distributions are that MM transitions are relatively flat, and that the contrast between M tones and other tones in terms of mean change is more consistent than HL and LH transitions.

Connell and Ladd [6] specifically mention HL and LH transitions as cases where the L and H tones are realised by falling and rising tones. In Table 2 we present the mean syllable gradients for different tones (and also with different preceding tone contexts), confirming that HL and LH cases exhibit the steepest gradients.

Tone and context	F0 gradient (Hz/s)
H	62.2
H-H	22.1
M-H	77.7
L-H	80.3
M	-21.3
H-M	-72.7
M-M	-12.1
L-M	20.3
L	-96.1
H-L	-154.7
M-L	-93.8
L-L	-53.5

Table 2. Mean F0 gradient within a syllable for different tones, including different preceding tone contexts.

As it seems clear that the perception of tone is in some instances reliant on the movement of F0 within the syllable depending on tonal context, we attempt to investigate the general properties of contours of tri-tones by calculating mean contours over three syllables sequences in the corpus as follows:

1. We extract all contours of three-syllable sequences from the corpus, labelling each according to tone sequence (e.g. HLH) to form a set of K contours for each context.

2. Each contour is resampled using cubic-spline interpolation to normalise lengths to N samples.

3. A *mean contour* is calculated from each set of contours (each context) as in (1), where p is the pitch value.

$$\bar{p_j} = \frac{1}{K}\sum_{i=1}^{K} p_{ij}, \text{where } 1 \leq j \leq N \qquad (1)$$

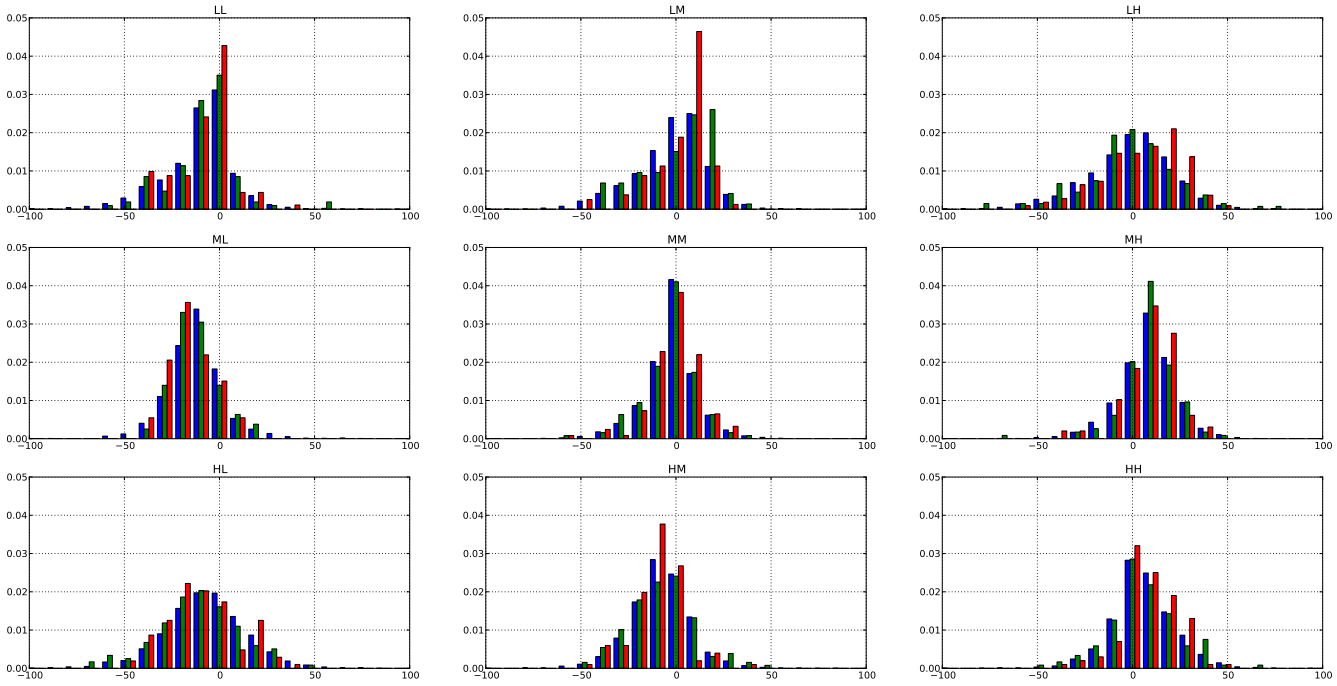

Fig. 3. Distributions of change in mean F0 for different tone transitions; blue bars are calculated over the entire corpus, while green and red bars are examples of a female and male speaker respectively (x is the change in mean pitch in Hertz and y the fraction of samples).

