TONE REALISATION FOR SPEECH SYNTHESIS OF YORÙBÁ

by

Daniel Rudolph van Niekerk

Thesis submitted for the degree Philosophiae Doctor (Information Technology)

at the

Vaal Triangle Campus of the North-West University

Promoter: Professor Etienne Barnard

May 2014

SUMMARY

TONE REALISATION FOR SPEECH SYNTHESIS OF YORÙBÁ

by

Daniel Rudolph van Niekerk

Promoter:	Professor Etienne Barnard
Faculty:	Economic Sciences and Information Technology (Vaal Triangle Campus)
University:	North-West University
Degree:	Philosophiae Doctor (Information Technology)
Keywords:	Speech synthesis, text-to-speech, intonation model, target approximation, tone language, Yorùbá, under-resourced language

Speech technologies such as text-to-speech synthesis (TTS) and automatic speech recognition (ASR) have recently generated much interest in the developed world as a user-interface medium to smartphones [1, 2]. However, it is also recognised that these technologies may potentially have a positive impact on the lives of those in the developing world, especially in Africa, by presenting an important medium for access to information where illiteracy and a lack of infrastructure play a limiting role [3, 4, 5, 6]. While these technologies continually experience important advances that keep extending their applicability to new and under-resourced languages, one particular area in need of further development is speech synthesis of African tone languages [7, 8].

The main objective of this work is acoustic modelling and synthesis of tone for an African tone language: Yorùbá. We present an empirical investigation to establish the acoustic properties of tone in Yorùbá, and to evaluate resulting models integrated into a Hidden Markov model-based (HMM-based) TTS system.

We show that in Yorùbá, which is considered a *register* tone language, the realisation of tone is not solely determined by pitch levels, but also inter-syllable and intra-syllable pitch dynamics. Furthermore, our experimental results indicate that utterance-wide pitch patterns are not only a result of cumulative local pitch changes (*terracing*), but do contain a significant gradual *declination* component. Lastly, models based on inter- and intra-syllable pitch dynamics using underlying linear pitch

targets are shown to be relatively efficient and perceptually preferable to the current standard approach in statistical parametric speech synthesis employing HMM pitch models based on context-dependent phones. These findings support the applicability of the proposed models in under-resourced conditions.

SAMEVATTING

TOONREALISERING VIR SPRAAKSINTESE VAN YORÙBÁ

deur

Daniel Rudolph van Niekerk

Promotor:	Professor Etienne Barnard
Departement:	Ekonomiese Wetenskappe en Inligtingtegnologie (Vaaldriehoekkampus)
Universiteit:	Noordwes-Universiteit
Graad:	Philosophiae Doctor (Inligtingtegnologie)
Sleutelwoorde:	Spraaksintese, teks-na-spraak, intonasiemodel, teiken-benadering, toontaal, Yorùbá, hulpbron-skaars taal

Spraaktegnologieë soos teks-na-spraaksintese (TTS) en outomatiese spraakherkenning (ASR) het onlangs heelwat belangstelling ontlok as gebruikerskoppelvlak tot slimfone [1, 2]. Die moontlikheid vir dié tegnologie om 'n positiewe bydrae as inligtingsmedium tot die lewensstandaard van mense in ontwikkelende streke te lewer, veral in Afrika waar ongeletterdheid en tekort aan basiese infrastruktuur 'n negatiewe rol speel, word ook herken [3, 4, 5, 6]. Hoewel volgehoue vooruitgang die toepaslikheid van dié tegnologie tot hulpbron-skaars tale voortdurend uitbrei, is die ontwikkeling van spraaksintese van Afrika-toontale een onderwerp wat verdere aandag benodig [7, 8].

Die hoofdoel van hierdie werk is die suksesvolle akoestiese modellering en sintese van toon vir 'n Afrika-toontaal: Yorùbá. Gevolglik word 'n empiriese ondersoek voorgelê om die akoestiese eienskappe van toon in Yorùbá vas te stel en modelle wat daaruit volg te evalueer binne 'n TTS stelsel wat gebruik maak van versteekte Markovmodelle (HMMe).

Ons resultate dui daarop dat in Yorùbá, wat beskou word as 'n *registertoontaal*, die uitdrukking van toon nie slegs afhanklik is van die vlakke van toonhoogte nie, maar ook die verloop van toonhoogte beide tussen en binne die bereik van sillabes. Verder word daar deur middel van modellering aangetoon dat toonhoogtepatrone oor die bereik van heel uitinge nie slegs 'n gevolg van plaaslike (inter-sillabe) toonhoogte veranderinge (*toon-terrasse*) is nie, maar wel 'n beduidende geleidelike *deklinasie*-komponent bevat. Ten slotte bevind ons dat modelle direk gebaseer op die inter- en intra-

sillabe verloop van toonhoogte deur middel van onderliggende lineêre teikenfunksies relatief doeltreffend is en perseptueel gunstig vergelyk met die huidige standaardbenadering in statisties-parametriese spraaksintesestelsels wat HMM toonhoogte modelle saamgestel uit konteks-afhanklike fone gebruik. Hierdie bevindinge ondersteun die toepaslikheid van die voorgestelde modelle onder hulpbron-skaars omstandighede.

ACKNOWLEDGEMENTS

To Professor Etienne Barnard I am most grateful. To him I owe not only the privilege and guidance to work on this topic, but also my understanding of scientific research and appreciation for the game of bridge. His patience, kindness and sense of purpose will remain inspiring to me throughout.

I also thank Oluwapelumi Giwa, Professor Marelie H. Davel, Professor Gerhard B. van Huyssteen, and Professor Brian K-W. Mak who have all directly enabled this work. Without their support it would certainly not have gotten off the ground or have been completed.

I am also indebted to all my past and present colleagues, especially at the HLT research group in the Meraka Institute of the CSIR in Pretoria and the North-West University in Potchefstroom.

To friends and family, especially my parents Pieter and Ildikó van Niekerk, who have grounded me in the principles of life and provided continuous encouragement. I am because we are.

To everyone who had faith in this: 2 Corinthians 5:7.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ASR	Automatic speech recognition
CS	Computer science
CSIR	Council for Scientific and Industrial Research
CV	Consonant-vowel
DSP	Digital signal processing
DP	Dynamic programming
DTW	Dynamic time warping
EM	Expectation maximisation
F0	Fundamental frequency
HLT	Human language technology
HMM	Hidden Markov model
HTK	Hidden Markov model toolkit
HTS	Hidden Markov model-based speech synthesis system
H	High (tone)
IF0	Intrinsic fundamental frequency
IPO	Institute for Perception Research
INTSINT	International transcription system for intonation
L	Low (tone)
ML	Maximum likelihood
MSE	Mean squared error
MFCC	Mel-frequency cepstral coefficient
M	Mid (tone)
MOMEL	Modélisation de Melodie
MSD-HMM	Multi-space probability distribution hidden Markov model
N	Nasal
NLP	Natural language processing
PENTA	Parallel encoding and target approximation
qTA	Quantitative target approximation
RMSE	Root mean squared error
Stem-ML	Soft template markup language
SVM	Support vector machine
TTS	Text-to-speech
ToBI	Tones and break indices
V	Vowel

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Speech technologies such as text-to-speech synthesis (TTS) and automatic speech recognition (ASR) have recently generated much interest in the developed world as a user-interface medium to smartphones, which represent ever smaller and more convenient devices for personal computing [1, 2]. It is also recognised that these technologies may potentially have a positive impact on the lives of those in the developing world, especially in Africa, by presenting an important medium for access to information where illiteracy and a lack of infrastructure play a limiting role [3, 4, 5, 6]. However, the development and application of TTS systems in the developing world has been challenging to date. On the one hand, the challenges of designing and implementing appropriate speech-based interfaces for users in this context calls not only for highly intelligible systems, but also for systems that instil a sense of familiarity in the target user group [11, 12] – requiring a significant degree of naturalness (similarity to human speech) and the ability to adapt voice characteristics rapidly to accommodate changes in persona and dialect. On the other hand, the lack of infrastructure, expertise and particularly basic language and speech resources presents significant engineering challenges and limits the quality of systems that can be built with current approaches [13, 14, 15]. One particular area in need of further development to address these challenges is speech synthesis of African tone languages, as the following section demonstrates.

1.1 PROBLEM STATEMENT

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 realised on each syllable. In such *register tone systems*, tone realisation to some extent relies on changes in pitch between consecutive syllables. Such systems stand 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, it is important for the development of speech technologies in these languages to understand the tone system in detail [16]. Developing systems such as TTS and ASR for tone languages 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. While deriving surface tone from text (point 1) is a significant linguistic challenge in many tone languages [16, 17], the focus of the current work is the problem of acoustic modelling and synthesis for tone realisation (point 2).

Increasingly powerful and efficient algorithms and models for speech and language processing have recently enabled the construction of successful corpus-based acoustic models for TTS systems in under-resourced environments [18, 19]. However, the construction of systems that adequately account for tone information continues to be a challenge, with basic systems often not incorporating tone information at all [8, 7]. This may result in degraded intelligibility as well as naturalness of resulting speech in various ways depending on the specific language [20]. The main acoustic correlate of tone is pitch, which is also known to have other significant linguistic and paralinguistic communicative functions, even in tone languages [21]. The modelling of pitch is an active research topic in the field of speech synthesis with researchers still proposing improved methods based on different theories and synthesis technologies [22, 23, 24, 25]. It is thus clear that the modelling of pitch for speech synthesis is a complex problem due to the multiplexing of parallel streams of information. Despite the importance and complexity of the problem however, little attention has been devoted to speech synthesis for register tone languages (particularly African languages) and consequently it is still difficult to construct reliable TTS systems for tone languages in this context [14, 20].

The focus of this work will be on the modelling and synthesis of pitch contours for an African tone language (Yorùbá) given limited resources, investigating approaches that are expected to generalise to other African tone languages. Yorùbá is a relatively well studied 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 [26]. Tones are marked explicitly on the orthography (shallow marking [27]), making automatic derivation of surface tone from text possible. These aspects of Yorùbá in particular make it an attractive model case for studying tone realisation in African tone languages.

1.2 RESEARCH QUESTIONS

Given the context provided and problem statement presented in the preceding sections, the following research questions are formulated:

1. What are the salient acoustic features (especially of pitch) attributable to the expression of tone in Yorùbá as manifested in general continuous utterances?

2. How can this be suitably modelled and applied in speech technologies (especially TTS systems) in typical under-resourced environments?

1.3 OVERVIEW OF THE STUDY

Given the above-mentioned research questions, this study presents:

- A detailed description of the acoustic properties (especially pertaining to pitch) associated with the expression of tone in utterances of Yorùbá with the aim of supporting the development of speech technologies.

- The development and evaluation of models and methods for the implementation of acoustic tone realisation in a speech synthesis system in under-resourced environments.

- A discussion on the potential application of the developed methods to other African tone languages and for various development scenarios in under-resourced contexts.

The study commences with a literature review and discussion including a basic overview of the Yorùbá tone system, current approaches to intonation modelling, and state-of-the-art implementations of prosody in TTS systems in Chapter 2. This serves to motivate the approaches followed during the empirical investigations in the remainder of the study. In Chapter 3 a descriptive investigation is conducted to confirm the phonetic properties of tone in Yorùbá as described in the linguistics literature using established methods from the speech technology field. In Chapter 4, this information is used in conjunction with approaches described in the literature to develop and analytically test the basis for appropriate intonation models using an analysis by modelling and synthesis methodology [28, 29]. In Chapter 5, the proposed models are refined, implemented and evaluated *in situ* for their validity and utility in reference to the stated objectives. Finally, Chapter 6 contains a summary and discussion of

the contributions and applicability of this study and highlights avenues for future work.

CHAPTER 2

BACKGROUND

The aims of this work as motivated and outlined in the previous chapter are relevant to the fields of text-to-speech synthesis, particularly acoustic modelling, and the related linguistic topics of prosody, intonation and phonetic description of tone. In this chapter aspects of these topics pertaining to the current work are concisely presented and discussed.

2.1 TEXT-TO-SPEECH SYNTHESIS

Text-to-speech synthesis is the process of converting written text into speech. The sub-processes involved in implementing such a process may be formulated in different ways and as a consequence the construction of TTS systems generally involves the integration of knowledge and techniques from various disciplines. While there are a number of ways to formulate the overall process of TTS, a common view makes a distinction between two fundamental processes; *text analysis* and *speech synthesis* [30, 31].

During text analysis, information relevant to the speech synthesis process is recovered from the input text. This process usually involves the application of techniques developed in the field of natural language processing (NLP). Although the details of this process vary widely between systems, depending amongst others on input language, system complexity and granularity of sub-components, text analysis systems are usually broadly responsible for *tokenisation*, *normalisation* and *phonetisation*.

Speech synthesis involves the use of the information produced by the text analysis process to produce acoustic signals representing speech. In modern systems this component relies on digital signal processing (DSP) techniques to generate acoustic signals and techniques from sub-fields of computer

science (CS) to represent speech signals. The exact synthesis algorithm used to generate the output acoustic signal depends on the form of acoustic units or models. In this regard systems are traditionally divided into so-called *rule-based* and *corpus-based* (also referred to as *knowledge-driven* and *data-driven*) systems. In reality these terms represent extremes of a continuum of modern approaches[1] where *rule-based* systems attempt to represent and synthesise speech with a compact set of parameters which may or may not be estimated directly from speech recordings (usually at the cost of synthesised speech quality), while *corpus-based* systems rely on powerful machine learning techniques and algorithms in an attempt to reproduce natural sounding speech (usually at the cost of model size and development data requirements). While the development of modern TTS systems began with purely *rule-based* approaches such as *formant synthesis* and *articulatory synthesis* where parameters were determined and set manually, current state-of-the-art systems invariably rely on *corpus-based* approaches supported by ever increasing availability of computing power and improvements in machine learning algorithms [30]. Of these *corpus-based* approaches two broad approaches; *statistical parametric* and *unit-selection synthesis* (a non-parametric *concatenative* approach) continue to compete for state-of-the-art results in large-scale TTS evaluations [32, 33].

In the following sections *unit-selection* and *statistical parametric synthesis* are presented in more detail to illustrate how different aspects of speech (especially prosody) are modelled and synthesised and the properties of the resulting synthesised speech are discussed.

2.1.1 Unit-selection synthesis

Early *concatenative* approaches to corpus-based speech synthesis involved carefully constructing minimal acoustic inventories (traditionally based on phone transition units or *diphones*) from specially designed speech corpora and splicing these together again during synthesis. These inventories contained all *phonemic* acoustic units with prosodic parameters (such as pitch and duration) implemented by adapting acoustic units using DSP based on explicit prosodic models. This approach was superseded by the unit-selection approach on the premise that natural sounding speech synthesis can be achieved by selecting and concatenating appropriate sub-word units obtained directly from a corpus of natural speech.

Based on this idea, the problem of synthesising a new utterance is viewed as a search over available acoustic units to select a sequence which minimises a *cost function* designed to determine the prop-

[1] based on constructing acoustic models from primarily deductive or inductive perspectives

erties of the output speech signal. An important formulation of this *cost function* is found in [34] where the cost is a combination of the *target cost* representing the mismatch between a candidate unit and the desired output unit (usually based on linguistic context) and the *concatenation cost* representing the mismatch between two consecutive units (usually based on a perceptually relevant acoustic distance measure). This allows the process of synthesis to be seen as determining the optimal unit sequence in a state transition network (the speech database) with *state occupation* and *transition costs* corresponding to *target* and *concatenation costs*. An exhaustive search through the database is then usually avoided during synthesis by applying a dynamic programming (DP) algorithm such as the Viterbi algorithm [35] to perform the search/optimisation process efficiently.

The advantage of the unit-selection approach lies in the ability to achieve high quality synthesis by relying directly on the properties of the underlying speech corpus. Synthesised speech quality generally improves with increases in the corpus size (due to better acoustic unit coverage), with state-of-the-art results achievable with large speech corpora (the most natural sounding systems continue to be based on unit-selection [32, 33]). This is a result of a substantial amount of work on various aspects such as database size and structure, improving the search and synthesis time, different methods for calculating and combining the target and concatenation costs, the size of units and its effect on synthesis quality and acoustic distance measures, amongst others. A good overview of active research threads is found in [19].

Approaches to modelling prosody within the unit-selection framework vary from fully *implicit models* to detailed *explicit models*. Implicit models incorporate simple contextual features into the cost function in an attempt to reconstruct prosodic patterns existing in the corpus. Explicit models directly determine pitch and duration values that may be integrated into the calculation of the *target cost* or used to constrain the search space appropriately. Whether implicit or explicit prosodic models are employed, high quality synthesised output is strongly dependant on the contents of the speech corpus with synthesis quality degrading when attempting to synthesise long sequences which do not naturally occur in the corpus. This lack of flexibility constitutes one of the major disadvantages of unit-selection synthesis [36]. The fact that multiple properties (acoustic parameters) of speech need to be jointly optimised by selecting a single (shared) unit sequence leads to a combinatorial explosion and poses a serious challenge when considering data requirements for the synthesis of varied prosody [37, 24]. These concerns are compounded when considering speech synthesis based on smaller corpora.

2.1.2 Statistical parametric synthesis

In the last decade an increasing amount of work has been done on *statistical parametric synthesis*, where speech corpora are used as basis for the estimation of statistical acoustic models [19]. This approach relies on deconstructing speech signals into fundamental parameters: *excitation* (including *pitch* and *voicing*), *duration* and *spectral envelope*. These parameters are then modelled individually, and new parameter sequences are generated from the resulting acoustic models and combined during synthesis. Pioneering work involved modelling and generating parameters using Hidden Markov Models (HMMs) [38, 39]. While other generative models, such as decision trees [40], have since been proposed, *HMM-based synthesis* remains the dominant approach and represents the state-of-the-art [32, 33].

Models are usually estimated from a speech corpus using the maximum likelihood (ML) criterion[2] as follows (from [19]):

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\{p(\boldsymbol{O}|\mathscr{W},\boldsymbol{\lambda})\} \tag{2.1}$$

where $\boldsymbol{\lambda}$ is a set of model parameters, $\boldsymbol{O}$ is a set of training data, and $\mathscr{W}$ is a set of word sequences corresponding to $\boldsymbol{O}$. These models, $\hat{\boldsymbol{\lambda}}$, are then used to generate speech parameters, $\boldsymbol{o}$, for a new word sequence, w, to maximise the output probabilities of the parameters:

$$\hat{\boldsymbol{o}} = \underset{\boldsymbol{o}}{\operatorname{argmax}}\{p(\boldsymbol{o}|w,\hat{\boldsymbol{\lambda}})\} \tag{2.2}$$

Conceptually, this is analogous to generating the expected value of parameters seen in the training set for distinct segments of speech. The estimation of HMM model parameters uses the *forward-backward algorithm* which is essentially a form of the *expectation maximisation* (EM) algorithm and in practice HMM *state distributions* are tied using decision trees as is common in acoustic modelling for speech recognition [42, 43, 44]. For speech synthesis however, detailed speaker-specific models of multiple parameter streams (excitation, duration and spectral envelope) are usually the result. This is achieved by using "full-context" phone models employing more contextual information than the context-dependant *triphone* models customary in speech recognition. Typically two preceding and two succeeding phones (*quinphones*) as well as syllable, word, phrase (or breath-group) and sentence context are employed to allow the modelling of longer-term patterns associated with prosody (e.g. in the *pitch* parameter) [45]. The generation of smooth trajectories using maximum likelihood from HMM *state output distributions* (Eq. 2.2) is achieved by incorporating the generated parameters of

[2] although other criteria have also been proposed [41]

dynamic features [46]. Finally, given the individually generated speech parameters, a speech signal is synthesised using a *vocoder*.

Important advantages of *HMM-based synthesis*, especially considering application in under-resourced environments are *robustness* and *flexibility*. The fact that speech parameters are generated by essentially averaging over instances in the corpus with an effective mechanism for dealing with data sparsity results in gradual quality degradation when data is limited or non-ideal. This is preferable over the more distinct synthesis artefacts that may be expected from *unit-selection synthesis* in these scenarios [47, 48]. Speech synthesis based on statistical models also provides possibilities for data sharing [49] and rapid development of application-specific systems by employing speaker-adaptive training [50] (a desirable property; see the introduction in Chapter 1). Challenges in statistical parametric synthesis stem from the fact that current modelling and reconstruction techniques incur a loss of *naturalness* in the resulting speech. Particular problems relate to, amongst others, the approximate nature of acoustic models, over-smoothing due to the averaging process during training, and the "vocoded quality" of speech due to inadequate or inaccurate modelling of *excitation* parameters. Important recent improvements dealing with over-smoothing and improving excitation modelling that have become standard,[3] include the modelling of global variance [23] and mixed excitation modelling and synthesis [51] respectively. A comprehensive discussion on recent advances and relevant threads of ongoing research can be found in [19].

Within the HMM-based synthesis framework, in order to capture prosody, pitch is modelled together with spectral features using multi-space probability distribution HMMs (MSD-HMMs), which are able to seamlessly deal with undefined segments (e.g. in the case of *unvoiced speech*). Duration models are commonly represented by Gaussian distributions from state occupancy probabilities obtained in the last iteration of embedded (forward-backward) re-estimation [38, 22].[4] While models of the different parameters are based on the same contextual features and state distributions are temporally aligned, decision tree HMM-state tying is done independently. Pitch models rely on supra-segmental contextual features to model longer term patterns such as pitch *declination* over a sentence, while *microprosody* associated with segmental interaction may be captured through the inclusion of segmental features. The fact that *pitch* and *duration* parameters are tied independently from spectral parameters makes more efficient modelling possible (compared to unit-selection), however, for high-quality synthesis of both *macro-* and *microprosodic* patterns, data requirements based on this approach also

[3] meaning that they are part of freely available open-source software implementations

[4] although other models have also been proposed [52]

increase rapidly.

2.2 PROSODY AND INTONATION

In the field of linguistics, *prosody* is concerned with the *rhythm*, *stress* and *intonation* of speech, which is perceived by the listener through a combination of changes in *tempo*, *loudness* and *pitch*. These perceptual features are physically measured in terms of segment *duration*, signal *intensity* and *fundamental frequency* (F0) respectively. The role of *prosody* in speech communication is to provide structure to and contextualise linguistic meaning and is manifested on a *suprasegmental* level. *Intonation*, in a broad sense,[5] refers to the use of pitch in speech communication and as such may carry *linguistic*, *paralinguistic* and *extralinguistic* information [53]. With regards to the linguistic functions of intonation, pitch patterns may have varying temporal scope (from global to local) and may be associated with various linguistic levels from *morphological* and *lexical* to *phrase*, *sentence* and *discourse*.

The surface form or realisation of intonation is a *pitch contour* influenced by the various communicative functions mentioned above, as well as *physiological* factors. The physical production of F0 is a direct result of the time-varying rate of vibration of the vocal cords, which impose physical constraints on the rate of pitch change [54] and may also be linked to patterns such as *declination* [55]. Additionally, certain speech segments are made without vocal cord vibration (i.e. without *voicing*) and the exact articulation of segments also influences physical conditions, thus having an effect on realised pitch. The resulting *fundamental frequency contour* is thus relatively smooth with interspersed segmental effects (*microprosody* due to for example plosive perturbation and *intrinsic F0* [56, 57]) and undefined (unvoiced) sections. It has however been shown that for perceptual purposes, the *pitch contour* may be considered to be continuous [30, 53].

Intonation models have to deal with the mapping of the various communicative functions to parameters representing patterns that form an appropriate pitch contour. If one includes the full set of possible influences and functions, this mapping is one-to-many, including significant speaker-specific variation [53]. Focussing on the mapping between various *linguistic functions* and surface form, there are contrasting theories and approaches to intonation and intonation modelling, with fundamental questions including:

[5] the term *intonation* may also be used more narrowly to refer only to matters of global pitch distribution [53], the use of the term *intonation* in this work corresponds with *speech melody* as used in [21]

1. What are the distinctive forms of different streams of information originating at different linguistic levels and how are these multiplexed to form the surface pitch contour?

2. What should the nature of models (and parameters) be and how should these be tied to linguistic elements?

An important distinction between models (question 1) is whether contours are seen as the result of a linear sequence of tone events (the *intonational phonology* or *tone sequence* view) or the combination of parallel patterns of differing temporal scope (the *additive* or *superpositional* view) [53]. Another significant distinction is whether model parameters are based on underlying mechanisms of speech or F0 production (*articulatory*) or directly describe the surface contour (*acoustic phonetic*) [21]. Furthermore, models may consider a finite set of forms that are symbolically represented according to linguistic theory (phonological categories), be fully stochastic and data-driven (possibly linguistic theory agnostic), and may differ in terms of functional form assumptions or simplifications (such as the *stylisation* proposed on perceptual grounds by the IPO model) [53].

Different theories of intonation assume particular mechanisms and various finite symbol sets. It is unclear how widely these theories are applicable, and whether perceptual empirical results are transferable across languages. Historically, work on intonation has often been done in the context of a single language or language family, with associated assumptions. In the next section, major intonation modelling frameworks will thus be briefly presented in order to highlight distinct assumptions made in different contexts and applications (and languages), particularly with reference to the questions presented above. The focus will thus be on model assumptions and mechanisms, rather than theory leading to linguistic functional (phonological category) notations such as ToBI [58], which is beyond the scope of this work.

2.2.1 Generative intonation modelling frameworks

In this section we discuss the following intonation models and frameworks that may be used to synthesise pitch contours:

- The Tilt model [59].

- The Fujisaki model [60].

- The Soft Template Markup-Language (Stem-ML) [61].
- The Modélisation de Melodie (MOMEL) method along with the International transcription system for intonation (INTSINT) [62].
- The Parallel Encoding and Target Approximation (PENTA) model [21].

The Tilt model of intonation is a framework for the acoustic phonetic modelling of intonation originally developed for English [59]. It models a sequence of non-contiguous intonation events (called pitch accents and boundary tones) using three continuous parameters: duration, amplitude and tilt (the tilt parameter determines the shape of the contour). Intonation events are anchored to syllables and global patterns such as *downtrend* are seen as resulting from a sequence of local event outcomes rather than a separate global phrase component. Automatic procedures are presented in order to detect, model and synthesise intonation events and classification of events into discrete symbols as in ToBI and INTSINT (described below) is avoided.

The command-response model, also known as the Fujisaki model, is motivated by speech production mechanisms, and was originally developed for Japanese [60]. This is a superpositional model where pitch contours are assumed to consist of two separate components: a slow-varying phrase component (modelling *declination* explicitly) and a rapidly-varying accent component. Each component is modelled as the result of second order linear systems with different excitation signals (or commands) and are added together in the log F0 domain to form the complete contour. While phrase components are contiguous, accent commands can occur freely and may thus be anchored to any appropriate linguistic item. Free parameters of the model are the temporal positions and magnitudes of impulses resulting in the phrase component and the temporal positions, magnitudes and durations of the step inputs resulting in the accent components. It has been shown that this model can be used to accurately represent and synthesise F0 contours in a number of languages [63]. While original applications of the model only used positive accent commands, causing positive excursions from the baseline phrase contour, it was shown that some languages including Mandarin Chinese and Swedish also require negative accent commands [63]. A procedure for automatically extracting model parameters from speech has also been developed [64].

The Soft Template Markup-Language is a phonetic intonation description and synthesis system [61]. It proposes a set of phonetic primitives (tags) containing attributes defining aspects of the F0 contour and how it is realised when embedded in continuous speech. The mechanisms controlling F0 realisa-

tion are motivated by the physiological process and the view that the surface form is a compromise between effort and communication clarity. Determining the surface form is a process of considering the interaction of tags in forward and reverse directions within definable scope, which attempts to allow for the effects of pre-planning by the speaker. The system was developed to be independent of theory and language and may be used in conjunction with different linguistic theories. As such, the exact parameterisation depends on details of the language and which tags are employed to model relevant phenomena.

The MOMEL and INTSINT methods, aiming to be universally applicable, were developed for the automatic analysis, representation and synthesis of intonation [62]. Firstly, using the system named MOMEL, microprosodic features are removed from pitch contours using smoothing and interpolation and prominent pitch targets are identified. Secondly, this sequence of pitch targets is quantised and represented as a sequence of symbols representing absolute and relative pitch targets (INTSINT). From such a discrete description, the continuous pitch contour may again be synthesised. The model thus represents a sequence of tone targets in discrete terms based on an acoustic analysis and links to linguistic items should thus be determined by the application.

The PENTA model proposes the study and modelling of intonation based on two central aspects, namely the need to encode communicative meanings and the influence of the physical production process [21]. Based on extensive empirical results, the model proposes four primitive parameters responsible for pitch realisation: local pitch targets (or *underlying form*), pitch range, articulatory strength and duration. Empirical evidence is presented from, amongst others, Mandarin and English to demonstrate how these primitive parameters may be employed to encode different communicative functions in parallel. A theory of "syllable-synchronized sequential target approximation" is also proposed and quantified in [65] to explain and implement the synthesis of complete pitch contours given these primitives. The model is thus different from the acoustic phonetic approaches described above in that it is based on *underlying form* instead of *surface form* and proposes that different languages use different, possibly complex, encoding schemes based on the proposed primitives. As an articulatory-oriented model it is distinct from the command-response and Stem-ML models in that it chooses to model the articulatory aspects in terms of the effects on the outcome of realising underlying forms (pitch target functions) and strict left-to-right (causal) assumptions respectively.

2.3 TONE IN YORÙBÁ

In linguistics, *tone* is the use of pitch in language to distinguish or inflect words; it may thus, more precisely, convey *lexical* or *grammatical* meaning. *Tone languages* may employ pitch in different ways, to various degrees and for different functions, that is, their *tone systems* may differ significantly. For example, with regards to function, East Asian tone languages such as Mandarin Chinese largely use tone for *lexical* distinction, while African tone languages often use tone for both *grammatical* (*syntactic*) and *lexical* distinction. Tone systems are also often distinguished theoretically as either *register tone systems*, using distinct pitch levels and inter-syllable contrasts, or *contour tone systems*, using distinct intra-syllable pitch movements, to encode meaning. In practice, however, it is more difficult to classify languages in this way and some languages, such as Cantonese, may use a combination of both mechanisms (levels and contours). Lastly, the extent to which tones are responsible for distinguishing meaning, the *functional load*[6] of tones in tone languages may vary.

Yorùbá is considered to have a register tone system with three distinct tones. These level tones (labelled High (H), Mid (M) and Low (L)) are associated with syllables, have a high *functional load* and are said to exhibit a *terracing* nature [26]. *Terracing* refers to an utterance-wide pattern where distinct 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*. Downstep and pitch resetting are pitch changes occurring in local contexts, while declination refers to a gradual lowering of pitch independent of local context. Previous investigations into the effects of these mechanisms on pitch contours suggest that the utterance-wide pitch contour in Yorùbá is largely dependent on a combination of local pitch changes (and therefore the tone sequence) and that gradual declination plays a relatively minor role [9, 66].

In addition to *tonemic* level tones, distinct intra-syllable pitch patterns in Yorùbá are *falling* and *rising* contours when L and H tones are realised after H and L tones respectively [9]. These tone realisation (phonetic) patterns and others, such as *dissimilative* H raising before L and *final lowering* where the pitch level is lowered in phrase-final positions despite tone identity, are likely to be perceptually important [9, 66, 67].

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` and the nasal-

[6] a linguistic concept analogous to *entropy* in *information theory*

isation of certain vowels followed by 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 [68]. 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) may also be used).

2.3.1 Related work on intonation modelling of Yorùbá

Recent work on the realisation of tone in Yorùbá for the development of a speech technology has been described by Ọdẹ́jọbí et al. Their work involved the development of two models (based on Stem-ML [61] and a novel rule-based approach) for the synthesis of F0 contours suitable for use in a speech synthesiser [68, 69, 14]. In this model each syllable is represented by a stylised pitch contour based on a third order polynomial parameterised by its peak and valley. Relative heights are then determined locally by phonological rules based on two-syllable contexts described in [70], thus considering co-articulation given the previous syllable, and globally using constraints motivated by assumed downtrend and implemented using a hierarchical data structure (S-Tree). The assumption of continued downtrend is then also used to combine sub-trees into a single structure preserving relative height in the case of multi-phrase utterances. Absolute values of pitch are then obtained using a sophisticated model based on the exponential decline of pitch (in Hertz) over the course of an utterance, with tone-specific asymptotes and parameters estimated from data using a fuzzy logic framework and taking into account the observations in [9, 66].

2.4 DISCUSSION

While more specific and detailed discussions motivating aspects of this work will be presented in each chapter, a brief discussion follows here based on the background given in this chapter with reference to the research questions proposed in Section 1.2.

Firstly, while a number of descriptive acoustic phonetic analyses provide insight into the tone system of Yorùbá, there is little quantitative information on tone realisation presented in the context of speech technology development. That is, while some of the surface forms of tones have been described in carefully designed studies, no attempts have been made at determining the reliability of these

observations in general. In particular, the inter- and intra-speaker variability of pitch patterns need to be investigated to determine reliable tone indicators in different tone contexts based on features that can automatically be extracted from general continuous utterances.

In Chapter 1, citing difficulties by researchers to construct reliable corpus-based TTS systems in this context, it was noted that pitch modelling for appropriate tone realisation has not received a sufficient amount of attention. Given the background presented here, we argue that ensuring correct acoustic realisation of tone using the data-driven approaches presented in Section 2.1 is non-trivial given limited speech databases, lack of quantitative information on tone realisation and the composite nature of surface pitch contours.

Furthermore, the prospect of developing large high-quality speech corpora suitable for TTS development (many hours of audio [32, 33]) in some under-resourced languages is hampered by a lack of available textual content and especially appropriately digitised text [71, 72]. Thus, even for commercially and politically important languages such as Yorùbá (with a large number of speakers) the development of such corpora is still a future goal, while for some "smaller" languages that have to compete directly with (technologically) established "world languages" such as English and French, the eventual development of resources of this proportion is not a certainty.

Also, while some of the intonation modelling frameworks presented (Section 2.2) have the potential to model tone realisation in this context, to date, very little work has been done on African tone languages within these frameworks [69, 73, 74, 75]. Consequently, successful pitch modelling in this context is still an open problem.

Regarding pitch modelling for Yorùbá in particular, we argue that while the specialised model developed in [14] may ease the data requirements necessary to build an appropriate intonation model, there are a few aspects that may benefit from further work:

- The stylised acoustic representation of tones with parameterisation in terms of peak and valley may not be optimal, especially in the case of M tones where it is noted that the peak and valley often have the same value [14].

- The phonological rules modelling the effects of local co-articulation are based on [70] and only takes into account the previous syllable [14]. The work of Akinlabi [67] suggests that a larger context may need to be considered.

- In our opinion, the models accounting for relative and absolute pitch heights seem to contain some redundancy, the fact that predictions from subsystems may diverge [14], suggests the possibility of a more constrained or simplified model.

The primary goal of this work is to support the development of "tone-aware" speech technologies for African tone languages in general and Yorùbá in particular. The focus is thus firstly on understanding and suitably representing and modelling the crucial aspect of tone realisation in this context. This however only constitutes one aspect of a complete intonation model, and an attempt will be made throughout this work to develop this aspect within the framework of a complete model. Thus, the outcome should be a model that robustly synthesises pitch for tone realisation while being complete in the sense that it also exhibits the natural patterns of "neutral" intonation (e.g. *downtrend*), to the degree that it may be integrated into a functional TTS system.

As motivated by the naturalness and flexibility requirements of potential applications (Chapter 1), the approaches followed in this work will be data-driven as far as possible in order to faithfully reproduce speaker-specific properties of intonation. However, the under-resourced context within which this work is done will serve to motivate the reduction of free parameters, and thereby data-requirements, where possible. Interpretable parameters and models will also be considered for their potential to be adapted relatively easily for other linguistic functions based on non-ideal or cross-language data, or even theoretical rule-based manipulation in future work. This will be important for the rapid development of speech synthesis systems in this context to support advanced applications such as speech dialogue and concept-to-speech systems considering the demanding data requirements of such applications using current automatic acoustic modelling techniques (Section 2.1).

CHAPTER 3

TONE REALISATION IN YORÙBÁ

In this chapter we attempt a general description of tone realisation based on statistical analysis of a multi-speaker speech corpus, relying on automatic text processing, phone alignment and acoustic feature extraction. The nature of acoustic features, particularly F0, is described in different tone contexts to quantify aspects such as speaker-specific variation and co-articulation in continuous utterances. These aspects need to be investigated for the development of appropriate acoustic models that may be used directly in speech technologies. The focus here is on investigating local acoustic patterns resulting from tone realisation and its interaction with other aspects of the utterance.

3.1 APPROACH

The investigation presented here is guided to some degree by previous acoustic analyses of Connell and Ladd [9] and Laniran and Clements [66] which aimed to systematically investigate linguistic concepts such as *downstep* and *high tone raising* alongside effects such as *declination*. The experimental setup, especially with regards to how the text is processed, is informed by the work of Ọdẹ́jọbí et al. (see Sections 2.3 and 2.3.1 in Chapter 2) and the analysis follows in a similar manner to work done by Barnard and Zerbian on Sepedi, a Southern Bantu language spoken in South Africa [76].

The above-mentioned studies relied on relatively small samples (3 to 4 speakers) based on carefully designed corpora in order to answer questions about tone realisation in specific utterance contexts. Here we attempt a more general analysis, part of which has been published previously in [77]; the F0 measurements have, however, been updated to be based on semitones instead of Hertz units. The use of a logarithmic scale results in F0 contour shapes becoming approximately invariant at different absolute pitch levels [28, 60, 65], which is suitable for the analyses involving averaging presented

later (e.g. see Section 3.3.1 and Figure 3.4) and corresponds to the way in which pitch is perceived. A linear scale would result in perceptually similar pitch patterns being stretched or contracted depending on the absolute pitch (or "key") of the speaker.

In the following section we describe our experimental setup and processing based on detail of the Yorùbá tone system described in Section 2.3. This is followed by experiments and results in Section 3.3. Finally, we conclude in Section 3.4 with a discussion, motivating work presented in the next chapter.

3.2 EXPERIMENTAL SETUP

3.2.1 Corpus alignment

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 Lagos, Nigeria and North-West University, South Africa. Each speaker recorded between 115 and 145 short utterances (sentences and sentence fragments) from the pool of selected sentences, amounting to about 5 minutes of audio per speaker. The audio is broadband, collected in Nigeria using a microphone attached to a laptop computer. In some cases significant background noise is present; data of one speaker was omitted from the 34 speakers considered 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 (Section 2.3). Similarly, a simple syllabification algorithm was implemented and syllable tones were obtained from the diacritics.

