Predicting utterance pitch targets in Yorùbá for tone realisation in speech synthesis

Daniel R. van Niekerk[a,b], Etienne Barnard[a]

[a] *Multilingual Speech Technologies, North-West University, Vanderbijlpark, South Africa.*
[b] *Human Language Technologies Research Group, Meraka Institute, CSIR, Pretoria, South Africa.*

Abstract

Pitch is a fundamental acoustic feature of speech and as such needs to be determined during the process of speech synthesis. While a range of communicative functions are attributed to pitch variation in speech of all languages, it plays a vital role in distinguishing meaning of lexical items in tone languages. As a number of factors are assumed to affect the realisation of pitch, it is important to know which mechanisms are systematically responsible for pitch realisation in order to be able to model these effectively and thus develop robust speech synthesis systems in under-resourced environments. To this end, features influencing syllable pitch targets in continuous utterances in Yorùbá are investigated in a small speech corpus of 4 speakers. It is found that the previous syllable pitch level is strongly correlated with pitch changes between syllables and a number of approaches and features are evaluated in this context. The resulting models can be used to predict utterance pitch targets for speech synthesisers (whether it be concatenative or statistical parametric systems), and may also prove useful in speech-recognition systems.

Keywords: Yorùbá, tone language, speech synthesis, fundamental frequency

1. Introduction

Increasingly powerful and efficient machine learning algorithms for speech and language processing have resulted in a suite of open source tools that have enabled the construction of successful corpus-based speech synthesis systems in under-resourced environments (Davel and Barnard, 2008; Zen et al., 2006). In many cases, acoustic models for a basic speech synthesiser in a new language can be constructed automatically from a relatively small corpus of speech recordings (less than 1 hour of audio) and little language-specific development; typically a phoneme set, small pronunciation dictionary or hand-written grapheme-to-phoneme rules and a simple syllabification algorithm will suffice.

Building such basic systems for tonal languages, however, requires additional resources. Tonal information in pronunciation resources needs to be available and linguistic processes affecting such tones in speech (e.g. tone sandhi) need to be modelled. Given this, acoustic properties of tonal speech must be understood. The main acoustic correlate of tone is pitch (measured as fundamental frequency or F_0), which is also known to have other significant linguistic and para-linguistic communicative functions (Xu, 2005). This multiplexing of information in the pitch feature poses a challenge to robust acoustic modelling, especially in under-resourced environments. For this reason, basic systems built in this context often do not include tone information (Louw et al., 2006; Ekpenyong et al., 2008), which may result in degraded intelligibility as well as naturalness of resulting speech in various ways depending on the specific language.

In this work, we focus on the problem of pitch modelling for a tone language with limited resources, taking an approach that we expect to generalise to other African register tone languages, with the eventual goal of enabling rapid development of robust speech synthesis systems. For this purpose we investigate syllable pitch levels in continuous utterances of Yorùbá. Yorùbá is a relatively well studied African tone language of which the linguistic details of the tone system have been thoroughly described. Three level tones, labelled High (H), Mid (M) and Low (L) are associated with syllables and have a high functional load (Courtenay, 1971). Tones are marked explicitly on the orthography, making automatic derivation of surface tone (i.e. tones that are realised after all linguistic processes have been applied) from text relatively straightforward. These aspects of Yorùbá in particular make it an attractive

proxy for studying tone realisation in African (register) tone languages. We thus investigate pitch changes in Yorùbá associated with tones specified in this fashion on the orthography.

Despite the fact that the language and tone system has been well studied, Yorùbá (along with most other African languages) is considered under-resourced with respect to speech and language technology development where large text and speech corpora are often used to build advanced language and acoustic models respectively. This fact is indicated by a report on language technology development for African languages by Adegbola (2009) and corroborated by approaches taken to develop specific language technologies such as automatic text diacritization (Adegbola and Odilinye, 2012) and speech synthesis systems (Ọdẹ́jọbí et al., 2008).

In the following section we provide more relevant details of the Yorùbá tone system, discuss related work and formulate and motivate the approach followed in this work. This is followed by a section where the effects of specific mechanisms and features on pitch change are measured and discussed. In Section 5 we evaluate models and features for predicting pitch targets in this context and in Section 6 we discuss results and future work.

2. Related work and current approach

Yorùbá is considered to have a three-tone register (level) tone system with a *terracing* nature. Terracing refers to an utterance-wide trend based on the fact that tones are not realised at fixed pitch levels, but at systematically decreasing levels through the course of an utterance, depending on the effects of mechanisms including *downstep*, *declination* and *pitch resetting* (see Figure 1). Distinct intra-syllable patterns occurring in Yorùbá are *falling* and *rising* pitch contours when L and H tones are realised after H and L tones respectively.

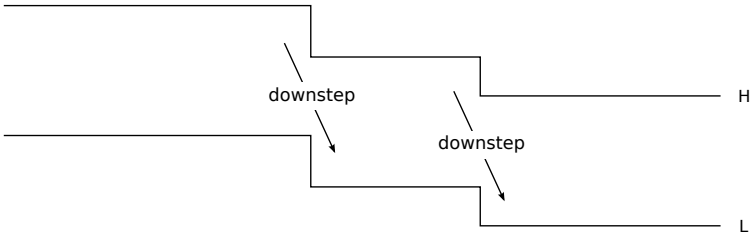

Figure 1: A simplified illustration of terracing in an utterance showing the lowering of pitch targets for H and L tones. This lowering effect is largely associated with *downstep* and as such is generally neither gradual over the course of an utterance nor independent of the tone sequence. Other factors potentially influencing pitch targets in utterances are excluded from this illustration.

For the generation of realistic pitch contours in continuous utterances, prosodic models often consider *short-term* (intra-syllable) and *long-term* (utterance-wide) pitch patterns or trends independently (Kochanski and Shih, 2003; Fujisaki et al., 1998). Short-term patterns, a direct result of the realisation of tones associated with each syllable by contrasting inter-syllable pitch levels or characteristic intra-syllable pitch movements, are combined with long-term patterns, usually associated with phrases or complete utterances. Such long-term trends are then modelled with distinct functions that may vary depending on the type of utterance (e.g. questions or statements etc.).

Recent work on the realisation of tone in Yorùbá has been described by Ọdẹ́jọbí et al. Their work involved the development and comparison of prosodic models for pitch synthesis based on the Stem-ML (Kochanski and Shih, 2003) framework and a system designed specifically for Yorùbá (Ọdẹ́jọbí et al., 2006; Ọdẹ́jọbí, 2007; Ọdẹ́jọbí et al., 2008). While this model considers intra-syllable and utterance-wide patterns independently, it builds the utterance-wide contour resulting from a sequence of pitch changes in local contexts using a recursive tree structure. Such a model relying on the cumulative effects of local pitch changes (rather than an independent phrase function) is plausible given a *terracing* tone system and is supported by the findings of Connell and Ladd (1990) and Laniran and Clements (2003) that the overall *downtrend* in utterances in Yorùbá seems to be dependent on the specific tonal content of the utterance.

Assuming thus that the utterance-wide pitch contour is largely a consequence of the cumulative effect of local pitch changes, we base our investigation on quantifying such pitch changes in different contexts in continuous utterances. To compactly quantify pitch changes between syllables we need to define relevant *pitch targets* where pitch values can be measured. Two candidates for pitch targets are found in Ọdẹ́jọbí et al. (2008) and the target approximation model proposed by Xu (2005). Ọdẹ́jọbí et al. (2008) assumes that each tone contour can be represented by exactly one peak

and one trough, while the target approximation model proposes static (or dynamic in the case of contour tones) targets that are reached asymptotically (i.e. towards the end of a syllable), depending on the effort exerted towards clear tone realisation. In earlier work (Van Niekerk and Barnard, 2012), we described the canonical contours observed in our Yorùbá corpus in different tonal contexts, confirming the following:

1. M tones generally have a relatively flat profile and H and L tones may have a rising or falling realisation over the course of a syllable.
2. Carryover assimilation seems more prominent or significant than anticipatory effects.
3. Extreme points (peaks or troughs) are generally realised late in a syllable, with a significant amount of variation of the exact turning points.

Due to the fact that some contours are difficult to represent as a peak and trough (point 1) and that points 2 and 3 seem to be in agreement with the assertions by Xu (2005), we adopt this model for our analysis of pitch targets and assume in the case of Yorùbá that we are to determine static (flat) pitch targets at different levels for each of the three tones (H, M and L).