From plots of these mean contours (Figure 4) we observe a few general properties of tone realisation of the three tones in different contexts. We note the following:

- L tones generally have a negative F0 gradient regardless of different contexts, with mean slightly higher when preceded by a H tone, in which case it generally has the steepest gradient. This is in line with the falling tone described in [6].
- M tones generally have a relatively flat to slightly declining contour, with notable exception when preceded by a H tone. Inspection of the MMM sequence seems to confirm the slight *backdrop declination* found by Connell and Ladd [6].
- H tones generally have steep positive gradients when preceded by L tones and peaks seem to be relatively raised when followed by L tones (rising tone [6]). Peaks are generally realised late in the syllable (possibly even in the following syllable), especially in the case of LHL sequences. It also seems to be the case that the final H in a HLH sequence is not generally raised to the same level as the initial H tone.
- Comparing the beginnings of these mean contours with their endings, we see indications that the tone of the current syllable has a greater influence on F0 pattern in the following syllable than in the preceding one. This is similar to the finding by Xu [10] that *carryover assimilation* is more significant than *anticipatory* effects in Mandarin tone realisation.

We performed two further investigations; plotting mean contours as in Figure 4, firstly for each individual speaker and secondly on the complete corpus after separating contours from different parts in an utterance (i.e. initial, medial or final). Mean contours determined for individual speakers generally followed the same patterns seen in Figure 4 (see Figure 5 for two examples; one female and one male speaker) with some variation in the location of peaks, presumably due to the effect of speech rate variation. Mean contours originating from different sections of an utterances seemed to vary primarily with regards to level, with little if any variation in shape or range of pitch change.

5. CONCLUSION AND FUTURE WORK

In this work we have presented some general properties of the acoustic realisation of tone in Yorùbá based on the analysis of a multi-speaker speech recognition corpus using automatically extracted phonemic alignments and F0 extraction. This demonstrates that certain phenomena described in the published literature can be observed through the analysis procedure presented here and that these phenomena seem to hold

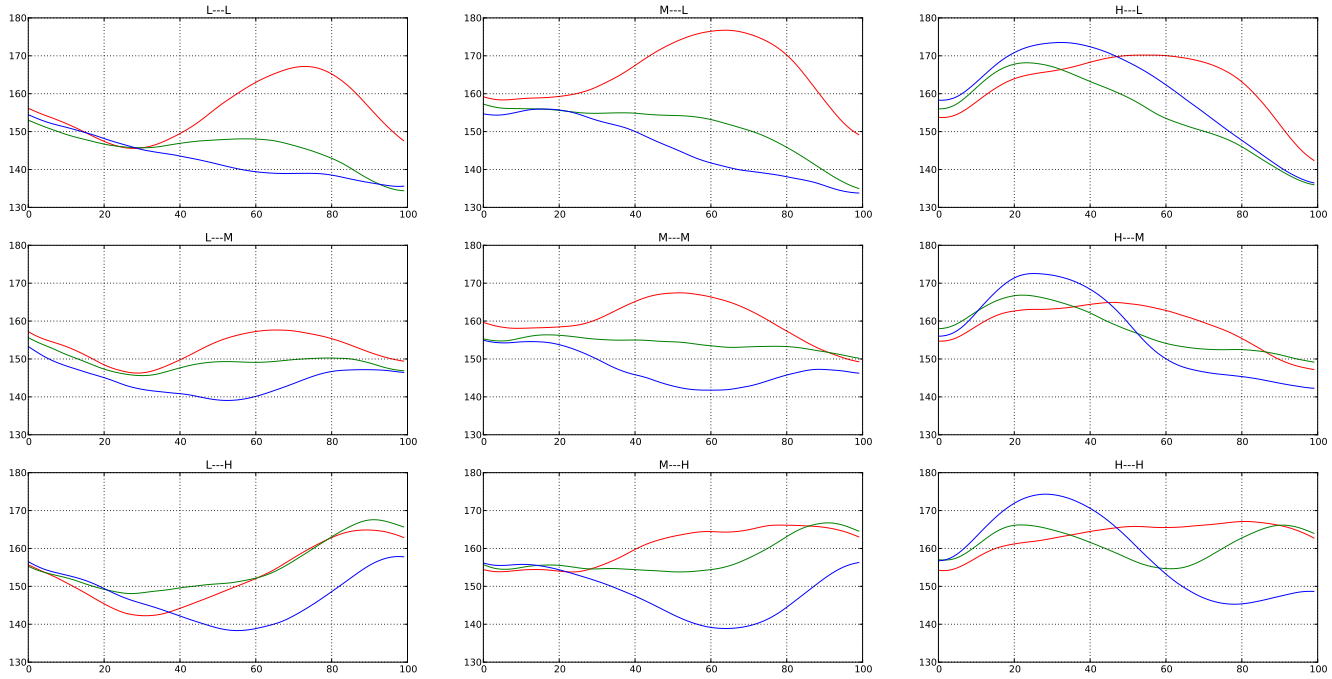

Fig. 4. Mean contours for three-syllable sequences with the different tones H (red), M (green) and L (blue) in different tonal contexts (x is the normalised time and y the pitch in Hertz).

in the case where a larger sample of speakers are considered. Finally we have highlighted features and presented "canonical contours" of local tone realisation in terms of F0 that may be used in tone recognition and synthesis algorithms.