Based on this information we performed automatic phonemic alignment of the audio by forced-alignment of Hidden Markov Models (HMMs) using HTK [43]. As is standard practice in phone alignment for corpus-based TTS development, speaker-specific HMMs were trained on the data to be aligned. Here we followed a procedure similar to the one described in chapter 3 of the HTK book [43]. Mel-frequency cepstral coefficients (MFCCs) with delta and acceleration coefficients were extracted using 10 ms windows with a frame shift of 5 ms as acoustic features using HTK's `HCopy` (see Table A.1 in Appendix A for the exact configuration). Considering the limited number of utterances available for each speaker, we followed a process of careful initialisation of the 5-state HMM models instead of the conventional "flat start" approach. This involved pooling data from the

manually aligned TIMIT corpus [78] into broad phone classes, initialising models using Viterbi re-estimation (using HTK's `HInit` and `HRest`) and copying these for all Yorùbá phonemes pooled in the same way (see Table A.2). After this initialisation procedure standard embedded (Baum-Welch) re-estimation was done (`HERest`). This was followed by a re-alignment where the decoder (`HVite`) was used to optionally insert pause models between words, after which further re-estimation was done. At this point monophone models were cloned to form triphones, re-estimated, and decision-tree clustered tied-state-triphone models were used during the final stage of forced alignment. Alignments were then post-processed to remove all pauses inserted that were shorter than 100 ms and insert a breath-group boundary marker where pauses were longer than 300 ms. While we did not re-investigate the parameters and procedure employed here during phone alignment in detail, the setup is based on extensive previous work reported in [79, 80] and used during the development of TTS systems for 11 languages in South Africa [71, 81]. Consequently, we expect relatively accurate phone alignments depending mainly on the accuracy of the transcriptions and grapheme-to-phoneme rules.

Lastly, in an attempt to discard samples where automatic alignment might have failed, especially due to mismatches between transcriptions and audio (Yorùbá is known to have significant cases of vowel assimilation and elision [9, 82]), we only retained utterances where all syllables have durations of more than 30 ms. 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 investigation.

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 (the tone counts were H: 12777, M: 10743, L: 11050).

3.2.2 Acoustic feature extraction

To extract F0 and intensity contours, we used *Praat* [83]. For the F0 contours we estimated values every millisecond using the autocorrelation method, and applied a small amount of smoothing to reduce measurement (estimation) noise; 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 narrower, more suitable, ranges (see Table B.1

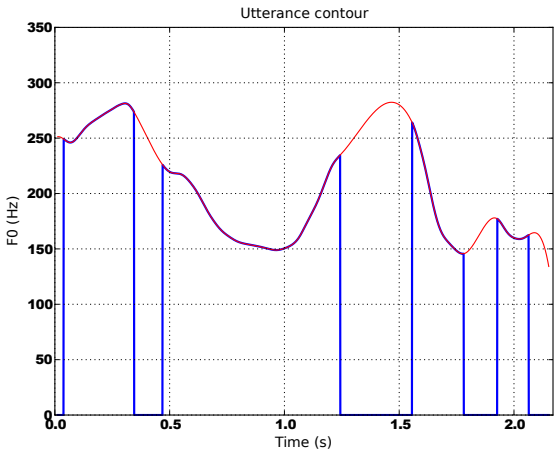

Figure 3.1: *Example of spline interpolation for an utterance F0 contour, the originally estimated contour is in* blue *with the interpolated contour in* red.

in Appendix B). Finally, all F0 values in Hertz were converted to semitones (relative to 1 Hz) before further processing:

$$F_{0_{st}} = 12\log_2 F_{0_{Hz}}$$

For each utterance contour extracted, we use cubic spline interpolation to obtain non-zero F0 values for unvoiced regions (see Fig. 3.1). For the intensity contours we similarly extracted values every 1 ms using Praat, setting the minimum pitch (used to determine the analysis window size) to the minimum pitch determined for the speaker.

3.2.3 Reliability of setup

For an indication of the reliability of the resulting data 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 the waveform, spectrograms and F0). For alignments, gross errors were counted when any syllable region did not correspond by at least 50% of its duration to the corresponding label. This occurred in the following ways:

1. A phone was predicted from the transcriptions without having been realised in the audio.

2. A sound was realised in the audio without having been predicted from the transcriptions.

3. A phone boundary for a predicted and realised phone was misplaced so that less than 50% of

either of the adjacent regions represented the corresponding predicted syllable.

Case 1 occurred either because of phone elision or assimilation, which is common in Yorùbá, or due to a transcription or pronunciation prediction error. Although assimilation occurs often to varying degrees, this was only rarely classified as a significant mismatch considering the complete syllable duration. Transcription errors and grapheme-to-phoneme conversion mismatches accounted for the majority of distinct mismatches. Case 2 occurred during speaker disfluencies or simple reading or transcription mismatches, these cases were also easily identified. Case 3 occurred rarely in the absence of a transcription mismatch, but sometimes adjacent to cases 1 or 2.

For F0 extraction, identifiable errors typically rendering more than 50% of the extracted values in a syllable erroneous resulted from either voicing or octave errors. Rates for errors identified in this way are reported in Table 3.1.

Alignment error rate	$\approx 5\%$
F0 error rate	$\approx 8\%$
Total number of syllables	355

Table 3.1: *Gross error rates observed in a small subset of the corpus.*

The main goal in this chapter is an investigation of tone realisation and reducing the error rates observed above by either selecting a more reliable subset or manual intervention may assist in this process. However, considering the greater aim of this work to support speech technology development (see Section 1.3), we stop at quantifying potential errors and continue the investigation, explicitly refraining from manual intervention or biasing the corpus towards more ideal utterances. Re-iterating, the only interventions that have been applied to the original corpus is the discarding of one speaker's data due to power line noise and the discarding of utterances where alignments indicate that the orthographic transcriptions are likely incorrect (described in Section 3.2.1).

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

3.3 EXPERIMENTAL RESULTS

In this section we present measurements on the corpus to characterise the acoustics of tone realisation. The initial investigations are organised in sections considering pitch, intensity and duration separately. This is followed by additional experiments regarding pitch.

3.3.1 General observations of pitch

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 [76]. Thus, whether it is the general trend that H $>$ M $>$ L with regard to mean absolute F0 values of syllables and whether mean change in F0 is a good indicator of tone, i.e. may be used to distinguish different tones. This is discussed more formally in Section 3.3.4 where we compare tone classification rates for different acoustic features. One advantage of using the mean F0 instead of more direct measurements such as maximum or minimum F0 is improved robustness considering the influence of both alignment and F0 estimation errors.

Determining the mean F0 for syllables based on interpolated contours results in 31 of the 33 speakers indeed having H $>$ M $>$ L (see Table B.1 in Appendix B), with typical distributions as shown in Fig. 3.2 (using speaker `08` as an example). Most speakers seem to have 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.

In [76], overlap between mean absolute F0 values was to some degree explained by the general trend of declining F0 within an utterance and the difference in mean F0 between syllables was found to be a good indicator of tone. The general trend of declining F0 (referred to as *downtrend*) is also found in our corpus: we performed a least-squares linear fit for each utterance with the result that 2940 of the 3435 utterances have a declining trend with a mean gradient of about -1.36 st/s over the complete corpus. Obtaining a robust utterance-wide estimate for expected F0 level due to *downtrend* is however reliant on understanding phenomena such as *downstep* and *pitch reset* [9, 66]. This is investigated explicitly in Chapter 4.

In Figure 3.3 we plot the distributions of change in mean F0 between syllables for different transitions (means are summarised in Table 3.2). Histograms obtained for single speakers are generally similar in nature to the corpus-wide distributions. Interesting features of these distributions are that MM transitions are relatively strongly centered around zero, and that the contrast between M tones and other tones in terms of the change in mean F0 is more consistent than HL and LH transitions.

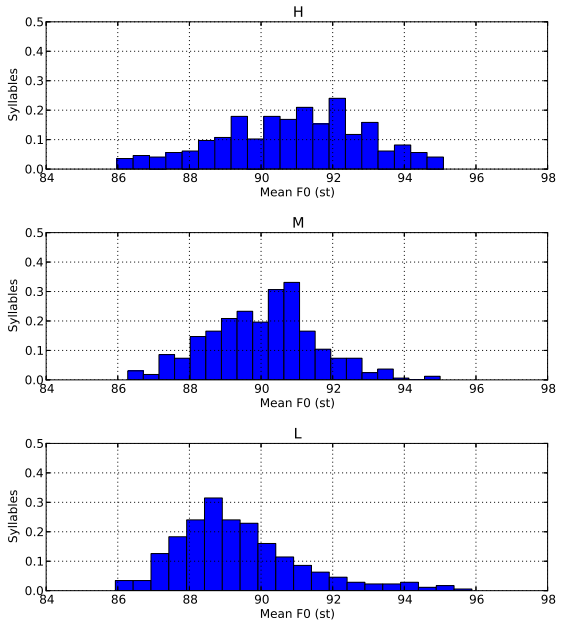

Figure 3.2: *Example of mean F0 distributions for syllables of each tone by a female speaker (*08*). The x and y axes indicate the mean F0 in semitones and fraction of all syllables respectively.*

Connell and Ladd [9] specifically mention HL and LH transitions as cases where the L and H tones are realised by falling and rising tones respectively. In Table 3.2 we present the mean intra-syllable gradients obtained by least squares linear fit for the three tones in different preceding tone contexts, confirming that H-L and L-H cases exhibit among the steepest gradients. While M-L and L-M also exhibit relatively steep gradients on average, H-L and L-H are noticeably steeper and we attempt to gain further insight into how these tones may be distinguished by investigating the contour shapes below and discriminative potential of acoustic features (including intra-syllable gradient) in Section 3.3.4.

One can also see that the change in mean F0 is in some cases less distinctive here than the intra-syllable gradient: see for example the overall small negative pitch change for LH transitions in Table 3.2, this is also reflected in the distributions in Figure 3.3. One reason is the fact that the mean is taken over the entire syllable and this aspect is revisited in more detail in Section 3.3.4.

As it seems possible that the perception of tone is in some instances reliant on the movement of F0 within the syllable depending on tone context, we attempt to investigate this property by considering the contours of tri-tones; calculating mean contours over three-syllable sequences in the corpus in a

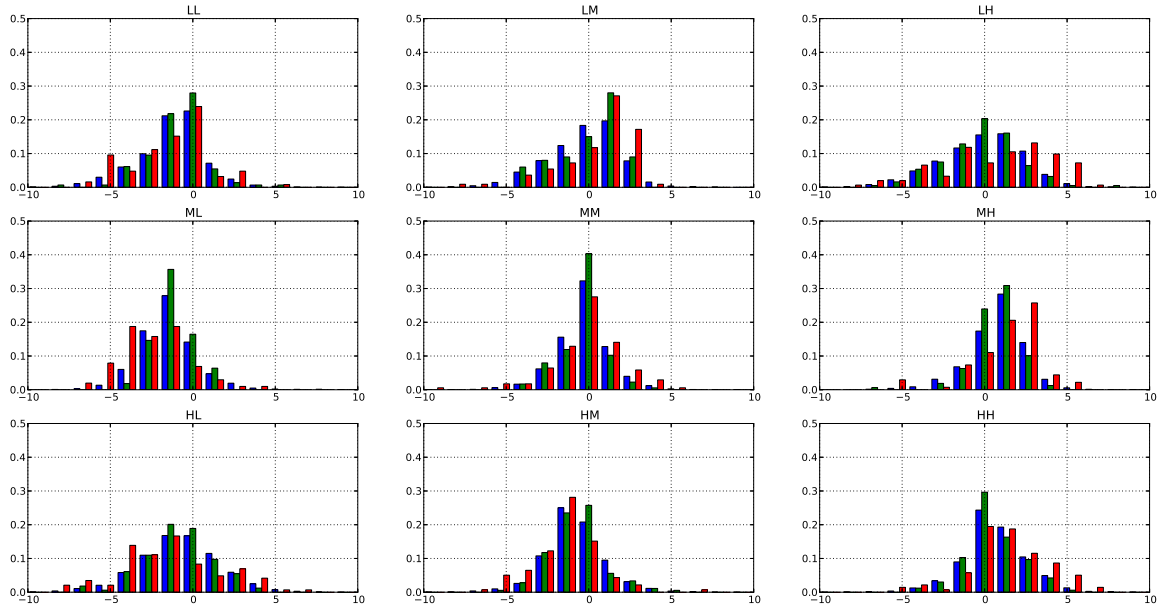

Figure 3.3: *Distributions of change in mean F0 between syllables for different tone transitions; blue bars are calculated over the entire corpus, while green and red bars are examples of a female (*`08`*) and male (*`23`*) speaker respectively. The x axis is the change in mean F0 in semitones and y the fraction of samples. Not all samples in the corpus fall into these ranges.*

manner similar to [84]:

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 N_c contours for each context c (27 contexts). These contours are extracted across word boundaries (the number of instances and source utterances are reported in Table B.2).

2. Each contour is resampled using cubic spline interpolation to normalise lengths to N_T samples (linear time normalisation).

3. A mean contour is calculated from each set of contours (in each context c), where x is the F0 value, i indexes the contour and j the normalised time instant:

$$\bar{x}_j = \frac{1}{N_c} \sum_{i=0}^{N_c-1} x_{ij} \quad \text{where } 0 \leq j < N_T \tag{3.1}$$

From plots of these mean contours (Fig. 3.4) we observe a few general properties of tone realisation

Tone and context	F0 gradient (st/s)	Number of syllables	Tone transition	Change in mean F0	Number of transitions
H	7.34	12777	*H	0.52	11621
N-H	10.20	1156	NH	—-	—-
H-H	2.65	3801	HH	0.72	3801
M-H	8.36	3667	MH	0.93	3667
L-H	9.93	4153	LH	-0.01	4153
M	-1.86	10743	*M	-0.38	9565
N-M	4.78	1178	NM	—-	—-
H-M	-7.66	3560	HM	-0.69	3560
M-M	-1.59	3611	MM	-0.23	3611
L-M	3.08	2394	LM	-0.14	2394
L	-11.32	11050	*L	-1.06	9949
N-L	-2.39	1101	NL	—-	—-
H-L	-16.84	3958	HL	-0.64	3958
M-L	-15.78	2393	ML	-1.75	2393
L-L	-5.012	3598	LL	-1.06	3598

Table 3.2: *Mean F0 gradient within a syllable and mean change in mean F0 between syllables for different tones, including different preceding tone contexts,* N *denotes "none", i.e. the initial syllable in each utterance, and* * *denotes any preceding tone.*

of the three tones in different contexts:

1. L tones generally have a negative F0 gradient regardless of different contexts, with mean higher when preceded by a H tone, in which case it generally has the steepest gradient. This agrees with the falling tone described in [9].

2. The mean contour for M tones is relatively flat to declining, with notable exception when preceded by a H tone. Inspection of the MMM sequence seems to confirm the "backdrop declination" found by Connell and Ladd [9].

3. H tones generally have steep positive gradients when preceded by L tones and peaks seem to be relatively raised when followed by L tones (the rising tone [9] and high raising [66] phenomena). Peaks are generally realised late in the syllable (possibly even in the following syllable, which is a phenomenon also noted in [67]), 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.

4. Comparing the beginnings of these mean contours with their endings, we see indications that

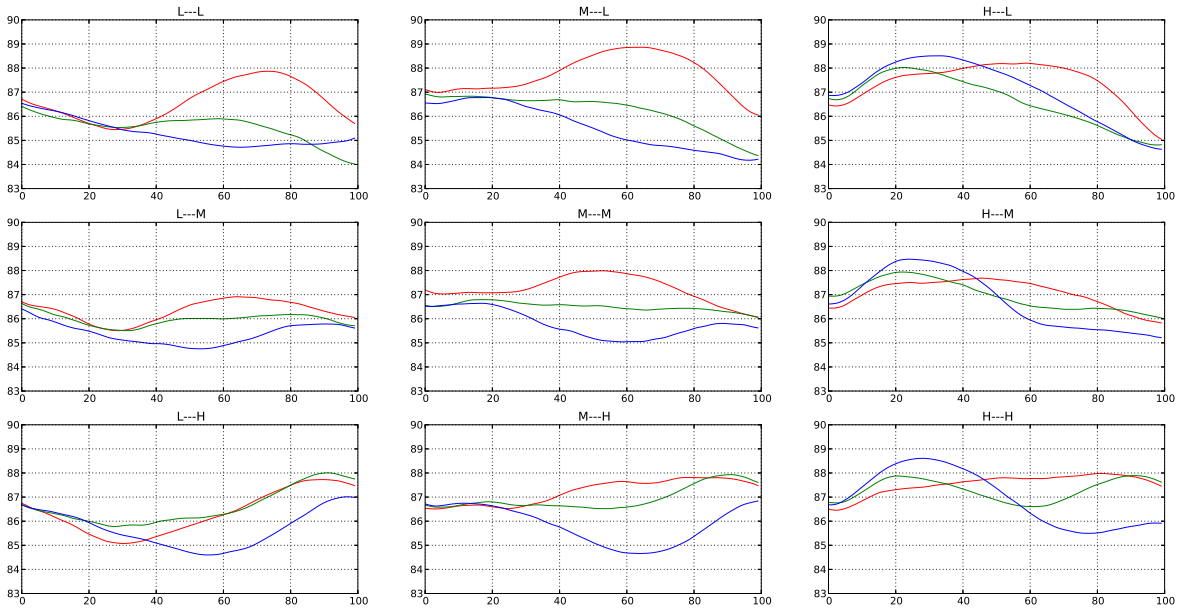

Figure 3.4: *Mean F0 contours for three-syllable sequences with the different tones H (red), M (green) and L (blue) in different tone contexts (x is the normalised time and y the F0 in semitones).*

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 [21] that "carryover assimilation" is more significant than "anticipatory effects" in Mandarin tone realisation.

Given the characterisation of F0 contours in different contexts (Fig. 3.4) we attempt to measure and visualise the variation of contours in these contexts. This is done by normalising the location of each contour by subtracting the mean from each contour (row-wise) to obtain a set of N_c contours $x_{n_{ij}}$ in each context,

$$x_{n_{ij}} = x_{ij} - \bar{x}_i \quad \text{where } 0 \leq i < N_c \text{ and } 0 \leq j < N_T \quad \text{and } \bar{x}_i = \frac{1}{N_T} \sum_{j=0}^{N_T-1} x_{ij} \tag{3.2}$$

and determining the standard deviation at each point j:

$$s_j = \sqrt{\frac{1}{N_c} \sum_{i=0}^{N_c-1} (x_{n_{ij}} - \bar{x}_{n_j})^2}, \text{where } 0 \leq j < N_T \text{ and } \bar{x}_{n_j} \text{ calculated as in Eq. 3.1} \tag{3.3}$$

The result may be plotted for each context as in Figure 3.5.

We firstly consider the source of variation caused by the embedding of these contours in the larger (utterance) context. By examining the initial and final segments of these plots, it is again evident that carryover assimilation is more significant in terms of the proportion of syllable duration affected than

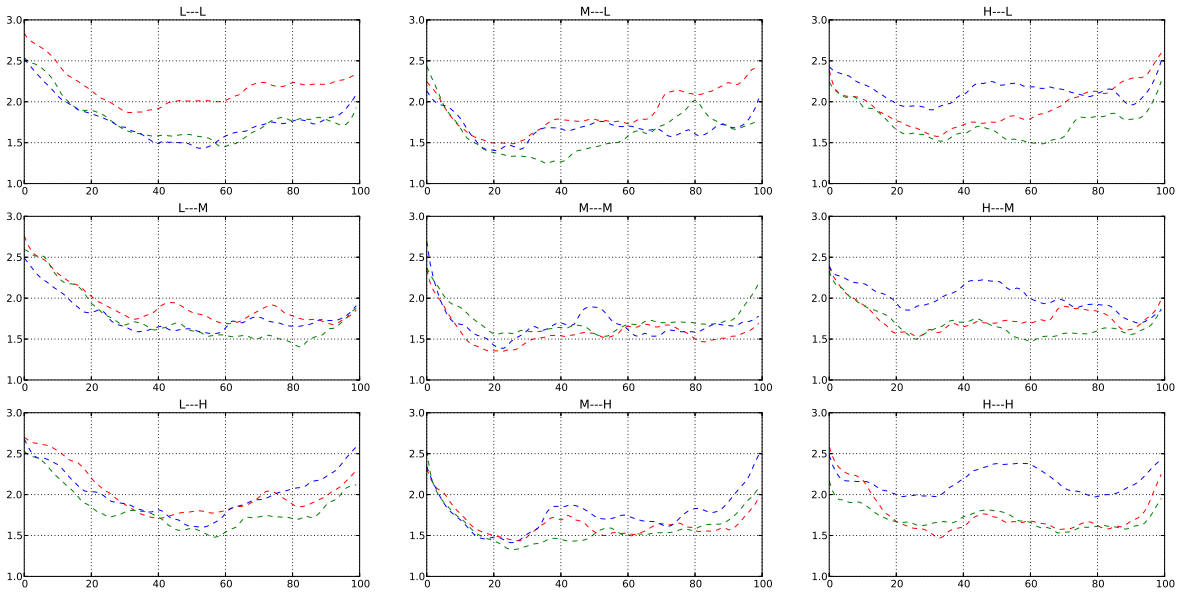

Figure 3.5: *Standard deviation contours for three-syllable sequences with the different tones H (red), M (green) and L (blue) in different tone contexts (x is the normalised time and y the F0 in semitones).*

anticipatory effects: The variance increases systematically towards the edges at the start and end of the window, with the extent in terms of duration being larger at the start of the window than at the end. This observation is in agreement with observation 4 made based on Figure 3.4. It is important to note however that systematic fine alignment inaccuracies resulting from the automated procedure might affect the variation measured here. Consequently, we would suggest (especially in the case of contours ending in M) that there is little concrete evidence for anticipatory effects in our corpus based on this analysis other than the interaction between H and L tones (e.g. dissimilative H tone raising [66]).

A second significant source of variation as measured here is a result of peak variation exacerbated by alignment inaccuracies and the inadequate linear time normalisation procedure (applied over the course of three syllables). The distribution of peaks and valleys illustrated in Figures B.1 and B.2 in Appendix B indicate significant variation, in part because of the general difference in durations of different syllable types; syllabic C, CV or V (see Fig. 3.8 showing duration distributions for CV and V). This is responsible to a significant extent for the overall higher variance in LHL and HLH contexts (where small temporal misalignments may result in higher variance as measured here than for less dynamic contexts) and also to some degree for the increase in variance in the final parts of contexts ending in H. Further attempts to reduce this source of variability by non-linear time normalisation

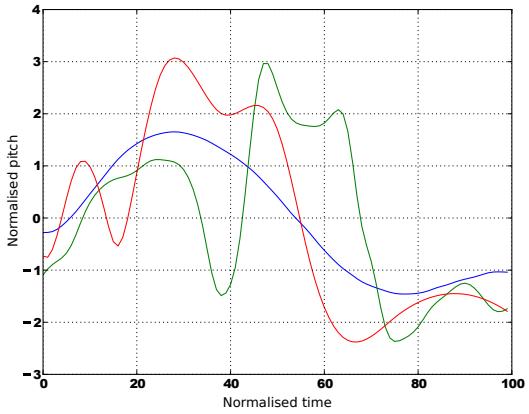

Figure 3.6: *Example of a contour (red) resulting from non-linear time normalisation based on DTW alignment of an original contour (green) against the reference (blue). This example is for an HLH sequence.*

based on dynamic time warping (DTW) alignment against the expected contours of Figure 3.4 (e.g. Fig. 3.6) did not result in robust results over all samples and syllable types and good estimates could not be obtained when partitioning the data based on syllable types due to the increased number of permutations.

Further investigation of the mean contours from different parts of utterances (i.e. initial, medial or final) resulted in patterns similar to those plotted in Figure 3.4, differing largely in height. It should be noted again however that the current corpus only represents short utterances ($\leq$ 10 words), thus limiting the scope of this analysis to approximately one breath-group.

3.3.1.1 Speaker-specific variation

The above views on the data may be repeated on the limited subsets of data from each specific speaker; plotting mean contours generally resulted in similar contours (see Fig. 3.7 for two examples; one female and one male speaker) with some variation in the location of peaks, possibly due to the effect of speech rate differences.[1] We also quantify the intra-speaker variation of all the three-syllable

[1] To clarify, we are not suggesting that speech rate variation may affect the alignment of peaks to syllables, but rather may affect the location of peaks in our linearly time-normalised three-syllable contours, e.g. due to other processes such as elision in faster speech.

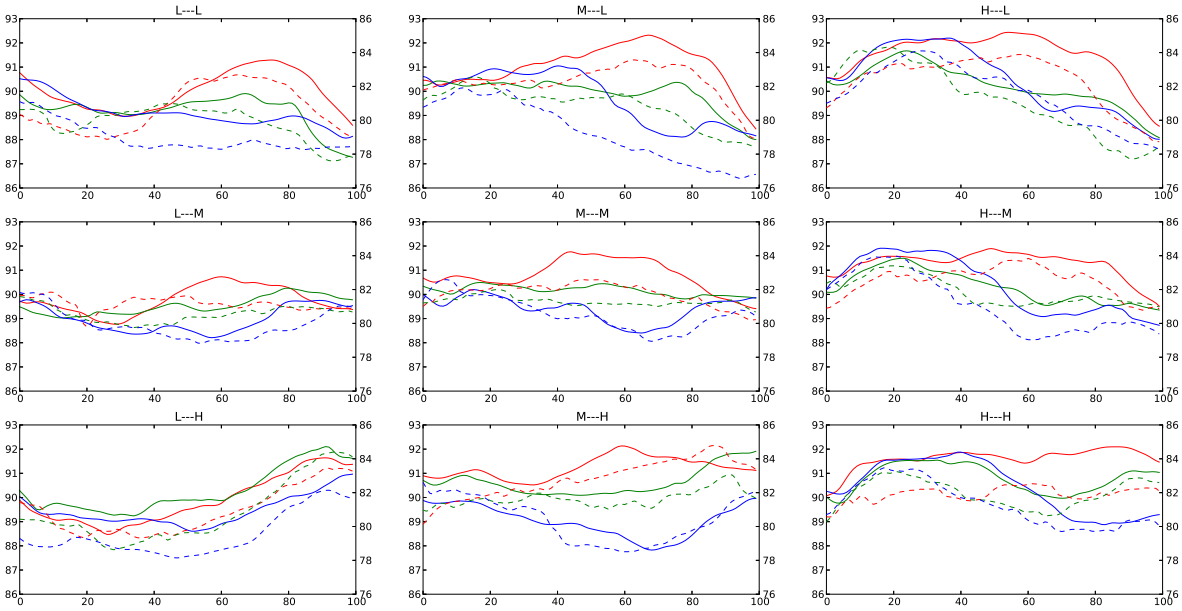

Figure 3.7: *Mean contours for three syllable sequences with the different tones H (red), M (green) and L (blue) in different tone contexts. The solid and dashed lines represent examples of a female (08) and male (23) speaker, with y-axis values indicated on the left and right respectively (x is the normalised time and y the F0 in semitones).*

contexts using Eq. 3.3. The results are summarised by taking the mean over j for each context

$$\bar{s} = \frac{1}{N_T} \sum_{j=0}^{N_T-1} s_j$$

and presented in Figures B.3 and B.4. These measurements make it possible to determine the relative intra-speaker variability and indicate that the pitch contours in tone contexts for speakers 02 and 13 are more variable than speakers 04 and 14 and speakers 20 and 29 are more variable than speakers 30 and 31 for the female and male speakers respectively. This measure of variability is sensitive to variation in dynamic range, temporal differences (phase and non-linear differences) and deviation from the prototypical (mean) contour.

In an attempt to quantify the relative degree to which each speaker's mean contours correspond to the overall mean contours, i.e. the relative "prototypicality" of each speaker's tone realisation in these contexts, we calculate the location normalised mean contours in each context (Eq. 3.2 followed by Eq. 3.1) for the speaker and over the corpus and range normalise the resulting contours:

$$\bar{x}_{r_j} = \frac{\bar{x}_{n_j}}{r} \quad \text{where } 0 \leq j < N_T \quad \text{and } r = \max_j \bar{x}_{n_j} - \min_j \bar{x}_{n_j}$$

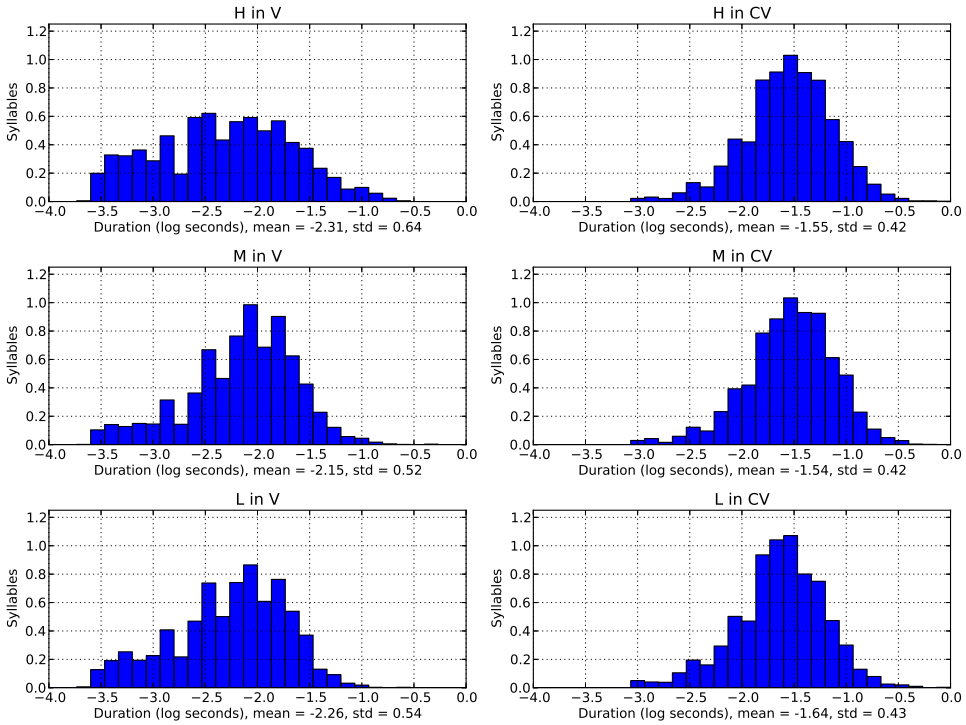

Figure 3.8: *Distribution of all syllable durations (in the log domain) for different tones in CV and V syllables. The means and standard deviations are given.*

These contours are then compared (speaker-specific against corpus-wide) by calculating the root mean squared error (RMSE) over the dynamic time warping (DTW) alignment of the contours. This procedure serves to reduce the variation due to temporal and dynamic range differences. The results are presented in Figs. B.5 and B.6. Here we can see that speakers `01`, `09`, and `15` differ most from the norm[2]. Some contexts also seem to exhibit more variation than others, especially contexts containing M tones. We note that the proportion of CV to V syllables is smaller in the case of M tones than H tones and may thus be the source of additional variation, especially in contexts with interacting H and M tones. These scores and inspection of some contours compared to the RMSEs calculated in this way confirm that most speakers' mean contours agree with the global patterns established in Figure 3.4.

[2]that is the tonal patterns exhibited by these speakers are least consistent with the mean contours derived over the whole corpus

3.3.2 General observations of duration

To examine the distribution of syllable durations associated with tones, we consider the different syllable types separately because of the expected differences in duration for different syllable types and non-uniform distribution of syllable types and tone combinations. Of the valid Yorùbá syllables; V, CV and C (the syllabic nasal), we inspect V and CV for different tones as we do not have enough samples to consider syllabic C.

Figure B.7 shows the mean durations for syllable type and tone combinations for each speaker. The first observation is that measurements are more variable in the case of V than in the case of CV. This is also evident in Figure 3.8 and illustrates the expected additional measurement noise due to the effects of vowel assimilation and elision on expected durations and possibly due to decreased alignment accuracy because of inadequate modelling – note the skew towards shorter syllables. Secondly, L tone syllables tend to be shorter than other syllables and M tone syllables tend to be relatively long.

For the purpose of modelling tone realisation, a detailed analysis of the effects of tone on instantaneous (or local) speech rate as continuous functions is desirable. A robust way of estimating such a continuous speech rate, however, requires an alignment of each utterance against a reference [85]. While the current corpus includes neither repeated utterances nor utterances considered to be of neutral speech rate, we experimented with creating such reference utterances. This was done by training an HMM-based statistical parametric model using the standard HTS setup [86, 87][3] for each speaker and performing DTW alignment of each utterance against a synthesised version[4] serving as our reference (essentially a maximum-likelihood estimate; see Section 2.1.2). Results from this analysis, however, proved too noisy to get detailed insights into the effects of tone realisation on speech rate.

Based on the current analysis it is clear that syllable duration in our corpus primarily depends on phonological aspects such as syllable structure, vowel assimilation and elision, with weak evidence for longer M syllables and shorter L syllables in general. The exact influence of tone on syllable duration requires further investigation; however, we speculate that the relationship may be related to the physical constraint of maximum possible speed of pitch change, where upward changes have been found to be slower than downward changes in general [54]. The exact causality might be dependent

[3] using the standard demonstration script for HTS version 2.2 available at: `http://hts.sp.nitech.ac.jp/`

[4] using the HTS engine version 1.05 available at: `http://hts-engine.sourceforge.net/`

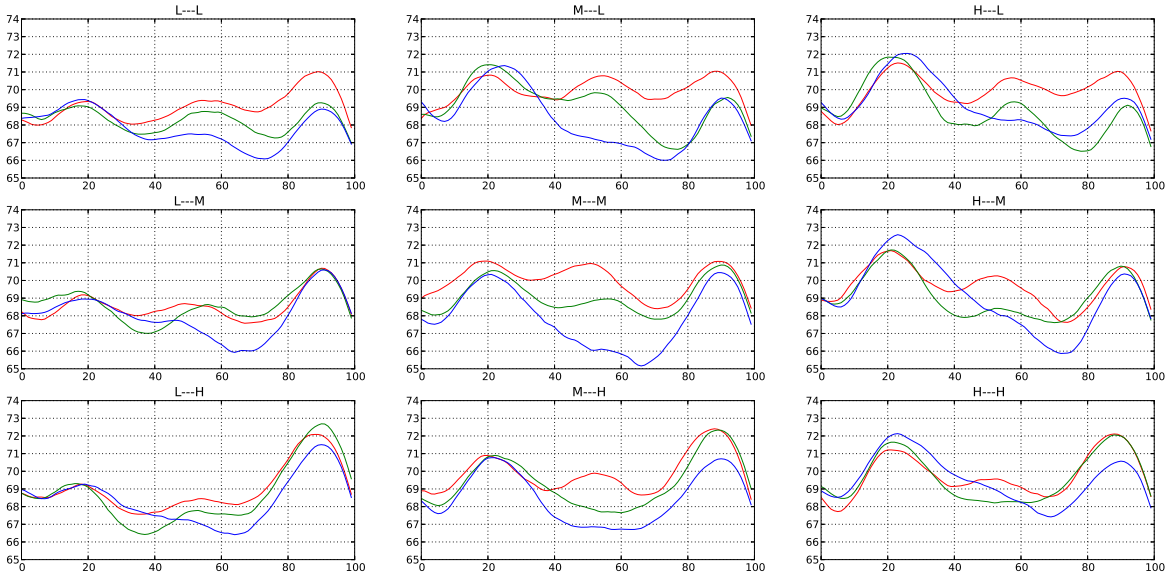

Figure 3.9: *Mean intensity contours for three-syllable sequences with the different tones H (red), M (green) and L (blue) in different tone contexts (x is the normalised time and y the intensity in decibels).*

on the message context where the speaker presumably has a choice between compromising perception (by missing perceptual pitch targets in maintaining a high speech rate) or temporal efficiency (reducing speech rate by lengthening syllable duration to achieve perceptual pitch targets). If this is the case, the trend in our data – shorter L tones and longer M and H tones – would be consistent with speakers keeping a relatively high speech rate without compromising perceptually important pitch movements.

3.3.3 General observations of intensity

As is the case with F0, the overall intensity declines over the course of an utterance (in 2904 of 3435 utterances with an average change of -2.69 dB/s). For this reason we also need to consider the dynamics of intensity in different tone contexts. To examine this we repeat the analysis of Eq. 3.1 on the sets of intensity contours; shown in Figure 3.9. We see that the intensity is often higher in H syllables and usually lower in L syllables. The valleys and peaks that are evident correspond with the general locations of onset and codas of syllables respectively. The absence of peaks in some L contexts suggests that L syllables are often de-emphasised.

Investigating the covariance of intensity and F0 shows that speakers differ regarding the degree to

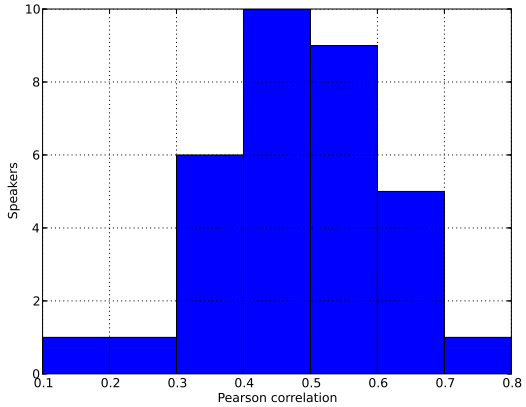

Figure 3.10: *Pearson correlation coefficients between the mean F0 in each syllable and the mean intensity in the syllable nucleus (measured in the vowel of the syllable) for different speakers.*

which intensity is varied with pitch (summarised by the correlation coefficients in Figure 3.10). Closer inspection shows that speakers may also be using intensity to a greater or lesser degree to indicate tone, especially in the case of H and L tones. Figure 3.11 shows scatter plots for different speakers, illustrating: little usage of intensity (speaker `08`), a stronger correlation over all syllables (speaker `23`), a stronger correlation for higher F0, especially for H tones (speaker `12`), and a stronger correlation for lower F0, especially for L tones (speaker `31`). Some of these results, such as cases suggesting L syllable de-emphasis and emphasised H tones may be tied to some extent to other causes than pure tone realisation, such as relative word prominence where raised intensities might be focussed on H syllables.

Thus, while tone realisation may primarily depend on F0, it seems possible that speakers can to some degree implement the necessary contrasts or perhaps enhance perception by using complementary change in intensity.