We continue in the following section by describing our experimental setup followed by details of our investigation in Section 4.

3. Experimental setup

We start our investigation by preparing a small speech recognition corpus for statistical analysis. This is done by performing phonetic alignment, careful F_0 extraction and estimation of pitch targets for each syllable. Details of this process are given in the following subsections.

3.1. Corpus preparation

The speech data used in this study consisted of a subset of 33 speakers from a speech recognition corpus currently under development at the University of Lagos, Nigeria and North-West University, South Africa. Each speaker recorded about 100 short utterances from the pool of selected sentences, amounting to about 5 minutes of audio per speaker. Audio is broadband, collected in Lagos, 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 F_0 estimation.

Literary or Standard Yorùbá has a fairly regular orthography with graphemes generally corresponding directly to underlying phonemes with the inclusion of a few simple digraphs (such as gb, the voiced labial-velar stop /g͡b/ and certain nasalised vowels indicated by a succeeding n, for example ọn refers to /ɔ̃/). 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 presentation of these language details can be found in Section 2 of Ọdẹ́jọbí et al. (2006). 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 (Connell and Ladd, 1990). 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).

For our 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 above. Syllable tones were obtained from the orthography (diacritics). Given this information, we performed automatic phonemic alignment of the audio by forced-alignment of Hidden Markov Models (HMMs) as described in Van Niekerk and Barnard (2009), treating each speaker's utterances independently. 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 F_0 contours, we used *Praat* (Boersma, 2001), specifically the autocorrelation method. Pitch ranges were determined for each speaker manually, by plotting histograms of F_0 samples extracted using the range 60 to 600 Hz and subsequently resetting and re-extracting contours for a narrower range to reduce the occurrence of octave errors. All contours are

converted to semitone units (relative to 1 Hz) before further processing. For an indication of the reliability of this process we randomly selected a small sample (one utterance from each speaker), manually determining and counting the number of gross errors in alignment and F_0 extraction. A total of 355 syllables were inspected in *Praat* (using spectrograms and F_0), counting gross errors when a significant part of the syllable is misrepresented (approximately 50% or more). The gross-error count in this sample combining alignment and F_0 extraction errors was 13%.

3.2. Pitch target estimation

From our corpus described above, we selected four speakers (two from each gender), for which we proceeded to extract pitch targets as described below. For these speakers we manually inspected alignments and F_0 extraction for correctness. Here we intervened by correcting transcriptions and alignments in the case of gross errors, refraining from editing phonetic boundaries extensively. If F_0 extraction was particularly unreliable or transcriptions were completely erroneous, we discarded the utterance (a total of 10 utterances were discarded in this way). Table 1 shows the resulting corpus statistics.

Speaker ID	Gender	F_0 range (Hz)	Number of utterances	Number of syllables		
				H	M	L
013	female	100 - 350	136	534	462	444
017	female	120 - 300	136	540	441	458
021	male	70 - 220	129	486	397	417
024	male	100 - 220	126	477	381	417

Table 1: Corpus statistics with syllable counts by tone reflected in the last three columns.

Assuming the target approximation model (see discussion in Section 2), observable pitch contours are a result of a speaker's efforts to reach a specific pitch target. Targets are defined as straight lines that may be static (for level tones such as in Yorùbá) or dynamic (for contour tones such as in Mandarin) within a syllable. Actual pitch contours approach, not necessarily reaching, these targets towards the end of syllables depending on carryover effects from the previous syllable and the amount of effort exerted by the speaker to clearly enunciate the current tone. We considered two methods of estimating such level pitch targets:

1. Determining the maximum, mean and minimum pitch values in the syllable nucleus for H, M and L tones respectively (considering point 1 in Section 2).
2. Estimating the underlying pitch targets via an analysis-by-synthesis method implemented by the *PENTAtrainer Praat* script based on the quantitative target approximation model described by Prom-On et al. (2009).

Estimates from (1), using smoothed interpolated contours to reduce measurement noise and (2), extracted as described below, led to very similar results. Targets based on estimate (2), however, exhibited less variance and we thus adopted this estimate for further analysis.

The quantitative target approximation model described in Prom-On et al. (2009) uses a simple linear equation (1) to describe pitch targets, with a third-order critically damped linear system (2) defining the resulting pitch contour approximating the target in a syllable:

$$x(t) = mt + b \tag{1}$$

$$f_0(t) = x(t) + (c_1 + c_2 t + c_3 t^2)e^{-\lambda t} \tag{2}$$

where m and b represent the gradient and height of the current syllable pitch target respectively and c_i are determined by initial conditions of F_0 and the current syllable target, thus modelling the carryover effect on F_0 by the previous syllable. λ represents the rate of target approximation. The relevant parameters that need to be determined for each syllable are thus m, b and λ. The *PENTAtrainer* script scans predefined ranges of these parameters for each syllable, searching for optimal values minimising the error between resulting synthesised and actual F_0 contours extracted with *Praat* (and interpolated to have values in unvoiced regions). For our purposes we firstly assume $m = 0$ in all syllables (level tones). For the target height parameter, b, we leave a broad search range ±20 semitones from the measured F_0

(the default value in *PENTAtrainer*), but practically restrict this by assuming relatively high values of λ. This assumes that speakers are being clear in expressing tones in their speech and the result is that estimated targets will not lie far from measurable extreme points in F_0 contours (as described in point 2 above). Although manual inspection of resulting pitch target estimates suggests that some errors do occur, especially when syllables are very short combined with slight alignment inaccuracies, the process seemed robust in general: Figure 2 shows an example of pitch targets extracted for an utterance in our corpus.

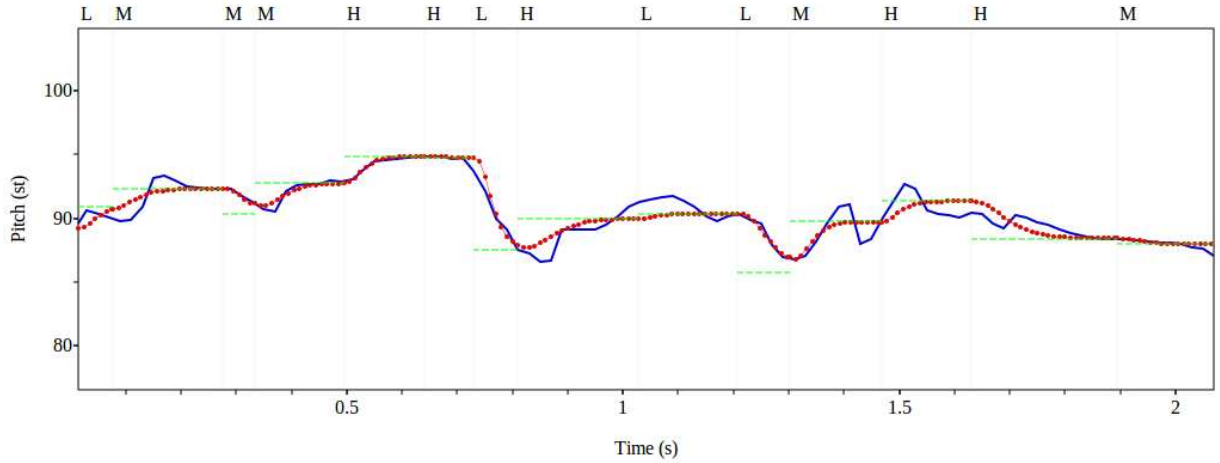

Figure 2: Example of pitch targets extracted from an utterance in our corpus; The original F_0 contour is represented by the solid line (blue), with estimated pitch targets indicated with dashed lines (green) and the resulting synthetic contour with connected dots (red).

4. Features of pitch change

In this section we consider the changes in pitch between syllables based on an analysis of the pitch targets extracted in Section 3.2. We attempt to establish the salient features that would enable us to predict pitch changes in an utterance context by considering how such features might provide information on the underlying mechanisms involved.