Further work is needed to understand the influence of speech rate variation on these "canonical contours" in particular as preliminary analysis of the distribution of peak positions within the sets of contours indicate significant variation (which might also be in part explained by fine alignment errors). Furthermore, in this work we have not explicitly investigated linguistic phenomena such as *downstep*, although the pattern observed in HLH contexts seems to indicate that *downstep* specifically might be measurable in our corpus. Future work should include an explicit investigation and quantification of phenomena such as *downstep* and *declination* involved in the general *downtrend* observed in the corpus. As the eventual aim is to enable speech technologies such as recognition and synthesis systems, we plan to experiment with and evaluate such systems based on the features presented here.

6. ACKNOWLEDGEMENTS

It is a pleasure to thank Prof. Brian Mak at the Hong Kong University of Science and Technology for his hospitality and constructive discussions during a visit to Hong Kong, where much of this work was done. The authors would also like to thank Oluwapelumi Giwa for assistance with regard to the speech corpus.

7. REFERENCES

[1] 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.

[2] G. Kochanski and C. Shih, "Prosody modeling with soft templates," *Speech Communication*, vol. 39, pp. 311–352, Feb. 2003.

[3] Ọ. A. Ọdẹ́jọ́bí, 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.

[4] Ọ. A. Ọdẹ́jọ́bí, "A Quantitative Model of Yorùbá Speech Intonation Using Stem-ML," *INFOCOMP Journal of Computer Science*, vol. 6, no. 3, pp. 47–55, 2007.

[5] Ọ. A. Ọdẹ́jọ́bí, S. H. S. Wong, and A. J. Beaumont, "A modular holistic approach to prosody modelling for Standard Yorùbá speech synthesis," *Computer Speech & Language*, vol. 22, no. 1, pp. 39–68, Jan. 2008.

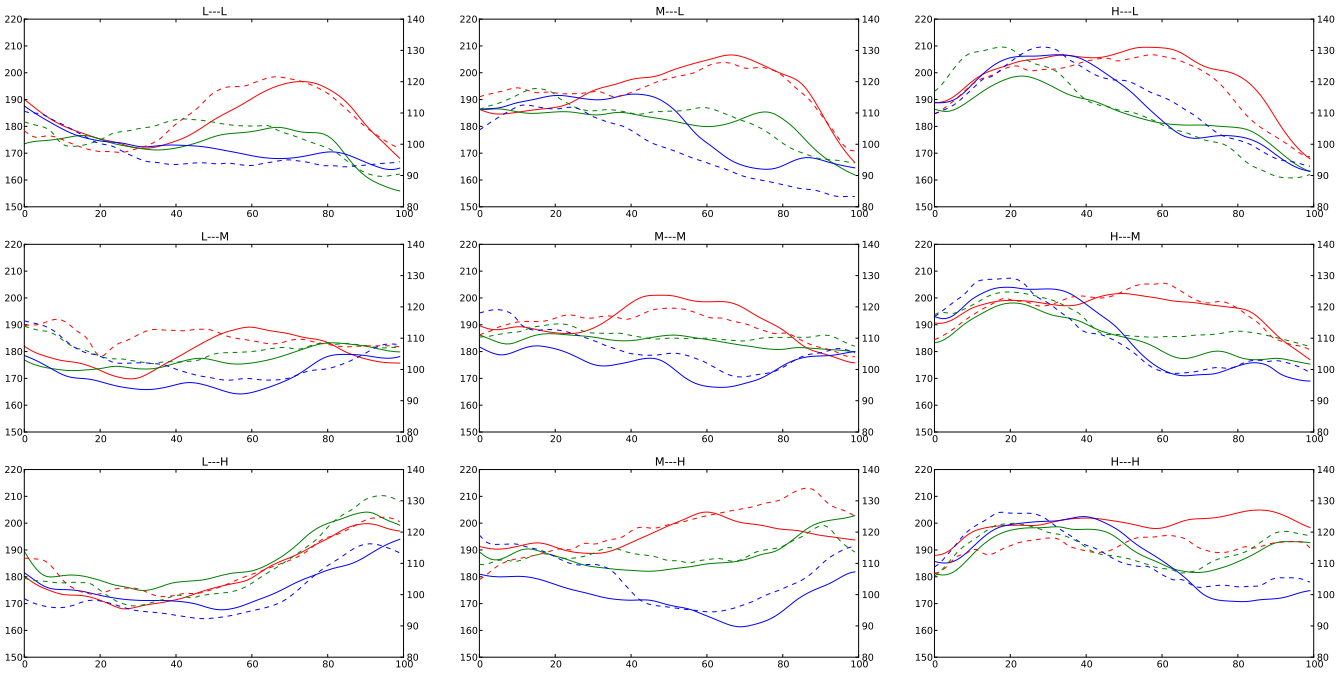

Fig. 5. Mean contours for three syllable sequences with the different tones H (red), M (green) and L (blue) in different tonal contexts. The solid and dashed lines represent examples of a female and male speaker, with y-axis values indicated on the left and right respectively (x is the normalised time and y the pitch in Hertz).

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

[7] 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.

[8] 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.

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

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