3.3.4 Tone indicators

In this section we attempt to systematically quantify and compare the relationship (level of association) between some of the properties of F0 noted in Section 3.3.1 and the extracted syllable tone labels. To do this we construct a simple Gaussian classifier, using a single Gaussian distribution estimated by maximum likelihood, to model the class-conditional densities of a given feature X for each

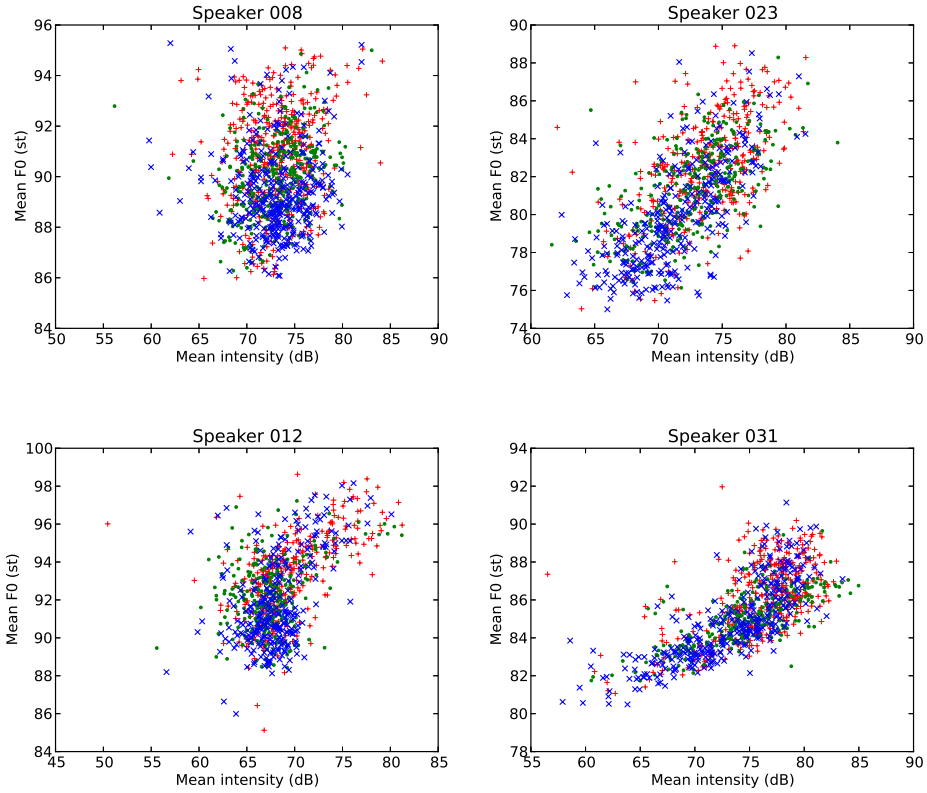

Figure 3.11: *Examples of the covariance between mean F0 and mean intensity in each syllable for four different speakers. Mean intensity was calculated in the syllable nucleus.*

class c in the set C:

$$X_c \sim \mathcal{N}(\mu_c, \sigma_c^2), \tag{3.4}$$

the set of classes C being the three tones ($C = \{H, M, L\}$). Despite the fact that the proposed classes are not equiprobable in our corpus, we use a simple decision rule dependent only on the class-conditional densities to perform classification. The decision rule with this simplification (equivalent to the Bayes decision rule with equal prior probabilities) is then

$$c = \underset{C}{\operatorname{argmax}}\{p(c|x)\} = \underset{C}{\operatorname{argmax}}\{p(x|c)\} \tag{3.5}$$

This simplifies the interpretation of the results where we are primarily interested in investigating the utility of the class-conditional distributions of each feature. Each experiment is performed on each speaker separately, results are based on the classification of all syllables by randomising and splitting the data set into 10 non-overlapping subsets and performing 10 "train-test" procedures in a cross-validation fashion – all results reported are averages for the procedure over 5 runs.

As a baseline we consider the discriminative value of the distributions of absolute syllable pitch level per tone (as illustrated in Figure 3.2) and compare classification rates between speakers, also illustrating the difference when taking the mean pitch over the entire syllable (`mean100`) compared to the mean in the latter 50% (`mean50`). As expected, taking the mean in the latter part of the syllable results in better separation between the classes (average classification rate over all speakers of 50.42% versus 45.53%, see detailed results in Figure B.8). It is also interesting to note here that we got higher classification rates on average for male speakers (19 to 36) than for female speakers (1 to 17), presumably due to the increased pitch range of female speakers.

Utterance context	Tone	Precision
Initial	H	0.71
	L	**0.34**
	M	0.51
Medial	H	0.50
	L	0.49
	M	0.60
Final	H	**0.34**
	L	0.58
	M	0.47

Table 3.3: *Classification results for tones in different utterance contexts using the mean F0 in the latter part of the syllable.*

Following up on this result we confirm that a significant component of the spread seen in Figure 3.2 is indeed due to *downtrend*. In the same experiment completed above we labelled syllables as utterance initial, medial and final when the syllable number fell into the first, middle or last third of the utterance length in terms of number of syllables. Classification rates for these subsets, shown in Table 3.3, indicate that H and L syllables are not successfully distinguished in the latter and earlier parts of the utterance respectively.

Following this result, we turn to the dynamics of F0 as a potentially more reliable indicator of tone considering *downtrend* [76]. Previous observations regarding intra- and inter-syllable F0 dynamics in Table 3.2 and Figure 3.3 respectively also suggest two relatively robust ways of quantifying this aspect:

- `lingrad`: the gradient of the linear least squares fit to the current syllable contour.

- `deltamean`: the difference in mean F0 between the last 50% of the previous and current syllable (for the initial syllable in each utterance, the initial F0 value of the current syllable is used instead of the previous mean).

We also investigated taking the difference between start and end F0 values of each syllable as a measure of intra-syllable dynamics with less promising results, possibly due to fine alignment inaccuracies and "carryover assimilation" rendering this measurement procedure unreliable.

In this experiment we estimate a distinct context-dependent distribution for each tone ($t \in \{H,M,L\}$) depending on the previous syllable tone ($p \in \{H,M,L,N\}$, where N means "none"), resulting in 12 distributions denoted pt (as in Table 3.2). During classification we assume that p is known and only discriminate between the three tone classes by applying the decision rule to the appropriate set of 3 distributions depending on p. We compared `lingrad` and `deltamean` against using the absolute mean (`mean50`). Detailed results are presented in Figures B.9 and B.10 and summarised in Figure 3.12 and Table 3.4 (The F1 score is the harmonic mean between precision and recall). The following observations are made:

- The tone of the utterance-initial syllable can be relatively successfully identified by absolute pitch measured and modelled in this way.
- The linear gradient is the most reliable indicator in H-L and L-H contexts.
- For transitions from M tones the change in pitch between syllables is a reliable feature.
- For transitions to M tones the absolute pitch distributions are relatively successful.
- The utterance-initial syllable, H tones and transitions from M tones are most easily distinguished in this experiment, while M tones are in general more difficult to classify correctly.

The observations here seem to suggest that speakers and listeners may employ multiple complementary cues in different contexts for the implementation and perception of tone. For example, in simple contexts such as the utterance-initial syllable (and presumably also in short or single-word utterances) the absolute pitch may be employed successfully. In more complex contexts with significant *downtrend* component the perception of tone may also rely on the F0 dynamics; the gradients of falling and rising realisations can presumably be used to distinguish between L-H and L-M and H-L and H-M transitions. The results also suggest that M tones are more reliably characterised by pitch level

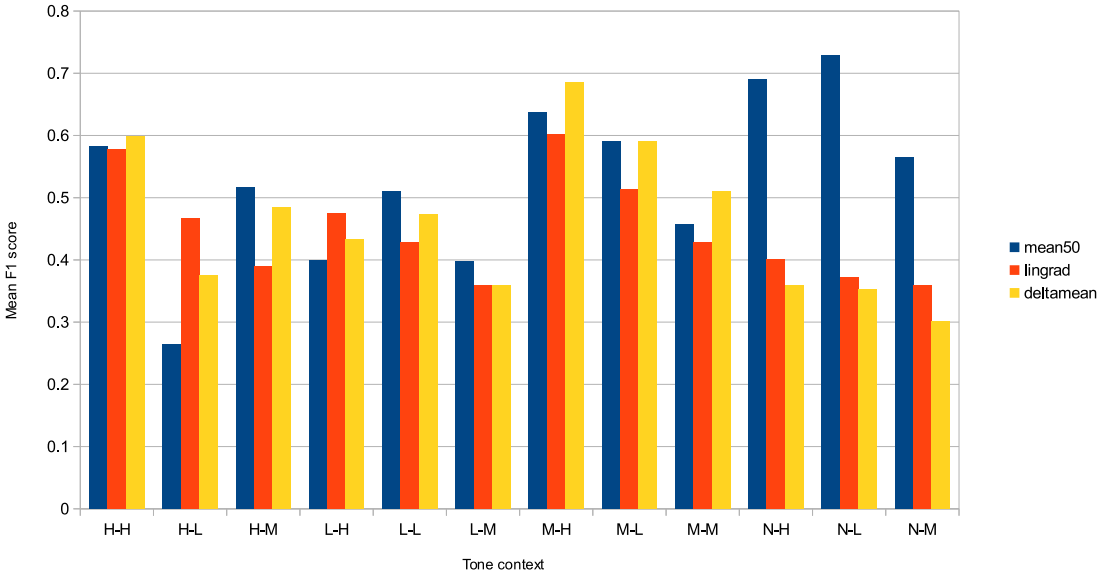

Figure 3.12: *Mean F1 scores over all speakers for the 12 distinct tone contexts modelled.*

and may then possibly act as a reference in terms of pitch level which might explain why transitions from M tones may be classified relatively well based on change in pitch level. A further experiment combining these three simple features of pitch by simply estimating multivariate distributions and repeating the classification procedure results in a mean classification rate of almost 56% overall with a best rate of almost 65% (speaker `32`) and lowest rate of about 43% (speaker `15`). The results here once again indicate that speakers' `01`, `09` and `15` data are the least consistent with the investigated tone indicators (compare Figure B.5 described in Section 3.3.1.1 and Figure B.10).

While features involving the dynamics of F0 may be more stable considering the effect of *downtrend* (Table 3.4), the absolute pitch nevertheless seems to be an important feature and classification error systematically related to the utterance context thus needs to be further investigated.

3.3.5 Variation in pitch contours

Having presented and characterised the mean contours in Section 3.3.1, we extend this work in this section by:

1. Investigating the nature of variation in tri-tone contexts, and

Utterance context	Tone	mean50	lingrad	deltamean	mean50 +lingrad +deltamean
Initial	H	0.69	0.54	0.53	0.67
	L	0.45	0.42	0.48	0.53
	M	0.52	0.45	0.48	0.51
Medial	H	0.54	0.57	0.56	0.60
	L	0.46	0.42	0.44	0.50
	M	0.61	0.47	0.52	0.57
Final	H	0.38	0.58	0.54	0.54
	L	0.54	0.42	0.44	0.53
	M	0.51	0.48	0.54	0.56

Table 3.4: *Classification results (precision) for tones in different utterance contexts when modelling distributions conditional on the previous tone.*

2. Extending the observation window; considering longer syllable sequences.

To identify distinct variations to the contours in each of the tri-tone contexts in Figure 3.4 we partition the set of contours into two disjoint sets using the standard *k-means* clustering algorithm as described below:

1. Consider the collection of N_c extracted contours of standard length N_T (as described in Section 3.3.1) in the tri-tone context c. Each contour is standardised resulting in x_{ij} where $0 \leq i < N_c$ and $0 \leq j < N_T$, by row-wise subtracting the mean and dividing by the standard deviation.

2. Determine the mean and standard deviations column-wise for this set to obtain $\bar{x}_j$ and s_j as in Figures 3.4 and 3.5:

$$\bar{x}_j = \frac{1}{N_c} \sum_{i=0}^{N_c-1} x_{ij} \qquad \text{where } 0 \leq j < N_T$$

$$s_j = \sqrt{\frac{1}{N_c} \sum_{i=0}^{N_c-1} (x_{ij} - \bar{x}_j)^2} \qquad \text{where } 0 \leq j < N_T$$

3. Initialise the K means x_{kj} (we used $K = 2$):

$$\bar{x}_{0j} = \bar{x}_j + s_j$$
$$\bar{x}_{1j} = \bar{x}_j - s_j$$

4. Assign each contour to a cluster based on the K means using the mean squared error (MSE):

$$S_i = \underset{k}{\text{argmax}}\{\frac{1}{N_T}\sum_{j=0}^{N_T-1}(x_{ij}-\bar{x}_{kj})^2\} \quad \text{where } 0 \leq i < N_c$$

5. Update the means as in (2) for each of the K partitions given by S_i.

6. Repeat the assign and update steps (4 and 5) until convergence.

The results of this procedure are presented in the first row of plots in Figures B.11, B.12 and B.13 comparing the mean calculated over all contours (as in Figure 3.4) with the 2 clusters after convergence. Row 2 and 3 in these figures show the results when performing another iteration of the above procedure with the 2 clusters separately. Attempts to use $K > 2$ resulted in clusters collapsing back to the overall mean in most contexts. Examining these contours we observe the following regularities:

- Variation in temporal alignment, possibly due to expected duration differences of constituent syllables.
- Systematic deviation at the start and end of contours probably due to interaction with different adjacent tones not within the tri-tone window.
- Variation in the extent of relative F0 excursions which may be variation between speakers or articulatory effort between utterances.
- Apparently random variation, especially in the second iteration.

Given the identified clusters of the tri-tone contours we attempt to determine the relative influence of the contextual factors we assume to be responsible for the observed variations. Based on the above list of observations we extract the following discrete features:

1. Speaker; the identity if the speaker, i.e. 33 values $V_X = \{1,...,33\}$
2. Previous tone; the identity of the previous tone or "none", i.e. 4 values $V_X = \{H,L,M,N\}$
3. Following tone; the identity of the following tone or "none", i.e. 4 values $V_X = \{H,L,M,N\}$
4. Syllable structure; whether syllables are "full", having an onset consonant, or "short", nucleus

only, i.e. 8 values $V_X = \{000, 001, 010, 011, ..., 111\}$

These features are then compared to the clusters identified with the k-means procedure by calculating the mutual information between the random variables associated with the feature labels X and cluster labels Y for the initial partitioning estimated over all N_c contours and for the second iteration over the relevant subsets:

$$\begin{aligned} H(Y) &= -\sum_{k=0}^{K-1} p_Y(k) \log_2 p_Y(k) \\ I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum_{v \in V_X} \frac{N_v}{N_c} H(Y_v) \end{aligned}$$

where H is the entropy function, N_v is the number of samples and Y_v the random variable associated with the subset where samples of X take on the value v.

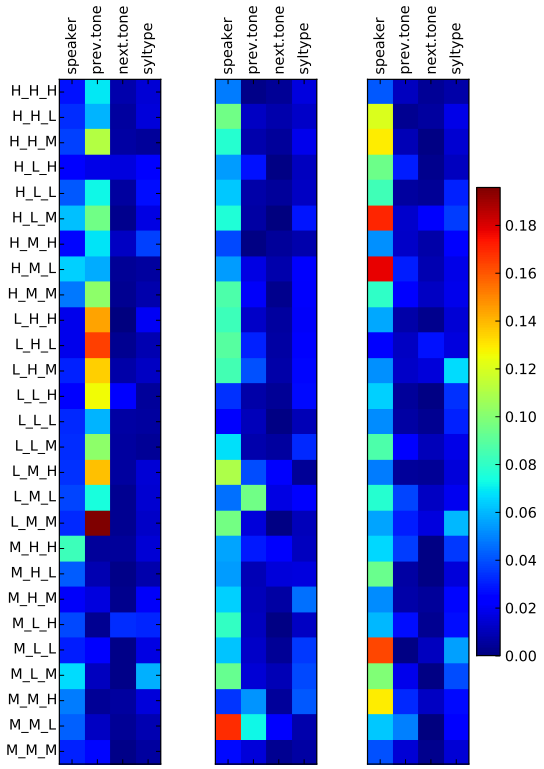

Figure 3.13: *Mutual information between discrete features representing speaker, previous tone, following tone and syllable structure and labels representing the k-means clusters for different tri-tones. The first plot (left) shows the association between the features and clusters identified in the first iteration of* k-means *with the the second and third plots for the second iteration based on the two clusters identified in the first iteration.*

The mutual information for the features in each tri-tone context for each partitioning is presented in Figure 3.13. The first plot, showing the mutual information for the first partitioning iteration, suggests that the most relevant contextual factor is the previous tone (for most tri-tones), followed by speaker. For the subsequent partitionings variations tend to be most strongly associated with differences in data for different speakers. In all cases the following tone has the lowest overall association with the clusters identified. Focussing on the results of the first partitioning; clusters identified for tri-tones starting with L had a relatively high association with the previous tone, followed by H and relatively low for M.

These results quantify the relative association between different tri-tone patterns and factors such as previous tone and speaker identity and supports previous assertions that contours are relatively uniform over different speakers and illustrates the typical variations to the patterns presented in Figure 3.4.

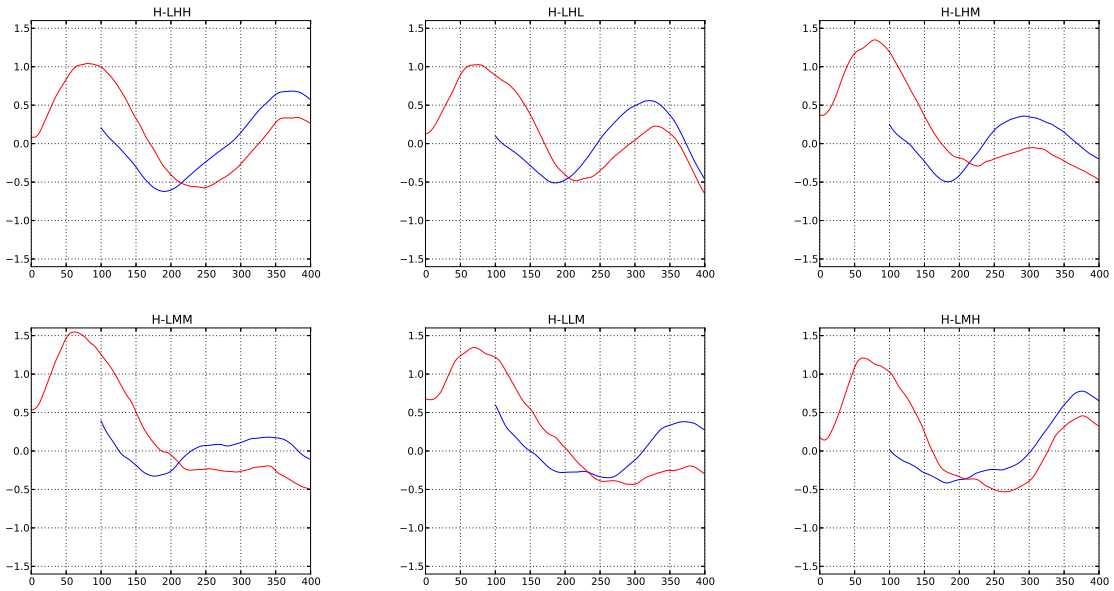

Figure 3.14: *Four-syllable contours with initial H tones that are distinct from three-syllable contours. The first row of plots illustrates the extent of carried over momentum and the second row illustrates variation presumably due to available additional lower pitch range.*

While we have thus far thoroughly presented and analysed the expected co-articulation of tones in sequences of up to three syllables, previous studies indicate that there is some evidence that the pitch contour pattern can be determined over longer sequences; in [66] some evidence was found for "pre-planning", and distinct pitch patterns may arise for longer sequences of repeating tones; [67] describes

a case in the Igbo language where sequences of repeating H's and L's may be realised differently when not interacting with other tones. Another potential source of such regularities in the pitch contour could be the partial "phonologisation" of patterns that may occur in frequent word sequences. While we do not investigate these possibilities thoroughly here, we extend our investigation to four-syllable sequences as our data allows, and specifically attempt to determine how and when these may differ from the tri-tone sequences presented thus far.

We start by extracting all four-syllable and three-syllable contours and normalising time linearly as done in Section 3.3.1. Each contour is standardised by subtracting its mean and dividing by its standard deviation as done before (in the case of four-syllable contours the mean and standard deviation is determined approximately, after time normalisation, over the relevant three-syllable portion). We then calculate the means over the standardised contours for each context and compare the three-syllable cases with the corresponding sections in the four-syllable mean contours – essentially each of the 27 tri-tones in the context of previous tones or following tones (162 contexts denoted in the following way: e.g. the tri-tone HLH preceded by H is denoted **H**-HLH or followed by H is HLH+**H**). Comparisons are done by calculating the RMSE between the corresponding three quarters of the four-syllable contours against the three-syllable contours. These contexts ordered from high to low according to RMSE are presented in Table B.3. Although this method for detecting variants is relatively crude, we are able to make a few observations by examining the general properties of cases with high RMSEs:

1. Four-syllable contours starting with a HL tone sequence have the highest RMSEs, firstly because of higher dynamic range and consistent offset in the second syllable due to the initial H tone, but also influencing the pattern of the contour up to the third and fourth syllable – see Figure 3.14 for examples. The influence on the following tri-tone contour may be explained by two distinct factors. Firstly there is the **carried over momentum** in the H-LH* contexts where turning points in the third syllable (H) are realised later – this influence only continues into the third syllable and secondly, a distinction as a result of **having more available lower pitch range**; in sequences of L tones the contour continues a constant decline instead of the evidence for "bottoming out" in the three-syllable contours – this influence extends into the fourth syllable.

2. Cases with repeating H and L tones were also inspected. The first row of plots in Figure 3.15 show some indications that the rise and fall associated with H and L tones are more evenly

distributed over two syllables for repeated tones (similar to the observation in [67] for Igbo). However, this pattern is less evident in the second and third row of plots; corresponding more closely to the three-syllable contours. The final row suggests that such a distribution of the pitch movement does not frequently occur (if at all) over four consecutive syllables. These patterns thus also seem to be conditional on available pitch range which would explain why more gradual falls may be realised when preceded by H's and similarly rises when preceded by L's.

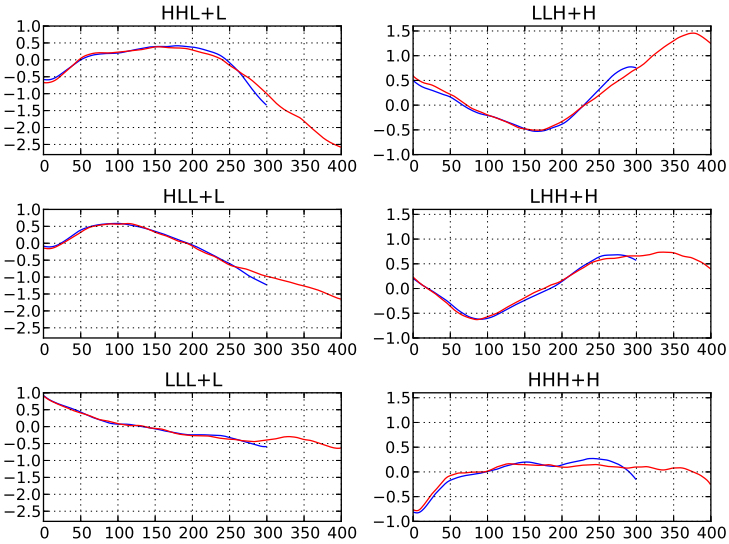

Figure 3.15: *Four-syllable contours with repeated H and L tones. In the first row it is evident that two-syllable sequences of L and H tones often result in a gradual falling or rising contour if pitch range is available. With diminishing evidence for such a distribution of pitch movement over three and four syllable extents (row 2 and 3).*

We did not detect any further distinct differences between three- and four-syllable contours beyond the expected noise due to alignment and other variation not accounted for. The question regarding the possibility that distinct patterns have been "phonologised" for particular words or word sequences is also beyond the scope of the current investigation.

3.4 CONCLUSION AND FURTHER WORK

In this chapter 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 automatic phonemic alignments, intensity and F0 estimates. This has confirmed certain phenomena described in the published

literature and quantified this in certain contexts. Furthermore, we have attempted here to determine to what degree the observed patterns hold for different speakers and tone contexts.

Regarding F0 we summarise the following observations:

1. H tones are raised higher before L tones ("dissimilative H tone raising") as reported in [66] – Figure 3.4.

2. L tones after H and H tones after L have distinctive falling and rising realisations as reported in [9]. This is shown in Table 3.2 and the contours in Figure 3.4 and may serve to distinguish these transitions from similar pitch movements associated with H-M and L-M transitions – Figure 3.12. We thus suggest that these falling and rising realisations are not merely phonetic in nature but may have a "phonemic" role in Yorùbá (in the terminology used in [21, 88], this is a case of true "target alternation" and not a purely "implementational variation" due to "carryover assimilation").

3. Falling and rising realisations (in HL and LH transitions) result in a significant amount of "carryover assimilation" into the following syllable (to a greater degree than other contexts) – Figures 3.4, B.11, 3.14.

4. H tones are not raised to the same level when occurring after a HL tone sequence (Figure 3.4). This "downstep" effect is likely caused by significant carried over momentum ("carryover assimilation") into the third syllable due to the previous falling HL realisation – Figure 3.14.

5. While it is difficult to make strong assertions regarding syllable alignment of pitch contours, because the linear time normalisation process could not be performed on individual syllables in our corpus without distorting the shapes of contours (due to pooling different syllable types), the results in Figures 3.4 and B.11 to B.13 seem to exhibit consistent changes in dynamics at approximate syllable boundary locations (i.e. at normalised times 33 and 67 in Figure 3.4). We thus suggest that these results are in agreement with assertions made in [89].

6. Carryover effects are more evident and extend further into the affected syllable than anticipatory effects – Figures 3.4, 3.5, 3.13. This is in agreement with [21].

7. Important anticipatory effects include "dissimilative H tone raising" when followed by L (as reported in [66]) – Figure 3.4, and weaker evidence for the distribution of pitch change to a

gradual falling or rising over sequences of L or H tones (as noted by [67] for Igbo) – Figure 3.15.

8. There is little evidence from our analyses for the gradual falling or rising of pitch mentioned in (7) over more than two syllables in our corpus and this effect is likely dependent on available pitch range, i.e. such a gradual rise is more evident from a relatively low pitch (or a low L tone) and similarly for the gradual falling pitch. If this condition is not met, the pitch contour "levels out" – Figure 3.15. This suggests a refinement to the contour synthesis strategy devised for repeated tones in Igbo in [67].

9. The tri-tone patterns presented in Figure 3.4 generally hold for different speakers (examples in Figure 3.7) with major variations most strongly associated with previous tone context, followed by speaker data differences – Figure 3.13.

10. Of the dynamic acoustic features investigated here, both inter-syllable pitch change and intra-syllable gradient are useful indicators of tone in different contexts – Figures 3.3 and 3.12.

11. Pitch levels also play a role in indicating tone as is evidenced by the classification rates for M tones following H or L tones and the utterance initial syllables – Figure 3.12.

Considering the realisation of intensity associated with tone we found that for some speakers signal intensity may be correlated to various degrees with F0, in some cases only for H or L tones. This may indicate that H or L tones may be realised or perception enhanced to some extent using emphasis or de-emphasis respectively, however these trends were not consistent over all speakers – Figure 3.11. In the case of duration, there is a weak trend in our corpus of L tones being shorter in duration and M tones longer – Figure 3.8. While tone seems to have a relatively minor or inconsistent influence in these acoustic parameters, the evidence suggests that it is likely beneficial to include tone identity when modelling these parameters for natural speech synthesis, depending on the speaker.

The focus of the remaining part of this work will thus be on pitch as this is likely the most perceptually relevant acoustic feature, as well as exhibiting a complex distribution which is not easily modelled in under-resourced contexts (see motivation in Chapter 2).

With speaker-specific tone classification rates using only class-conditional distributions of between 43% and 65% in our corpus, we have identified some reliable local acoustic features and tone contexts associated with tone realisation. It is however clear that extended tone contexts and information re-

garding available pitch range would be beneficial to more accurately characterise and classify tones in different utterance contexts. It is thus imperative that further work is done to investigate the utterance-wide pitch *downtrend*. This includes determining and quantifying the effects of reported linguistic phenomena such as *downstep* and *final lowering* alongside the physiologically motivated trend of *declination* on the utterance pitch contour. This will be the subject of the next chapter with the goal of predicting pitch movements associated with tone in continuous utterances.

CHAPTER 4

UTTERANCE PITCH TARGETS IN YORÙBÁ

Yorùbá is considered to have a *terracing* register tone system. This refers to a pattern where distinct level tones are not realised at a fixed pitch, but at systematically decreasing pitch through the course of an utterance [26]. While most languages exhibit a gradual lowering of pitch throughout an utterance known as *declination* [55], in Yorùbá the implementation of tones in different contexts are considered to have a significant effect on the utterance-wide or *global* pitch distribution, either accelerating or slowing the total measurable *downtrend*. Previous work studying downtrend in Yorùbá consider declination to be only a relatively minor factor [9, 66], describing a number of effects which may have an influence on the global pitch distribution or overall downtrend:

- *Downstep*: A discrete downward shift in pitch. In Yorùbá, obligatory or "automatic" downstep is said to occur in the HLH tone context [9, 66].
- *Pitch resetting*: A discrete upward shift in pitch especially of H tones when pitch levels have dropped to the point of making tone contrast difficult [66].
- *Anticipatory initial H tone raising*: The fact that H tones may sometimes be raised at the start of an utterance in anticipation of expected downtrend [66].
- *Final lowering*: A lowering of pitch in the final syllable to a baseline value for the utterance [9].
- *Declination*: Gradual lowering of pitch throughout the course of an utterance [9, 66].

For speech synthesis purposes the significance and effect of these concepts need to be investigated and quantified in order to incorporate them into a complete intonation model considering both the

critical local patterns to effect tone realisation (affecting word *intelligibility*) and global distribution that affect the perception of information-structure and the overall *naturalness* of the speech.

While intonation models employed in speech synthesis systems may vary as reviewed in Section 2.2 depending on fundamental assumptions such as *superposition* or contiguous *tone sequences*. In corpus-based speech synthesis intonation models (i.e. *statistical parametric* and *unit-selection* systems) the global pitch is usually modelled based on the syllable position in *breath-group* and *breath-group* position in the *utterance*. Thus breath-groups may have distinct pitch distributions and the average pitch usually also declines over breath-groups in an utterance.

In this chapter we focus on the general pitch distribution in short utterances in the hope that results may be generalised in a systematic way to multi-breath-group utterances in future (see discussion in Section 4.4).

4.1 APPROACH

The investigation presented here will start by re-contextualising an analysis of utterance pitch patterns in terms of changes in relevant syllable pitch targets first presented in [90] in the following section. This is followed by work on predicting pitch targets in this corpus, and in particular testing the significance of *downstep* and *declination* in such models (Section 4.3). Finally, results are summarised and avenues for further work involving *pitch resetting* and longer multi-breath-group utterances are proposed in Section 4.4.

4.2 CHANGES IN SYLLABLE PITCH TARGETS FOR TONES IN UTTERANCE CONTEXT

Motivated by the results in [76] and Chapter 3, and also noting that the change in pitch may exhibit relatively stable distributions, we attempt here to investigate these distributions in various tone and utterance contexts.

In order to quantify change in pitch we need to define "pitch targets", including a procedure to derive targets from F0 contours. Two potential candidates for pitch targets are found in the work of Ọdéjọbí [14] and the target approximation model proposed by Xu [21]. Ọdéjọbí assumes that each syllable contour can be represented by exactly one peak and one valley, while the target approximation model

proposes an underlying linear target function that is reached gradually without oscillation towards the end of the syllable depending on the effort exerted by the speaker. In the previous chapter we described the typical contours observed in our Yorùbá corpus in different tone contexts, observing the following of interest here:

1. M tones often have a relatively flat or monotonic profile and H and L tones may have monotonically rising or falling realisations over the course of a syllable. The selection of a third order polynomial to describe syllable contours may thus not be the simplest plausible representation; it is noted in [14] that the peak and valley often have the same value in the case of M tones. This is supported by the results in Section 3.3.4.

2. Carryover assimilation is more prominent than anticipatory effects and extreme points (peaks or valleys) associated with the current syllable tone are generally realised late in a syllable or even in the following syllable, with a significant amount of variation of the exact turning points relative to syllable alignments (see Figures B.1 and B.2). This suggests that a sophisticated alignment model would be necessary to model contours accurately based on peaks and valleys. In [14] a heuristic alignment model was implemented, however this was found to be unsatisfactory by the authors of that work.

Given these observations and others (in Chapter 3) we adopt the assumption of syllable synchronisation argued for in [21], consequently the target approximation model is adopted for our analysis of pitch targets here. In this chapter we limited the investigation to static linear pitch targets, i.e. with a gradient of zero (see Section 4.2.2).

We continue in the following two sections by describing our corpus and experimental setup followed by details of our investigations from Section 4.2.3 onward.

4.2.1 Corpus

From our corpus described in Chapter 3, we selected four speakers (two from each gender), for which we proceeded to extract pitch targets as described in the following subsection. For these speakers we manually inspected alignments and F0 extraction for correctness. Here we intervened by correcting transcriptions and alignments in the case of gross errors (see Section 3.2.3), refraining from editing phone boundaries extensively. If F0 extraction was particularly unreliable or transcriptions were

completely erroneous, we discarded the utterance (a total of 10 utterances were discarded in this way). Table 4.1 shows the resulting corpus properties.

Speaker ID	Gender	F0 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 4.1: *Manually verified corpus properties with syllable counts by tone reflected in the last three columns.*

4.2.2 The quantitative target approximation model

In the target approximation model, observable contours are a result of gradually reaching underlying target functions specified for each syllable. Targets are usually linear functions that may be static (i.e. with a gradient of zero), consistent with the prototypical definition of register tones, or dynamic, to implement contour tones as occur in languages such as Mandarin and Cantonese. Actual contours approach, not necessarily reaching, target functions towards the latter part of syllables depending on carryover effects from the previous syllable and the effort exerted by the speaker. We considered two methods of estimating such pitch targets:

1. Estimating targets by direct measurement: determining the maximum, mean and minimum pitch values in the syllable nucleus for H, M and L tones respectively (based on the observations in Chapter 3, especially Figure 3.4).

2. Estimating the underlying (abstract) pitch targets via the analysis-by-synthesis method described in [65] as implemented by the authors of that work in the *PENTAtrainer* tool.[1]

Estimates from (1), using smoothed interpolated contours to reduce measurement noise and (2), extracted as described below, led to comparable results. Targets based on estimate (2), however, exhibited less variance and the process is independent of tone labels, leading us to adopt this estimate for further analysis.

[1] Available at: `http://www.phon.ucl.ac.uk/home/yi/PENTAtrainer1/`

The quantitative target approximation model described in [65] uses a simple linear equation to describe pitch targets:

$$x(t) = at + b \tag{4.1}$$

with a third-order critically damped linear system defining the resulting pitch contour approximating the target in a syllable:

$$f_0(t) = x(t) + (c_1 + c_2 t + c_3 t^2)e^{-\lambda t} \tag{4.2}$$

where a and b represent the gradient and height of the current syllable pitch target respectively and c_i are determined by initial conditions of F0 and the current syllable target (see Eqs. 3, 4 and 5 in [65]). The carryover effect on F0 by the previous syllable is thus considered, and λ represents the rate of target approximation, reflecting speaker effort. The relevant parameters that need to be determined for each syllable are thus a, b and λ. The *PENTAtrainer* script implements an exhaustive search over predefined ranges of these parameters for each syllable, finding optimal values by minimising the error between resulting synthesised and actual F0 contours extracted with *Praat* (and interpolated to have values in unvoiced regions). For the current experiment we set $a = 0$ in all syllables since we are investigating height. For the target height parameter, b, we leave a broad search range ± 20 semitones from the measured F0 (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 F0 contours (such as the proposed measurements in point (1) described above). Although manual inspection of resulting pitch target estimates indicate that inaccuracies do occur, especially when syllables are very short combined with slight alignment inaccuracies, the process seemed generally robust; Figure 4.1 shows an example of pitch targets extracted for an utterance in our corpus.

4.2.3 Initial observations

In Figure 4.2 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 (in semitones) emerges on average, however a significant amount of variation is present in all cases (tones and speakers) as one would expect if the realisation of tone has a significant influence on the absolute pitch distribution. Closer inspection also indicates evidence for pitch resetting – compare the moving average with the linear fit. We summarise again the expected sources of variability:

1. *Downstep* and varying rates of downtrend expected due to the influence of tone realisation in

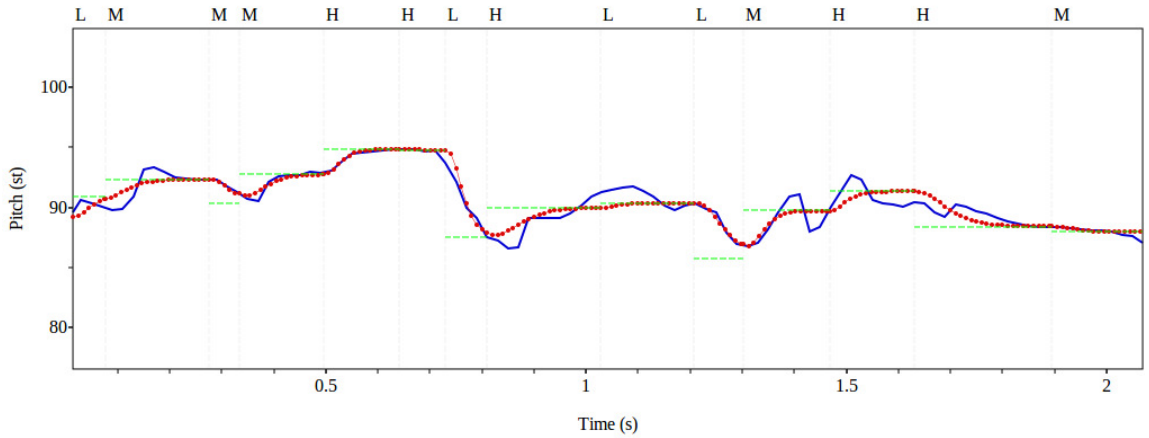

Figure 4.1: *Example of pitch targets extracted from an utterance in our corpus; The original F0 contour is represented by the solid line (blue), with estimated pitch targets indicated with dashed lines (green) and the resulting synthetic contour with connected dots (red). The tones indicated are obtained from the text (diacritics).*

other tone contexts [9, 66].

2. Utterance-initial variation; anticipatory raising of H tones and the "start-up" effect resulting in lower than usual L tones [66].

3. Final lowering – can result in the final syllable, whether H, M or L, being realised lower than usual [9].

4. Declination.

5. Pitch resetting [66].

6. Syllable duration.

7. Intrinsic F0.

8. The realisation of word prominence or emphasis, which may increase the dynamic range of pitch movement [21].

9. Assimilation of syllables potentially causing false measurements (see final note in Section 4.2.2).

10. Possible tone sandhi effects not accounted for – tone is assumed to be shallowly marked on the orthography.

11. Possible changes in speaker effort (the λ parameter discussed in Section 4.2.2).

12. Errors in estimation due to F0 estimation inaccuracies.

Given the current experimental setup, it is not possible to consider all of these causes and points 8 to 11 are thus not explicitly investigated while we attempt to determine the nature of points 1 to 7. We discuss these in the following subsections.