4.1. Initial observations

In Figure 3 we present the distribution of pitch targets for the three tones for each of our speakers. When targets for all utterances are combined in this way, a linear downtrend (measuring in semitones) seems to emerge, however a significant amount of variation is present in all cases (tones and speakers). We expect several sources to contribute to this variability:

1. Local *downstep* and different rates of downtrend expected due to the *terracing* nature and differences in length and tone sequences between utterances (Connell and Ladd, 1990; Laniran and Clements, 2003).
2. Anticipatory raising of pitch for specific tone sequences (Laniran and Clements, 2003).
3. Pitch resetting (Laniran and Clements, 2003).
4. Syllable duration.
5. Intrinsic F_0.
6. The realisation of word focus or emphasis, which is expected to increase the dynamic range of pitch movement according to Xu (2005).
7. Assimilation of syllables potentially causing false measurements (see notes in Section 3.2).
8. Possible tone sandhi effects not accounted for - tone is assumed to be shallowly marked on the orthography.
9. Possible changes in speaker effort (the λ parameter discussed in Section 3.2).
10. Errors in estimation due to F_0 estimation inaccuracies.

Given the current experimental setup, it is difficult to consider all of these causes and points 6 - 9 are thus not explicitly investigated while we attempt to determine the predictability of points 1 - 5. We discuss each of these in the following subsections.

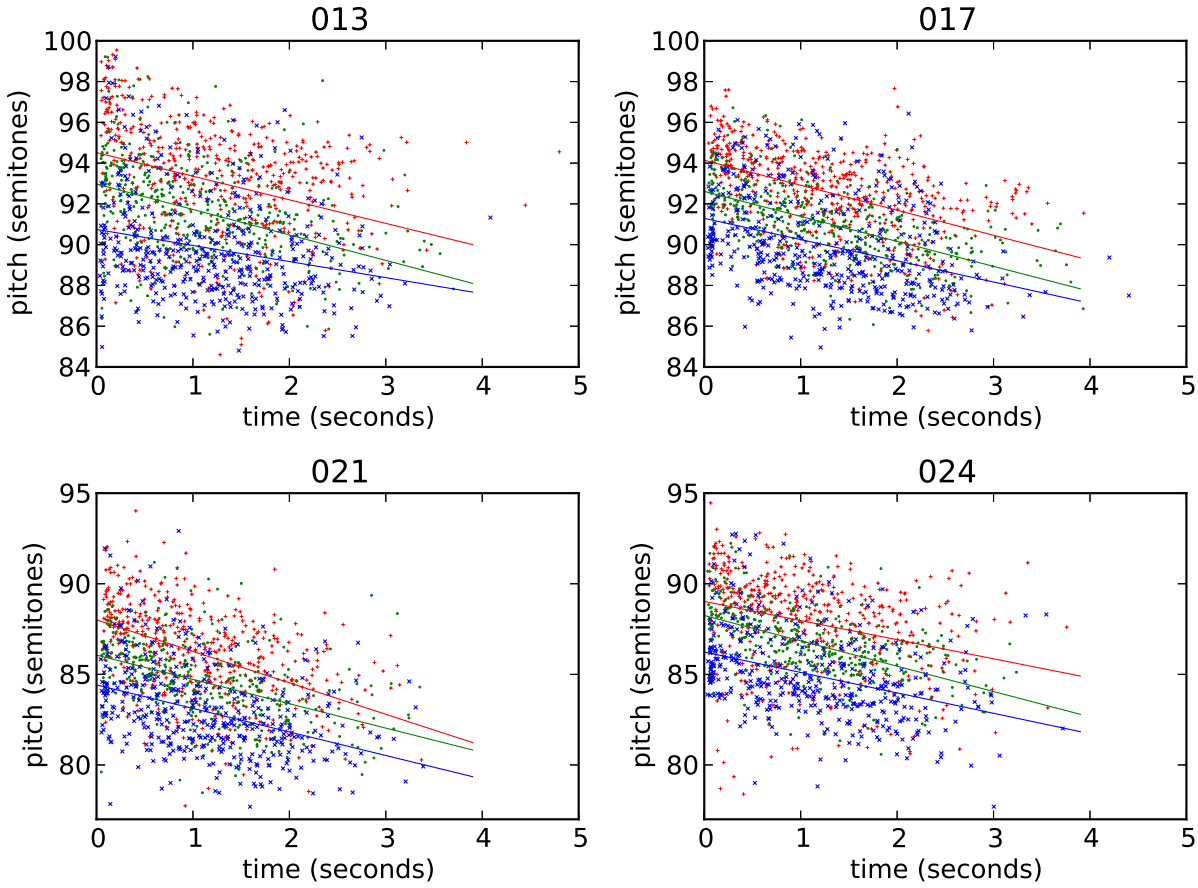

Figure 3: Pitch targets extracted for each speaker (speakers 013 and 017 are female, with 021 and 024 male). The tones H, M and L are represented by red (+), green (.) and blue (x) respectively, with a linear fit plotted for each. Times correspond to the central instant of each syllable.

4.2. Tonal context

Considering point 1 above: The phonological concepts identified by Connell and Ladd (1990); Laniran and Clements (2003) to have a significant effect on utterance pitch contours are:

- *Initial tone level*: utterance-initial Hs can be higher than usual and initial Ls lower than usual.
- *Final lowering*: can result in the final syllable, whether H, M or L, being realised lower than usual.
- *Downstep*: the occurrence of a HLH sequence results in lower subsequent pitch targets.
- *Declination*: a sequence of like tones generally exhibits a slight tone-specific declination in pitch.

Based on these descriptions a context of at least four syllables would be required to uniquely quantify relative pitch target shifts, i.e. know the identities of the two syllables preceding and following a specific transition (e.g. $HH \rightarrow LH$, $NH \rightarrow LH$ or $HH \rightarrow LN$, where the start and end of the utterance are represented by N). Figures 4 and 5 present the mean changes in syllable pitch targets for our 4 speakers in different contexts.

While the confidence intervals are fairly wide for most cases, we are able to make a few observations:

- Contexts of only two syllables behave consistently between speakers and are generally distinct.
- We see significant evidence for *downstep* in the H-LH context in all speakers.

- Pitch changes at the start and end of utterances vary and are often significantly different from the corresponding general two-syllable context.

Due to the fact that a phonological effect such as *downstep* seems to be directly dependent on tonal context it seems fair to conclude that at least the previous syllable context is important when considering pitch target change. Isolated other contexts are also shown to be significantly different from the corresponding general two-syllable context. However, none of these seem to hold across all speakers and we are thus reluctant to suggest that syllable context is the best way to describe these observed differences.

4.3. Pitch range

At this point we need to consider how a sequence of pitch changes behaves over the course of an utterance. Inspection of our corpus suggests that there are larger changes in pitch at the onset of utterances; also, that pitch changes are contracted later in utterances, with periodic pitch resetting in longer utterances. However, despite the observed trend in Figure 3 and the fact that models such as Fujisaki et al. (1998) have an explicit model dependent on utterance position, relying on syllable position or time since start-of-utterance to estimate such trends is problematic as pitch resets do not seem to be predictable in this way (Laniran and Clements, 2003).

Considering that these mechanisms (pitch change contraction and resetting) serve to manage pitch range usage, we investigate the relation between pitch changes and previous pitch level (Figures 6 – 9). It is evident that there is a strong linear relationship between the pitch change and previous pitch level in all speakers.

- The subplots for transitions to specific tones indicate differences in height as expected (e.g. most transitions to L tones have negative pitch changes, with transitions to H mostly positive).
- Transitions to the M tone seem to be more homogeneous than for the other two tones (e.g. differences in the originating tone does not seem to carry much extra information regarding pitch change not already contained in the previous pitch level).
- Transitions to the H tone especially seem to contain more distinct distributions that may be associated with the originating tone (and even the broader context as suggested by previous results e.g. the H-LH context).
- Most pitch changes lie within -5 to 5 semitones.

While this relationship between previous pitch level and pitch target changes might suggest that speakers are in fact attempting to realise absolute pitch targets, it is clear from the gradients of linear fits that this is not fully achieved in general (with speaker 013 the closest in realising constant pitch targets).

It is possible that speakers are attempting to achieve fixed pitch targets for each tone, but are limited in the amount of energy that may be exerted. This might explain why certain contexts provide information about pitch changes probably capturing effects of overshoot or under-articulation (e.g. L-HM for speaker 021 in Figure 5 etc.).

4.4. Syllable duration

Another factor which could have a systematic effect on pitch change is syllable duration. Xu and Sun (2000) report that it takes 125-141 ms to raise pitch by 3.6 - 6.3 semitones and we find that the pitch changes for all of our speakers largely lie between -5 to 5 semitones. This suggests that syllable duration could have a limiting effect on pitch change, especially when pitch has to change from L to H. However, the distribution of pitch changes versus syllable duration does not seem to indicate a strong relationship between these two parameters (for our four speakers, the Pearson correlation coefficients between pitch change and syllable duration lies in the range 0.16 to 0.22). A general limiting of pitch change is evident for short syllables (more so for upward pitch shifts than downwards), but this does not constitute a major contribution to the observed pitch changes.

Tone	Vowel	013	017	021	024
H	a	92.69 ± 0.50	92.29 ± 0.32	86.03 ± 0.44	87.59 ± 0.43
	ɛ	93.96 ± 0.57	92.49 ± 0.43	85.77 ± 0.70	88.41 ± 0.65
	ɔ	92.26 ± 0.74	92.35 ± 0.46	86.12 ± 0.58	87.06 ± 0.63
	e	93.06 ± 0.59	92.96 ± 0.37	85.73 ± 0.49	87.75 ± 0.63
	o	93.47 ± 0.78	92.55 ± 0.45	86.49 ± 0.68	88.44 ± 0.64
	i	92.87 ± 0.47	92.48 ± 0.31	85.94 ± 0.36	87.44 ± 0.48
	u	93.57 ± 0.58	92.33 ± 0.42	85.82 ± 0.60	87.79 ± 0.74
M	a	91.29 ± 0.45	90.99 ± 0.26	84.54 ± 0.38	86.57 ± 0.33
	ɛ	91.40 ± 0.69	90.52 ± 0.36	84.27 ± 0.52	87.22 ± 0.82
	ɔ	91.73 ± 0.44	91.10 ± 0.23	84.37 ± 0.36	86.55 ± 0.33
	e	90.97 ± 0.56	91.34 ± 0.44	84.63 ± 0.50	86.23 ± 0.71
	o	91.48 ± 0.41	90.95 ± 0.32	84.88 ± 0.54	86.83 ± 0.43
	i	91.28 ± 0.48	90.94 ± 0.36	84.68 ± 0.43	86.99 ± 0.31
	u	92.22 ± 0.80	91.25 ± 0.61	84.96 ± 0.50	86.42 ± 0.79
L	a	90.03 ± 0.39	90.03 ± 0.33	83.01 ± 0.34	85.14 ± 0.36
	ɛ	90.53 ± 0.81	89.97 ± 0.55	82.94 ± 0.72	85.24 ± 0.56
	ɔ	90.58 ± 0.83	90.30 ± 0.56	82.97 ± 0.79	85.19 ± 0.65
	e	90.47 ± 0.70	90.19 ± 0.64	83.28 ± 0.60	85.60 ± 0.90
	o	89.92 ± 0.71	90.36 ± 0.64	83.98 ± 0.66	85.16 ± 0.73
	i	89.84 ± 0.50	89.84 ± 0.42	82.95 ± 0.48	84.71 ± 0.43
	u	89.78 ± 0.59	90.41 ± 0.86	82.95 ± 0.73	84.92 ± 0.87

Table 2: Mean F_0 (in semitones) for syllables with different tones and vowels (vowels are ordered increasing in height). These values were calculated for utterances where the linear trend was removed. The 95% confidence intervals are indicated.

4.5. Intrinsic F_0

According to Whalen and Levitt (1995) the phenomenon known as *intrinsic pitch* (IF_0) is a universal phonetic effect associated with vowels and occurring in all languages to a greater or lesser degree. This refers to the tendency of high vowels such as [i] and [u] to have higher fundamental frequencies than low vowels such as [a]. Although Connell (2002) argues that this effect may be constrained in some tone languages under specific circumstances, the effect is largely confirmed for Yorùbá in other studies cited by Whalen and Levitt (1995); Connell (2002).

We investigated this in our corpus by removing the linear (downward) trend in each utterance and determining the mean pitch level for each vowel and tone by each speaker (Table 2). This was done by subtracting the linear least-squares fit estimated using all syllables' pitch targets (H, M and L) for each utterance individually and adding back the mean. For speakers 013 and 024 we observed greater differences across different vowels than speakers 017 and 021; however, we could not verify a consistent gradient of increasing IF_0 with increasing vowel height. The results measured do confirm that there is more measurable variation in F_0 across vowels for H tones than M or L, which is consistent with findings by Whalen and Levitt (1995). Further investigation into the values presented here would have to start with verification of the actual speaker pronunciations (we did not investigate possible dialectal differences in each speaker that might affect these results). The measurements obtained suggest that we should consider vowel identity as a potentially useful feature towards predicting pitch changes.

5. Pitch target prediction

In the previous section we presented an analysis of features and possible mechanisms affecting pitch change in our corpus. In this section we propose a number of models for the prediction of pitch targets in utterances and evaluate the features presented. Specifically, we aim to evaluate the following:

1. Effective ways of predicting pitch targets: we evaluate different regression models, attempting to predict pitch target values directly and by means of predicting pitch target changes (deltas).

2. The utility of features investigated in the previous section, specifically in this context (i.e. given a relatively small number of speech samples).
3. Whether the selected models proposed here adequately model the aspects of pitch targets observed in the previous section (e.g. the downtrend seen in Figure 3 and deltas for different tone transitions in Figures 4 and 5).

This is done by considering the 10-fold cross-validation error measured on pitch target values for each speaker given specific model and feature combinations. For tuning model meta-parameters we performed 10-fold cross-validation on the training set of each fold before re-estimating on the complete training set and predicting the test set. For testing purposes we assume the pitch target value is known for the first syllable and only generate and evaluate predictions from the second syllable of each utterance onwards. (This is done in order to have comparable results between models predicting targets directly or via deltas.)

5.1. Initial models

We start by proposing two models based on the observations in Section 4. The first model is based on the linear declination (observed in Figure 3) for each tone (H, M and L), and predicts this target value based on the current syllable tone and syllable utterance position (i.t.o. normalised utterance time). Such a model does not account for local dynamics such as *downstep* and pitch resetting, but maintains basic pitch contrasts between syllables that are assumed important for tone perception. Applying the cross-validation process described above resulted in the error rates presented in Table 3: `lint`.

The second model predicts pitch deltas between syllables based on the linear relationship between previous pitch level and pitch change (Figures 6 – 9). For the implementation of this model we determine a linear fit for samples in different tonal contexts. Thus, for each tonal context (e.g. H-LH) we have:

$$\Delta F_0 = aF_{0p} + b \tag{3}$$

where ΔF_0 is the predicted pitch change to the current syllable in this context and F_{0p} is the pitch level of the previous syllable. Parameters a and b are estimated for each tonal context instance provided a pre-determined minimum number of samples (`minsamples`) exist. Syllable context features used were (in specific order): target tone (`tt`), previous tone (`pt`), pre-previous tone (`ppt`) and following tone (`ft`). If `minsamples` were not available, more general contexts were used for estimation (by removing contextual information in reverse order, starting with `ft`). The process of cross-validation described above often resulted in a relatively large value for `minsamples`, leading to models with few distinct contexts (in the majority of cases only the target tone). The cross-validation error for this model using the `tt` feature is presented in Table 3: `lind`. Results are compared and discussed in Section 5.3.

5.2. Additional features

To further investigate the utility of features discussed in Section 4, we experimented with two additional model types; regression trees (Breiman et al., 1984), and support vector machines (SVM) (Chang and Lin, 2011) implemented in the *scikit-learn* software package (Pedregosa et al., 2011). Both decision trees and SVMs have been succesfully applied to problems of acoustic modelling of speech (Young et al., 1994; Peng and Wang, 2005).

Different feature combinations were evaluated using cross validation as described above. For tree-based models we used mean-squared-error criterion implemented in *scikit-learn* and estimated the meta-parameter controlling the minimum number of samples required to split a node (`minsamples`) by internal cross-validation on each training set. For SVM-based models we used the radial basis function kernel with meta-parameters C and ϵ determined by training-set cross validation and $\gamma = 1/N_f$ where N_f is the number of features. Categorical features were represented using "one-hot" binary coding with the absence of a category represented by zeros and continuous features represented by floating point values (normalised to range [0.0, 1.0] for SVM training). Models based on predicting pitch targets directly as well as deltas were evaluated. Features investigated are `tt`: target tone, `up`: utterance position, `pt`: previous tone, `ppt`: pre-previous tone, `pl`: previous pitch level, `ft`: following tone, `d`: syllable duration and `v`: base vowel. Results of these experiments for the most competitive model and feature combinations are reported in Table 3 and discussed in the next section.