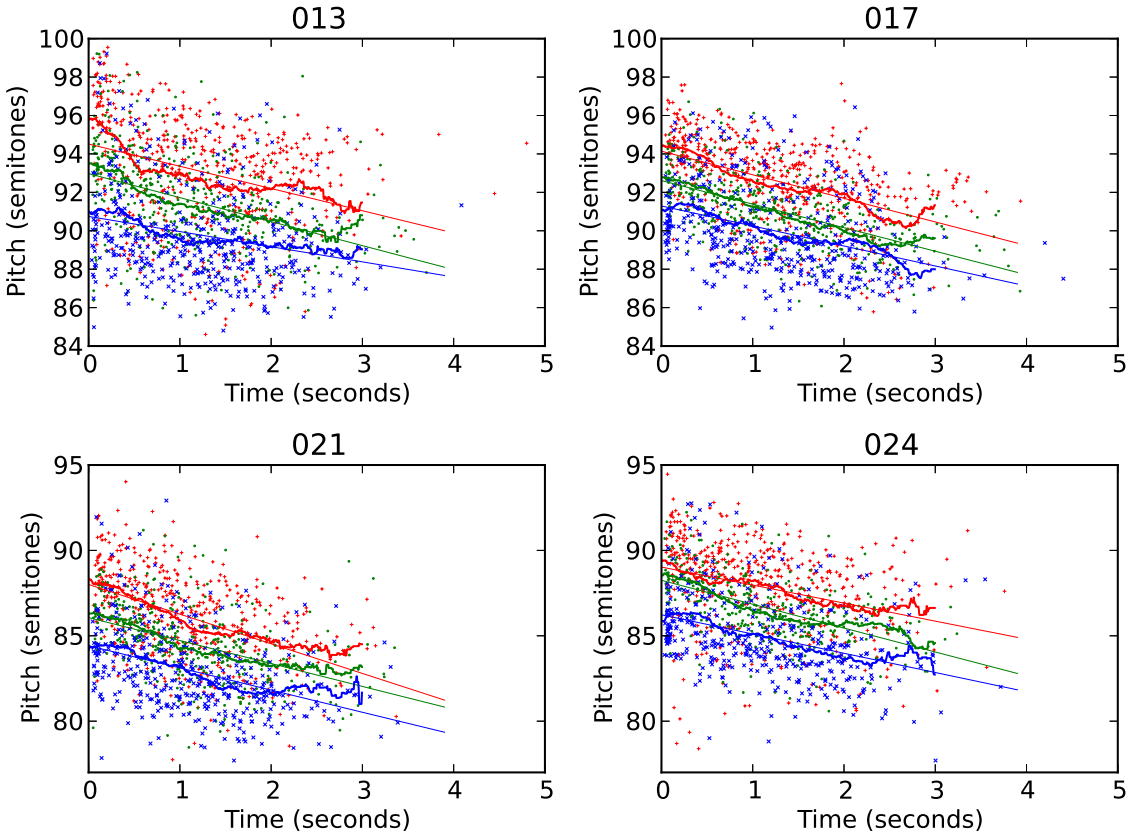

Figure 4.2: *Pitch targets extracted for all syllables of each speaker (speakers 013 and 017 are female, with 021 and 024 male). The tones H, M and L are represented by red (+), green (.) and blue (x) respectively, with a linear fit and moving average within a 500 ms window plotted for each. Times for individual points correspond to the central instant of each syllable.*

4.2.4 Local changes in pitch targets in tone and utterance contexts

Considering points 1 to 3 above regarding utterance contexts that significantly affect changes in pitch between syllables and findings in Chapter 3 regarding the importance of tone context, we expect that a local window of at least four syllables would be required to quantify inter-syllable changes in pitch targets, i.e. tone labels of the two syllables preceding and following a transition, including whether syllables are utterance-initial or final (e.g. HH to LH, NH to LH or HH to LN, where the start and end of the utterance are represented by N). Figures 4.3 and 4.4 present the mean changes in syllable pitch targets for our 4 speakers in different contexts.

While the variance is relatively high, we are able to make a few reliable observations:

- Transitions between syllables in two-syllable contexts (i.e. where only the preceding and following syllable labels are known) behave consistently between speakers and are generally distinct.

- We see significant evidence for the local effects of *downstep* in the H-LH context in all speakers based on the pitch targets extracted in this experiment.

- Changes in pitch 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 significant effect such as *downstep* is dependent on the tone context, we conclude that the tone labels for at least the previous two syllables before a transition are needed to accurately predict local pitch target changes.

Some 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 tone context alone is sufficient to describe these observed differences. Nevertheless, we note some further observations with possible interpretations:

1. In the H-LH context the mean pitch change measured is negative for all speakers. Closer inspection of these cases suggest that these heights may be weakly positive if measurements are more appropriately extracted (i.e. without the zero gradient constraint), nevertheless, we ascribe this to significant carryover effects due to the preceding raised H tone, see point (2)

discussing evidence of H-raising. This interpretation is supported by a similar pattern in the H-LM context which also exhibits a smaller pitch change than other contexts. This effect is also symmetrical to some degree, although weaker in this case; L-HL contexts exhibit a smaller than expected fall in pitch and likewise in L-HM contexts. In these contexts involving strongly falling and rising contours, the mean transitions to M tone is more likely to be distinct from the norm. This supports the hypothesis of weak articulatory effort described in [88] (see also the discussion in Section 5.5.1).

2. H-raising is evident by comparing contexts such as LH+L and MH+L, where these generally have larger positive pitch changes than contexts followed by other tones (this effect is least significant in the case of speaker `24`).

3. Transitions from M tones have the most uniform distributions (corresponding with results in Section 3.3.4), probably largely due to the fact that range limiting has a smaller effect when changing from a central value in the speakers' register (see Section 4.2.5).

4. The transitions LL and HH are most variable, due to range limiting factors (Section 4.2.5) and possibly patterns involving gradual rises and falls (Section 3.3.5) also motivated in the next point.

5. Gradual falls and rises seem to be evident to some degree when comparing HL+L and LH+H contexts with the norm. This effect is least significant in the case of speaker `21`, and may thus be speaker specific or dependent on other aspects of the data.

6. Utterance-initial syllables show distinct patterns, such as the "start-up" effect where L tones start lower than usual (see N-LL contexts with positive pitch changes) and high starting values for initial H tones (see N-HL contexts).

In this section we have quantified pitch target changes in specific contexts, confirming significantly distinct changes in pitch associated with certain tone contexts (as suggested by the results presented in Chapter 3) and utterance contexts such as initial and final syllables. The fact that distribution of these changes in pitch targets seem to correspond to some degree between speakers suggests a model based on change in pitch in local tone and utterance contexts. It is clear however that a significant amount of unexplained variation remains. In the next section we investigate pitch range.

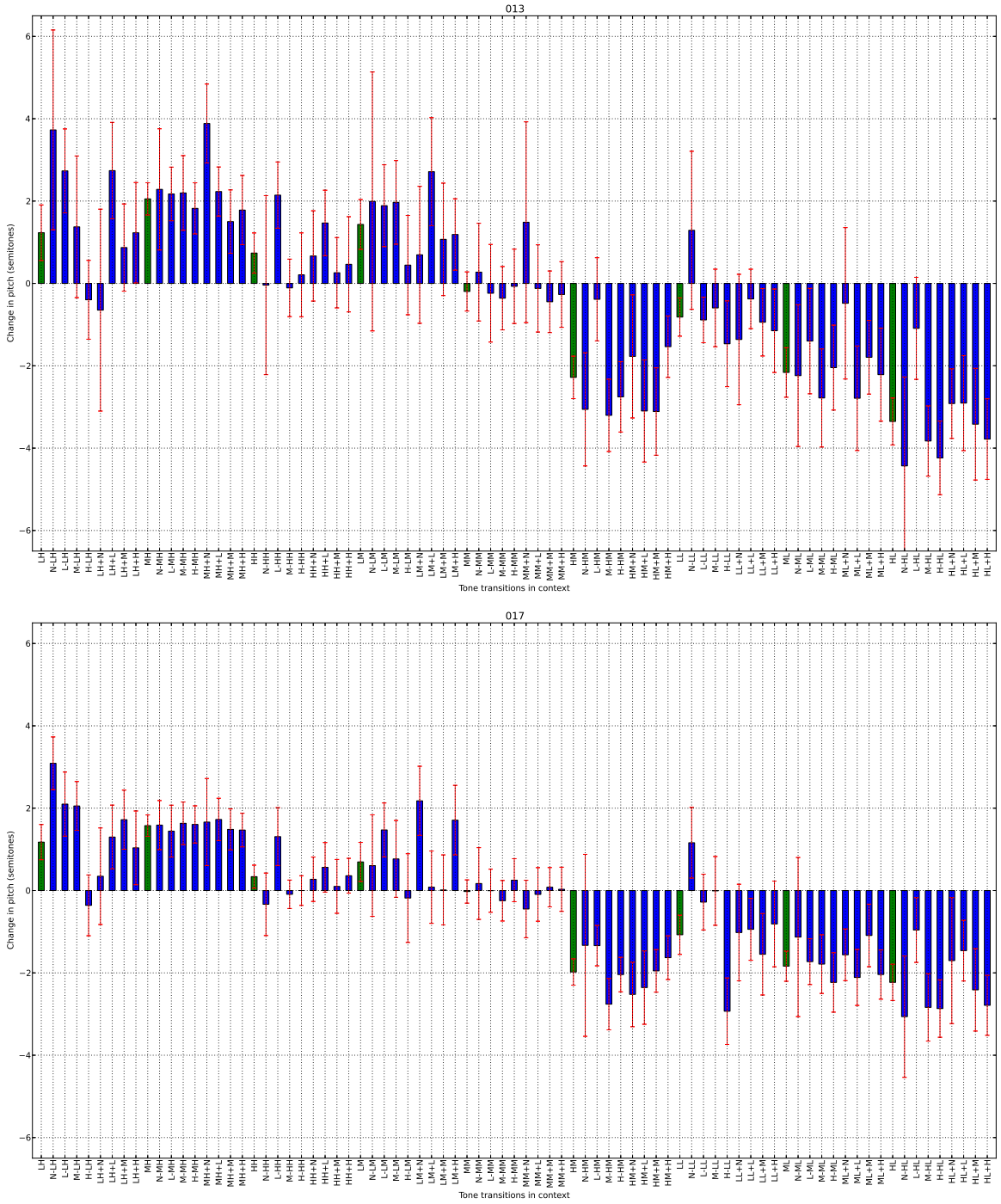

Figure 4.3: *Mean changes in pitch between syllables in different contexts, for speakers 013 and 017; preceding contexts are denoted by a "-" and succeeding contexts by a "+". H, M and L represent High, Mid and Low tones, with N representing the utterance boundary. Error bars denote the 95% confidence interval.*

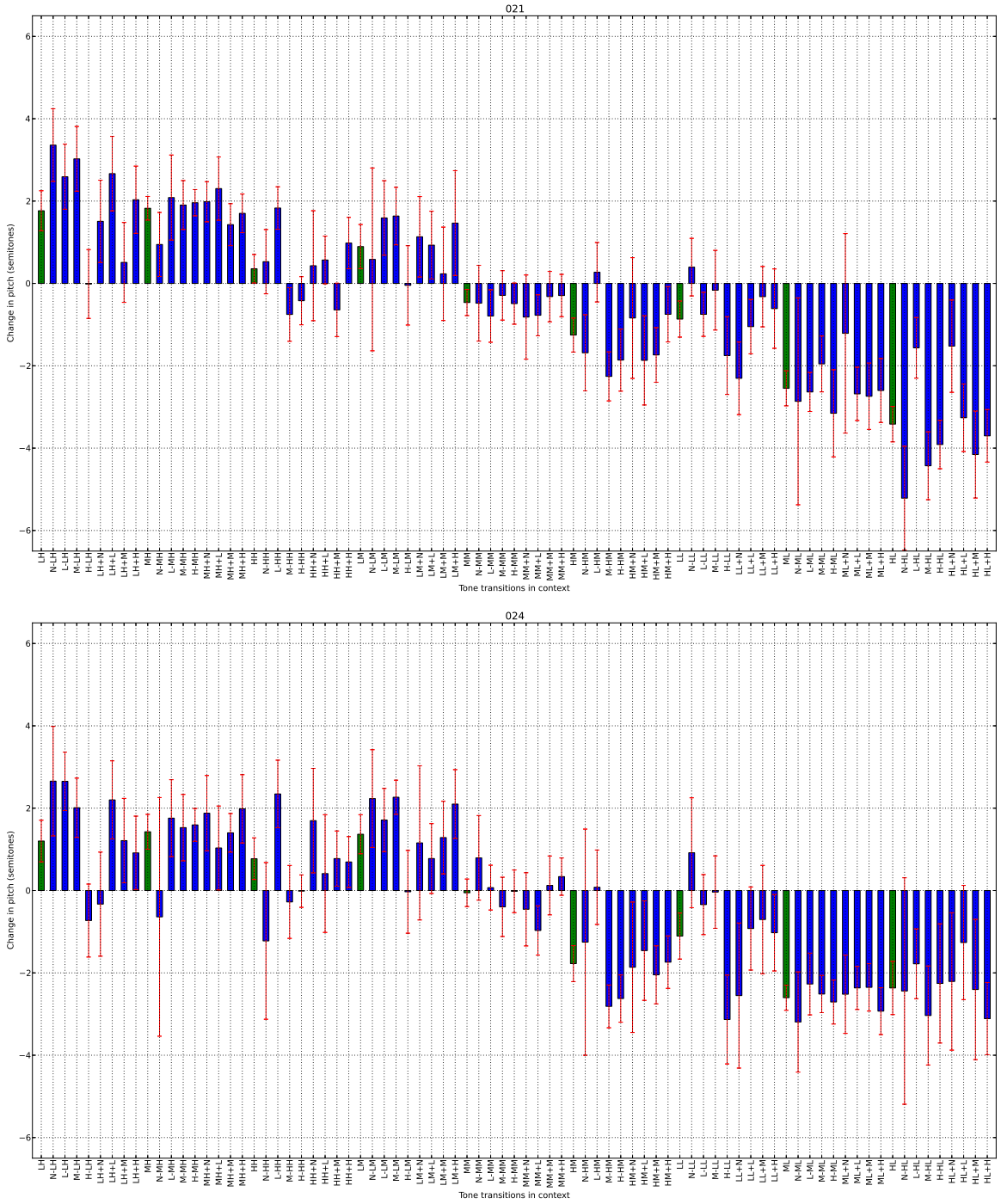

Figure 4.4: *Mean changes in pitch between syllables in different contexts for speakers 021 and 024; preceding contexts are denoted by a "-" and succeeding contexts by a "+". H, M and L represent High, Mid and Low tones, with N representing the utterance boundary. Error bars denote the 95% confidence interval.*

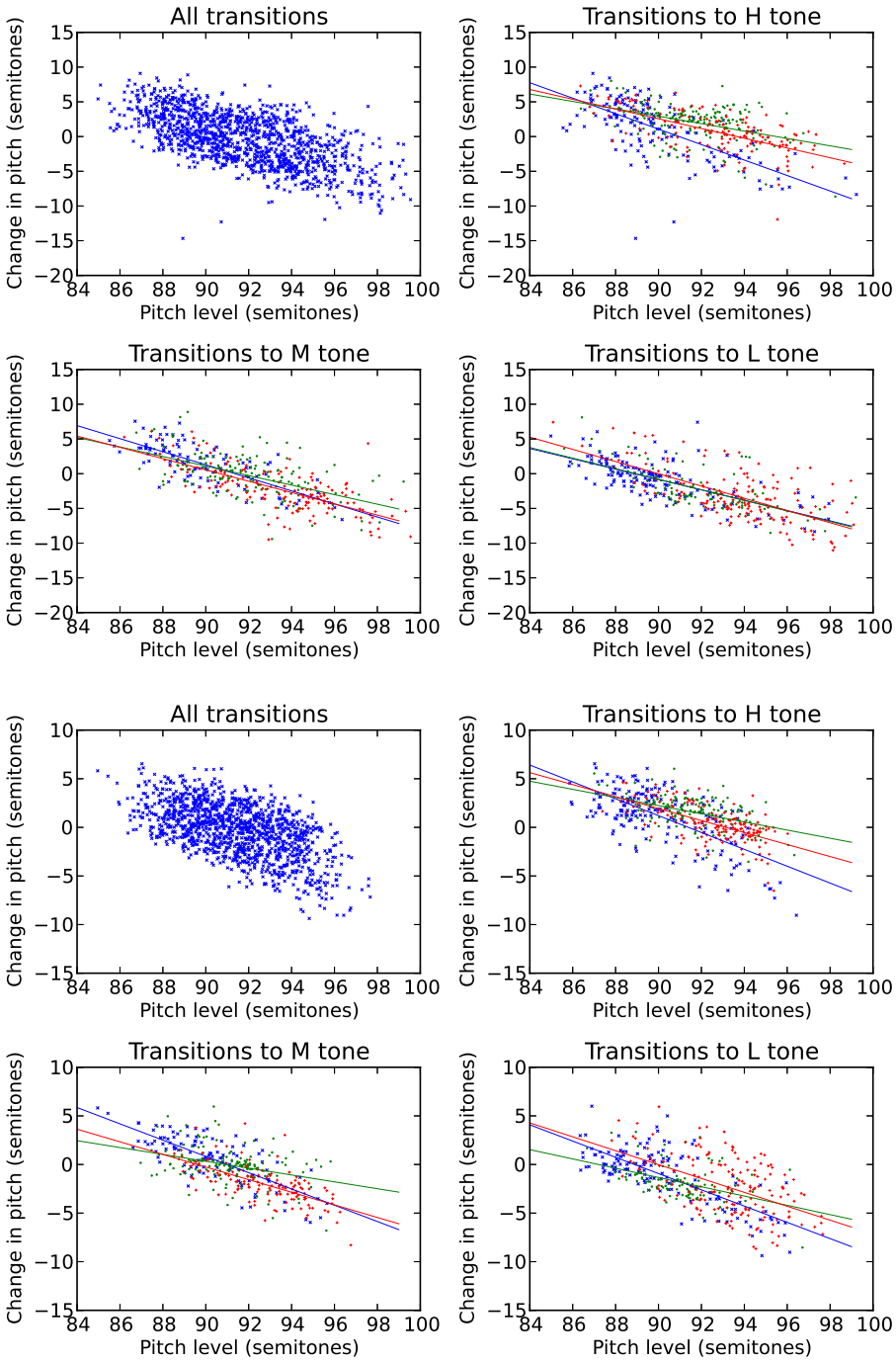

Figure 4.5: *Female speakers, 013 (top four plots) and 017 (bottom four plots): Changes in pitch for targets in consecutive syllables. Subplot 1 (top left) shows all transitions, with subplots 2 to 4 showing transitions to H, M and L tones respectively. In subplots 2 to 4 blue (x), green (.) and red (+) represent transitions from L, M and H tones respectively.*

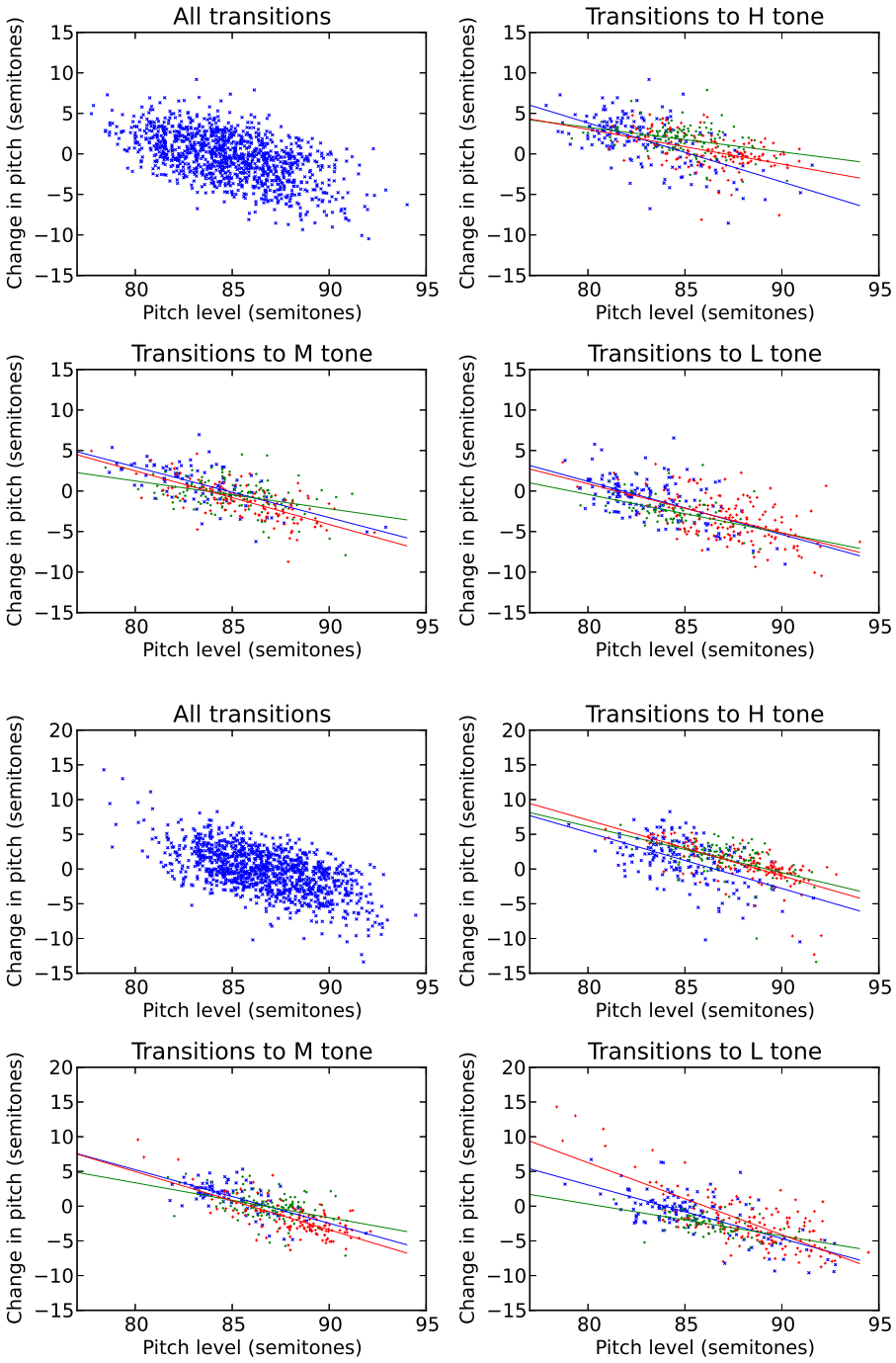

Figure 4.6: *Male speakers, 021 (top four plots) and 024 (bottom four plots): Changes in pitch for targets in consecutive syllables. Subplot 1 (top left) shows all transitions, with subplots 2 to 4 showing transitions to H, M and L tones respectively. In subplots 2 to 4 blue (x), green (.) and red (+) represent transitions from L, M and H tones respectively.*

4.2.5 Pitch range

While the inter-syllable change in pitch might be an appropriate basis for predicting pitch targets in local tone and utterance contexts, one aspect that remains to be determined is the mechanism for pitch range conservation. That is, how a sequence of changes in pitch associated with local contexts for an arbitrary utterance is implemented within a finite pitch range. Inspection of utterances in our corpus suggests that there are larger changes in pitch at the onset of utterances; also, that pitch changes may be contracted later in utterances, with periodic pitch resetting in longer utterances. This pitch conservation effect however does not seem to be a simple function of utterance position (i.e. either syllable number or time) such as the linear trend in Figure 4.2 or the phrase component in superpositional models such as the Fujisaki model [63].

Considering that these mechanisms (pitch change contraction and resetting) serve to manage available pitch range, the effect should be a function of pitch. We thus investigate the relationship between pitch changes and previous pitch level (Figures 4.5 and 4.6). It is evident that there is a strong linear relationship between the pitch change and previous pitch level in all speakers and we also make the following additional observations:

- 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).

- The distribution for transitions to the M tone seems to be more homogeneous than for the other two tones: there is a lower variance in the pitch-change dimension at specific values of the current pitch. Differences in the originating tone do not seem to carry as much extra information regarding pitch change not already contained in the previous pitch level, compared to the other tones.

- 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 the range -5 to 5 semitones.

These results suggest that local pitch changes are scaled to accommodate pitch range limits. Combined with the results in the previous section, this suggests that pitch targets in utterance contexts may

be broadly modelled based on scaled pitch changes between syllables in appropriate local contexts. With this in mind, we investigate further contextual information that may be useful in characterising local pitch changes in the following sections and combine all observations towards a partial pitch target model in Section 4.3.

4.2.6 Syllable duration

Another factor which could have a systematic effect on pitch change is syllable duration. Xu [54] reports 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 mostly 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). Some 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 (see Figure C.1 in Appendix C). These observations correspond with the results over the entire corpus presented in Section 3.3.2 and supports the possibility that syllable duration in our corpus might be more strongly influenced by the requirements of tone realisation than vice versa.

4.2.7 Intrinsic F0

The phenomenon known as *intrinsic pitch* (IF0) is a universal phonetic effect associated with vowels and occurring in all languages to a greater or lesser degree [56]. 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 [57] 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 in [56, 57].

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 4.2). 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 IF0 with increasing vowel height. The results measured do confirm

that there is more measurable variation in F0 across vowels for H tones than M or L, which is consistent with findings by [56]. It is possible that the expected level of noise associated with the process of pitch target extraction used here is masking the pitch differences expected due to vowel identity and IF0. Further investigation would require a more carefully controlled experimental setup, ensuring vowel instances occur in a diverse set of utterance and tone contexts. In our corpus this 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 do not provide sufficient evidence for a systematic influence of vowel height, but they do suggest that we should consider vowel identity as a potentially useful feature towards predicting pitch targets (see Section 5.5.2).

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 4.2: *Mean F0 (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.*

4.3 PREDICTING UTTERANCE PITCH TARGETS

The observations in Sections 4.2.3, 4.2.4 and 4.2.5 provide insight into the distribution of pitch height targets resulting from the systematic effects of utterance position, tone context and pitch range respectively:

1. An approximately linear trend of pitch decline, measured in semitones, emerges on average

over the course of an utterance for all tones (compare the moving average and linear function estimates in Figure 4.2) and the rate of decline for different tones seems to be similar.

2. The effect of tone context can be quantified in terms of pitch changes between syllables and the effects of extended context expected based on the analysis of carryover and anticipatory effects in Section 3.3.5 are shown to be distinct compared to simple transition contexts and consistent across speakers. We also confirm systematic effects due to utterance position such as the "start-up" effect (see N-LL contexts in Figures 4.3 and 4.4as an example) and distinct pitch changes in the final syllable. These local pitch changes still exhibit a large amount of variation.

3. A strong correlation between current syllable pitch and following syllable pitch is observed suggesting that speakers manage pitch range by scaling pitch changes and effecting pitch resets when the pitch level is close to the lower limit.

These observations suggest a model based on pitch changes between syllables and taking into account pitch scaling as follows (from Figures 4.5 and 4.6):

$$x_i = x_{i-1} + (a_c x_{i-1} + b_c)$$

which may be simplified by redefining the quantity a_c to

$$x_i = a_c x_{i-1} + b_c \tag{4.3}$$

where x is the pitch height target and i indexes the syllable and a and b are estimated in different tone contexts c (see Figures 4.3 and 4.4 for tone context examples). Such a model was implemented in an earlier experiment in [90], included in Section C.1 in Appendix C. However, it is clear that the general downtrend is not adequately modelled in this way (Table C.2) – a simplified example of the pitch distribution resulting from such a model with hypothetical parameters can be seen in Figure 4.7. Thus, while this model may well describe the local dynamics, a mechanism is required to adjust the parameters over the course of an utterance.

4.3.1 Considering downtrend

The downward trend of pitch in utterances has been described and explained in different terms, especially in work on tone languages. Different authors have used terms including *downdrift*, *downtrend*, *downstep* and *declination* to refer, sometimes inconsistently, to downward shifts in pitch assumed to

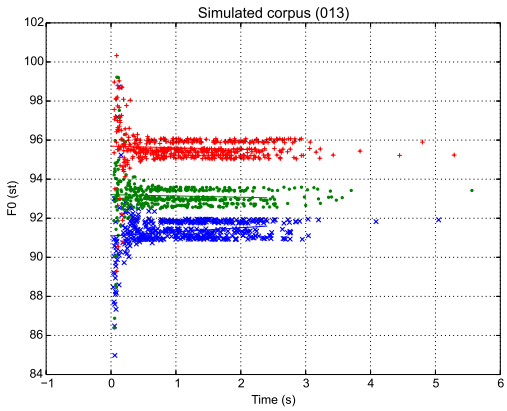

Figure 4.7: *A simulation of the model defined in Eq. 4.3 using hypothetical parameters and simple tone contexts ($c \in \{H,M,L\}$). Initial pitch values, syllable times and tone sequences are taken from the utterances of speaker 013 (compare Figure 4.2). The tones H, M and L are represented by red (+), green (.) and blue (x) respectively. Times for individual points correspond to the central instant of each syllable.*

be due to different causes or to refer to a collection of such effects. Key concepts that need to be considered to describe the overall downtrend in tone languages such as Yorùbá are:

- *Declination*: the gradual lowering of pitch throughout the course of an utterance. Different causes have been identified for this pattern, including physiological and linguistic factors [55]. The observation that declination may have a role in conveying information structure and the findings that the difference in utterance onset and offset frequencies are often relatively constant irrespective of utterance length, with little variation in offset frequencies, agree with descriptions of "final lowering" to a baseline frequency in the last syllable of Yorùbá utterances [9].

- *Terracing*: refers to a more discrete lowering of pitch throughout the course of an utterance due to the occurrence of certain local events such as *downstep* [26, 9, 66]. It is assumed that the utterance-wide pitch contour in this case is largely dependent on a combination of local pitch changes (and therefore the tone sequence) and that gradual declination plays a relatively minor role [9, 66].

Given the physiological grounds for declination one would expect that such a gradual effect be present also in Yorùbá, and this was found to a limited degree in the data analysed in [9, 66]. However,

recent work has also suggested that a declining utterance-wide pattern may be used communicatively to distinguish statements from other utterance types in Yorùbá and is thus perceptually important [91].

A key question considering the discussion of downtrend in Yorùbá presented above concerns an appropriate independent variable or events at which "register lowering" occurs.

Previous work describing terracing asserts that register lowering primarily occurs because of *downstep*. In Yorùbá, "automatic" (or obligatory) *downstep* is said to occur in HLH contexts, indicated by the significantly lower pitch level of the second H tone. The local realisation of the *downstep* phenomenon in HLH contexts is likely caused by a combination of anticipatory local dissimilative H tone raising and carryover assimilation into the the second H in HLH sequences [84]. The expected effects on pitch contours according to this description have been confirmed in HLH contexts in Chapter 3; that is, a significantly lower pitch for the second H than for the first, as well as a higher than average pitch for the initial H compared to three-syllable contours starting with H. However, it is asserted that the pitch level associated with the second H tone results in a "ceiling", effectively changing the pitch "register" (or range) for subsequent tones [9].

In the work of Connell and Ladd, three possible patterns describing the change in register are discussed (see Figure 4.8): Firstly, a discrete lowering of the entire pitch register occurring at downsteps, affecting the pitch levels of all tones. Secondly, a discrete narrowing of the pitch range occurring at downsteps, affecting only the pitch levels of H tones. Lastly, a gradual resetting reversing the effects of downstep, resulting in little net downtrend over the course of an utterance. Although the exact nature of register change was not established, it is suggested that some register lowering does take place during downstep but that in some cases this effect is reversed as in Figure 4.8c. The authors maintain, however, that the majority of downtrend happens in discrete steps during downstep and that "backdrop declination" exists but accounts for a small component in the total downtrend.

Subsequent work by Laniran and Clements [66] found that a "register shift" or "key lowering" of all tone levels (Figure 4.8a) does not occur, showing that L tone levels are not affected by downstep. The authors do however support downstepping of H tone levels, essentially describing a pattern as in Figure 4.8b, with periodic resetting of H tones. Based on this view, an exponential decline function (measuring pitch in Hertz) is fitted to H tone pitch levels with the number of downsteps having occurred being the independent variable and a good fit is found. Similarly to Connell and Ladd,

"background downdrift" in sequences of like tones are noted, but no strong evidence for global, utterance-level declination is found.

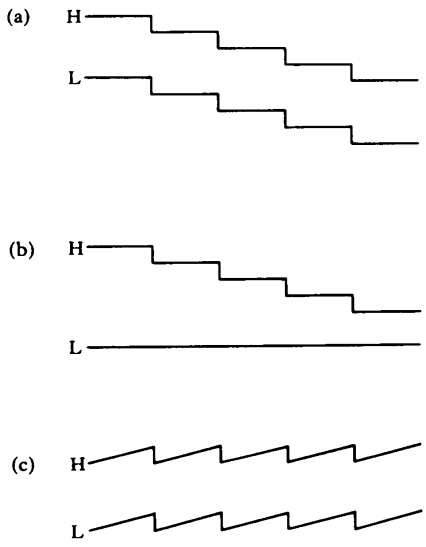

Figure 4.8: *Three possible patterns for downtrend in Yorùbá according to Connell and Ladd (reproduced from [9]). In (a) the entire pitch register is shifted downward by downstep. In (b) the pitch range is reduced systematically by downstep (affecting only H tone pitch level). In (c) downstep shifts the entire pitch register downwards which is gradually reset before the next downstep.*

Regarding downtrend, our own observations (Figure 4.2) suggest that all tone pitch levels are lowered over the course of an utterance (corresponding most closely with Figure 4.8a); the average downtrend observed is approximately linear (pitch measured in semitones) and rates of decline are similar for all tones.

Given these observations, we propose to investigate the modelling of downtrend given the model of local dynamics presented in Eq. 4.3, essentially attempting to determine an appropriate independent variable to capture utterance context.

We perform 10-fold cross-validation experiments (similar to Section C.1), taking the pitch target of the first syllable as known and predicting the targets by applying Eq. 4.3 over the subsequent syllable sequence in each utterance (model: `lind`). Parameters are estimated in different discrete contexts. Local contexts consider the target tone `tt`, previous tone `pt`, pre-previous tone `ppt` and following tone `nt` motivated by the results in Section 4.2.4. The context specification, `tt,pt,ppt,nt`, where $\texttt{tt} \in \{H,L,M\}$ and $\texttt{pt;ppt;nt} \in \{H,L,M,N\}$ thus consists of a maximum of $3 \times 4^3 = 192$

contexts. The complexity of the resulting regression model is controlled by a single meta-parameter, `minsamples`, which determines the minimum number of samples that are required to estimate a distinct set of parameters for a context c. Parameters are estimated during estimation for more general ("back-off") contexts by removing context fields in reverse order specified. The context field ordering is thus important and is motivated at this point by the results obtained in Section 3.3.5. In our experiments, the `minsamples` meta-parameter is determined using training-set-internal 3-fold cross-validation to minimise the RMSE.

Firstly, we repeat the experiment presented in Section C.1 using only local contexts, observing the resulting RMSE and average linear gradient over utterances. The results in the first part of Table 4.3 are comparable with model "lind" in Tables C.1 and C.2 (note however that values in Table 4.3 are averages over 5 iterations). To this we add an explicit context field for the utterance final syllable, $\texttt{uf} \in \{True, False\}$, high up in the order to reliably capture "final lowering", noting a general improvement in the resulting downtrend. The overall downtrend for pitch targets under these models and features, however, still fall significantly short of the measured downtrend in our corpus.

We thus need to extend the context fields to include utterance context. Based on the above discussion there are two plausible alternatives for the independent variable of time:

1. Regular intervals over the course of an utterance; in this case, it is likely that a normalised version of utterance position is appropriate [55]. We implemented this by including a categorical feature capturing utterance context. Two resolutions were considered: $\texttt{uc} \in \{0,1,2\}$ and $\texttt{uc5} \in \{0,1,2,3,4\}$.

2. Intervals based on the number of previous downsteps. Here, we counted the occurrences of HLH sequences since the start of the utterance [9, 66], in our corpus the maximum number of downsteps occurring in any particular utterance was four: $\texttt{ds} \in \{0,1,2,3,4\}$.

The results in Table 4.3 show that a significant decrease in RMSE can be achieved by including the utterance context feature defined at regular intervals, with the average gradient approaching the measured values for all speakers. Using the occurrence of downsteps as a measure of downtrend however is not successful in this experiment, with the RMSEs and average gradients not significantly better than experiments including only local contexts.

Model	Features	Speakers							
		013		017		021		024	
		RMSE	Gradient	RMSE	Gradient	RMSE	Gradient	RMSE	Gradient
Lind	tt	2.56	-0.36	1.99	-0.40	2.24	-0.73	2.31	-0.41
Lind	tt,uf	2.55	-0.43	1.93	-0.51	2.21	-0.76	2.26	-0.52
Lind	tt,pt,ppt,nt	2.50	-0.37	1.93	-0.46	2.17	-0.85	2.18	-0.58
Lind	**tt,uf,pt,ppt,nt**	**2.52**	**-0.40**	**1.89**	**-0.60**	**2.15**	**-0.86**	**2.19**	**-0.70**
Lind	uc,tt	2.48	-0.95	1.80	-1.06	2.06	-1.44	2.15	-1.08
Lind	uc,tt,pt,ppt,nt	2.43	-0.93	1.75	-1.07	2.06	-1.45	2.11	-1.11
Lind	uc,tt,uf,pt,ppt,nt	2.48	-0.92	1.77	-1.06	2.08	-1.44	2.14	-1.09
Lind	uc5,tt	2.48	-0.96	1.78	-1.13	2.07	-1.45	2.15	-1.14
Lind	uc5,tt,pt,ppt,nt	2.46	-0.95	1.74	-1.10	2.06	-1.43	2.11	-1.13
Lind	uc5,tt,uf,pt,ppt,nt	2.48	-0.93	1.77	-1.11	2.07	-1.43	2.12	-1.14
Lind	ds,tt	2.53	-0.48	1.93	-0.61	2.19	-0.86	2.25	-0.53
Lind	ds,tt,pt,ppt,nt	2.52	-0.49	1.89	-0.58	2.15	-0.93	2.18	-0.60
Lind	ds,tt,uf,pt,ppt,nt	2.52	-0.49	1.89	-0.59	2.16	-0.90	2.20	-0.57
Lint	tt	2.52	-1.02	1.83	-1.14	2.13	-1.46	2.22	-1.14
Lint2	**tt,pt,ppt,nt**	**2.41**	**-0.97**	**1.69**	**-1.14**	**2.01**	**-1.52**	**2.04**	**-1.20**
Actual samples			**-1.03**		**-1.16**		**-1.53**		**-1.19**

Table 4.3: *Mean RMSE and linear utterance trends of syllable pitch height targets predicted over complete utterances for the repeated cross-validation experiments (5 iterations).*

Lastly, we compare these results with direct regression models of the utterance pitch targets (i.e. not in terms of the local dynamics), based on an independent variable of time defined at regular intervals (as in (1) above):

- `lint`: A baseline model where declination is assumed linear but may be tone specific. For each of the three tones a linear function is fitted over the normalised utterance position based on syllable index (analogous but not identical to the linear trends plotted in Figure 4.2).

- `lint2`: A model where a shared gradient is assumed, estimated over all the data as a linear function of normalised utterance position in terms of syllable index, with distinct intercepts estimated for different tone contexts implemented with the "back-off" mechanism as for the `lind` models.

The RMSE results for `lint` (Table 4.3) seem to confirm the importance of taking into account the local dynamics in extended tone contexts. The `lint2` model resulted in the lowest measured RMSE which further supports the informal observations regarding a constant linear declination over

all tones.

A discussion of the results in this section is continued below.

4.3.2 Discussion

In this section we have defined a simple regression model to predict pitch height targets in utterances based on the following mechanisms motivated by the results in Section 4.2.4 and Chapter 3:

1. A linear model of local (inter-syllable) dynamics that should account for: anticipatory H-raising; the local (carryover) effect of downstep (H-LH); short anticipatory patterns such as rises and falls over like-tone sequences; pitch range limit effects; and final lowering.

2. A simple prior-knowledge-based tree-like mechanism to control model complexity using ordered categorical context features and a simple "back-off" mechanism to more general contexts (motivated by results in Section 3.3.5).

Using this model we investigated the modelling of downtrend based on regular intervals and downstep, finding the former more successful. Further experiments implementing simple linear declination models add further support for gradual downtrend. While we have not explicitly quantified the relationship between sequences of tones and downtrend, the results presented suggest that the net effect is a gradual fall in pitch, and downstep is likely largely a local effect that may be modelled appropriately by the local dynamics model proposed in Eq. 4.3. This makes it unlikely that downtrend is simply a result of the cumulative effects of local dynamics, which is in agreement with the conclusions found in [91].

Despite the the fact that the simple linear declination model (`lint2`) resulted in the lowest RMSE, given the results in Section 4.2 and considering the underlying causes of the distribution of height targets in local contexts, we argue that a model based on local dynamics is more likely to ensure that aspects important to tone realisation are preserved during model estimation (Section 3.3.4).

Based on the observations and results here, a more complete model might assume a tone-independent linear declination with the local dynamics model capturing the fact that pitch targets may deviate

locally from this "baseline":

$$\begin{aligned} x_n &= a_t x_{n-1} + c_{n-1} + b_t \\ c_n &= c_{n-1} + d \end{aligned} \tag{4.4}$$

However, further work is needed to better understand the nature of parameters in such a model over the course of an utterance.

4.4 CONCLUSION AND FURTHER WORK

In this chapter we have considered the distribution of pitch height targets over the course of an utterance in a corpus of four speakers. Based on previous observations we have selected the target approximation model [21, 65] as a basis for our analysis in Section 4.2, attempting to quantify pitch target distributions meaningfully with the following results:

1. Informal observation of the distribution of extracted pitch targets shows an approximately linear downtrend emerging over the course of utterances.

2. Transitions between syllables in two-syllable contexts (i.e. where the preceding and following syllable labels are known) behave consistently between speakers and are generally distinct.

3. We see significant evidence for the local effects of *downstep* in the H-LH context in all speakers.

4. Changes in pitch at the start and end of utterances vary and are often significantly different from the corresponding general two-syllable context.

5. A strong correlation between current syllable pitch and following syllable pitch is observed suggesting that speakers manage pitch range by scaling pitch changes and effecting pitch resets when the pitch level is close to the lower limit.

6. We did not find a significant correlation between syllable length and pitch change, suggesting that syllable duration in our corpus might be more strongly influenced by the requirements of tone realisation than vice versa.

7. We could not find a significant systematic effect of vowel height on the pitch targets extracted from our corpus – vowel identity in general only seems to be a weak effect here.

Furthermore, in Section 4.3 we investigated the modelling of downtrend in the context of a model based on the above observations, observing the following:

1. Regression models based on (normalised) regular intervals over the course of an utterance perform better than models based on the occurrence of downstep.

2. Models assuming linear downtrend resulted in the lowest RMSE observed in our experiments.

Consequently, we find it unlikely that downtrend is simply a result of the cumulative effects of local dynamics, and propose a simple dynamic model allowing for both constant tone-independent declination as well as distinct local dynamics for further work on this topic. Immediate questions that need to be answered towards such a model involve pitch resetting (i.e. the management of pitch range on the utterance level) and initial conditions, (i.e. pitch targets for utterance and breath-group-initial syllables). While we believe that the simple conclusions regarding downtrend presented here are sufficiently supported by evidence from the limited corpus used here, a more detailed investigation should ideally be conducted on a more suitable corpus containing longer utterances representing full sentences and with diversity in terms of length and number of breath-groups.

In the following chapter we focus on generating and evaluating appropriate intonation contours in a complete TTS system based on the target approximation model adopted here and observations in Chapter 3.

CHAPTER 5

PITCH MODELLING FOR YORÙBÁ TEXT-TO-SPEECH SYNTHESIS

Increasingly powerful and efficient algorithms and models for speech and language processing have recently enabled the successful construction of corpus-based acoustic models for TTS systems in under-resourced environments [18, 19]. However, the construction of systems that adequately account for tone information continues to be a challenge, with basic systems often not incorporating tone information at all [8, 7]. This may result in degraded intelligibility and naturalness of resulting speech in various ways depending on the language [20].

Regarding corpus-based acoustic modelling in under-resourced environments, difficulties stem from the practical challenges of developing high-quality text and speech corpora. Firstly, the development of balanced text corpora is often both more expensive and technically challenging than in the case of well-resourced languages, due to lack of diversity in digitised content (available on the Internet) and challenges with the standardisation of existing digitised text [72]. This is compounded by the additional variable: tone. Secondly, the flexibility of the acoustic modelling process implemented in HMM-based techniques such as HTS, while reducing the phonetic expertise required to build realistic systems, is sensitive to both the quality and the quantity of speech recordings. Thus to build a high-quality (even successful) system from a limited text corpus, a professional speaker experienced with language technology is often needed. This is especially true in the case of tone languages, with pitch being one of the more variable acoustic features, thus limiting the practical reach (applicability) of resulting systems due to inevitable mismatches in dialect, persona, gender, etc. (see Chapter 1).

5.1 APPROACH

In this work we have thus far primarily sought to characterise tone realisation in terms of pitch with the goal of implementing an *efficient* intonation model for application in less than ideal circumstances. In this chapter we use these observations to develop a tone-aware HMM-based TTS system for Yorùbá using HTS [86] and an intonation model based on the quantitative target approximation (qTA) approach [65] suitable for use in such a system. A comparison is made between the effectiveness of these two techniques by means of analytical and perceptual evaluations with the aim of establishing the applicability of these approaches in a real-world scenario. This should provide practical insight to developers of TTS systems for tone languages in under-resourced environments.

In the following section we describe the corpus development process and resulting properties. This is followed by a description of the development of the TTS system, including parameter extraction, the handling of breath groups as well as considerations when modelling pitch using qTA and HTS. Finally we present analytical results on our test set as well as a perceptual evaluation in the form of a preference test.

5.2 CORPUS DEVELOPMENT

A small single-speaker speech corpus was developed to support the construction of a general purpose TTS system. Text was sourced from the Yorùbá language version of *Wikipedia*[1] retrieving the database dump dated 2013-05-25. The raw text data was cleaned up semi-automatically to remove foreign language content (especially English), standardise the representation of diacritics and punctuation, and sentencise the result. The result was then divided into "clean" and "normal" subsets based on simple heuristics including the detection of digits, English words and the presence of diacritics. These subsets amounted to approximately 11400 lines containing 145000 tokens and 11800 lines containing 156000 tokens respectively. Both sets were phonetised where possible (depending on the validity of the orthography) using the grapheme-to-phoneme rules developed in Section 3.2.1. Examining the diphone and tri-tone frequencies in both sets, we found that the second set did not contribute any additional diphones to the set of more frequently occurring units (containing more than 10 samples) and thus selected the "clean" set as basis for text selection, particularly because we assume that a significant fraction of these will be of foreign origin in this case. Although it is known

[1] `http://yo.wikipedia.org`

that these distributions generally have large numbers of legitimate rarely occurring units [37], such units are nevertheless likely to be the minority of the low-frequency items in the "normal" set. This procedure solves a significant problem when selecting a very limited subset from a noisy corpus using a greedy selection algorithm [71]: the greedy selection algorithm not only selects text containing a lot of foreign words, but these selections also tend to be smaller, less readable fragments.

Text selection proceeded for diphone coverage [92], with the stopping criterion of at least 10 samples of each diphone in the target set, identified from the "clean" text above by considering all diphones having at least 10 samples. The selected set contained 640 sentences (or fragments) consisting of 13654 tokens. This set underwent a process of professional proofreading and editing (two iterations) to:

- correct spelling and diacritics,
- naturalise, replace or mark foreign words where possible, and
- ensure proper sentencisation and syntactically complete sentences.

After post-editing the text contained 794 sentences and 13632 tokens,[2] still containing a small number of foreign words and names. To this the Yorùbá translation of the *Universal Declaration of Human Rights* document[3] (109 sentences and 2455 tokens) was added to serve as a pristine test set. Remaining foreign words and names marked above and tokens containing invalid graphemes were processed by the TTS text-analysis components as English tokens.

The sentences were randomised and recorded in a professional studio environment over the course of two consecutive days using a sampling rate of 44.1 kHz and 16-bit sample precision by a male first-language Yorùbá speaker. The resulting recordings were conservatively post-processed to reduce traces of reverberation using the "deverb filter" implemented in the freely available *Postfish* software.[4] Phonemic alignment of utterances and transcriptions, including the marking of pauses and breath-groups, was done exactly as in Section 3.2.1 using HTK with the parameters as in Appendix A. Lastly, utterances and orthographic transcriptions were inspected for mismatches by flagging outliers in a process of training and resynthesis using HTS and the HTS engine (more details in Section 5.4)

[2] as a result of 547 insertions and 393 deletions on the sentence level, and 1782 insertions and 1804 deletions on the token level.

[3] Dated 1998-11-12 and retrieved from: `http://www.ohchr.org/EN/UDHR/Pages/Language.aspx?LangID=yor`

[4] Available at: `http://svn.xiph.org/trunk/postfish`

and comparing the cepstral distance [93] as described in [94]. Manual corrections were made to inaccuracies detected in the transcriptions in this way by removing and correcting words that were inserted, deleted or mispronounced. Five iterations of this process were performed until no mismatches were observed in the top 150 words ordered by decreasing cepstral distance.

After the recording and post-processing stages, a total of 788 and 109 utterances were contained in the "train" and "test" sets respectively, with a subset of the train set ("clean") determined where all tokens could be processed by the Yorùbá text-analysis components (resulting in valid tone labels). The final corpus properties are summarised in Table 5.1.

Corpus	Utterances	Breath-Groups	Words	Syllables				Phones	Duration (mins.)
				H	M	L	N	excl. `pau`	excl. `pau`
train	788	1704	13778	8650	9471	7753	408	45863	86
train clean	654	1377	11171	7234	7564	6543	0	37248	70
test	109	276	2543	1609	1382	1753	0	8183	15

Table 5.1: *Corpus properties with syllable counts by tone (*`N` *indicates "None", mostly resulting from foreign words or names that were not processed by the Yorùbá text-analysis components). The number of phones and corpus duration exclude pauses.*

5.3 SYSTEM

5.3.1 Pitch extraction

The speaker's pitch range was determined by extracting F0 for all utterances using Praat in the range 50 to 600 Hz and using a histogram to examine the resulting distribution. Consequently, the range was reduced to **50 to 250 Hz** and all values were re-extracted and converted to semitones at millisecond intervals as before. No smoothing was applied in this case and for the HTS system pitch was extracted in exactly the same way except for converting to log F0 (instead of semitones) as is standard practice.

For extracting the qTA parameters, *height*, *gradient* and *strength* (see Section 4.2.2), we implemented the analysis-by-synthesis process as described in [65] and applied this to each breath-group separately as follows:

1. Find the starting F0 value by taking the mean over valid F0 values in the first quarter of the

breath-group-initial syllable and set the velocity and acceleration to zero.

2. Iterate over syllables and parameter ranges finding the best parameter set by minimising the RMSE between the actual and contours resulting from Eq. 4.2 as described in [65]. For cases where no valid F0 values could be found in a syllable (due to voicing estimation errors), a linear interpolation was used.

3. The initial conditions for each syllable is determined by fitting a cubic spline to the result and using this to estimate the pitch and derivatives at the syllable boundary (this is used to calculate c_i in Eq. 4.2). If a pause exists before the following syllable, the initial velocity and acceleration are set to zero for that syllable.

The process of extracting target parameters is quite sensitive to the placement of syllable boundaries – sometimes finding unexpected parameters given small misalignments – if the search space is left unconstrained. Hence, the search space was limited according to previous observations about tone contexts as described below. One example of such an unexpected fit occurs if a syllable boundary is placed slightly late in an HL transition: an unconstrained search may then find a negative gradient for the H tone based on the portion of the contour that should be associated with the L tone. In Yorùbá in particular, this may happen more frequently than expected in contexts where the syllable boundary between repeated (same) vowels with alternate H and L tones is sometimes almost exclusively defined by pitch movement. This scenario may cause difficulties during the automatic alignment process which does not consider F0. Due to this observation we also experimented briefly with alignments based on HTS models, which include F0 as a feature, but did not observe consistently better placed syllable boundaries in these vowel-vowel contexts and a comparison of resynthesis mean squared error did not indicate gains.

Thus, the qTA parameter search ranges were constrained based on the observations in Chapter 3, assuming that positive and negative gradients are significant in determining H and L tones respectively and that M tones are suitably represented with a flat pitch target (see Section 3.3.4). In applying constraints to the parameter search space, we initially experimented with three sets; “minimal”, “moderate” and “full” representing different degrees of constraint especially of the gradient and strength parameters. Cross-validation experiments as described in Section 5.5 indicated that small gains in both RMSE and correlation are possible if the feature extraction process is more constrained. However, inspection of some synthesised contours suggested that the small gains in RMSE and corre-

lation measures may not translate to perceptible improvement for smaller time intervals, but that over-constraining the search space may negatively affect the naturalness of synthesised speech utterances as a result of reduced variation over longer time intervals. We therefore chose to continue this investigation specifically using the least constrained set.

The details of the minimally constrained search space as implemented are as follows by parameter:

- **Height**:

  $$x \in linspace(a = 67.73, b = 95.59, c = 100),$$

  where *linspace* generates a linearly spaced vector including endpoints a and b in semitones with size c (the number of intervals is thus $c-1$), resulting in a resolution of about 0.28 st. This pitch range corresponds exactly to the limits used during F0 extraction (50 to 250 Hz) and was used for all tone contexts.

- **Gradient**: The absolute limits on gradient were set at $-60 \leq g \leq 60$ st.s^{-1} with a resolution of 4 st.s^{-1}, loosely corresponding to the findings on maximum speed of pitch change in [54]. Ranges were also limited more specifically for different tone contexts based on the assumptions outlined above. Constraints for H, L and M tone syllables were set as follows:

  $$g_H \in linspace(0, 60, 16)$$
  $$g_L \in linspace(-60, 0, 16)$$
  $$g_M \in \{0\}$$

- **Strength**: Here we simply applied a single set of constraints over all tone contexts. The non-zero minimum strength limits the potential distance between targets and the actual F0 contour and the maximum constraint corresponds to the maximum value used in [65] to allow for fully realising targets in short syllables. The resolution is set to 5 s^{-1}:

  $$\lambda \in linspace(40, 120, 17)$$

Given the extracted F0 and qTA parameters we proceed to discuss the considerations and implementation of pitch modelling using qTA and HTS in the follow subsections.

5.4 PITCH MODELLING AND SYNTHESIS USING HTS

Here, as is the case throughout this work, we used the HTS toolkit version 2.2 [86, 87] in combination with the HTS engine version 1.05. A brief overview of the theory of HMM-based acoustic modelling is presented in Section 2.1.2 and a more detailed description can be found in [19]. We used the standard training procedure with the following exceptions, additions and considerations highlighted:

- F0 extraction did not make use the default tool in the HTS demonstration script, but *Praat* as described above.

- We implemented mixed excitation as described in [51] with appropriate modifications to the training script to model the extra speech parameter streams representing bandpass voicing strengths in five frequency bands.[5] Synthesis is performed by a slightly modified version of the HTS engine[6] supporting mixed excitation speech generation and pulse dispersion post-filtering.

- Relying on the HTS engine for synthesis means that our acoustic models employ a single mixture per HMM state.

- We make use of models with a global variance [23], and parameter sharing implemented using decision trees is based on the standard minimum description length (MDL) criterion [42].

Keeping the training and synthesis process standard, we experimented with four sets of features including tone context to different degrees:

1. **No tone** (`none`): The "standard" HTS phone and positional features with the exclusion of "guess-part-of-speech" ("gpos"), which distinguishes frequently occurring grammatical words and categories.

2. **Immediate context** (`pt,tt,nt`): The standard features, including features identifying the current, previous and following syllable tone.

3. **Preceding context** (`ppt,pt,tt`): Includes features identifying the current, previous and preprevious syllable tone.

[5] See: `http://mary.opendfki.de/wiki/HMMVoiceCreation`

[6] Available here: `http://sourceforge.net/projects/me-hts-engine`

4. **Extended context** (`ppt,pt,tt,nt`): Includes features for the current, previous, pre-previous and following syllable tone.

For each of the tone features, questions were added to allow state clustering of tones in all possible contexts using the standard decision tree models.

To investigate the effect of different feature sets analytically, we performed 10-fold cross-validation experiments on the clean training set, calculating the root mean squared error (RMSE) and Pearson correlation between synthesised and actual F0 contours. The experiment was repeated for 3 different random partitionings of the data by utterance (A, B and C) and to obtain meaningful RMSE and correlation values we found it necessary to smooth the synthesised contours. This was done using cubic smoothing splines, and measures were calculated only for non-zero (voiced) sections in both the reference and target contours.

The results are presented in Table 5.2, giving an indication of the squared error and correlation baselines without tone information and showing significant improvements in both measures when tone information is included. Comparing the systems with different amounts of tone information, the system with **extended context** features performed marginally better and most consistently. This and previous observations regarding the importance of tone context lead us to adopt this system for further comparison here.

Tone context	RMSE				Correlation			
	A	B	C	Mean	A	B	C	Mean
`none`	3.436	3.418	3.422	3.425	0.491	0.495	0.497	0.494
`pt,tt,nt`	3.206	**3.149**	3.178	3.178	0.570	**0.585**	0.576	0.577
`ppt,pt,tt`	3.203	3.164	3.178	3.182	0.573	0.579	**0.577**	0.577
`ppt,pt,tt,nt`	**3.178**	3.171	**3.174**	**3.174**	**0.579**	0.580	0.573	**0.577**

Table 5.2: *Root mean squared errors and correlations for the HTS cross-validation experiments. A, B and C refer to independent experiment iterations using different random partitionings, with means in the shaded columns. The best values in each column are indicated in bold.*

5.5 PITCH MODELLING AND SYNTHESIS USING QTA

5.5.1 Synthesis algorithm

In order to synthesise appropriate contours using qTA, we need to predict plausible sets of values including *height*, *gradient* and *strength* for each syllable. As mentioned in Section 5.3.1 we assume both the *height* and *gradient* parameters to be crucial to tone realisation. This is supported by earlier contour synthesis experiments (see Section D.1 in Appendix D). While we have not investigated the status of the *strength* parameter in this regard, we assume it at least important to the overall naturalness of synthesised speech, e.g. in "higher level" (semantic) contexts the requirement of strong tone realisation might normally be relaxed due to unambiguous sentence or discourse context of the word.

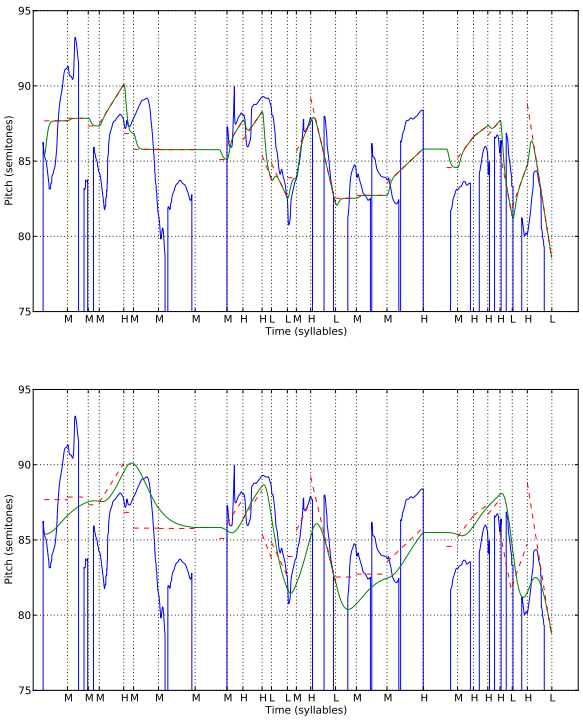

Figure 5.1: *Examples of synthesised contours (green) from predicted height and gradient targets (red) compared to the original unseen F0 contour (blue) in the training corpus. The top figure (a) illustrates the result for high values of the strength parameter. The second figure (b) illustrates the result given the same targets and the strength limiting synthesis algorithm proposed here using a low value of* $10\ s^{-1}$ *for the minimum strength.*

An appropriate regression model would need to consider the intra- and inter-syllable relationships between these parameters in order to ensure the continuity (or smoothness) of the resulting contour. In the absence of sufficient knowledge to construct such a complex model, we begin by assuming a strong specification of *height* and *gradient*; constructing a simple set of regression models estimating and predicting *height* and *gradient* independently in different tone contexts and considering the consequences on resulting contours. Figure 5.1a illustrates a potential problem: unnatural sharp acceleration especially at syllable boundaries which may be perceptually troublesome.

Initial attempts at reducing such artefacts using regression models for *strength* conditional on the *height* and *gradient* parameters while improving the result marginally were unsatisfactory. This leads to a heuristic solution aimed directly at the observed problem: the requirement is essentially to limit sharp acceleration in the opposite direction of height targets, without limiting the velocity of pitch movements required for the implementation of steep targets and natural effects of inertia at syllable boundaries. This was implemented by systematically limiting the *strength* in an iterative synthesis process (Eq. 2 in [65]) until no acceleration in the opposite direction of the height target is present or the minimum strength is reached. For this implementation we again used cubic splines to estimate acceleration and limited the strength in steps of 5 s^{-1}.

An example of the result of this synthesis algorithm (Figure 5.1b) suggests a potential criticism: flat targets (M tones) will generally be implemented with a low strength due to the deceleration required to approach such targets. While *height* seems to be the best indicator of M tones based on results in this work, the strength requirements have not been investigated here. In some works it has been suggested that the M tone is essentially "targetless" (an earlier work by Akinlabi is cited in [67]), while in work on the target approximation model the concept of a "targetless" syllable is disputed, instead it is suggested that low strength may be a distinct feature of certain tones, with specific reference to the neutral tone in Mandarin and M tone in Yorùbá [88] (see also the observations in Section 4.2.4). Experimental examples from Mandarin in [88] supporting this hypothesis resemble the synthesised examples for the first few M tone sequences in Figure 5.1b where targets are eventually approached after two or three syllables. The synthesis algorithm proposed here may thus be considered plausible. Given the initial concern we briefly experimented with a slightly modified algorithm to limit the strength based on acceleration if the gradient target is steep and limit the velocity when gradient targets are relatively flat. This adds at least one additional parameter if one simplifies the set of meta-parameters by setting the velocity limit equal to the gradient threshold at which the limiting mechanism is selected. In the absence of a suitable way to determine these extra parameters (see the

reservations expressed below regarding using RMSE and correlation directly), we adopt the simpler algorithm here.

Having already proposed a synthesis algorithm with a strong influence on the *strength* parameter, we extend the heuristic approach to completely determine the parameter by predicting a maximum strength value for each syllable and relying on the systematic constraint of acceleration to determine the eventual value. Such an approach is at least partially supported by the observation that the *strength* parameter values found during the analysis-by-synthesis process seem to be more variable depending on constraints on the *height* and *gradient* parameters than vice versa [65], and inspection of the distribution in our corpus where a majority of values are found to be at either the minimum or maximum values of the search range (i.e. 5772 and 9325 syllables from a total of 21341 at a minimum and maximum respectively). This reduces the required number of parameters for *strength* to the two meta-parameters: minimum and maximum.

The strength meta-parameters were determined using cross-validation experiments on the training set, measuring RMSE and correlation (details of the experimental setup follow). The results are presented in Table 5.3. It was found that varying the maximum parameter beyond 100 s^{-1} had a limited effect on contours while the minimum parameter was more significant, broadly determining the "smoothness" of contours. Direct optimisation using RMSE and correlation was not suitable as this led to overly high values of the minimum *strength*. We thus used the measurements as an approximate guide, selecting the lowest value before "over-smoothing" resulted in a significant increase in RMSE and decrease in correlation (see highlighted fields in Table 5.3).

5.5.2 Regression models

To model syllable targets in utterance context we initially experimented with including and excluding breath-group information obtained using the heuristic based on the length of detected pauses employed during alignment (Section 3.2.1). The main concern was that this information would prove particularly unreliable in this corpus due to a perceived decrease in reading fluency caused by the variable quality of the text. This uncertainty led us to abandon further attempts at developing models such as described in Chapter 4, which require accurate information of this nature.[7] Cross-validation error rates based on the models and features described below were comparable when including and

[7] the corpus developed here should however be able to support further research of this nature once breath-groups have been determined reliably.

Model	Strength min/max	Syllable features	RMSE				CORR			
			A	B	C	Mean	A	B	C	Mean
Mean (eval. height)	50/120	bl53,si510,tt,uf,pt,ppt,nt	2.940	2.935	2.931	2.936	0.550	0.549	0.550	0.550
Mean (eval. height)	40/120	bl53,si510,tt,uf,pt,ppt,nt	2.936	2.932	2.927	2.931	0.552	0.550	0.552	0.551
Mean (eval. height)	30/120	bl53,si510,tt,uf,pt,ppt,nt	2.942	2.940	2.935	2.939	0.550	0.546	0.550	0.549
Mean (eval. height)	**20/120**	**bl53,si510,tt,uf,pt,ppt,nt**	**2.998**	**2.997**	**2.993**	**2.996**	**0.529**	**0.525**	**0.529**	**0.527**
Mean (eval. height)	10/120	bl53,si510,tt,uf,pt,ppt,nt	3.402	3.398	3.395	3.398	0.424	0.421	0.423	0.423
Mean (eval. height)	20/100	bl53,si510,tt,uf,pt,ppt,nt	2.998	2.997	2.992	2.996	0.529	0.525	0.529	0.528
Mean (eval. height)	20/110	bl53,si510,tt,uf,pt,ppt,nt	2.998	2.997	2.993	2.996	0.529	0.525	0.529	0.528
Mean (eval. height)	50/120	bl53,si56,bgi5,tt,uf,pt,ppt,nt	2.959	2.958	2.957	2.958	0.543	0.542	0.541	0.542
Mean (eval. height)	40/120	bl53,si56,bgi5,tt,uf,pt,ppt,nt	2.956	2.954	2.954	2.954	0.544	0.544	0.542	0.543
Mean (eval. height)	30/120	bl53,si56,bgi5,tt,uf,pt,ppt,nt	2.964	2.962	2.962	2.963	0.541	0.540	0.539	0.540
Mean (eval. height)	**20/120**	**bl53,si56,bgi5,tt,uf,pt,ppt,nt**	**3.026**	**3.022**	**3.022**	**3.023**	**0.519**	**0.520**	**0.518**	**0.519**
Mean (eval. height)	10/120	bl53,si56,bgi5,tt,uf,pt,ppt,nt	3.459	3.463	3.453	3.458	0.413	0.412	0.413	0.413
Mean (eval. height)	20/100	bl53,si56,bgi5,tt,uf,pt,ppt,nt	3.025	3.022	3.021	3.023	0.520	0.519	0.518	0.519
Mean (eval. height)	20/110	bl53,si56,bgi5,tt,uf,pt,ppt,nt	3.025	3.022	3.022	3.023	0.520	0.519	0.518	0.519

Table 5.3: *Root mean squared errors and correlations using the independent means model for the cross-validation experiments considering different minimum and maximum strength constraints (in s^{-1}) as well as features including and excluding breath-group information. In the first section features were determined in utterance context and in the second section in breath-group context. Bold rows show the results for the adopted strength meta-parameters used in further experiments, with red fields indicating the significant reduction in performance due to "over-smoothing". A, B and C refer to independent experiment iterations using different random partitionings, with means in the shaded columns.*

excluding breath-group information (see Table 5.3). Consequently, we chose to include breath-group information in subsequent experiments, as was done during HTS modelling.

Two mechanisms were tested to model distinct contexts:

1. The simple "knowledge-based" tree implementation, first described in Section 4.3.1, using ordered categorical features with a back-off mechanism motivated by previous observations of the relative importance of utterance and tone contexts to subdivide samples for estimation. This mechanism employed a minimum-samples-per-context meta-parameter to control model complexity.

2. Independent modelling of the parameters using standard regression trees [95] as implemented in *Scikit-learn* [96] using a combination of categorical and numerical features, also using minimum-samples-per-leaf as meta-parameter.

The categorical features employed included target tone (`tt`), utterance final syllable (`uf`), previous

tone (pt), pre-previous tone (ppt) and following tone (nt), with the following utterance contexts categorical in the case of (1) and numerical for (2): length of the utterance or breath-group (bl) based on 5-syllable chunks into 3 states (bl53), i.e. states for 1 to 5, 6 to 10 and > 10 syllables, syllable index in utterance or breath-group (si) similarly using 5-syllable chunks and having up to 6 states (si56) and breath-group index (bgi) allowing for up to 5 states (bgi5).

Models based on mechanisms (1) and (2) were defined and estimated as follows:

1. To determine the minimum samples (minsamples) meta-parameter 3-fold cross-validation was performed to select a value resulting in a minimum RMSE of the *height* parameter. The same value of minsamples was used for both the *height* and *gradient* parameter estimates. For the estimation of regression models in each context we experimented with two simple models:

   (a) Taking the mean of each parameter independently (mean), and

   (b) A joint model of *height* and *gradient* (linR2) using multiple linear regression given the previous syllable values estimated with least squares (values were scaled to the same range). This model is an extension of the linear regression local dynamics model proposed in Eq. 4.3 to two variables.

2. For the standard tree models (tree) we optimised the regression trees for *height* and *gradient* independently with 3-fold cross-validation also using a MSE criterion.

Results given these models in a 10-fold cross-validation experiment as described in Section 5.4 are presented in Table 5.4.

Model	Syllable features	RMSE				Correlation			
		A	B	C	Mean	A	B	C	Mean
Mean (eval. height)	bl53,si56,bgi5,tt,uf,pt,ppt,nt	3.026	3.022	3.022	3.023	0.519	0.520	0.518	0.519
LinR2 (eval. height)	bl53,si56,bgi5,tt,uf,pt,ppt,nt	3.069	3.063	3.078	3.070	0.507	0.509	0.502	0.506
Tree (indep.)	**bl,si,bgi,tt,uf,pt,ppt,nt**	**2.959**	**2.955**	**2.960**	**2.958**	**0.545**	**0.544**	**0.546**	**0.545**

Table 5.4: *Root mean squared errors and correlations for the qTA cross-validation experiments. A, B and C refer to independent experiment iterations using different random partitionings, with means in the shaded columns.*

While the difference in squared errors and correlations obtained for the three models are likely not perceptually significant, these results suggest that (given the synthesis algorithm implemented) the height and gradient targets may successfully be modelled as independent parameters.

Given the competitive results obtained using the tree model, we repeat the cross-validation experiment once more using this model to investigate the utility of including vowel identity (`vow`, with 8 possible states including "no vowel") and syllable onset voicing (`voi`, with 3 possible states including "no onset" for vowel-only or syllabic-nasal cases) features in this context. This is motivated by the concept of intrinsic F0 [56] and the reported interaction between consonants and tone due to respective influence on pitch [97]. Results are presented in Table 5.5.

Model	Syllable features	RMSE				CORR			
		A	B	C	Mean	A	B	C	Mean
Tree (indep.)	bl,si,bgi,tt,uf,pt,ppt,nt	2.959	2.955	2.960	2.958	0.545	0.544	0.546	**0.545**
Tree (indep.)	bl,si,bgi,tt,uf,pt,ppt,nt,vow	2.957	2.954	2.953	**2.955**	0.545	0.545	0.542	0.544
Tree (indep.)	bl,si,bgi,tt,uf,pt,ppt,nt,voi	2.974	2.968	2.972	2.971	0.539	0.539	0.537	0.538
Tree (indep.)	bl,si,bgi,tt,uf,pt,ppt,nt,voi,vow	2.975	2.968	2.976	2.973	0.538	0.539	0.537	0.538

Table 5.5: *Root mean squared errors and correlations for the qTA cross-validation experiments including vowel and onset voicing features. A, B and C refer to independent experiment iterations using different random partitionings, with means in the shaded columns.*

Here again, as in Section 4.2.7, we fail to measure a significant effect by considering vowel identity. Neither did our simple specification of syllable onset voicing have any measurable effect. Further work is needed to gain insight in this regard.

5.6 RESULTS

5.6.1 Analytical tests

Having defined complete intonation models using HTS and qTA process in the previous sections, we attempt to compare these models for their efficiency in modelling tone contours by measuring the RMSE and correlation on the unseen test set given portions of the training set for estimation by executing 5 iterations of the following experiment:

1. Given the complete "clean" training set, create 6 subsets that contain 100%, 75%, 50%, 25%, 10% and 5% of the original number of utterances by iteratively randomly discarding utterances

so that each smaller set is a subset of the previous set.

2. Estimate models using HTS and qTA models for each subset. For the qTA models, we tested the simple back-off mechanism as well as the independent tree models and features as presented in Table 5.4. For HTS we used the **extended context** features.

3. Synthesise and calculate RMSEs and correlations against the original speech samples over the entire held-out "test" set.

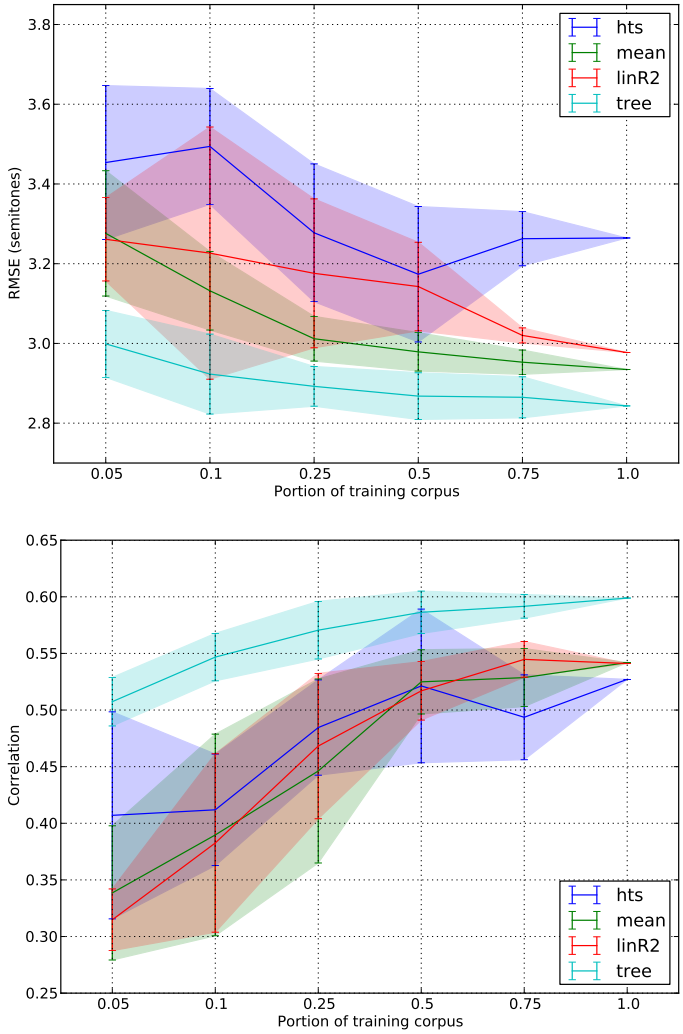

Figure 5.2: *Root mean squared errors and correlations on the held-out test set for models estimated from portions of the clean training set. Plots show the mean of 5 iterations using different randomly selected subsets and error bars show the 95% confidence intervals.*

The results are plotted in Figure 5.2. While the interpretation of these results is not straightforward, owing to unknown level of perceptual significance, we make the following observations:

- A lower RMSE may be achieved in general using the any of the target approximation models; however, the HTS models are relatively competitive in modelling appropriate pitch movement captured by the correlation measure; the simple TA models are presumably less successful at preserving appropriate pitch movement for the smaller subsets.

- Being a more flexible model, the HTS results are clearly relatively sensitive to the data and significant variation is present throughout the range tested here. The test set contained mostly relatively long utterances and the training set a mixture of lengths, which may explain why some selections at 50% may have resulted in models with a lower RMSE and better correlation than using the full corpus.

- The modelling of *height* and *gradient* using independently estimated trees was most consistent with better performance on both measures over the entire range tested.

Despite the approximate nature of the measures employed, we suggest that the consistency and margin achieved by the tree-based qTA model indicates a more accurate intonation model, better able to model and preserve the acoustics of tone in modest to very limited speech corpora such as tested here.

5.6.2 Perceptual test

While the results in the previous section strongly support the independent tree-based qTA model, the possibility remains that these simpler models do not account for all the crucial aspects of pitch or that the synthesis algorithm results in perceptually disturbing artefacts that are not penalised by the analytical measures we have employed. This concern motivates the implementation of a simple perceptual experiment with the aim of determining whether the qTA model is broadly beneficial in an HMM-based TTS system as developed here.

We thus set up a simple preference test comparing the HTS synthesised samples with samples when replacing F0 before vocoding with contours generated by the tree-based qTA model (thus leaving all other parameters generated by the tone-aware HTS system identical). Based on the results in especially Section 3.3.3, we assume it important to include tone features also for the modelling of

other acoustic parameter streams. Note, despite the fact that the qTA model does not determine voicing, this information is contained in the bandpass strengths already modelled, thus no voicing information was needed from the HTS MSD-HMM F0 model in samples using the qTA model.

For the perceptual experiment we selected 30 sample pairs from the "test" set (Table 5.1) by dividing the utterances into three sets based on length; "short", "medium" and "long" having from 1 to 10, 11 to 20 and more than 20 words and randomly selecting ten utterances from each set. The resulting properties for this perceptual test set are presented in Table 5.6.