Model	Type	Features	013		017		021		024	
			RMSE	Std	RMSE	Std	RMSE	Std	RMSE	Std
meant	target	tt	2.66	3.96	2.10	2.51	2.39	3.00	2.40	3.24
meant	target	tt,pt	2.55	3.69	1.97	2.37	2.27	2.92	2.24	3.08
meand	delta	tt,pt	4.20	5.22	2.99	3.57	3.12	3.77	3.61	4.45
meand	delta	tt,pt,ppt	3.71	5.16	2.28	2.82	2.63	3.58	3.21	4.46
lint	target	tt	2.53	3.70	1.84	2.30	2.13	2.81	2.23	3.23
lind	delta	tt	2.62	3.85	2.02	2.45	2.26	3.04	2.39	3.37
svm	target	tt,pt,up,pl	2.84	3.96	1.81	2.25	2.10	2.86	2.61	3.78
svm	target	tt,pt,ppt,up,pl	2.56	3.70	1.98	2.45	2.22	2.95	2.33	3.50
svm	delta	tt,pt,ppt,pl	2.54	3.69	2.06	2.52	2.27	3.03	2.88	4.00
svm	delta	tt,pt,ppt,up,pl	2.58	3.68	1.98	2.45	2.21	2.98	2.47	3.55

Table 3: Root mean square errors (RMSE) with standard deviations (Std) for the most competitive models and feature combinations. Results for regression tree models are not included here.

5.3. Discussion

In Table 3 we compare the cross validation root-mean-squared-error (RMSE) of the most competitive models and feature combinations with two baseline predictions (`meant`):

1. Predicting the mean F_0 observed per tone (e.g. H, M or L).
2. Predicting the mean F_0 observed per tone in context, where the previous tone is taken into account (e.g. LH, MH or HH given the target tone H).

We noted that the error rate generally decreased when *utterance position* and *previous pitch level* features were added, especially for the prediction of targets and deltas respectively. The inclusion of previous tone features seemed to decrease the error in general, with features such as following tone, syllable duration and vowel identity having variable effect on measured error. It is possible that the utility of these features, specifically syllable duration and following tone, is dependent on the speech rate (e.g. in faster speech one might find that syllable duration can be exploited due to its potentially constraining effect on pitch change and the following tone might affect the speech due to anticipatory effects). Overall, SVMs seemed to perform best, especially with the inclusion of continuous variable features (`up` and `pl`). Although the best error rates achieved are not significantly different from the best baseline approach considered (`meant` with `tt` and `pt` features), further investigation reveals that the nature of predictions vary in the degree to which short-term and long-term patterns are preserved. Table 4 shows the linear estimates of downtrend measured over all utterances. Models not considering utterance position (i.e. `meant` and `lind`) underestimate the overall downtrend (that is, the pitch values towards the end of utterances tend to be too high). Similarly, it can be shown that the inclusion of previous tone information (`pt` and `ppt`) is important towards preserving the patterns observed in Figures 4 and 5.

Model	Type	Features	013	017	021	024
meant	target	tt,pt	-0.28	-0.25	-0.46	-0.32
lint	target	tt	-1.08	-1.12	-1.50	-1.16
lind	delta	tt,pt,ppt	-0.45	-0.20	-0.81	-0.55
svm	target	tt,pt,ppt,up,pl	-0.90	-1.30	-2.33	-1.37
svm	delta	tt,pt,ppt,up,pl	-0.89	-1.29	-1.72	-1.21
Actual samples			-1.03	-1.16	-1.53	-1.19

Table 4: Linear downtrend estimates (in semitones per second) for different models and feature combinations compared to actual samples.

6. Discussion and future work

In this work, we attempted to model relevant pitch changes in Yorùbá for the purpose of speech synthesis or speech recognition. Towards this goal we have extracted pitch targets relying on automatic methods of pronunciation

prediction, phonetic segmentation, pitch contour and target extraction and performed a statistical analysis of relevant features available to a typical speech-processing system. In this context, with about 5 minutes of speech per speaker, we have confirmed a number of previously reported phenomena such as *downstep* and found limited evidence for others such as *intrinsic* F_0.

Based on our analysis in Section 4 we proposed and evaluated a number of models and features for predicting syllable pitch targets in this context. The best of these models results in pitch targets that preserve both local pitch changes that are assumed to be important for communicative function and long-term trends that presumably affect the perceived naturalness of synthesised utterances. Such models may be used in unit-selection type synthesis systems directly to predict pitch targets, which may be integrated into a target cost function or as a component in an explicit contour synthesis algorithm (e.g. in Ọdẹ́jọbí et al. (2008) or Xu (2005)). While the work in this paper focussed on the analysis of speech of neutral prosody, we hope that the approach taken will also be applicable when considering more expressive speech. This might be implemented either by additional models that adapt neutral pitch targets (an example can be found in Tao et al. (2006)) or by including expression features when estimating models such as presented in this work.

In line with our goals of enabling "tone-aware" synthesis systems in under-resourced contexts, future work will involve development and testing of explicit F_0 contour generation for integration into HMM-based synthesis systems. Evaluation of such systems will enable us to understand which of the aspects modelled here, local pitch changes and long-term trends, are important with regards to intelligibility and naturalness of synthesised speech, thus affording us an opportunity for refinement of these models. Similarly, these models will be evaluated for their ability to improve the performance of HMM-based speech-recognition systems. We are also interested in extending these efforts to other languages of similar nature.

7. References

Adegbola, T., 2009. Building capacities in human language technology for African languages. In: Proceedings of the First Workshop on Language Technologies for African Languages. pp. 53–58.

Adegbola, T., Odilinye, L. U., May 2012. Quantifying the effect of corpus size on the quality of automatic diacritization of Yorùbá texts. In: The Third International Workshop on Spoken Language Technologies for Under-resourced Languages. Cape Town, South Africa, pp. 48–53.

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

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

Chang, C.-C., Lin, C.-J., May 2011. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2 (3), 27:1–27:27.

Connell, B., 2002. Tone languages and the universality of intrinsic F0: evidence from Africa. Journal of Phonetics 30, 101–129.

Connell, B., Ladd, D. R., 1990. Aspects of pitch realisation in Yoruba. Phonology 7 (1), 1–29.

Courtenay, K., 1971. Yoruba: a terraced-level language with three tonemes. Studies in African Linguistics 2 (3), 239–255.

Davel, M., Barnard, E., 2008. Pronunciation prediction with Default&Refine. Computer Speech and Language 22, 374–393.

Ọdẹ́jọbí, O. A., 2007. A Quantitative Model of Yorùbá Speech Intonation Using Stem-ML. INFOCOMP Journal of Computer Science 6 (3), 47–55.

Ọdẹ́jọbí, O. A., Beaumont, A. J., Wong, S. H. S., 2006. Intonation contour realisation for Standard Yorùbá text-to-speech synthesis: A fuzzy computational approach. Computer Speech & Language 20 (4), 563–588.

Ọdẹ́jọbí, O. A., Wong, S. H. S., Beaumont, A. J., Jan. 2008. A modular holistic approach to prosody modelling for Standard Yorùbá speech synthesis. Computer Speech & Language 22 (1), 39–68.

Ekpenyong, M., Urua, E.-A., Gibbon, D., 2008. Towards an unrestricted domain TTS system for African tone languages. International Journal of Speech Technology 11 (2), 87–96.

Fujisaki, H., Ohno, S., Wang, C., November 1998. A command-response model for F0 contour generation in multilingual speech synthesis. In: The Third ESCA/COCOSDA Workshop on Speech Synthesis. Jenolan Caves House, Blue Mountains, NSW, Australia, pp. 26–29.

Kochanski, G., Shih, C., Feb. 2003. Prosody modeling with soft templates. Speech Communication 39, 311–352.

Laniran, Y. O., Clements, G. N., 2003. Downstep and high raising: interacting factors in Yoruba tone production. Journal of Phonetics 31 (2), 203–250.