Corpus	Utterances	Breath-groups	Words	Syllables			Phones	Duration (mins.)
				H	M	L	excl. `pau`	excl. `pau`
Test subset	30	77	515	334	278	340	1639	3

Table 5.6: *Properties for the synthesised test set with syllable counts by tone. The number of phones and duration exclude pauses.*

The preference test was conducted via a simple HTML form with respondents presented with the original text and able to listen to samples without limitation. The question to listeners was "Which sample do you prefer?" Respondents answered this by simply selecting either of the samples or "no preference". The results of this evaluation completed by 7 respondents[8] are presented in Table 5.7, showing the number of utterances preferred for each model (partitioned by utterance length).

Utterance length (words)	HTS	qTA	No preference	Total	χ^2
$n \leq 10$	3	51	16	70	41.782
$10 < n \leq 20$	17	16	37	70	0.008
$20 < n$	22	13	35	70	2.064
All	42	80	88	210	11.527

Table 5.7: *Perceptual preference.*

According to McNemar's test statistic using the chi-squared distribution with 1 degree of freedom and Yates' continuity correction, the 95% confidence level is given by $\frac{(|b-c|-0.5)^2}{b+c} \geq 3.841$. These results confirm that the qTA model is indeed applicable and suggests that it is broadly preferred over HTS in this context. Our interpretation of the results are as follows:

[8] respondents were linguistically naïve first language speakers of Yorùbá

- Firstly, as expected, listeners found it more difficult to judge longer utterances (the χ^2 values do not indicate a significant difference here). This may be because neither of the systems are faultless and longer utterances may include more cases of conflicting quality contrasts.

- In the short utterances listeners clearly preferred samples synthesised using the qTA model. Closer inspection would suggest that this preference stems from more accurate modelling of tone, which is easier to perceive when an utterance consists of only a few words. See examples in Figure 5.3, where we argue that the tones in the final two syllables in both utterances are better represented by the output of the qTA model.

- While the results obtained for medium and long utterances are not significantly in favour of the qTA model as in the case of short utterances, the results suggest that the modelling of downtrend over utterances is comparable to the HTS model. In Figure 5.3 the effect of utterance context can be seen by comparing the contours between the two utterances over the first 5 syllables. The tone sequence is identical and the contour is similar in form (more consistently so than the HTS example). However, in the longer utterance the second and third syllables are higher and in the short utterance the final L ends lower (at approximately 95 Hz). The final L in the longer utterance also eventually ends at approximately 95 Hz which agrees with assertions in [55] regarding the relative stability of utterance offset frequency.

- The fact that the HTS model more directly reflects the underlying training data leads to more variation in the realised pitch contours. While this may be detrimental to tone realisation in certain contexts, it may also result in certain words being perceived as more prominent. We suspect that this property likely contributes positively to perceived quality especially in longer sentences.

These observations support the applicability of the qTA model developed here. Further work will be discussed in the following section.

5.7 CONCLUSION AND FUTURE WORK

In this chapter we have developed and evaluated intonation models for Yorùbá TTS based on HTS and qTA in a real-world scenario. This was done by developing an HMM-based TTS system for Yorùbá, including the development of new text and speech corpora for this purpose.

For the qTA model, suitable regression models for the prediction of *height* and *gradient* were evaluated based on work in Chapter 3 and 4 and a synthesis algorithm was proposed for the heuristic specification of *strength*. Analytical and perceptual results showed respectively that a model based on independent regression trees is a promising option for speech synthesis development in under-resourced conditions and that the models and synthesis algorithm are applicable in synthesisers, with comparable results to the widely used HTS implementation and with improved tone realisation in some cases.

As well as being a more accurate and efficient model under data constraints, a working intonation model based on simple relevant parameters provide developers of TTS systems more opportunity for intervention for the implementation of higher level prosodic effects using little additional data or even directly based on theory (e.g. in the PENTA framework [21]) or the robust modelling of tone in the face of noisy or non-ideal data by employing stricter constraints based on prior knowledge during feature extraction. The current implementation modelling *height* and *gradient* independently is also an ideal vehicle for further work on practical dynamic models especially to predict height targets, potentially reducing the data requirements for model estimation further by removing the need for utterance context indices. This will be discussed further in the following chapter.

Immediate further work should include more detailed perceptual tests in the form of "Blizzard-style" naturalness and intelligibility tests [32, 33]. This may result in more insight into the appropriateness of the current modelling of M tones, and also allow for a comparison with other efforts to develop tone-aware synthesisers for African tone languages such as Ibibio [20].

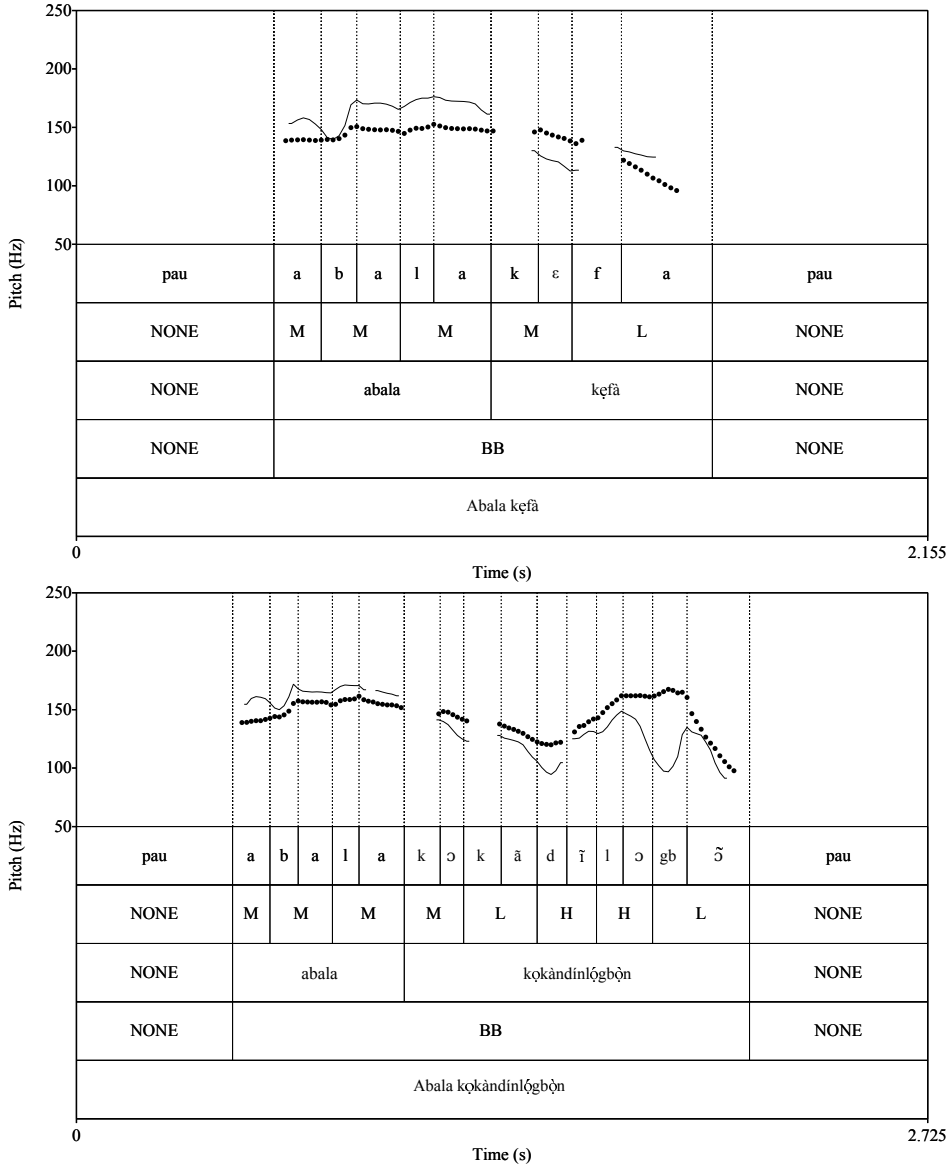

Figure 5.3: *Examples of HTS (solid line) and qTA (dotted line) pitch contours for two synthesised utterances from the perceptual test set where respondents unanimously preferred the qTA samples. In both utterances the final two syllables are perceptually distinct and correspond more clearly with the patterns uncovered in Chapter 3 and 4 in the case of qTA samples. Comparing the qTA contours over the first five syllables, with identical tone sequence, distinct downtrends can be seen.*

CHAPTER 6

CONCLUSION

This work represents a step towards developing more accurate pitch models for African tone languages with application in speech technology. With this broad goal in mind we have focussed on tone realisation in Yorùbá and considered the development of an efficient pitch model applicable in TTS systems in under-resourced environments. We believe that this and other attempts to build natural sounding TTS systems using compact data-driven models[1] are sensible approaches to contribute both towards fundamental knowledge in the field of speech processing and the development of speech technologies in under-resourced languages. In the latter case, the development of speech and language technologies seems to suffer from a "chicken-and-egg" problem: powerful and flexible corpus-based techniques rely on large amounts of digital resources and the large-scale generation of this content is reliant on technological support. The development of efficient (compact) models with incremental improvements may thus provide an opportunity to overcome this problem [4].

As stated in Chapter 1, we approached this problem by formulating the following questions:

1. What are the salient acoustic features (especially of pitch) attributable to the expression of tone in Yorùbá as manifested in general continuous utterances?

2. How can this be suitably modelled and applied in speech technologies (especially TTS systems) in typical under-resourced environments?

In the next section we discuss our approaches to answering these questions and highlight resulting contributions. This is followed in Section 6.2 by a discussion of identified shortcomings and suggested avenues for further work.

[1] e.g. data-driven implementations of traditionally *rule-based* approaches such as formant synthesis systems [98].

6.1 SUMMARY OF APPROACHES AND CONTRIBUTIONS

In Chapter 3 a multi-speaker speech corpus was analysed using automatic speech processing techniques to investigate and quantify aspects of tone realisation such as speaker-specific variation and co-articulation in continuous utterances. This included measurements of pitch, intensity and duration, but focussed on local pitch patterns with a temporal scope of up to four syllables. Pitch patterns were identified using a general statistical analysis of acoustic properties of pitch in and between syllables and contours in larger contexts guided by observations in the literature (Section 3.3.1). The association of these properties with tone labels derived from the text and variation of contours in different contexts and speakers was quantified by classification (Section 3.3.4) and clustering (Section 3.3.5) experiments respectively.

Our contributions in this chapter may be summarised as follows:

1. A key result is the finding that the intra-syllable pitch gradient is a consistent indicator of tone in HL and LH transitions (point (2) in Section 3.4), distinguishing these contexts from HM and LM transitions. We thus suggest that these falling and rising realisations are not merely a phonetic result of co-articulation but have a "phonemic" role in Yorùbá (using the terminology in [21, 88], this is a case of true *target alternation* and not just an *implementational variation* due to carryover assimilation). This and other results (see points (10) and (11) in Section 3.4), indicate that pitch level, change in pitch and gradient are important in tone realisation, which departs somewhat from traditional characterisation of both *register* and *contour* tone systems.

2. Furthermore, we have characterised typical pitch contours in general utterances in different tone contexts and investigated variation in these contours over speakers and extended tone contexts. The results have identified and quantified the most important contextual and speaker-specific variations, which directly informs the development of speech technologies in general.

In Chapter 4 the relationship between inter-syllable pitch target changes and downtrend was investigated for four speakers by examining the results of models based on local pitch changes (analysis by modelling [28]).

Our contributions in this chapter may be summarised as follows:

1. An important result is the relatively stable distributions of the change in pitch between syllables

in certain tone contexts and a strong correlation between previous and current pitch. This suggests a linear model for predicting the pitch in the following syllable based on the current pitch level. Such a model captures local dynamics and considers constraints in pitch range.

2. Neither models considering only local dynamics nor models based on utterance context determined by the number of downsteps adequately represent the observed downtrend in our corpus. We found models based on tone-independent linear declination to be the most appropriate. This leads us to conclude that *downstep* is a largely local phenomenon.

In Chapter 5 a tone-aware HMM-based TTS system was developed to test the applicability of the findings in chapters 3 and 4 directly using the quantitative target approximation approach to synthesise complete contours in a real-world case. By comparing the proposed pitch model analytically and perceptually against the standard HMM-based acoustic modelling approach implemented in HTS (similarly to what has been recently described for Ibibio [20], another African tone language), we aimed to determine:

1. The relative efficiency of the tested pitch models given the typical data quantity and quality constraints in under-resourced environments, and

2. The applicability of the proposed target approximation pitch model in an HMM-based synthesiser.

Our contributions in this chapter may be summarised as follows:

1. A Yorùbá TTS corpus was developed which should be able to support further research on the questions identified from chapters 4 and 5 (see the following section).

2. A pitch model based on quantitative target approximation [65] using tree regression models predicting pitch *height* and *gradient* independently, combined with a synthesis algorithm determining *strength* heuristically, was shown to be relatively efficient and applicable in an HMM-based TTS system. Analytical results suggest that such a model may be applicable even when data quantity is very limited, while perceptual results indicated preference for such a model over the standard HTS approach for short utterances.

We discuss further questions related to the implementation of tone-aware HMM-based pitch models as well as work needed on the target approximation model in the next section.

6.2 FURTHER APPLICATIONS AND FUTURE WORK

While the work completed during this study was focussed on the questions formulated above, the results obtained suggest a number of avenues for further work.

In Chapter 3, the work in Section 3.3.4 may be extended to general tone recognition experiments considering normalisation for speaker differences, extended tone context (Section 3.3.5), normalisation for downtrend (as in [99] for Cantonese) based on the results in Chapter 4, and recognition of tone sequences in utterance context (e.g. using a Viterbi search) for eventual integration into tone-aware ASR systems. We suggest that this may be a relatively fruitful endeavour, compared to building tone-aware acoustic models for say Mandarin, considering the lack of available text resources to build strong language models.

In Chapter 4, Section 4.3, we proposed a model based on the local pitch dynamics. While further work suggests at least an additional variable considering downtrend, we argue that a model based on local dynamics is more likely to ensure that aspects important to tone realisation are preserved during model estimation (Section 3.3.4). We thus suggest further research on modelling Yorùbá intonation starting with Eq. 4.4. Towards this goal, immediate questions that remain unanswered involve pitch resetting (i.e. the management of pitch range on the utterance level) and the determination of initial conditions for utterances and breath-groups. These questions should ideally be investigated using a more suitable corpus containing longer utterances representing full sentences and with diversity in terms of length and number of breath-groups (such as the corpus subsequently developed in Chapter 5).

Following on from Chapter 5, additional work should include more detailed perceptual tests in the form of "Blizzard-style" naturalness and intelligibility tests [32, 33]. This may result in more insight into the appropriateness of the current modelling of M tones, and also allow for a comparison with other efforts to develop tone-aware synthesisers for African tone languages such as Ibibio [20]. The TTS corpus developed here may also be applied towards further research on the local dynamics model as proposed above. Regarding the HMM-based pitch model, inspection of some of the samples where users overwhelmingly preferred the target approximation model (see the last two syllables of Figure 5.3b) suggest a potential improvement: HMMs for pitch models should perhaps be defined for syllable units rather than phone units. Lastly, while the preference test results were not statistically significant in the case of longer sentences, the slight but seemingly consistent shift in preference nevertheless

warrants further investigation in future work. Closer inspection of individual preferences suggest that samples generated using HMMs contain more variation, thus, certain words may be perceived as more prominent, while the target approximation model samples may be perceived as more neutral. While word prominence is not explicitly modelled in the HMM-based samples, the strong reliance of the system on the underlying data may reproduce such patterns appropriately in some cases leading to perceptually favourable samples. We thus suggest that this aspect should be considered for explicit modelling in the target approximation model, which we suspect will improve the perceived quality of longer utterances. This work, however, should be undertaken in combination with the analysis of other acoustic features such as intensity and duration.

6.2.1 Application to other African languages

While the focus of our empirical work in this study has been on Yorùbá, the approach followed should be applicable to other languages with similar tone systems. Specifically the methodology used in Chapter 3 and specifically sections 3.3.4 and 3.3.5 may be used to rapidly determine considerations for pitch target extraction. Following this, we suggest that the process followed in Chapter 5 may be reasonable for similar languages, with the exception of the constraints applied during the analysis-by-synthesis process for target extraction.

Two other African languages which may be immediately explored for the development of tone-aware TTS systems based on this study are Sotho and Ibibio, where significant work has already been completed on deriving surface tone from underlying tone [16, 17].

REFERENCES

[1] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, "Your Word is my Command: Google Search by Voice: A Case Study," *Advances in Speech Recognition*, pp. 61–90, 2010.

[2] T. Mozer, "Speech's Evolving Role in Consumer Electronics... From Toys to Mobile," *Mobile Speech and Advanced Natural Language Solutions*, pp. 23–34, Jan. 2013.

[3] E. Barnard, M. H. Davel, and G. B. van Huyssteen, "Speech technology for information access: a South African case study," in *AAAI Symposium on Artificial Intelligence*, 2010, pp. 22–24.

[4] E. Barnard, J. Schalkwyk, C. van Heerden, and P. J. Moreno, "Voice search for development," in *Proceedings of INTERSPEECH*, Makuhari, Japan, September 2010, pp. 282–285.

[5] V. de Boer, P. de Leenheer, A. Bon, N. B. Gyan, C. van Aart, C. Guéret, W. Tuyp, S. Boyera, M. Allen, and H. Akkermans, "RadioMarché: distributed voice-and web-interfaced market information systems under rural conditions," *Advanced Information Systems Engineering*, pp. 518–532, 2012.

[6] R. Tucker and K. Shalonova, "The Local Language Speech Technology Initiative – localisation of TTS for voice access to information," in *SCALLA Conference*, Nepal, 2004.

[7] J. A. Louw, M. Davel, and E. Barnard, "A general-purpose IsiZulu speech synthesizer," *South African journal of African languages*, vol. 2, pp. 1–9, 2006.

[8] M. Ekpenyong, E.-A. Urua, and D. Gibbon, "Towards an unrestricted domain TTS system for African tone languages," *International Journal of Speech Technology*, vol. 11, no. 2, pp. 87–96, 2008.

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

[10] A. Akinlabi and M. Liberman, "The tonal phonology of Yoruba clitics," *Clitics in Phonology, Morphology and Syntax*, vol. 36, pp. 31–62, 2000.

[11] A. Sharma Grover and E. Barnard, "The Lwazi community communication service: design and piloting of a voice-based information service," in *Proceedings of the 20th international conference companion on World Wide Web*. New York, NY, USA: ACM, 2011, pp. 433–442.

[12] A. S. Grover, O. Stewart, and D. Lubensky, "Designing interactive voice response (IVR) interfaces: localisation for low literacy users," in *Proceedings of the Computers and Advanced Technology in Education*, vol. 673, no. 29, 2009, pp. 328–335.

[13] T. Adegbola, "Building Capacities in Human Language Technology for African Languages," in *Proceedings of the First Workshop on Language Technologies for African Languages*, 2009, pp. 53–58.

[14] O. A. Ọdẹ́jọbí, 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.

[15] R. Tucker and K. Shalonova, "Supporting the Creation of TTS for Local Language Voice Information Systems," in *Proceedings of INTERSPEECH*, 2005, pp. 453–456.

[16] D. Gibbon, E.-A. Urua, and M. Ekpenyong, "Problems and solutions in African tone language text-to-speech," in *ISCA Workshop on Multilingual Speech and Language Processing*, 2006, pp. 1–14.

[17] M. Raborife, S. Ewert, and S. Zerbian, "An African Solution for an African Problem: A step towards perfection," in *Proceedings of the Twenty-First Annual Symposium of the Pattern Recognition Association of South Africa*, Stellenbosch, South Africa, Nov. 2010, pp. 225–230.

[18] M. Davel and E. Barnard, "Pronunciation prediction with Default&Refine," *Computer Speech and Language*, vol. 22, pp. 374–393, 2008.

[19] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[20] M. Ekpenyong, E.-A. Urua, O. Watts, S. King, and J. Yamagishi, "Statistical parametric speech synthesis for Ibibio," *Speech Communication*, vol. 56, pp. 243–251, 2014.

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

[22] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *Proceedings of ICASSP*, vol. 1, 1999, pp. 229–232.

[23] T. Toda and K. Tokuda, "A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.

[24] J. van Santen, A. Kain, E. Klabbers, and T. Mishra, "Synthesis of prosody using multi-level unit sequences," *Speech Communication*, vol. 46, no. 3, pp. 365–375, 2005.

[25] G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black, "Accent Group modeling for improved prosody in statistical parametric speech synthesis," in *Proceedings of ICASSP*, 2013, pp. 6890–6894.

[26] K. Courtenay, "Yoruba: a terraced-level language with three tonemes," *Studies in African Linguistics*, vol. 2, no. 3, pp. 239–255, 1971.

[27] S. Bird, "Strategies for Representing Tone in African Writing Systems," *Written Language & Literacy*, vol. 2, no. 1, pp. 1–44, 1999.

[28] Y. Xu, "Speech prosody: a methodological review," *Journal of Speech Sciences*, vol. 1, no. 1, pp. 85–115, 2011.

[29] D. Hirst, "The analysis by synthesis of speech melody: from data to models," *Journal of Speech Sciences*, vol. 1, no. 1, pp. 55–83, 2011.

[30] P. Taylor, *Text-to-Speech Synthesis*, 1st ed. Cambridge University Press, Mar. 2009.

[31] T. Dutoit, "High-Quality Text-to-speech Synthesis: An Overview," *Journal of Electrical and Electronics Engineering Australia*, vol. 17, pp. 25–36, 1997.

[32] S. King and V. Karaiskos, "The Blizzard Challenge 2011," in *Blizzard Challenge 2011*, Turin, Italy, 2011. [Online]. Available: http://www.festvox.org/blizzard/blizzard2011.html

[33] S. King, "The Blizzard Challenge 2012," in *Blizzard Challenge 2012*, Portland, Oregon, USA, 2012. [Online]. Available: http://www.festvox.org/blizzard/blizzard2012.html

[34] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings of ICASSP*, vol. 1, Atlanta, Georgia, USA, 1996, pp. 373–376.

[35] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *Information Theory, IEEE Transactions on*, vol. 13, no. 2, pp. 260–269, 1967.

[36] A. W. Black, "Perfect synthesis for all of the people all of the time," in *Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on*, 2002, pp. 167–170.

[37] B. Möbius, "Rare events and closed domains: Two delicate concepts in speech synthesis," *International Journal of Speech Technology*, vol. 6, no. 1, pp. 57–71, 2003.

[38] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis," in *Proceedings of EUROSPEECH*, Budapest, Hungary, 1999, pp. 2347–2350.

[39] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proceedings of ICASSP*, vol. 3, 2000, pp. 1315–1318.

[40] A. W. Black, "CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling," in *Proceedings of INTERSPEECH*, Pittsburgh, Pennsylvania, USA, 2006, pp. 1762–1765.

[41] Y. J. Wu and R. H. Wang, "Minimum generation error training for HMM-based speech synthesis," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, Toulouse, France, 2006, pp. 89–92.

[42] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modelling for speech recognition," *Journal of the Acoustic Society of Japan*, vol. 21, no. 2, pp. 79–86, 2000.

[43] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Veltchev, and P. Woodland, *The HTK Book (for HTK Version 3.3)*. Cambridge University Engineering Department, 2005. [Online]. Available: http://htk.eng.cam.ac.uk/

[44] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the workshop on Human Language Technology*, 1994, pp. 307–312.

[45] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *Proceedings of the IEEE Workshop on Speech Synthesis*, 2002, pp. 227–230.

[46] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proceedings of ICASSP*, vol. 1, 1995, pp. 660–663.

[47] J. Yamagishi, Z. Ling, and S. King, "Robustness of HMM-based speech synthesis," in *Proceedings of INTERSPEECH*, Brisbane, Australia, 2008, pp. 581–584.

[48] D. R. van Niekerk, E. Barnard, and G. Schlünz, "Perceptual evaluation of corpus-based speech synthesis techniques in under-resourced environments," in *Proceedings of PRASA*, Stellenbosch, South Africa, 2009, pp. 71–75.

[49] G. K. Anumanchipalli and A. W. Black, "Adaptation techniques for speech synthesis in under-resoured languages," in *The Second International Workshop on Spoken Language Technologies for Under-resoured languages*, Penang, Malaysia, 2010, pp. 51–55.

[50] J. Yamagishi, T. Nose, H. Zen, Z. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive HMM-based text-to-speech synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 6, pp. 1208–1230, 2009.

[51] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Proceedings of EUROSPEECH*, Aalborg, Denmark, 2001, pp. 2263–2266.

[52] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," in *Proceedings of INTERSPEECH*, Jeju Island, Korea, 2004, pp. 1393–1396.

[53] A. Botinis, B. Granström, and B. Möbius, "Developments and paradigms in intonation research," *Speech Communication*, vol. 33, no. 4, pp. 263–296, 2001.

[54] Y. Xu and X. Sun, "How fast can we really change pitch? Maximum speed of pitch change revisited," in *The Sixth International Conference on Spoken Language Processing*, Beijing, China, 2000, pp. 666–669.

[55] A. Cohen, R. Collier, and J. t'Hart, "Declination: Construct or Intrinsic Feature of Speech Pitch?" *Phonetica*, vol. 39, no. 4-5, pp. 254–273, 1982.

[56] D. Whalen and A. G. Levitt, "The universality of intrinsic F0 of vowels," *Journal of Phonetics*, vol. 23, no. 3, pp. 349–366, 1995.

[57] B. Connell, "Tone languages and the universality of intrinsic F0: evidence from Africa," *Journal of Phonetics*, vol. 30, pp. 101–129, 2002.

[58] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," in *Proceedings of the International Conference on Spoken Language Processing*, vol. 2, 1992, pp. 867–870.

[59] P. Taylor, "Analysis and synthesis of intonation using the tilt model," *The Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1697–1714, 2000.

[60] H. Fujisaki and S. Ohno, "Comparison and assessment of models in the study of fundamental frequency contours of speech," in *Proceedings of the ESCA workshop on Intonation: Theory, Models and Applications*, 1997, pp. 131–134.

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

[62] D. J. Hirst, "Form and function in the representation of speech prosody," *Speech Communication*, vol. 46, no. 3, pp. 334–347, 2005.

[63] H. Fujisaki, S. Ohno, and C. Wang, "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, November 1998, pp. 26–29.

[64] H. Mixdorff, "A novel approach to the fully automatic extraction of Fujisaki model parameters," in *Proceedings of ICASSP*, vol. 3, 2000, pp. 1281–1284.

[65] S. Prom-on, Y. Xu, and B. Thipakorn, "Modeling tone and intonation in Mandarin and English as a process of target approximation," *The Journal of the Acoustical Society of America*, vol. 125, pp. 405–424, 2009.

[66] 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.

[67] A. Akinlabi and M. Liberman, "Tonal complexes and tonal alignment," in *Proceedings of the North East Linguistic Society*, vol. 31, Georgetown University, 2001, pp. 1–20.

[68] O. A. Ọdẹ́jọbí, 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.

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

[70] J. M. Hombert, "Perception of tones of bisyllabic nouns in Yoruba," *Studies in African Linguistics Los Angeles, Cal.*, vol. 7, pp. 109–121, 1976.

[71] K. Calteaux, F. de Wet, C. Moors, D. R. van Niekerk, B. McAlister, A. Sharma Grover, T. Reid, M. Davel, E. Barnard, and C. van Heerden, "Lwazi II Final Report: Increasing the impact of speech technologies in South Africa," Council for Scientific and Industrial Research, Pretoria, South Africa, Tech. Rep. 12045, February 2013.

[72] K. P. Scannell, "Statistical unicodification of African languages," *Language Resources and Evaluation*, vol. 45, no. 3, pp. 375–386, 2011.

[73] L. Mohasi, H. Mixdorff, T. Niesler, and S. Zerbian, "Analysis of Sesotho Tone using the Fu-

jisaki Model," in *Tonal Aspects of Languages-Third International Symposium*, 2012, pp. 1–6.

[74] N. Govender, E. Barnard, and M. Davel, "Pitch modelling for the Nguni languages," *South African Computer Journal*, vol. 38, pp. 28–39, 2007.

[75] J. A. Louw and E. Barnard, "Automatic intonation modeling with INTSINT," in *Proceedings of the Pattern Recognition Association of South Africa*, 2004, pp. 107–111.

[76] E. Barnard and S. Zerbian, "From tone to pitch in Sepedi," in *The Second International Workshop on Spoken Language Technologies for Under-resourced Languages*, 2010, pp. 29–34.

[77] D. R. van Niekerk and E. Barnard, "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, May 2012, pp. 54–59.

[78] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "The DARPA TIMIT acoustic-phonetic continuous speech corpus, NIST order number PB91-100354," February 1993.

[79] 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.

[80] D. R. van Niekerk, "Automatic speech segmentation with limited data," Master's thesis, North-West University, Potchefstroom Campus, South Africa, 2009.

[81] Meraka Institute, "Lwazi Project Final Report," CSIR, Pretoria, South Africa, http://www.meraka.org.za/lwazi, Tech. Rep., November 2009.

[82] A. Akinlabi, "Yorùbá Sound System," in *Understanding Yoruba Life and Culture*, Lawal N. S. and Sadiku M., Ed. Trenton, NJ: Africa World Press Inc., 2004, pp. 453–468.

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

[84] Y. Xu, "Effects of tone and focus on the formation and alignment of f_0 contours," *Journal of Phonetics*, vol. 27, no. 1, pp. 55–105, 1999.

[85] S. Ohno, M. Fukumiya, and H. Fujisaki, "Quantitative analysis of the local speech rate and its application to speech synthesis," in *Proceedings of the Fourth International Conference on Spoken Language, ICSLP '96*, vol. 4, 1996, pp. 2254–2257.

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

[87] H. Zen, K. Oura, T. Nose, J. Yamagishi, S. Sako, T. Toda, T. Masuko, A. W. Black, and K. Tokuda, "Recent development of the HMM-based speech synthesis system (HTS)," in *Proceedings of th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Sapporo, Japan, 2009, pp. 121–130.

[88] Y. Xu, "Tone in Connected Discourse," in *Encyclopedia of Language & Linguistics (Second Edition)*, K. Brown, Ed. Oxford: Elsevier, 2006, pp. 742–751.

[89] Y. Xu, "Articulatory constraints and tonal alignment," in *Proceedings of the 1st International Conference on Speech Prosody*, Aix-en-Provence, France, 2002, pp. 91–100.

[90] D. R. van Niekerk and E. Barnard, "Predicting utterance pitch targets in Yorùbá for tone realisation in speech synthesis," *Speech Communication*, vol. 56, pp. 229–242, 2014.

[91] E. Fajobi, "The Nature of Yoruba Intonation: A New Experimental Study," in *Yoruba Creativity: Fiction, Language, Life and Songs*, T. Falola and A. Genova, Eds. New Jersey: Africa World Press Inc., 2005, pp. 183–221.

[92] J. P. van Santen and A. L. Buchsbaum, "Methods for optimal text selection," in *Proceedings of EUROSPEECH*, Rhodes, Greece, September 1997, pp. 553–556.

[93] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer Voice Quality of New Languages Calibrated with Mean Mel Cepstral Distortion," in *The First International Workshop on Spoken Language Technologies for Under-resoured languages*, Hanoi, Vietnam, 2008, pp. 63–68.

[94] D. R. van Niekerk, "Experiments in rapid development of accurate phonetic alignments for TTS in Afrikaans," in *Proceedings of PRASA*, Vanderbijlpark, South Africa, 2011, pp. 144–149.

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

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

[97] M. M. Bradshaw, "A crosslinguistic study of consonant-tone interaction," Ph.D. dissertation, Ohio State University, 1999.

[98] G. K. Anumanchipalli, Y.-C. Cheng, J. Fernandez, X. Huang, Q. Mao, and A. W. Black, "KLATTSTAT: Knowledge-based parametric speech synthesis," *Proceedings of the Seventh ISCA Workshop on Speech Synthesis*, pp. 63–68, 2010.

[99] G. Peng and W. S.-Y. Wang, "Tone recognition of continuous Cantonese speech based on support vector machines," *Speech Communication*, vol. 45, no. 1, pp. 49–62, Jan. 2005.

[100] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.

[101] D. R. van Niekerk and E. Barnard, "Generating fundamental frequency contours for speech synthesis in Yorùbá," in *Proceedings of INTERSPEECH*, Lyon, France, September 2013, pp. 1027–1031.

APPENDIX A

HMM-BASED PHONE ALIGNMENT

```
TARGETKIND: MFCC_0_D_A_Z
TARGETRATE: 50000.0
WINDOWSIZE: 100000.0
USEHAMMING: T
PREEMCOEF: 0.97
NUMCHANS: 26
CEPLIFTER: 22
NUMCEPS: 12
ENORMALISE: F
```

Table A.1: `HCopy` *configuration details including the window type, filterbank and pre-emphasis settings.*

TIMIT phones	**broad class**	**Yorùbá phones**
pau	long pause	pau
bcl, dcl, gcl, kcl, pcl, tcl	short pause	pau_cl
b, d, g, dx, q	voiced plosives	b, d, g, $\widehat{gb}$
p, t, k	unvoiced plosives	k, $\widehat{kp}$, t
jh	voiced affricates	dʒ
s, sh, f, th, hh, hv	unvoiced fricatives	f, h, s, ʃ
m, n, ng, nx	nasals	m, n
l, r, w, y	approximants	j, l, r, w
ih, eh, ae, aa, ah, uh, ix	short vowels	a, ã, e, ɛ, ɛ̃, i, ĩ, o, ɔ, ɔ̃, u, ũ

Table A.2: *Broad phone class mappings for the TIMIT and Yorùbá phonesets. During training for alignment, an HMM for each broad class is initialised with the corresponding TIMIT speech data using Viterbi re-estimation (HTK's* `HInit` *and* `HRest`*). Initial broad phone models are then copied for the corresponding Yorùbá phones and trained on Yorùbá speech data using embedded (Baum-Welch) re-estimation (*`HERest`*).*

APPENDIX B

DETAILED RESULTS FOR CHAPTER 3

Speaker	Gender	Pitch range (Hz)	Pitch range (st)	Mean pitch (st)		
				H	M	L
01	female	70 - 180	73.55 - 89.90	79.1	79.3	78.7
02	female	70 - 300	73.55 - 98.75	89.0	88.4	87.6
03	female	120 - 350	82.88 - 101.41	91.3	89.8	87.2
04	female	120 - 350	82.88 - 101.41	92.3	91.6	91.1
05	female	120 - 350	82.88 - 101.41	93.2	92.7	92.3
06	female	120 - 350	82.88 - 101.41	91.9	91.1	90.8
08	female	120 - 350	82.88 - 101.41	90.9	90.1	89.3
09	female	70 - 180	73.55 - 89.90	79.6	79.8	79.3
10	female	100 - 300	79.73 - 98.75	90.5	89.7	89.3
11	female	120 - 350	82.88 - 101.41	93.5	92.4	92.2
12	female	120 - 350	82.88 - 101.41	93.3	92.2	92.4
13	female	100 - 350	79.73 - 101.41	92.5	91.8	90.8
14	female	120 - 300	82.88 - 98.75	90.8	90.3	90.9
15	female	70 - 180	73.55 - 89.90	79.1	79.3	78.8
16	female	120 - 350	82.88 - 101.41	94.5	93.9	93.8
17	female	120 - 300	82.88 - 98.75	92.1	91.2	90.6
19	male	80 - 220	75.86 - 93.38	84.4	83.7	83.7
20	male	60 - 220	70.88 - 93.38	81.4	80.4	79.8
21	male	70 - 220	73.55 - 93.38	85.6	84.7	83.5
22	male	90 - 270	77.90 - 96.92	87.5	86.7	86.4
23	male	70 - 200	73.55 - 91.73	82.4	81.3	80.0
24	male	100 - 220	79.73 - 93.38	87.3	86.6	85.5
26	male	70 - 280	73.55 - 97.55	87.4	86.4	85.7
27	male	70 - 270	73.55 - 96.92	86.4	85.4	84.8
28	male	70 - 250	73.55 - 95.59	84.9	84.3	83.7
29	male	70 - 220	73.55 - 93.38	83.3	82.4	81.2
30	male	100 - 240	79.73 - 94.88	85.9	84.5	84.6
31	male	80 - 240	75.86 - 94.88	86.1	85.0	84.8
32	male	60 - 210	70.88 - 92.57	83.7	82.4	81.7
33	male	70 - 210	73.55 - 92.57	83.3	82.3	81.5
34	male	80 - 250	75.86 - 95.59	85.0	84.6	83.8
35	male	60 - 250	70.88 - 95.59	84.3	83.4	82.7
36	male	100 - 300	79.73 - 98.75	89.0	87.8	86.4

Table B.1: *Speaker properties summary.*

Tri-tone contexts	Number of instances	Number of source utterances
HHH	1193	788
HHL	1143	1020
HHM	949	895
HLH	1539	1208
HLL	1170	1046
HLM	852	802
HMH	1397	1122
HML	778	728
HMM	1010	930
LHH	1225	1094
LHL	1343	1049
LHM	1050	969
LLH	1169	1046
LLL	1357	828
LLM	732	676
LMH	721	668
LML	681	601
LMM	699	652
MHH	981	895
MHL	1068	966
MHM	1188	958
MLH	965	875
MLL	687	647
MLM	544	489
MMH	1032	945
MML	677	638
MMM	1480	712

Table B.2: *Three-syllable sequences extracted from the corpus.*

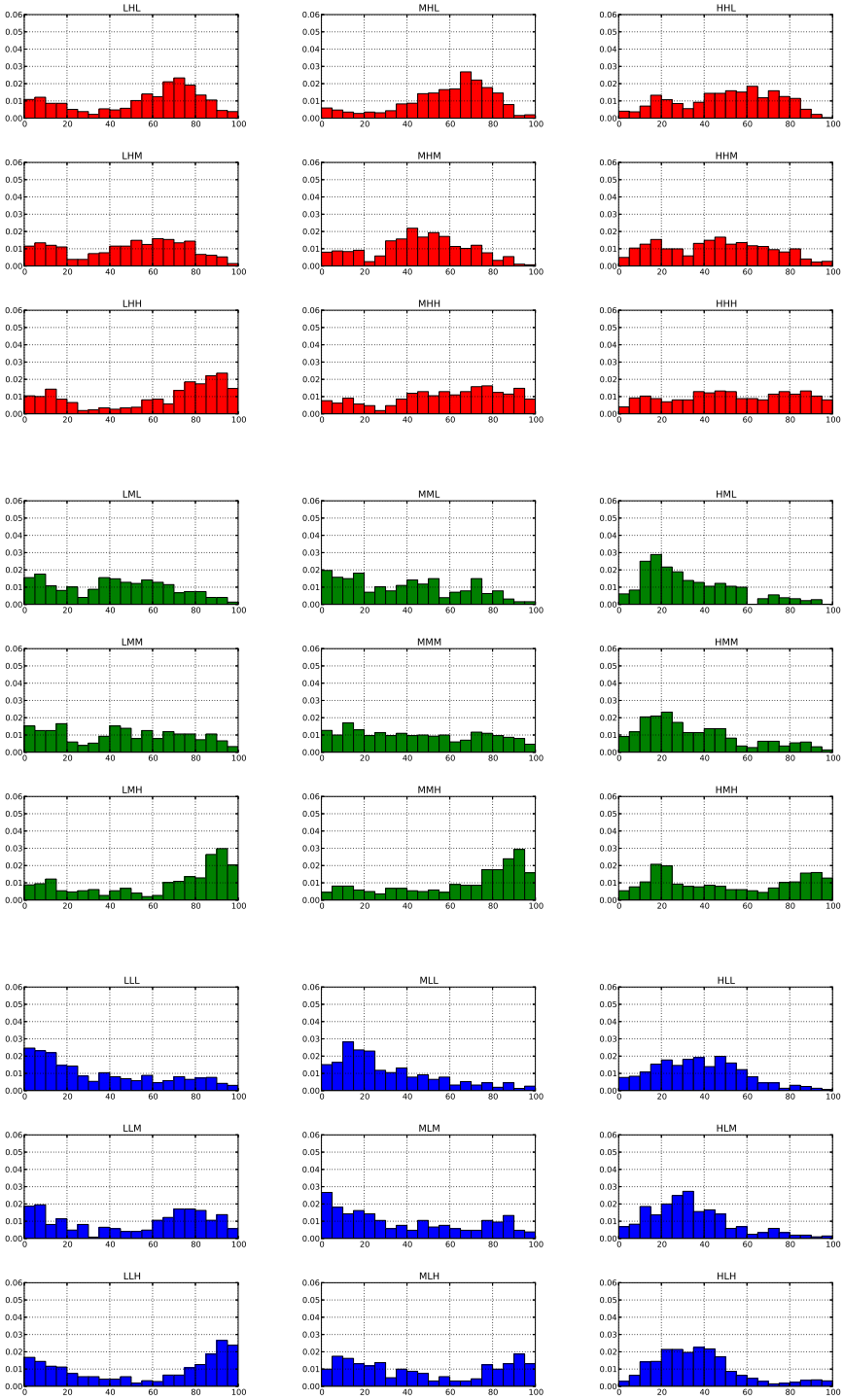

Figure B.1: *Distributions of peaks in syllables (i.e. turning points in the contour where the turning point is at a maximum value for the contour) for H (red), M (green) and L (blue) tones in context. The x-axis represents the normalised time and the y-axis the proportion of all samples.*

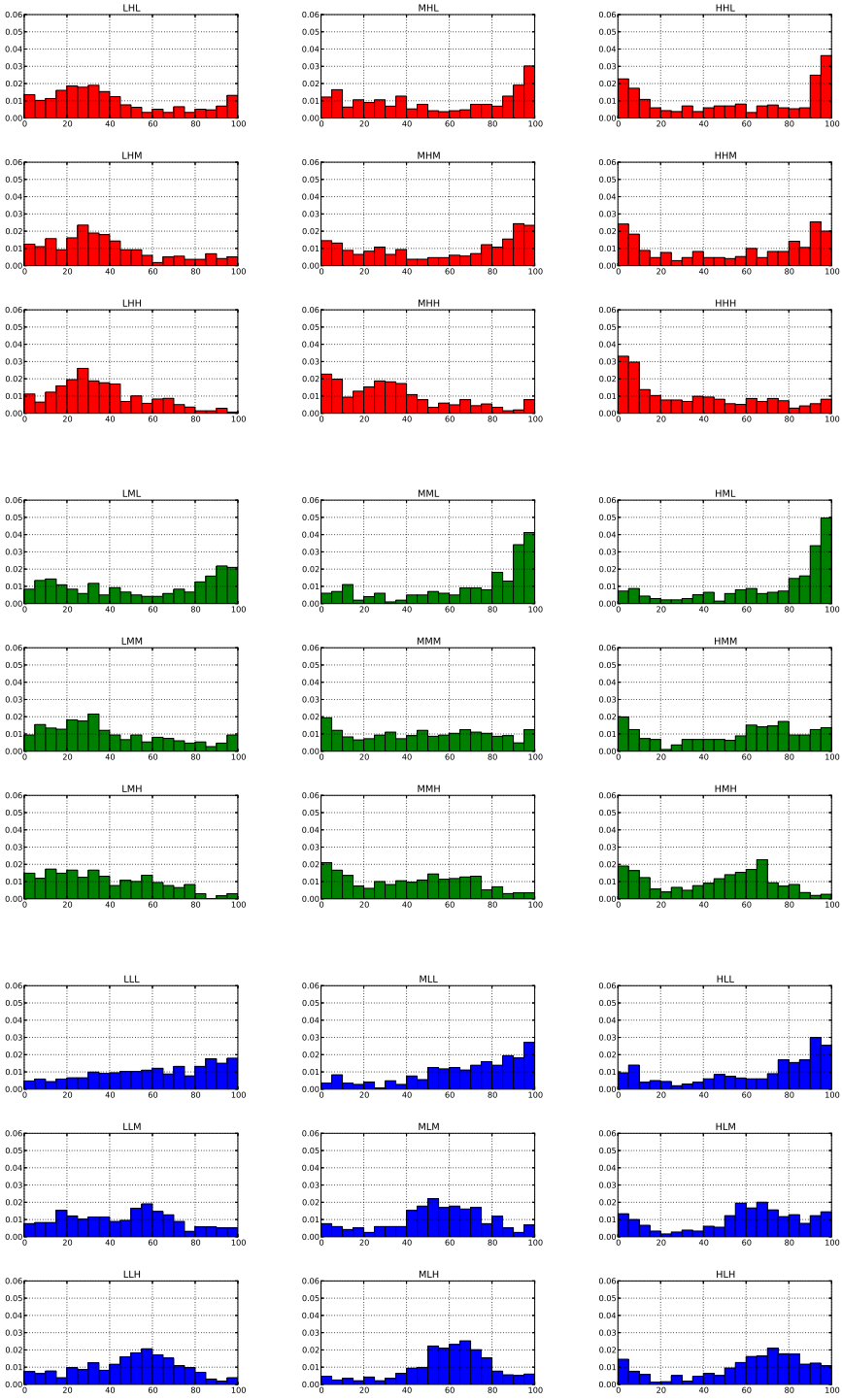

Figure B.2: *Distributions of valleys in syllables (i.e. turning points in the contour where the turning point is at a minimum value for the contour) for H (red), M (green) and L (blue) tones in context. The x-axis represents the normalised time and the y-axis the proportion of all samples.*

CONTEXT	01	02	03	04	05	06	08	09	10	11	12	13	14	15	16	17
HHH	1.55	2.14	1.69	1.03	1.12	1.20	1.35	1.66	1.42	1.62	1.67	2.47	1.11	1.84	1.14	1.18
HHL	1.66	2.87	1.70	1.23	1.45	1.45	1.66	1.57	1.69	1.19	1.76	1.82	1.11	1.86	1.14	1.49
HHM	1.50	2.09	1.59	1.08	1.30	1.11	1.57	1.65	1.36	1.42	1.37	2.29	1.01	1.63	1.22	1.45
HLH	1.51	2.56	2.37	1.61	1.56	1.56	2.07	1.63	2.00	1.94	2.03	2.35	1.25	1.71	1.70	1.84
HLL	1.47	3.32	2.17	1.49	1.61	1.67	2.11	1.57	1.89	1.69	2.01	1.99	1.28	1.57	1.35	1.62
HLM	1.47	2.40	1.96	1.40	1.39	1.41	1.64	1.41	1.57	1.53	1.85	2.27	1.25	1.71	1.46	1.43
HMH	1.47	2.09	1.53	1.20	0.99	0.98	1.48	1.63	1.46	1.58	1.72	2.04	1.06	1.72	1.29	1.40
HML	1.25	1.83	1.40	1.27	1.38	1.31	1.42	1.65	1.35	1.54	1.87	2.08	1.02	1.48	1.34	1.35
HMM	1.54	1.74	1.63	1.08	1.37	1.24	1.49	1.55	0.97	1.49	1.82	1.93	1.34	1.42	1.41	1.38
LHH	1.42	1.77	2.11	1.38	1.82	1.93	1.85	1.65	1.77	1.48	2.08	2.35	1.18	1.60	1.32	1.94
LHL	1.43	2.58	2.26	1.50	1.80	1.73	1.87	1.53	1.85	1.90	2.24	2.38	1.39	1.55	1.82	1.85
LHM	1.29	2.44	1.86	1.60	1.72	1.50	1.59	1.31	1.73	1.76	1.83	2.20	1.34	1.31	1.57	1.62
LLH	1.23	2.68	1.84	1.75	1.75	1.67	1.65	1.05	1.82	1.76	2.29	2.23	1.27	1.36	1.76	1.91
LLL	1.40	2.31	1.71	1.49	1.62	1.29	1.65	0.92	1.64	1.26	1.84	1.79	0.83	1.18	1.51	1.65
LLM	0.85	1.51	1.69	1.43	1.65	1.39	1.37	1.14	1.69	1.45	2.05	1.88	1.34	1.14	1.34	1.69
LMH	1.28	1.84	1.59	1.32	1.73	1.63	1.63	1.67	1.44	1.49	1.92	1.93	1.15	1.57	1.33	1.39
LML	1.05	1.62	2.22	1.08	1.14	1.65	1.32	1.10	1.83	1.39	1.79	2.04	1.45	1.42	1.33	1.45
LMM	1.08	1.94	1.65	1.31	1.43	1.63	1.24	1.06	1.20	1.37	1.61	2.16	1.22	1.17	1.11	1.40
MHH	1.52	1.95	1.55	1.10	1.27	0.75	1.34	1.47	1.18	1.16	1.54	1.94	1.05	1.65	0.98	1.13
MHL	1.73	2.49	1.60	1.28	1.23	1.35	1.50	1.42	1.41	1.79	1.48	1.81	0.98	1.55	1.15	1.39
MHM	1.41	1.93	1.37	1.14	1.08	1.10	1.35	1.68	1.35	1.49	1.75	1.85	0.99	1.64	1.09	1.44
MLH	1.35	1.60	1.73	1.48	1.45	1.42	1.39	1.32	1.85	1.46	1.96	2.18	1.24	1.47	1.46	1.61
MLL	0.96	1.71	1.68	1.58	1.14	1.16	1.47	1.21	1.17	1.32	1.95	1.85	1.04	1.19	1.40	1.19
MLM	0.88	1.96	1.48	0.88	1.29	1.45	1.20	1.07	1.56	1.32	1.59	1.83	1.20	1.41	1.17	1.39
MMH	1.47	1.73	1.44	1.01	1.17	1.27	1.21	1.43	1.41	1.57	1.75	1.93	1.03	1.52	1.20	1.24
MML	1.18	2.15	1.59	1.59	1.12	0.97	1.32	1.13	1.17	1.15	1.71	1.87	1.33	1.14	1.24	1.27
MMM	1.29	2.16	1.68	0.97	1.32	1.13	1.48	1.41	1.59	1.48	1.74	2.29	0.89	1.38	1.34	1.39

Figure B.3: *Summary of standard deviations (Eq. 3.3) in different contexts for different* **female** *speakers.*

CONTEXT	19	20	21	22	23	24	26	27	28	29	30	31	32	33	34	35	36
HHH	1.59	1.69	1.48	2.00	2.41	1.88	1.60	1.91	1.60	2.43	1.14	1.34	2.25	1.92	1.78	2.04	1.45
HHL	1.63	1.94	1.63	1.99	2.21	2.30	2.10	1.74	1.98	2.76	1.58	1.40	2.29	2.21	1.65	2.24	1.96
HHM	1.35	2.23	1.43	2.16	2.25	1.63	1.86	1.70	1.64	2.39	0.94	1.18	2.28	1.92	1.44	1.70	1.49
HLH	1.57	2.51	2.02	2.29	2.65	2.35	2.20	2.11	2.00	2.62	1.56	1.66	2.48	2.46	2.06	2.53	2.19
HLL	1.62	2.36	1.92	2.71	2.40	2.17	2.11	2.03	1.97	2.61	1.57	1.61	2.46	2.27	1.74	2.41	2.41
HLM	2.00	2.18	2.04	2.45	2.31	2.03	2.31	2.03	2.00	2.44	1.33	1.35	2.06	1.94	1.74	2.60	1.65
HMH	1.30	2.19	1.25	1.76	2.01	1.63	1.79	1.95	1.85	2.11	1.10	1.33	2.17	2.10	1.40	1.72	1.79
HML	1.58	2.27	1.85	1.79	1.82	1.68	1.82	2.00	1.67	1.73	1.14	1.11	1.53	1.83	1.43	2.10	1.62
HMM	1.37	1.67	1.69	1.92	1.75	1.60	1.63	1.77	1.97	1.93	1.04	1.15	1.89	2.08	1.86	1.80	1.54
LHH	1.89	2.61	1.80	2.07	2.19	2.16	2.34	2.05	1.95	2.63	1.42	1.42	2.17	2.16	1.67	2.47	2.13
LHL	1.77	2.69	2.19	2.45	2.82	2.16	2.55	2.31	2.03	2.44	1.71	1.75	2.52	2.32	2.02	2.65	2.37
LHM	1.52	2.44	1.78	2.18	2.17	2.03	1.77	2.01	1.95	2.55	1.41	1.65	1.87	2.16	1.75	2.57	2.02
LLH	1.87	2.40	1.94	2.12	1.91	2.03	2.22	1.94	2.03	2.41	1.65	1.65	1.94	1.96	1.73	2.69	2.12
LLL	1.59	2.47	1.49	2.19	1.82	1.90	1.62	1.60	1.75	1.87	1.22	1.31	1.62	2.10	1.79	2.51	2.06
LLM	1.64	2.19	1.48	2.03	2.13	1.49	1.42	1.51	1.87	2.17	1.55	1.34	1.26	1.96	1.42	2.39	1.90
LMH	1.51	1.86	1.73	1.94	2.41	1.77	1.65	1.64	1.61	2.29	1.21	1.31	1.98	2.02	1.81	2.21	1.63
LML	1.72	2.40	1.30	1.76	1.62	1.54	2.26	1.53	1.63	2.35	1.28	1.42	1.36	1.76	1.70	2.39	1.85
LMM	1.64	2.73	1.70	1.91	1.93	1.52	2.24	1.56	1.92	1.82	1.09	1.42	1.67	1.76	1.54	2.29	1.50
MHH	1.38	1.70	1.20	1.49	1.89	1.44	1.62	1.75	1.58	1.91	1.48	1.09	2.42	2.00	1.45	1.67	1.22
MHL	1.57	2.21	1.41	2.30	2.41	2.03	2.01	1.65	1.81	2.33	1.32	1.16	2.29	2.06	1.77	2.27	1.76
MHM	1.24	2.07	1.26	1.46	2.30	1.53	1.56	1.72	1.22	1.42	0.99	1.03	1.95	1.81	1.42	1.70	1.46
MLH	1.66	2.03	1.61	2.12	1.85	1.48	1.84	1.79	1.52	2.41	1.30	1.12	1.46	1.55	1.94	2.19	1.97
MLL	1.74	2.23	1.40	1.48	1.64	1.33	1.76	1.63	1.94	1.93	1.37	1.40	1.09	1.69	1.66	2.09	1.70
MLM	1.51	2.43	1.24	1.54	1.77	1.22	1.87	1.73	1.64	1.80	1.01	1.56	1.32	1.51	1.36	2.49	1.57
MMH	1.55	1.54	1.41	1.68	2.23	1.42	1.58	1.79	1.60	1.70	1.09	1.03	1.97	1.97	1.48	1.44	1.12
MML	1.21	1.55	1.30	1.28	1.88	1.42	2.14	1.36	1.86	2.11	1.00	0.93	1.55	1.39	1.73	1.98	1.96
MMM	1.48	2.24	1.57	1.99	2.19	1.62	1.26	1.77	1.69	2.18	0.82	1.08	2.54	1.77	1.66	1.66	1.49

Figure B.4: *Summary of standard deviations (Eq. 3.3) in different contexts for different* **male** *speakers.*

CONTEXT	01	02	03	04	05	06	08	09	10	11	12	13	14	15	16	17
HHH	0.119	0.047	0.014	0.039	0.014	0.017	0.022	0.161	0.020	0.041	0.022	0.061	0.017	0.096	0.020	0.029
HHL	0.130	0.035	0.017	0.028	0.018	0.014	0.011	0.095	0.010	0.025	0.014	0.019	0.020	0.069	0.042	0.009
HHM	0.099	0.027	0.011	0.013	0.019	0.033	0.023	0.216	0.013	0.038	0.035	0.020	0.035	0.086	0.018	0.036
HLH	0.127	0.026	0.033	0.020	0.013	0.028	0.018	0.034	0.008	0.030	0.008	0.018	0.026	0.076	0.018	0.018
HLL	0.011	0.047	0.007	0.011	0.011	0.017	0.018	0.049	0.010	0.018	0.008	0.021	0.018	0.061	0.020	0.014
HLM	0.043	0.038	0.009	0.015	0.032	0.012	0.018	0.130	0.033	0.042	0.015	0.020	0.027	0.056	0.021	0.017
HMH	0.132	0.077	0.065	0.045	0.027	0.041	0.052	0.050	0.047	0.031	0.056	0.053	0.039	0.092	0.024	0.045
HML	0.031	0.017	0.010	0.013	0.018	0.012	0.007	0.043	0.011	0.009	0.020	0.014	0.009	0.025	0.026	0.028
HMM	0.063	0.018	0.014	0.018	0.018	0.022	0.017	0.267	0.018	0.020	0.031	0.010	0.029	0.061	0.048	0.039
LHH	0.095	0.013	0.011	0.018	0.012	0.012	0.015	0.121	0.016	0.029	0.017	0.010	0.008	0.067	0.037	0.025
LHL	0.068	0.034	0.009	0.021	0.017	0.022	0.021	0.050	0.019	0.026	0.025	0.014	0.019	0.054	0.016	0.047
LHM	0.065	0.070	0.035	0.122	0.063	0.119	0.030	0.077	0.022	0.076	0.045	0.027	0.062	0.067	0.025	0.057
LLH	0.023	0.036	0.006	0.027	0.017	0.037	0.029	0.024	0.041	0.058	0.018	0.024	0.011	0.086	0.040	0.013
LLL	0.029	0.048	0.011	0.021	0.018	0.008	0.015	0.051	0.009	0.086	0.030	0.015	0.010	0.069	0.017	0.022
LLM	0.067	0.017	0.015	0.028	0.082	0.013	0.060	0.074	0.034	0.084	0.046	0.023	0.043	0.048	0.040	0.038
LMH	0.071	0.056	0.011	0.020	0.062	0.029	0.013	0.173	0.030	0.047	0.024	0.014	0.042	0.081	0.042	0.037
LML	0.037	0.056	0.020	0.079	0.053	0.028	0.023	0.055	0.038	0.016	0.027	0.233	0.022	0.045	0.033	0.031
LMM	0.126	0.125	0.022	0.064	0.030	0.036	0.122	0.041	0.089	0.047	0.065	0.058	0.089	0.060	0.057	0.069
MHH	0.338	0.049	0.057	0.077	0.072	0.028	0.095	0.205	0.038	0.035	0.056	0.043	0.078	0.202	0.012	0.029
MHL	0.043	0.031	0.015	0.011	0.009	0.059	0.015	0.045	0.025	0.020	0.037	0.020	0.047	0.009	0.026	0.013
MHM	0.039	0.026	0.019	0.011	0.021	0.015	0.014	0.205	0.014	0.049	0.037	0.020	0.012	0.058	0.026	0.028
MLH	0.062	0.070	0.058	0.040	0.029	0.022	0.012	0.046	0.020	0.046	0.067	0.043	0.088	0.074	0.046	0.033
MLL	0.022	0.021	0.010	0.029	0.017	0.013	0.033	0.046	0.016	0.043	0.068	0.015	0.020	0.036	0.023	0.018
MLM	0.017	0.069	0.042	0.021	0.072	0.022	0.057	0.035	0.048	0.067	0.089	0.062	0.079	0.056	0.055	0.038
MMH	0.212	0.027	0.091	0.092	0.042	0.017	0.049	0.178	0.023	0.060	0.017	0.040	0.108	0.190	0.041	0.043
MML	0.022	0.099	0.034	0.064	0.029	0.009	0.027	0.036	0.013	0.032	0.038	0.017	0.039	0.023	0.043	0.045
MMM	0.165	0.033	0.030	0.162	0.046	0.038	0.056	0.120	0.067	0.057	0.036	0.066	0.028	0.090	0.096	0.031

Figure B.5: *RMSEs between DTW-aligned speaker-specific and corpus-wide mean contours for* **female** *speakers.*

CONTEXT	19	20	21	22	23	24	26	27	28	29	30	31	32	33	34	35	36
HHH	0.111	0.048	0.010	0.036	0.056	0.020	0.033	0.078	0.038	0.041	0.016	0.020	0.035	0.038	0.083	0.019	0.026
HHL	0.017	0.011	0.018	0.063	0.012	0.016	0.014	0.018	0.018	0.036	0.020	0.024	0.019	0.020	0.010	0.010	0.020
HHM	0.024	0.064	0.020	0.052	0.031	0.030	0.042	0.013	0.023	0.057	0.019	0.031	0.063	0.050	0.077	0.039	0.017
HLH	0.035	0.027	0.008	0.016	0.022	0.019	0.011	0.037	0.039	0.029	0.014	0.019	0.021	0.041	0.017	0.015	0.028
HLL	0.028	0.023	0.021	0.019	0.016	0.023	0.016	0.010	0.013	0.022	0.023	0.011	0.031	0.013	0.016	0.026	0.022
HLM	0.036	0.029	0.012	0.015	0.044	0.048	0.013	0.012	0.045	0.070	0.011	0.009	0.038	0.058	0.018	0.016	0.016
HMH	0.041	0.061	0.050	0.049	0.035	0.057	0.042	0.035	0.076	0.047	0.058	0.035	0.033	0.042	0.060	0.035	0.055
HML	0.018	0.023	0.010	0.018	0.026	0.012	0.012	0.030	0.015	0.012	0.014	0.009	0.013	0.015	0.012	0.009	0.018
HMM	0.069	0.038	0.014	0.050	0.041	0.027	0.033	0.026	0.036	0.033	0.017	0.018	0.036	0.079	0.029	0.081	0.022
LHH	0.021	0.047	0.019	0.009	0.032	0.010	0.010	0.019	0.013	0.018	0.010	0.009	0.014	0.016	0.020	0.021	0.009
LHL	0.088	0.018	0.016	0.042	0.015	0.054	0.019	0.013	0.014	0.035	0.023	0.014	0.016	0.017	0.021	0.017	0.016
LHM	0.036	0.029	0.019	0.022	0.048	0.039	0.061	0.029	0.020	0.062	0.034	0.024	0.032	0.053	0.044	0.017	0.027
LLH	0.048	0.030	0.023	0.009	0.077	0.023	0.012	0.018	0.045	0.018	0.017	0.028	0.011	0.025	0.041	0.030	0.012
LLL	0.046	0.017	0.006	0.012	0.033	0.021	0.024	0.014	0.013	0.032	0.009	0.010	0.019	0.018	0.015	0.027	0.029
LLM	0.069	0.074	0.039	0.059	0.023	0.057	0.089	0.054	0.087	0.045	0.043	0.018	0.029	0.060	0.081	0.058	0.013
LMH	0.060	0.032	0.044	0.010	0.010	0.019	0.036	0.013	0.058	0.020	0.013	0.020	0.027	0.025	0.025	0.050	0.018
LML	0.027	0.032	0.053	0.034	0.044	0.012	0.023	0.023	0.040	0.073	0.018	0.039	0.016	0.093	0.027	0.029	0.084
LMM	0.051	0.065	0.087	0.059	0.033	0.215	0.044	0.023	0.063	0.094	0.025	0.028	0.079	0.016	0.037	0.075	0.041
MHH	0.061	0.090	0.038	0.043	0.028	0.030	0.049	0.050	0.077	0.019	0.031	0.021	0.100	0.020	0.035	0.021	0.020
MHL	0.030	0.012	0.012	0.071	0.040	0.036	0.014	0.022	0.028	0.059	0.030	0.028	0.018	0.015	0.017	0.022	0.024
MHM	0.055	0.036	0.022	0.033	0.026	0.016	0.061	0.027	0.026	0.051	0.042	0.021	0.053	0.024	0.033	0.017	0.066
MLH	0.115	0.027	0.023	0.049	0.024	0.039	0.030	0.029	0.033	0.036	0.120	0.023	0.036	0.031	0.024	0.015	0.021
MLL	0.040	0.029	0.018	0.026	0.012	0.039	0.022	0.012	0.032	0.016	0.021	0.024	0.011	0.036	0.010	0.013	0.045
MLM	0.059	0.120	0.018	0.064	0.018	0.038	0.021	0.021	0.024	0.035	0.028	0.154	0.038	0.105	0.024	0.029	0.030
MMH	0.075	0.039	0.037	0.043	0.050	0.039	0.064	0.027	0.105	0.139	0.026	0.032	0.039	0.055	0.049	0.022	0.023
MML	0.051	0.054	0.012	0.036	0.029	0.018	0.030	0.019	0.034	0.033	0.053	0.021	0.029	0.058	0.047	0.041	0.048
MMM	0.035	0.096	0.028	0.065	0.031	0.050	0.073	0.108	0.031	0.107	0.053	0.071	0.080	0.168	0.076	0.079	0.035

Figure B.6: *RMSEs between DTW-aligned speaker-specific and corpus-wide mean contours for* **male** *speakers.*

SPEAKER	CV			V		
	H	M	L	H	M	L
01	0.210 (394)	0.227 (195)	0.204 (232)	0.118 (47)	0.126 (123)	0.119 (162)
02	0.247 (305)	0.234 (221)	0.204 (150)	0.118 (36)	0.136 (112)	0.129 (118)
03	0.235 (368)	0.235 (214)	0.221 (224)	0.146 (40)	0.142 (103)	0.129 (150)
04	0.245 (318)	0.253 (191)	0.240 (182)	0.113 (30)	0.146 (108)	0.117 (126)
05	0.187 (340)	0.191 (176)	0.191 (187)	0.087 (39)	0.121 (105)	0.101 (124)
06	0.225 (336)	0.220 (206)	0.200 (198)	0.128 (33)	0.125 (128)	0.127 (132)
08	0.235 (371)	0.246 (228)	0.220 (212)	0.125 (55)	0.133 (144)	0.109 (139)
09	0.204 (304)	0.206 (181)	0.189 (168)	0.091 (30)	0.113 (103)	0.103 (120)
10	0.188 (317)	0.189 (180)	0.181 (173)	0.104 (34)	0.110 (105)	0.105 (129)
11	0.172 (257)	0.181 (169)	0.167 (162)	0.076 (27)	0.101 (102)	0.100 (099)
12	0.204 (297)	0.204 (198)	0.184 (167)	0.083 (25)	0.121 (100)	0.102 (123)
13	0.241 (366)	0.238 (217)	0.221 (236)	0.133 (47)	0.124 (150)	0.123 (124)
14	0.212 (286)	0.218 (143)	0.209 (161)	0.113 (34)	0.116 (073)	0.108 (085)
15	0.234 (332)	0.254 (211)	0.218 (188)	0.129 (40)	0.141 (122)	0.123 (121)
16	0.243 (382)	0.234 (211)	0.211 (218)	0.114 (43)	0.132 (129)	0.120 (145)
17	0.249 (416)	0.244 (237)	0.236 (224)	0.139 (44)	0.141 (132)	0.127 (174)
19	0.210 (375)	0.218 (216)	0.190 (211)	0.094 (37)	0.119 (118)	0.097 (122)
20	0.279 (379)	0.275 (284)	0.258 (216)	0.146 (47)	0.160 (140)	0.143 (153)
21	0.218 (389)	0.225 (229)	0.204 (206)	0.122 (35)	0.133 (120)	0.115 (156)
22	0.207 (265)	0.203 (184)	0.191 (197)	0.096 (40)	0.118 (098)	0.107 (104)
23	0.296 (331)	0.302 (232)	0.256 (178)	0.150 (25)	0.143 (111)	0.136 (133)
24	0.222 (348)	0.226 (197)	0.200 (182)	0.112 (45)	0.122 (108)	0.116 (132)
26	0.215 (308)	0.220 (183)	0.211 (156)	0.124 (31)	0.132 (102)	0.119 (110)
27	0.267 (427)	0.275 (238)	0.236 (225)	0.158 (44)	0.151 (148)	0.121 (171)
28	0.223 (418)	0.221 (263)	0.204 (259)	0.107 (47)	0.139 (133)	0.127 (156)
29	0.223 (325)	0.230 (216)	0.202 (188)	0.126 (36)	0.121 (111)	0.117 (126)
30	0.276 (321)	0.261 (207)	0.217 (203)	0.140 (38)	0.142 (106)	0.124 (130)
31	0.184 (328)	0.178 (179)	0.178 (193)	0.088 (50)	0.107 (109)	0.095 (140)
32	0.256 (315)	0.260 (172)	0.229 (209)	0.147 (41)	0.141 (095)	0.119 (120)
33	0.232 (375)	0.233 (199)	0.209 (227)	0.113 (42)	0.126 (122)	0.113 (141)
34	0.227 (319)	0.232 (226)	0.213 (205)	0.120 (38)	0.133 (115)	0.124 (148)
35	0.248 (417)	0.255 (223)	0.232 (247)	0.160 (41)	0.161 (123)	0.134 (169)
36	0.260 (387)	0.269 (220)	0.255 (227)	0.148 (39)	0.149 (116)	0.146 (132)
ALL	**0.231 (11416)**	**0.234 (6846)**	**0.213 (6611)**	**0.121 (1280)**	**0.132 (3814)**	**0.119 (4414)**

Figure B.7: *Mean durations of syllables (in seconds) for different speaker, syllable type and tone combinations (number of instances are indicated in parentheses). The unequal distribution of tones over the different syllable types may be due the tonotactic restriction where the H tone generally only occurs in word-initial position in consonant-initial words [10]. This restriction, however, presumably only applies to polysyllabic words (examples of vowel-only words with H tone are presented in [10]). Counting all the word-initial syllables of polysyllabic words for different syllable types and tones resulted in* **CV***: H: 3128, M: 1005, L: 1133 and* **V***: H: 70, M: 3071, L: 3384. Inspection of the few cases with word-initial V and H syllables revealed some words appearing to be of foreign origin (e.g. "álífábẹ́ẹ́ti"), with other cases possibly being due to typographical errors.*

SPEAKER	mean100					mean50				
	H	L	M	MEAN	CORRECT %	H	L	M	MEAN	CORRECT %
01	0.01	0.50	0.41	0.31	37.29	0.09	0.54	0.44	**0.36**	40.71
02	0.47	0.44	0.49	0.47	47.15	0.53	0.51	0.49	**0.51**	50.91
03	0.53	0.04	0.20	0.26	37.01	0.55	0.05	0.19	**0.26**	38.15
04	0.49	0.46	0.45	0.47	46.72	0.60	0.50	0.49	**0.53**	53.24
05	0.45	0.39	0.44	0.43	43.05	0.55	0.44	0.46	**0.48**	48.71
06	0.51	0.38	0.50	0.46	47.26	0.59	0.45	0.48	**0.51**	51.08
08	0.54	0.47	0.50	0.50	50.13	0.63	0.58	0.46	**0.56**	55.79
09	0.26	0.50	0.40	0.39	40.66	0.31	0.54	0.46	**0.44**	45.43
10	0.48	0.40	0.49	0.46	46.25	0.56	0.47	0.50	**0.51**	51.19
11	0.51	0.42	0.10	0.35	40.10	0.56	0.49	0.10	**0.39**	44.49
12	0.48	0.24	0.51	0.41	44.33	0.59	0.13	0.54	**0.42**	48.66
13	0.47	0.51	0.41	0.46	46.29	0.56	0.61	0.39	**0.52**	52.61
14	0.46	0.13	0.45	0.35	39.75	0.54	0.09	0.45	**0.36**	41.98
15	0.05	0.44	0.44	0.31	36.18	0.21	0.46	0.47	**0.38**	40.39
16	0.54	0.33	0.49	0.45	47.19	0.64	0.38	0.54	**0.52**	54.22
17	0.53	0.49	0.44	0.49	48.80	0.60	0.53	0.48	**0.54**	54.03
19	0.46	0.29	0.43	0.39	40.99	0.57	0.32	0.49	**0.46**	48.00
20	0.54	0.46	0.45	0.48	48.66	0.62	0.51	0.42	**0.52**	52.72
21	0.50	0.45	0.44	0.46	46.31	0.55	0.45	0.47	**0.49**	49.34
22	0.46	0.43	0.53	0.47	48.08	0.57	0.51	0.57	**0.55**	55.41
23	0.50	0.51	0.44	0.48	47.80	0.58	0.58	0.49	**0.55**	54.46
24	0.50	0.52	0.47	0.49	49.39	0.58	0.59	0.48	**0.55**	55.01
26	0.51	0.38	0.44	0.44	44.81	0.58	0.44	0.47	**0.50**	50.11
27	0.54	0.46	0.44	0.48	48.10	0.64	0.58	0.45	**0.56**	56.13
28	0.43	0.46	0.51	0.47	47.17	0.55	0.52	0.49	**0.52**	52.20
29	0.52	0.54	0.49	0.52	51.48	0.57	0.61	0.53	**0.57**	56.68
30	0.57	0.28	0.51	0.45	48.29	0.65	0.40	0.55	**0.53**	54.81
31	0.54	0.43	0.48	0.48	48.97	0.62	0.51	0.49	**0.54**	54.02
32	0.55	0.51	0.41	0.49	48.96	0.64	0.61	0.47	**0.57**	57.54
33	0.51	0.50	0.48	0.50	49.78	0.58	0.58	0.48	**0.55**	54.69
34	0.38	0.51	0.38	0.42	43.03	0.45	0.57	0.35	**0.46**	46.03
35	0.52	0.51	0.12	0.39	44.92	0.57	0.55	0.13	**0.41**	47.97
36	0.47	0.37	0.52	**0.46**	47.56	0.51	0.25	0.53	0.43	47.28
MEAN	**0.46**	**0.42**	**0.43**	**0.44**	**45.53**	**0.54**	**0.47**	**0.45**	**0.49**	**50.42**

Figure B.8: *Results in the table show classification results for two experiments; when pitch level is represented by the mean over the entire syllable (*`mean100`*) and over the final 50% of the syllable duration (*`mean50`*). Results for the three tones are reported in terms of the F1 score for each speaker with the mean of the three values and overall percentage of correct classifications included. Bold entries in the "mean" column indicate the larger of the values between* `mean100` *and* `mean50`*. Shading in the last column illustrates the relative correct classification rates between speakers.*

SPEAKER	mean50												
	H-H	H-L	H-M	L-H	L-L	L-M	M-H	M-L	M-M	N-H	N-L	N-M	CORRECT %
01	0.40	0.39	0.38	0.65	0.66	0.27	0.02	0.58	0.54	0.34	0.43	0.48	47.00
02	0.56	0.38	0.49	0.32	0.57	0.47	0.67	0.53	0.37	0.62	0.58	0.56	50.27
03	0.61	0.09	0.56	0.50	0.14	0.33	0.59	0.07	0.21	0.82	0.82	0.52	44.79
04	0.53	0.11	0.60	0.45	0.61	0.36	0.72	0.63	0.53	0.82	0.79	0.60	55.77
05	0.57	0.11	0.56	0.49	0.29	0.35	0.61	0.67	0.47	0.72	0.77	0.71	50.67
06	0.62	0.02	0.56	0.46	0.54	0.42	0.64	0.64	0.49	0.69	0.72	0.59	52.95
08	0.67	0.39	0.51	0.38	0.58	0.41	0.72	0.67	0.50	0.76	0.89	0.66	56.83
09	0.40	0.40	0.43	0.18	0.52	0.36	0.38	0.60	0.54	0.33	0.65	0.37	44.10
10	0.63	0.07	0.57	0.12	0.61	0.35	0.70	0.56	0.49	0.81	0.83	0.72	52.69
11	0.53	0.21	0.53	0.50	0.46	0.33	0.64	0.52	0.47	0.56	0.54	0.09	47.98
12	0.50	0.05	0.57	0.45	0.21	0.46	0.74	0.68	0.50	0.78	0.63	0.57	51.52
13	0.65	0.56	0.42	0.11	0.64	0.36	0.67	0.59	0.42	0.78	0.69	0.66	54.10
14	0.49	0.06	0.53	0.52	0.08	0.41	0.67	0.62	0.25	0.70	0.65	0.56	47.28
15	0.34	0.34	0.46	0.20	0.53	0.43	0.35	0.51	0.50	0.33	0.55	0.35	42.25
16	0.59	0.01	0.64	0.64	0.39	0.31	0.72	0.62	0.57	0.88	0.75	0.53	55.42
17	0.63	0.18	0.54	0.34	0.60	0.47	0.69	0.64	0.37	0.77	0.73	0.62	53.95
19	0.59	0.01	0.56	0.44	0.48	0.40	0.63	0.46	0.44	0.70	0.64	0.40	48.75
20	0.64	0.42	0.52	0.51	0.58	0.38	0.68	0.56	0.38	0.67	0.73	0.59	54.57
21	0.59	0.24	0.47	0.35	0.63	0.35	0.65	0.66	0.42	0.73	0.73	0.56	51.62
22	0.55	0.30	0.63	0.51	0.56	0.35	0.69	0.63	0.55	0.72	0.77	0.72	55.92
23	0.63	0.51	0.44	0.34	0.56	0.45	0.61	0.65	0.43	0.57	0.77	0.54	52.85
24	0.69	0.44	0.52	0.39	0.60	0.47	0.68	0.73	0.50	0.70	0.82	0.61	57.83
26	0.59	0.28	0.48	0.43	0.52	0.42	0.64	0.53	0.47	0.62	0.68	0.15	49.82
27	0.60	0.14	0.52	0.55	0.72	0.36	0.73	0.62	0.50	0.79	0.79	0.55	56.70
28	0.61	0.30	0.50	0.34	0.52	0.45	0.66	0.65	0.52	0.65	0.68	0.62	52.42
29	0.63	0.47	0.53	0.35	0.69	0.46	0.65	0.55	0.46	0.69	0.79	0.60	55.98
30	0.58	0.12	0.54	0.59	0.55	0.49	0.79	0.65	0.58	0.82	0.74	0.71	57.61
31	0.64	0.10	0.55	0.35	0.62	0.41	0.76	0.57	0.60	0.67	0.87	0.58	55.41
32	0.67	0.34	0.54	0.60	0.66	0.36	0.71	0.70	0.47	0.81	0.81	0.75	59.71
33	0.70	0.53	0.50	0.45	0.56	0.41	0.63	0.64	0.53	0.75	0.73	0.70	56.84
34	0.56	0.44	0.44	0.06	0.57	0.38	0.58	0.63	0.32	0.64	0.74	0.57	48.53
35	0.55	0.32	0.49	0.26	0.50	0.42	0.69	0.68	0.08	0.70	0.86	0.73	50.80
36	0.69	0.44	0.47	0.32	0.11	0.45	0.72	0.46	0.60	0.80	0.82	0.69	52.78
MEAN	**0.58**	**0.26**	**0.52**	**0.40**	**0.51**	**0.40**	**0.64**	**0.59**	**0.46**	**0.69**	**0.73**	**0.57**	**52.29**