Louw, J. A., Davel, M., Barnard, E., 2006. A general-purpose IsiZulu speech synthesizer. South African journal of African languages 2, 1–9.

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

Peng, G., Wang, W. S.-Y., Jan. 2005. Tone recognition of continuous Cantonese speech based on support vector machines. Speech Communication 45 (1), 49–62.

Prom-On, S., Xu, Y., Thipakorn, B., 2009. Modeling tone and intonation in Mandarin and English as a process of target approximation. The Journal of the Acoustical Society of America 125, 405.

Tao, J., Kang, Y., Li, A., 2006. Prosody conversion from neutral speech to emotional speech. IEEE Transactions on Audio, Speech, and Language Processing 14 (4), 1145–1154.

Van Niekerk, D. R., Barnard, E., September 2009. Phonetic alignment for speech synthesis in under-resourced languages. In: Proceedings of INTERSPEECH. Brighton, UK, pp. 880–883.

Van Niekerk, D. R., Barnard, E., May 2012. Tone realisation in a Yorùbá speech recognition corpus. In: The Third International Workshop on Spoken Language Technologies for Under-resourced Languages. Cape Town, South Africa, pp. 54–59.

Whalen, D., Levitt, A. G., 1995. The universality of intrinsic F0 of vowels. Journal of Phonetics 23 (3), 349–366.

Xu, Y., 2005. Speech melody as articulatorily implemented communicative functions. Speech Communication 46, 220–251.

Xu, Y., Sun, X., 2000. How fast can we really change pitch? Maximum speed of pitch change revisited. In: The Sixth International Conference on Spoken Language Processing. Beijing, China, pp. 666–669.

Young, S. J., Odell, J. J., Woodland, P. C., 1994. Tree-based state tying for high accuracy acoustic modelling. In: Proceedings of the workshop on Human Language Technology. pp. 307–312.

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

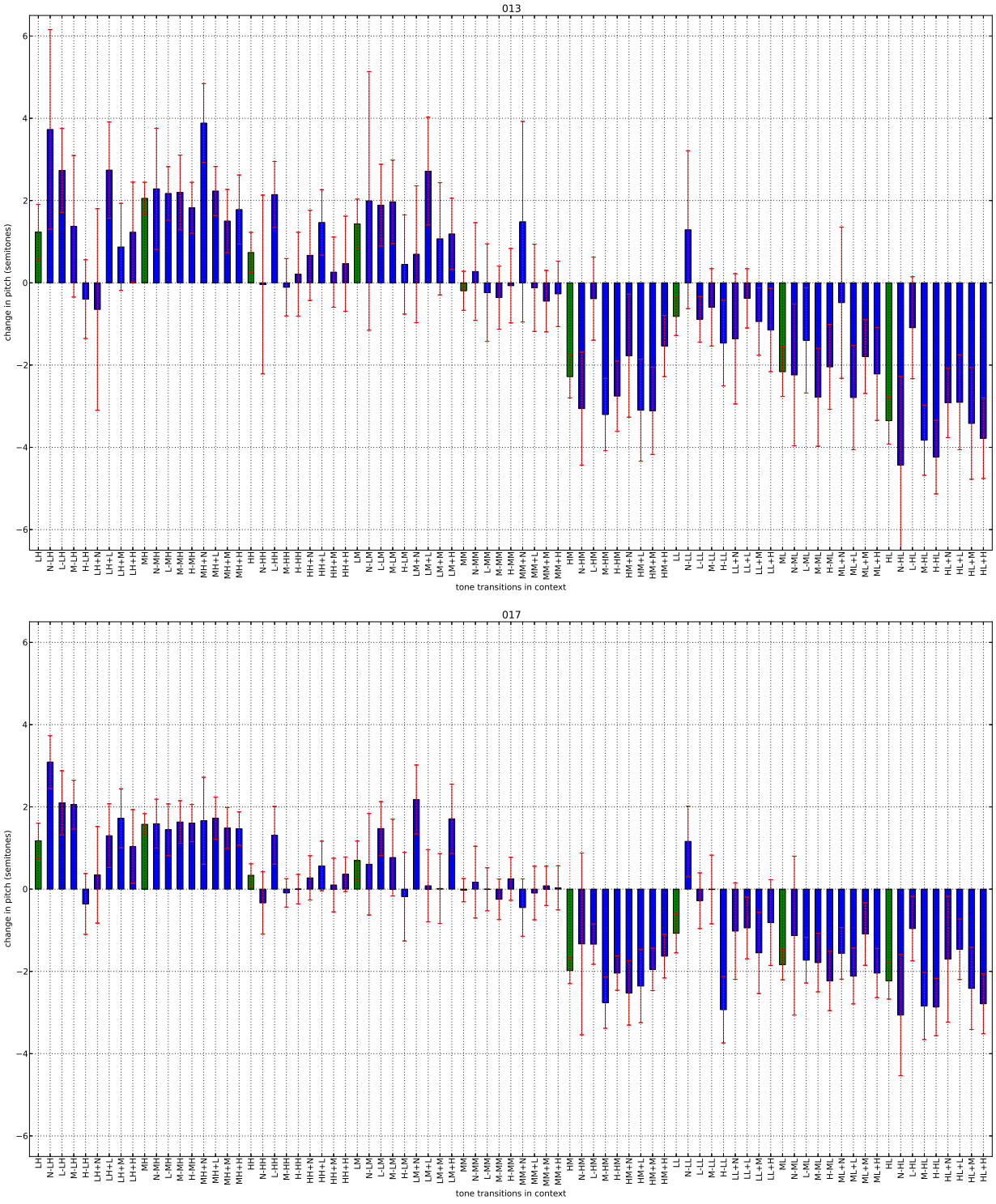

Figure 4: Mean pitch changes between syllables in different contexts, for speakers 013 and 017; preceding contexts are denoted by a ”-” and succeeding contexts by a ”+”. H, M and L represent High, Mid and Low tones, with N representing the utterance boundary. Error bars denote the 95% confidence interval.

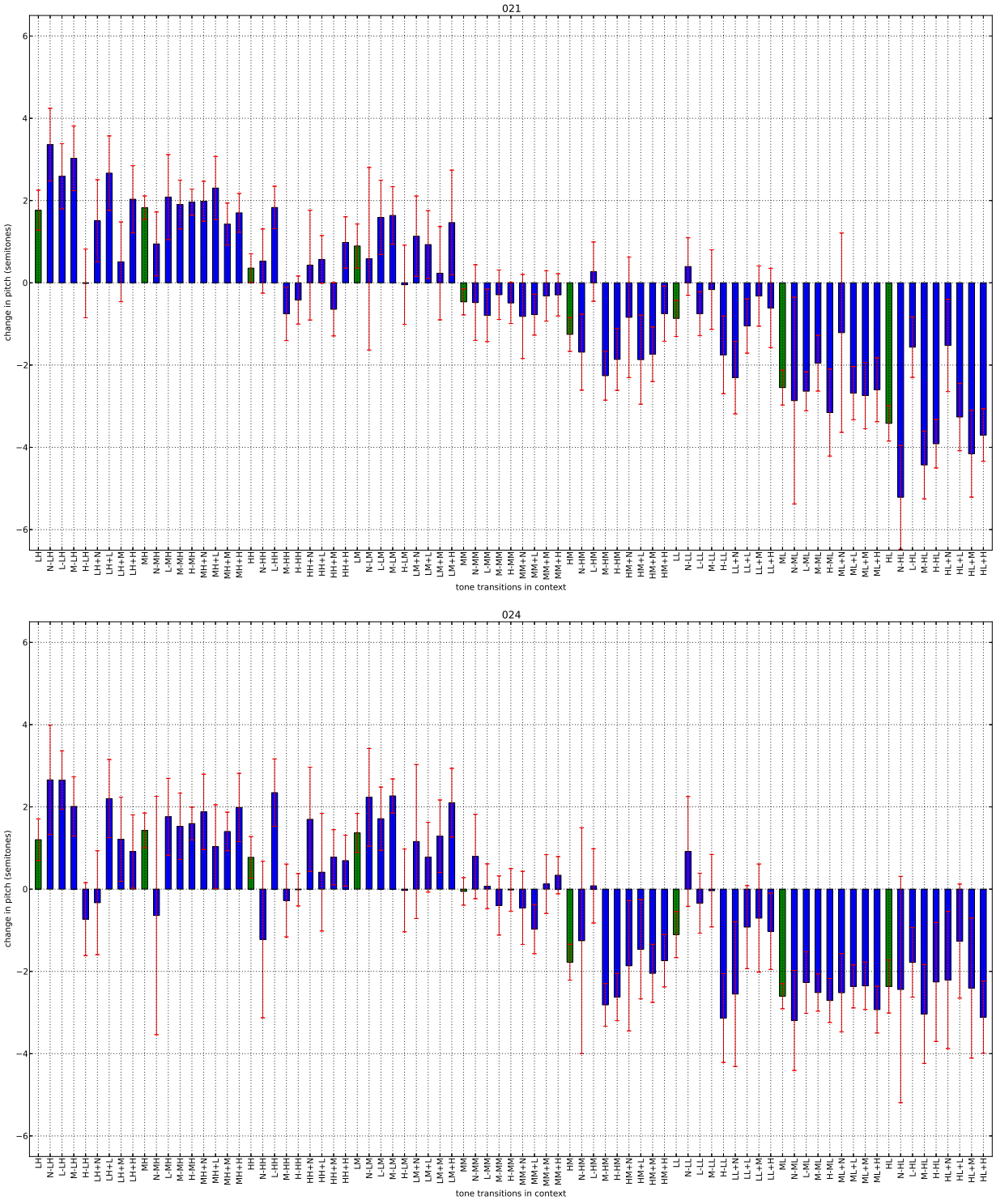

Figure 5: Mean pitch changes between syllables in different contexts for speakers 021 and 024; preceding contexts are denoted by a ”-” and succeeding contexts by a ”+”. H, M and L represent High, Mid and Low tones, with N representing the utterance boundary. Error bars denote the 95% confidence interval.

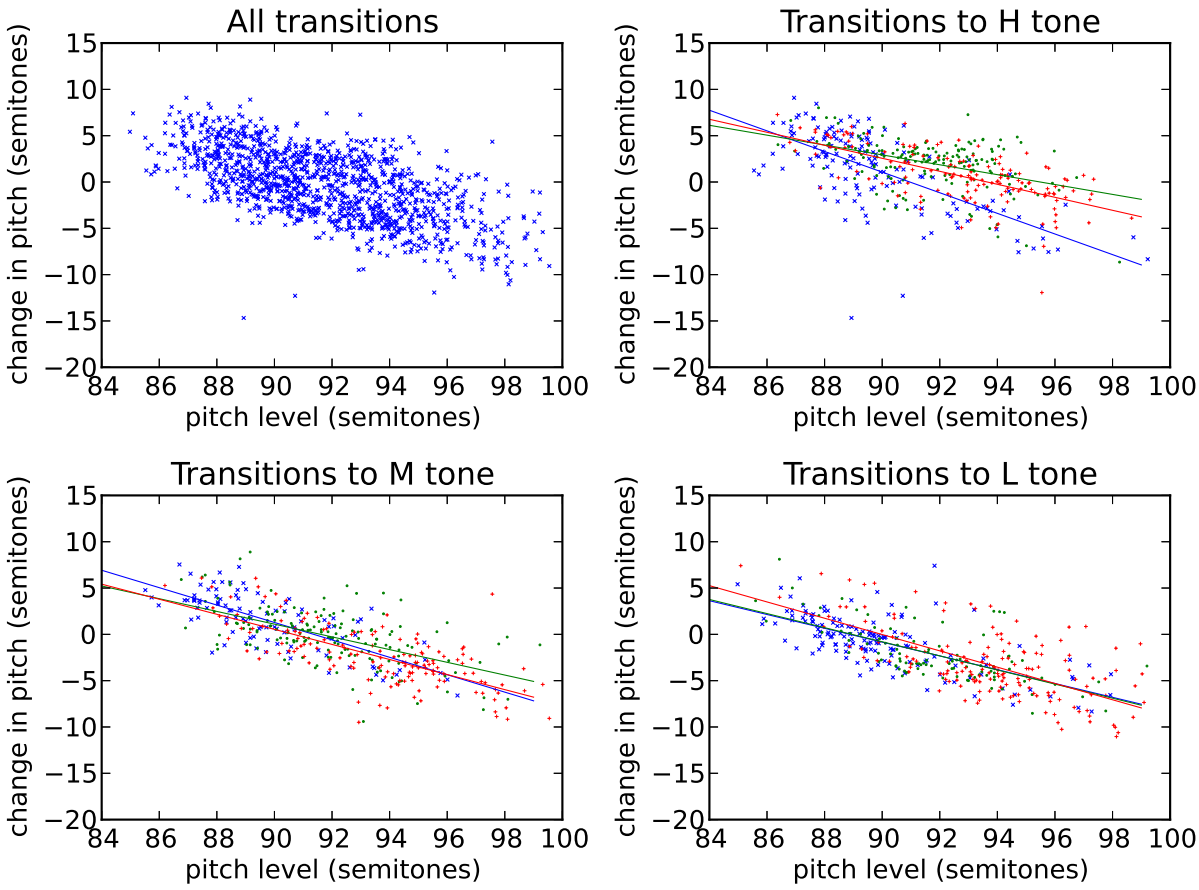

Figure 6: Speaker 013: Changes in pitch for targets in consecutive syllables. Subplot 1 shows all transitions, with subplots 2 4 showing transitions to H, M and L tones respectively. In subplots 2 4 blue (x), green (.) and red (+) represent transitions from L, M and H tones respectively.

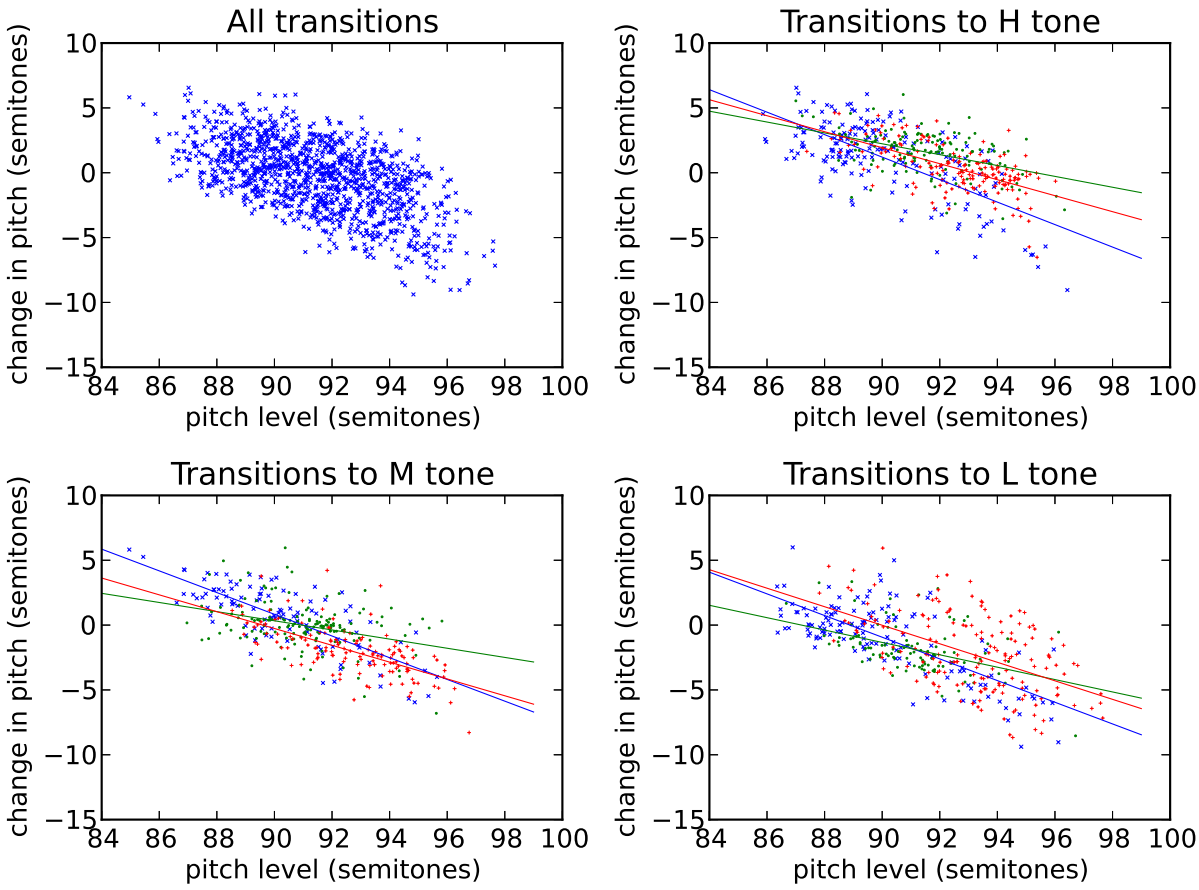

Figure 7: Speaker 017: Changes in pitch for targets in consecutive syllables. Subplot 1 shows all transitions, with subplots 2 4 showing transitions to H, M and L tones respectively. In subplots 2 4 blue (x), green (.) and red (+) represent transitions from L, M and H tones respectively.

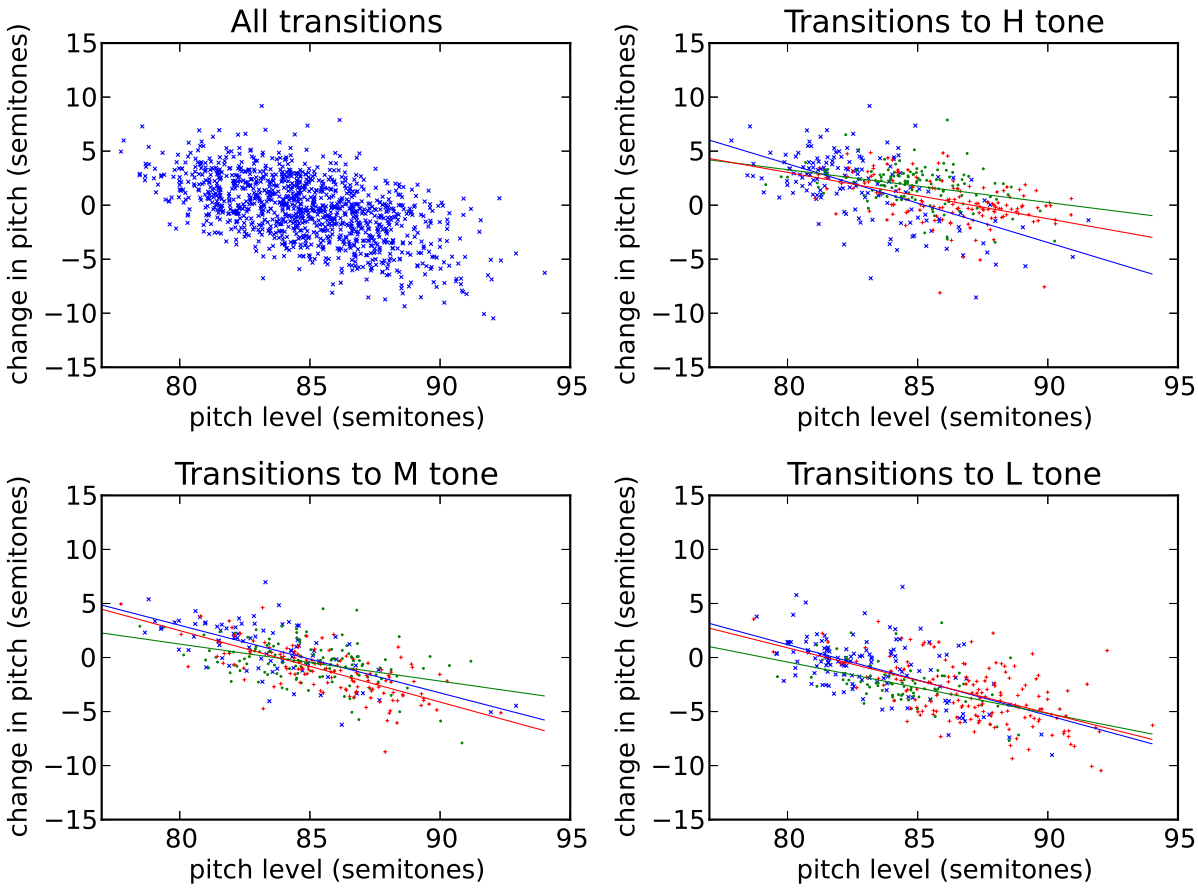

Figure 8: Speaker 021: Changes in pitch for targets in consecutive syllables. Subplot 1 shows all transitions, with subplots 2 4 showing transitions to H, M and L tones respectively. In subplots 2 4 blue (x), green (.) and red (+) represent transitions from L, M and H tones respectively.

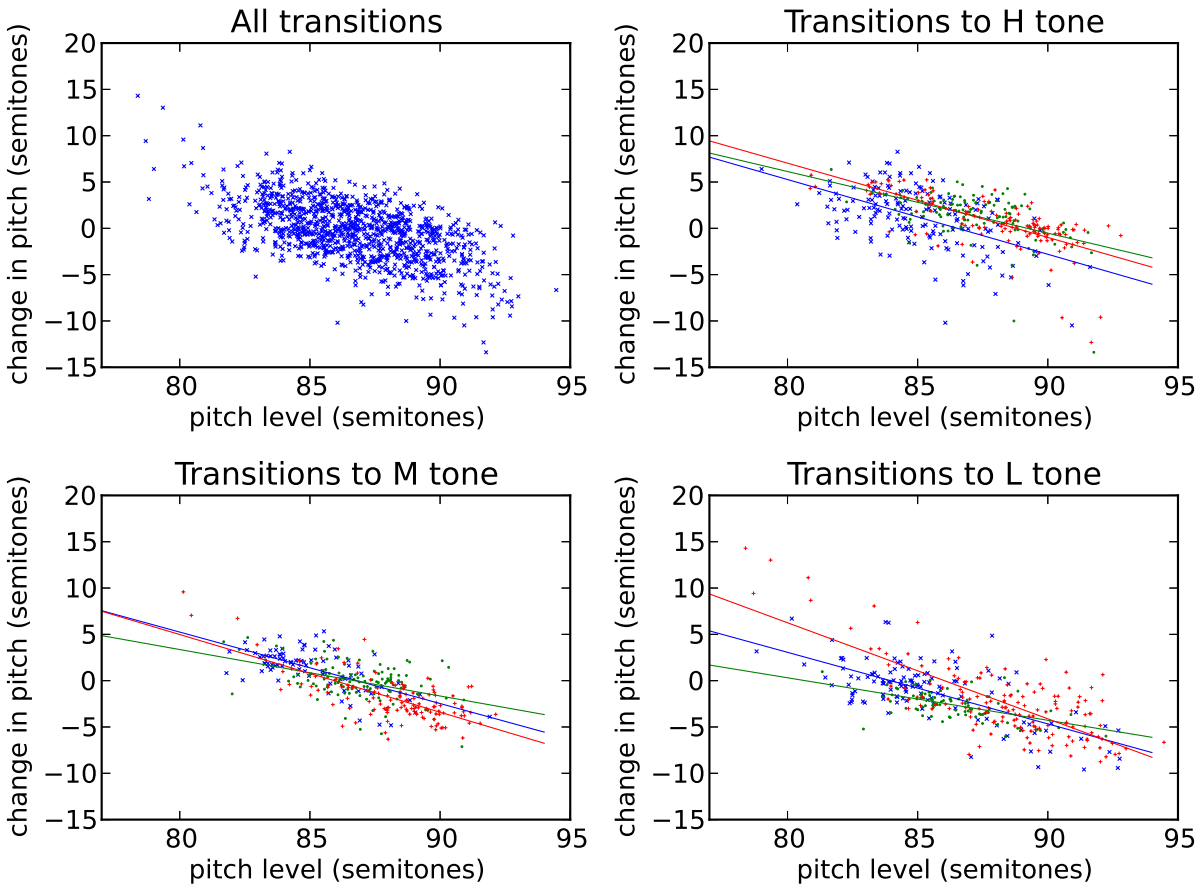

Figure 9: Speaker 024: Changes in pitch for targets in consecutive syllables. Subplot 1 shows all transitions, with subplots 2 4 showing transitions to H, M and L tones respectively. In subplots 2 4 blue (x), green (.) and red (+) represent transitions from L, M and H tones respectively.