SPEAKER	lingrad												
	H-H	H-L	H-M	L-H	L-L	L-M	M-H	M-L	M-M	N-H	N-L	N-M	CORRECT %
01	0.51	0.31	0.17	0.62	0.65	0.18	0.32	0.54	0.37	0.31	0.39	0.11	44.26
02	0.52	0.40	0.51	0.29	0.33	0.40	0.40	0.41	0.30	0.44	0.40	0.13	39.54
03	0.71	0.53	0.53	0.62	0.35	0.39	0.59	0.06	0.21	0.21	0.17	0.45	47.78
04	0.58	0.44	0.50	0.45	0.57	0.33	0.62	0.53	0.54	0.37	0.39	0.47	50.22
05	0.65	0.43	0.53	0.58	0.41	0.37	0.72	0.65	0.34	0.51	0.48	0.51	53.27
06	0.63	0.49	0.36	0.54	0.47	0.38	0.75	0.59	0.60	0.42	0.33	0.55	53.38
08	0.63	0.49	0.52	0.43	0.43	0.40	0.61	0.50	0.56	0.40	0.46	0.45	50.67
09	0.34	0.35	0.27	0.17	0.53	0.16	0.28	0.47	0.33	0.25	0.17	0.34	34.17
10	0.71	0.52	0.47	0.23	0.59	0.35	0.72	0.58	0.49	0.52	0.40	0.17	52.27
11	0.61	0.42	0.37	0.44	0.25	0.40	0.64	0.44	0.46	0.53	0.12	0.05	45.22
12	0.44	0.39	0.47	0.55	0.22	0.45	0.65	0.41	0.40	0.51	0.25	0.47	46.10
13	0.64	0.58	0.23	0.59	0.55	0.14	0.44	0.49	0.48	0.68	0.46	0.43	49.87
14	0.61	0.48	0.14	0.65	0.08	0.33	0.69	0.49	0.31	0.51	0.15	0.24	47.48
15	0.18	0.39	0.48	0.36	0.36	0.43	0.33	0.49	0.35	0.48	0.09	0.32	38.68
16	0.68	0.36	0.54	0.69	0.28	0.34	0.72	0.56	0.54	0.67	0.59	0.20	54.85
17	0.63	0.50	0.22	0.43	0.40	0.39	0.69	0.63	0.22	0.51	0.30	0.24	48.08
19	0.57	0.28	0.31	0.57	0.31	0.12	0.65	0.32	0.25	0.30	0.41	0.46	43.42
20	0.56	0.34	0.18	0.44	0.53	0.36	0.56	0.52	0.23	0.39	0.33	0.20	43.03
21	0.43	0.40	0.45	0.60	0.62	0.39	0.66	0.65	0.47	0.23	0.14	0.48	50.03
22	0.68	0.58	0.35	0.50	0.48	0.35	0.71	0.65	0.61	0.33	0.33	0.45	53.07
23	0.63	0.53	0.33	0.35	0.25	0.44	0.61	0.56	0.53	0.08	0.48	0.13	47.01
24	0.48	0.48	0.48	0.41	0.56	0.43	0.42	0.55	0.53	0.49	0.44	0.51	48.26
26	0.59	0.52	0.39	0.66	0.38	0.40	0.64	0.59	0.46	0.43	0.52	0.07	51.23
27	0.59	0.48	0.30	0.62	0.47	0.44	0.63	0.52	0.61	0.48	0.35	0.21	51.13
28	0.63	0.43	0.40	0.58	0.38	0.28	0.58	0.44	0.57	0.14	0.32	0.55	48.98
29	0.59	0.46	0.25	0.34	0.68	0.38	0.55	0.44	0.61	0.34	0.27	0.54	50.79
30	0.69	0.51	0.54	0.58	0.34	0.43	0.74	0.53	0.66	0.28	0.54	0.52	55.72
31	0.64	0.50	0.54	0.39	0.50	0.40	0.70	0.75	0.66	0.60	0.56	0.24	55.71
32	0.57	0.66	0.51	0.46	0.65	0.38	0.64	0.61	0.14	0.52	0.61	0.61	54.31
33	0.59	0.61	0.17	0.47	0.44	0.42	0.56	0.52	0.53	0.38	0.44	0.51	48.85
34	0.59	0.63	0.46	0.49	0.48	0.44	0.65	0.67	0.11	0.09	0.63	0.33	52.15
35	0.54	0.53	0.50	0.28	0.44	0.36	0.63	0.61	0.11	0.24	0.17	0.50	45.54
36	0.65	0.41	0.42	0.29	0.17	0.41	0.75	0.19	0.55	0.58	0.57	0.39	48.09
MEAN	**0.58**	**0.47**	**0.39**	**0.47**	**0.43**	**0.36**	**0.60**	**0.51**	**0.43**	**0.40**	**0.37**	**0.36**	**48.58**

Figure B.9: *Results in the table show results for two classification experiments; when modelling tones with 12 distributions using the absolute pitch (*`mean50`*) and linear gradient within the current syllable (*`lingrad`*). Results are reported in terms of the F1 score for each speaker and context, with overall percentage of correct classifications included. Shading illustrates the relative classification rates between speakers within each experiment.*

SPEAKER	deltamean												
	H-H	H-L	H-M	L-H	L-L	L-M	M-H	M-L	M-M	N-H	N-L	N-M	CORRECT %
01	0.43	0.10	0.31	0.64	0.48	0.23	0.26	0.55	0.54	0.32	0.39	0.15	40.83
02	0.53	0.31	0.48	0.23	0.57	0.46	0.68	0.49	0.43	0.38	0.09	0.57	47.76
03	0.69	0.43	0.56	0.64	0.24	0.29	0.62	0.08	0.27	0.21	0.18	0.46	46.67
04	0.61	0.17	0.60	0.43	0.56	0.34	0.75	0.58	0.61	0.48	0.43	0.48	53.20
05	0.64	0.28	0.57	0.54	0.38	0.34	0.72	0.71	0.52	0.47	0.27	0.57	53.09
06	0.60	0.31	0.48	0.43	0.52	0.40	0.79	0.67	0.61	0.18	0.47	0.09	51.91
08	0.63	0.39	0.55	0.34	0.52	0.42	0.76	0.67	0.60	0.55	0.14	0.12	53.49
09	0.19	0.35	0.40	0.42	0.60	0.36	0.30	0.50	0.35	0.38	0.35	0.34	39.45
10	0.68	0.23	0.58	0.01	0.60	0.36	0.80	0.61	0.64	0.43	0.53	0.24	53.28
11	0.62	0.30	0.51	0.51	0.10	0.33	0.59	0.45	0.58	0.46	0.08	0.08	45.34
12	0.53	0.19	0.49	0.49	0.32	0.41	0.76	0.60	0.38	0.48	0.42	0.11	47.58
13	0.63	0.55	0.33	0.17	0.61	0.39	0.70	0.57	0.34	0.68	0.45	0.22	50.62
14	0.62	0.39	0.13	0.57	0.07	0.35	0.81	0.60	0.16	0.37	0.19	0.51	47.15
15	0.27	0.40	0.27	0.06	0.48	0.37	0.37	0.47	0.50	0.09	0.46	0.23	36.55
16	0.63	0.17	0.62	0.59	0.34	0.22	0.73	0.57	0.60	0.26	0.54	0.16	51.17
17	0.67	0.31	0.55	0.32	0.46	0.42	0.72	0.67	0.56	0.44	0.26	0.33	51.51
19	0.49	0.34	0.49	0.37	0.36	0.40	0.64	0.44	0.51	0.20	0.19	0.42	44.47
20	0.64	0.43	0.40	0.59	0.60	0.11	0.64	0.50	0.09	0.52	0.40	0.24	48.74
21	0.66	0.37	0.49	0.50	0.62	0.37	0.78	0.65	0.59	0.21	0.13	0.58	54.79
22	0.66	0.29	0.50	0.41	0.53	0.36	0.72	0.63	0.60	0.30	0.29	0.47	51.45
23	0.74	0.44	0.57	0.47	0.46	0.49	0.70	0.69	0.55	0.11	0.58	0.23	55.59
24	0.61	0.38	0.53	0.30	0.48	0.42	0.69	0.68	0.59	0.16	0.16	0.43	50.20
26	0.58	0.52	0.52	0.61	0.52	0.22	0.75	0.60	0.69	0.44	0.09	0.46	55.06
27	0.64	0.55	0.34	0.63	0.64	0.34	0.77	0.65	0.69	0.19	0.18	0.42	56.53
28	0.55	0.26	0.50	0.37	0.57	0.33	0.69	0.61	0.54	0.13	0.39	0.07	47.64
29	0.66	0.50	0.60	0.40	0.64	0.40	0.67	0.60	0.38	0.36	0.43	0.26	53.65
30	0.70	0.43	0.59	0.53	0.49	0.45	0.83	0.68	0.71	0.23	0.17	0.45	57.43
31	0.64	0.42	0.54	0.45	0.52	0.33	0.73	0.78	0.66	0.64	0.64	0.21	56.19
32	0.68	0.52	0.59	0.54	0.67	0.40	0.72	0.74	0.58	0.58	0.61	0.19	59.65
33	0.63	0.60	0.36	0.48	0.57	0.40	0.71	0.68	0.64	0.46	0.49	0.10	54.51
34	0.57	0.54	0.46	0.52	0.46	0.36	0.74	0.70	0.60	0.19	0.62	0.40	54.05
35	0.66	0.40	0.53	0.38	0.49	0.37	0.70	0.67	0.11	0.31	0.57	0.05	49.84
36	0.70	0.48	0.54	0.33	0.13	0.39	0.78	0.42	0.60	0.59	0.43	0.32	50.79
MEAN	**0.60**	**0.37**	**0.48**	**0.43**	**0.47**	**0.36**	**0.69**	**0.59**	**0.51**	**0.36**	**0.35**	**0.30**	**50.61**

SPEAKER	mean50+lingrad+deltamean												
	H-H	H-L	H-M	L-H	L-L	L-M	M-H	M-L	M-M	N-H	N-L	N-M	CORRECT %
01	0.45	0.28	0.33	0.60	0.63	0.31	0.36	0.61	0.51	0.48	0.47	0.41	46.47
02	0.61	0.41	0.55	0.45	0.42	0.44	0.64	0.48	0.34	0.53	0.51	0.45	49.26
03	0.74	0.62	0.61	0.61	0.33	0.38	0.61	0.10	0.22	0.57	0.31	0.47	51.28
04	0.67	0.41	0.62	0.48	0.60	0.34	0.71	0.57	0.58	0.74	0.74	0.56	57.87
05	0.68	0.46	0.61	0.55	0.42	0.41	0.73	0.72	0.49	0.61	0.67	0.65	57.91
06	0.67	0.52	0.59	0.57	0.47	0.44	0.78	0.67	0.63	0.54	0.62	0.40	58.84
08	0.69	0.48	0.60	0.44	0.52	0.45	0.74	0.71	0.61	0.78	0.75	0.59	59.97
09	0.46	0.36	0.44	0.38	0.55	0.35	0.38	0.57	0.45	0.39	0.62	0.35	44.80
10	0.75	0.51	0.65	0.33	0.62	0.36	0.77	0.59	0.59	0.76	0.80	0.56	60.11
11	0.61	0.38	0.47	0.50	0.42	0.42	0.56	0.47	0.52	0.59	0.54	0.32	49.39
12	0.52	0.48	0.61	0.56	0.19	0.48	0.75	0.65	0.49	0.76	0.52	0.60	55.30
13	0.67	0.61	0.40	0.57	0.63	0.38	0.66	0.59	0.51	0.73	0.57	0.64	57.90
14	0.59	0.41	0.34	0.63	0.05	0.41	0.75	0.61	0.35	0.65	0.34	0.55	50.84
15	0.42	0.35	0.49	0.36	0.50	0.41	0.36	0.49	0.46	0.46	0.53	0.33	43.37
16	0.73	0.48	0.66	0.66	0.39	0.33	0.75	0.66	0.62	0.84	0.69	0.19	60.53
17	0.71	0.47	0.57	0.41	0.58	0.44	0.73	0.66	0.45	0.78	0.67	0.62	58.12
19	0.58	0.37	0.45	0.56	0.48	0.32	0.63	0.49	0.33	0.67	0.66	0.52	50.31
20	0.63	0.42	0.39	0.48	0.58	0.40	0.65	0.57	0.26	0.66	0.59	0.24	50.62
21	0.65	0.45	0.53	0.53	0.66	0.41	0.75	0.68	0.51	0.67	0.47	0.60	58.02
22	0.70	0.58	0.61	0.45	0.42	0.40	0.72	0.66	0.66	0.50	0.61	0.58	57.46
23	0.73	0.52	0.54	0.45	0.46	0.50	0.68	0.74	0.58	0.46	0.69	0.41	57.67
24	0.73	0.47	0.59	0.46	0.63	0.48	0.68	0.73	0.61	0.60	0.77	0.59	60.63
26	0.67	0.53	0.51	0.57	0.58	0.39	0.71	0.59	0.62	0.46	0.53	0.16	56.69
27	0.67	0.50	0.54	0.60	0.64	0.44	0.78	0.71	0.69	0.77	0.48	0.50	61.84
28	0.66	0.44	0.53	0.53	0.55	0.35	0.67	0.59	0.60	0.27	0.56	0.31	53.84
29	0.67	0.51	0.53	0.47	0.70	0.40	0.65	0.58	0.62	0.65	0.77	0.60	59.56
30	0.72	0.57	0.65	0.57	0.47	0.47	0.80	0.61	0.68	0.77	0.57	0.62	62.23
31	0.70	0.58	0.63	0.38	0.63	0.45	0.77	0.74	0.71	0.74	0.83	0.34	63.09
32	0.73	0.64	0.64	0.62	0.70	0.32	0.72	0.71	0.41	0.80	0.83	0.77	64.58
33	0.73	0.63	0.47	0.48	0.54	0.44	0.67	0.65	0.62	0.68	0.67	0.64	59.05
34	0.64	0.60	0.50	0.52	0.47	0.47	0.67	0.72	0.34	0.50	0.70	0.42	55.37
35	0.63	0.57	0.56	0.31	0.51	0.38	0.67	0.75	0.11	0.70	0.84	0.73	54.75
36	0.73	0.49	0.52	0.26	0.34	0.44	0.76	0.46	0.61	0.75	0.78	0.64	55.12
MEAN	**0.65**	**0.49**	**0.54**	**0.49**	**0.50**	**0.41**	**0.67**	**0.61**	**0.51**	**0.63**	**0.63**	**0.50**	**55.84**

Figure B.10: *Results in the table show results for two classification experiments; when modelling tones with 12 distributions using the change in pitch between the current and previous syllable (*`deltamean`*) and a combination of features:* `mean50`, `lingrad` *and* `deltamean`*. Results are reported in terms of the F1 score for each speaker and context, with overall percentage of correct classifications included. Shading illustrates the relative classification rates between speakers within each experiment.*

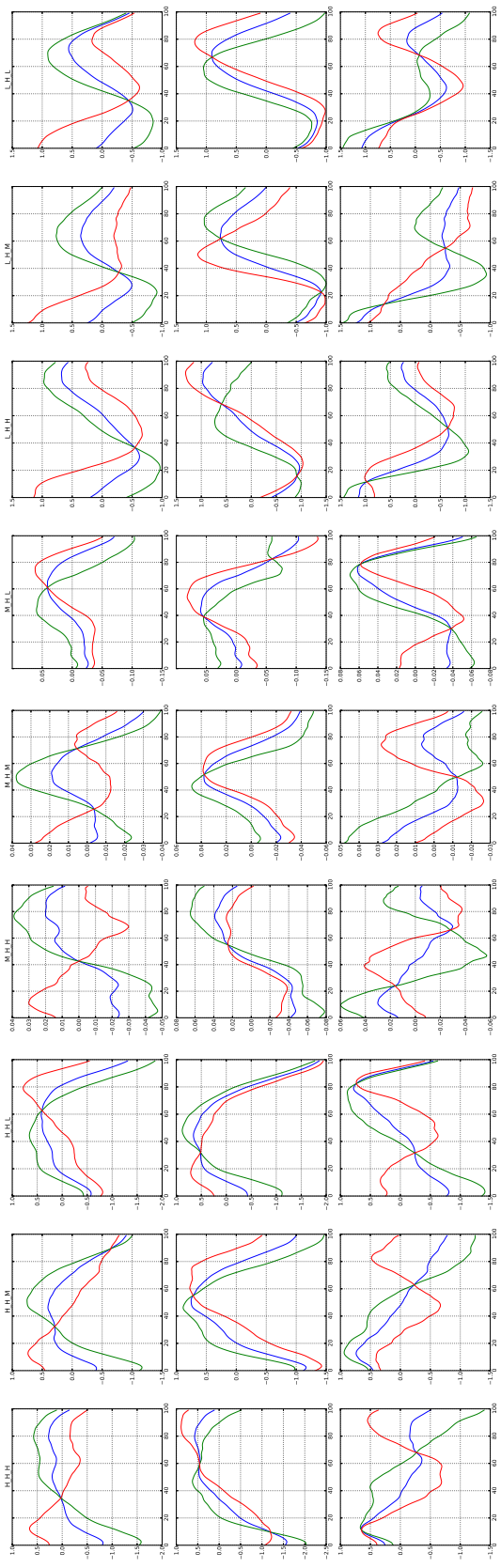

Figure B.11: *Tri-tone contours with* **H as the central tone** *identified using k-means clustering as described in Section 3.3.5. In each plot, blue contours are the mean over all the tri-tone samples, with red and green the resulting clusters. The first row of plots show the first iteration of clustering with the second and third rows the second iteration starting with the clusters identified in the first iteration. Blue contours in the second and third rows thus correspond to green and red contours in the first row respectively.*

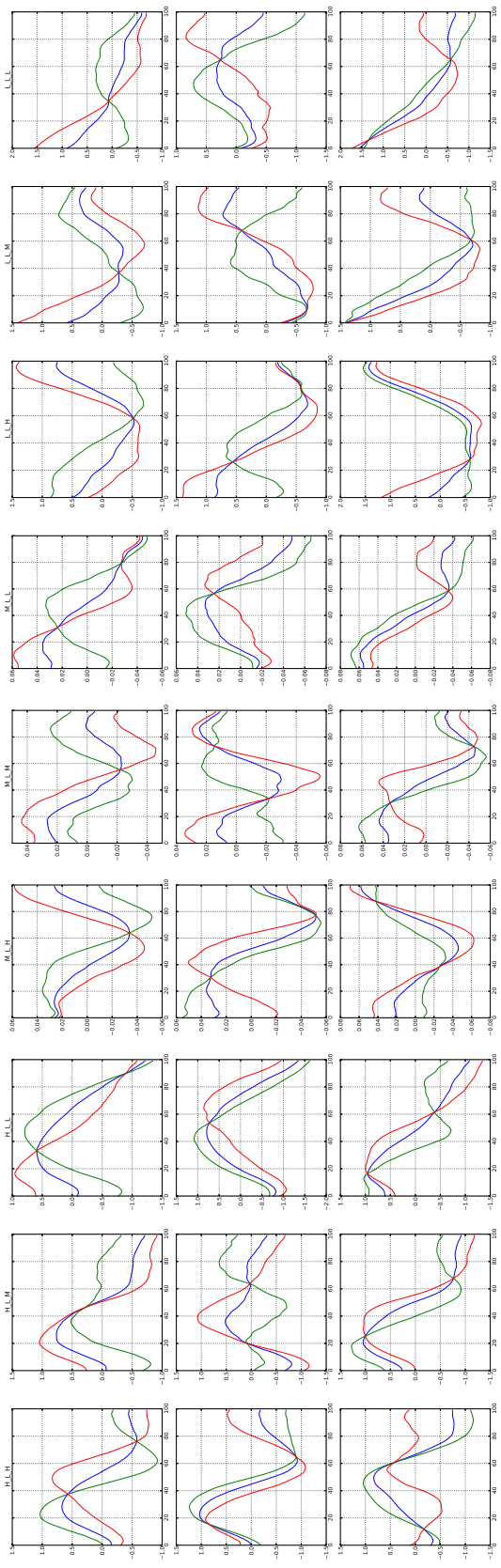

Figure B.12: *Tri-tone contours with* **L as the central tone** *identified using k-means clustering as described in Section 3.3.5. In each plot, blue contours are the mean over all the tri-tone samples, with red and green the resulting clusters. The first row of plots show the first iteration of clustering with the second and third rows the second iteration starting with the clusters identified in the first iteration. Blue contours in the second and third rows thus correspond to green and red contours in the first row respectively.*

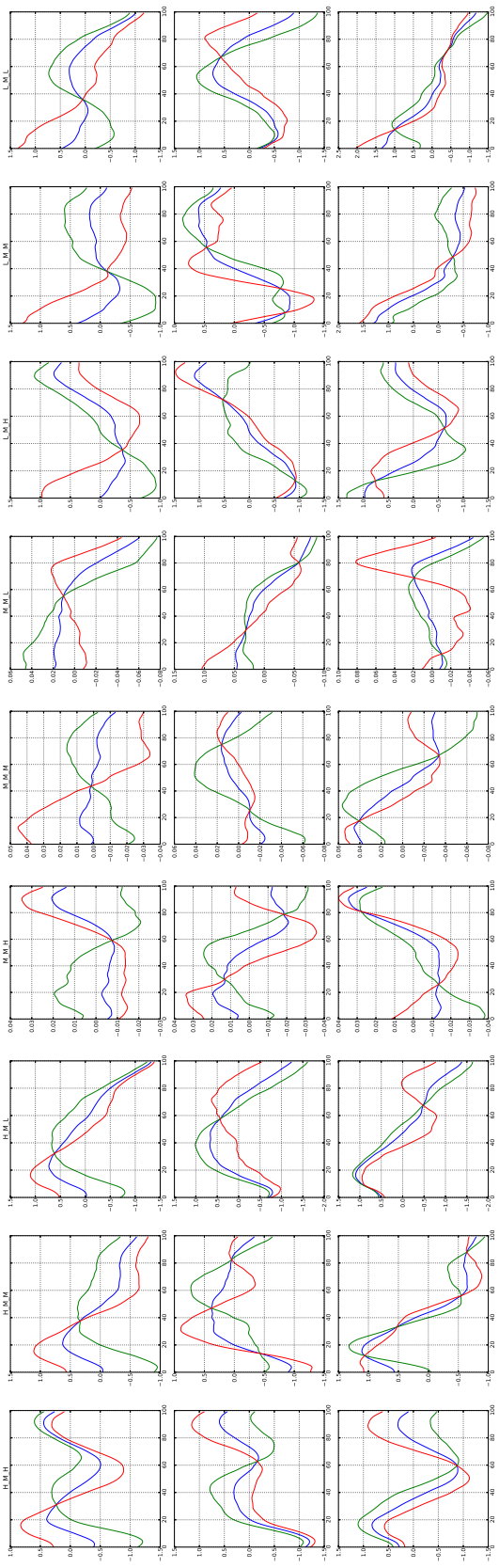

Figure B.13: *Tri-tone contours with* **M as the central tone** *identified using k-means clustering as described in Section 3.3.5. In each plot, blue contours are the mean over all the tri-tone samples, with red and green the resulting clusters. The first row of plots show the first iteration of clustering with the second and third rows the second iteration starting with the clusters identified in the first iteration. Blue contours in the second and third rows thus correspond to green and red contours in the first row respectively.*

Tone context	RMSE	Tone context	RMSE	Tone context	RMSE	Tone context	RMSE
H-LMM	0.49	H-MHH	0.17	LML+H	0.12	HLH+L	0.08
H-LHM	0.47	MHL+L	0.17	MMH+H	0.12	M-MHH	0.08
H-LLH	0.46	M-HMM	0.17	M-LML	0.12	MMH+L	0.08
H-LHH	0.46	M-LLM	0.17	MLM+L	0.12	LLH+H	0.08
H-LLM	0.45	LLL+M	0.17	LMM+H	0.12	M-MMM	0.08
H-LHL	0.45	L-HLM	0.17	MLH+L	0.11	HMM+H	0.08
H-LMH	0.42	HLL+M	0.16	HHM+H	0.11	M-LLH	0.08
H-LML	0.37	LHL+M	0.16	M-HHL	0.11	M-MLM	0.08
L-HHM	0.31	MHL+M	0.16	LHM+L	0.11	MML+L	0.08
H-HMM	0.29	L-HML	0.16	HHL+M	0.11	MLH+H	0.08
L-HMH	0.28	LHL+L	0.16	LLM+H	0.11	HHH+H	0.08
H-HHM	0.26	MMH+M	0.15	LHH+M	0.11	LML+L	0.07
H-MMM	0.25	M-HLM	0.15	MHM+M	0.11	HML+L	0.07
H-HLH	0.25	L-HLH	0.15	M-HLL	0.11	L-MLL	0.07
L-HMM	0.25	M-HMH	0.15	LLM+M	0.11	LMH+L	0.07
H-HLL	0.25	H-MLL	0.15	MMM+H	0.11	LLM+L	0.07
L-HHH	0.24	MML+M	0.14	L-MLH	0.11	HLL+L	0.07
H-MHM	0.24	L-LML	0.14	L-MMH	0.10	MMM+L	0.07
L-LLH	0.24	LML+M	0.14	HHL+L	0.10	HHM+L	0.07
H-MMH	0.24	MHH+M	0.14	M-LLL	0.10	LMM+M	0.07
L-LHM	0.24	L-LLL	0.14	HHH+L	0.10	HLH+M	0.06
L-LHH	0.23	L-LLM	0.14	MHL+H	0.10	M-MLL	0.06
L-LHL	0.23	HHL+H	0.14	LHM+H	0.10	HLM+M	0.06
H-HHH	0.23	M-LHM	0.13	L-MHH	0.10	HMM+M	0.06
L-LMM	0.23	HMH+M	0.13	MML+H	0.10	MLL+L	0.06
H-LLL	0.22	L-MLM	0.13	HHM+M	0.10	LMM+L	0.06
H-HMH	0.22	LLH+L	0.13	MLH+M	0.10	MLM+M	0.06
H-HLM	0.21	M-HHM	0.13	MLL+H	0.09	MHH+L	0.06
L-HHL	0.21	M-LHL	0.13	LHL+H	0.09	MMM+M	0.05
L-LMH	0.21	M-LMH	0.13	M-MHL	0.09	M-MLH	0.05
H-HHL	0.20	L-MHL	0.13	LLH+M	0.09	HLM+L	0.05
H-MHL	0.19	LMH+M	0.13	HML+H	0.09	HMM+L	0.05
M-HML	0.19	HML+M	0.12	LMH+H	0.09	HLH+H	0.05
H-MLH	0.18	HHH+M	0.12	MLL+M	0.09	MHM+L	0.05
L-HLL	0.18	HMH+L	0.12	MHH+H	0.09	M-MMH	0.05
H-MML	0.18	M-HLH	0.12	M-MML	0.09	LLL+L	0.05
H-MLM	0.18	LLL+H	0.12	M-LMM	0.08	HLM+H	0.04
L-MMM	0.18	MLM+H	0.12	HLL+H	0.08	LHH+H	0.04
L-MHM	0.18	M-MHM	0.12	L-MML	0.08	LHH+L	0.04
M-LHH	0.18	MHM+H	0.12	LHM+M	0.08		
H-HML	0.18	M-HHH	0.12	HMH+H	0.08		

Table B.3: *Root mean squared errors between four-syllable contours and corresponding three-syllable contours.*

APPENDIX C

ADDITIONAL RESULTS FOR CHAPTER 4

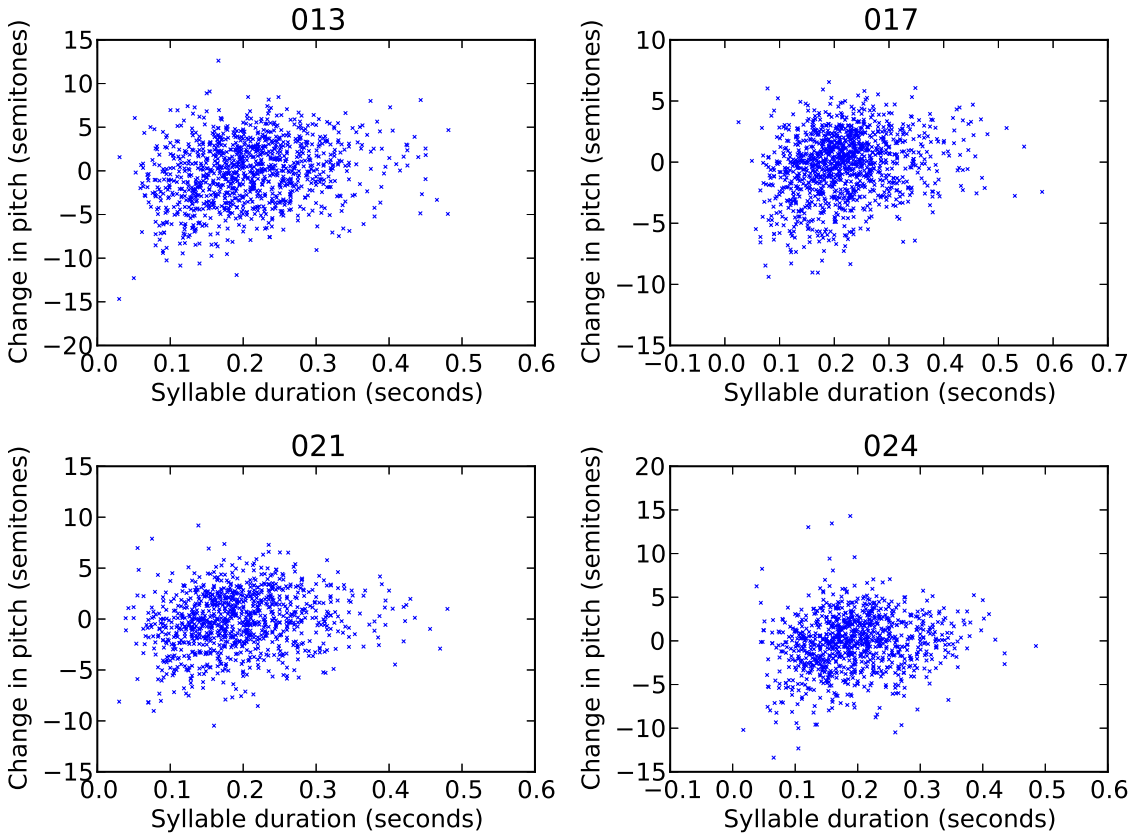

Figure C.1: *Pitch target changes versus syllable duration for the four speakers.*

C.1 INITIAL PITCH TARGET PREDICTION EXPERIMENT

The following is the initial pitch target prediction experiment published in [90].

In the Section 4.2 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 4.2 and deltas for different tone transitions in Figures 4.3 and 4.4).

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 onward. (This is done in order to have comparable results between models predicting targets directly or via deltas.)

C.1.1 Initial models

We start by proposing two models based on the observations in Section 4.2. The first model is based on the linear declination (observed in Figure 4.2) 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 C.1: `lint`.

The second model predicts pitch deltas between syllables based on the linear relationship between previous pitch level and pitch change (Figures 4.5 and 4.6). 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:

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

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 C.1: `lind`. Results are compared and discussed in Section C.1.3.

C.1.2 Additional features

To further investigate the utility of features discussed in Section 4.2, we experimented with two additional model types; regression trees [95], and support vector machines (SVM) [100] implemented in the *scikit-learn* software package [96]. Both decision trees and SVMs have been successfully applied to problems of acoustic modelling of speech [44, 99].

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 C.1 and discussed in the next

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 C.1: *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.*

section.

C.1.3 Discussion

In Table C.1 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

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 C.2: *Linear downtrend estimates (in semitones per second) for different models and feature combinations compared to actual samples.*

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 C.2 shows the linear estimates of downtrend measured over all utterances. Models not considering utterance position (i.e. `meant` and `lind`) under-estimate 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 4.3 and 4.4.

APPENDIX D

ADDITIONAL RESULTS FOR CHAPTER 5

D.1 INITIAL PITCH CONTOUR SYNTHESIS EXPERIMENTS

We include a presentation on methods for modelling and synthesising F0 contours compared to pitch modelling with the HMM-based speech synthesis system HTS under extremely limited data conditions first presented in [101]. Evaluation is done by comparing ten-fold cross validation squared errors on the small corpus described in Section 4.2.1. We show that the target-based methods are relatively effective at modelling and generating F0 contours in this context, achieving lower error rates than HTS modelling. While we conclude that pitch modelling based on qTA is promising in this context, it is clear from this work that in addition to height, pitch targets also need to contain gradient information (this has since been shown conclusively in Chapter 3).

D.1.1 Pitch contour generation

In this section we propose methods for contour modelling and synthesis. The proposed methods may be seen as belonging to two broad approaches; target-based methods (D.1.1.1) and methods based on extracting and using contours directly (template-based methods D.1.1.4). These methods are compared with a baseline HMM-based approach described in Section D.1.1.5.

D.1.1.1 Target-based methods

The target-based methods assume a pitch value or underlying target function associated with each syllable being responsible for the encoding of tone. Given this assumption, we can formulate the problem of modelling and synthesis of pitch contours as shown in Figure D.1.

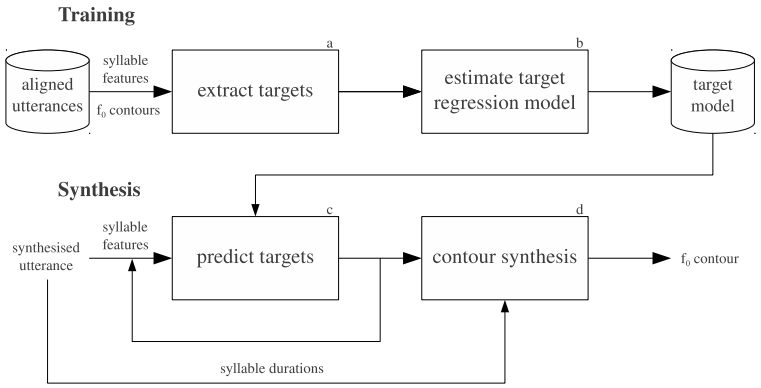

Figure D.1: *F0 model estimation and synthesis using target-based methods.*

During the training stage, utterances that have been annotated by a TTS system and phonetically aligned are used as input data. Target extraction or estimation is performed by taking syllables (any features of the syllable that are available and may be relevant for target prediction) and F0 contours as input and outputting parameters of targets for each syllable. A regression model is then estimated to predict target parameters given syllable features. For synthesis, an utterance would be constructed by the TTS system, generating the features required by the target model (including syllable durations). Targets are predicted and a contour synthesis algorithm is used to convert the target specification into a complete F0 contour.

In Sections D.1.1.2 and D.1.1.3 we describe the specifics of two target-based methods evaluated in this paper.

D.1.1.2 Quantitative target approximation method

This method is based directly on the parallel encoding and target approximation model presented in [21], where targets in each syllable are represented by a linear function. We implemented the quantitative target approximation (qTA) model as described in [65], including the analysis-by-synthesis method for parameter extraction. In this model, each syllable is described by three parameters; target height (b), target slope (m) and articulatory effort or approximation rate (λ).

In this work we reduced the number of parameters per syllable describing the target function to one (the target height b). This is done by assuming the slope parameter m to be zero for all syllables (i.e. static or level targets consistent with the theoretical expectations of a register tone language) and pre-estimating the λ parameter and keeping it constant during training. Another value that needs to be

estimated is the *initial pitch*, this is the pitch level at the beginning of an utterance from whence pitch movements commence and is required during target extraction and contour synthesis (Figure D.1a and b)).

Syllable features employed in the target regression model were the *pre-previous syllable tone*, *previous syllable tone*, *current syllable tone*, whether the syllable is *first in the current breath group*, whether the syllable is *last in the current breath group* and the *previous target* height. These features were investigated in [90] and the *previous target* height was found to be particularly important.

The complete training and synthesis (testing) process for this method can thus be described as follows:

1. Estimate the *initial pitch* by taking the mean F0 in the first half of initial syllables in all training utterances.

2. Pre-estimate λ by performing extraction and resynthesis (Figure D.1a and d) of the training set for a range of fixed λ values and choosing the value with the lowest mean squared error (MSE).

3. Figure D.1a: Extract feature vectors and target heights for all syllables in the training set using the analysis-by-synthesis method described in [65].

4. Figure D.1b: Estimate a regression model for syllable target heights (*target model*). We used support vector machines (SVMs) as in [90].

5. Obtain syllable features and durations from TTS natural language processing (NLP) modules and duration models. For testing we used durations from phone alignments of test utterances directly.

6. Figure D.1c: Predict syllable targets given the *initial pitch* and *target model*. Predicted targets are fed back as input for prediction of subsequent syllable targets (*previous target* feature).

7. Figure D.1d: Synthesise contour as described in [65] with fixed λ and slope parameters, *initial pitch*, predicted target heights and syllable durations as inputs.

D.1.1.3 Point target method

This method represents a simplification of the qTA model described above. A single pitch target value is extracted from each syllable directly by obtaining the maximum, minimum and mean F0 values for H, L and M tones respectively. Values extracted in this way may be interpreted as approximations of the underlying targets (as in the qTA model) and would correspond theoretically, assuming level target functions, if the speaker employed sufficient articulatory effort (larger values of λ).

In this paper we extract and model only the pitch targets, discarding the time instants at which they occur. These extreme points in the case of H and L tones will generally be realised late in the syllable, in the syllable nucleus, and often at the syllable boundary (from observations in [21, 77, 67]), while M tones generally exhibit a level contour [77]. During synthesis, we use these observations heuristically to define time instants where predicted pitch targets should be realised. Points are placed at the end of syllables (only one point for H and L syllables) and an additional point at the same pitch at 1/3rd of the syllable duration for M (thus 2 points for M syllables). To produce a complete contour we linearly interpolate these points as suggested in [67] and apply smoothing. As in the previous method we need to determine an *initial pitch* value which is here only used during synthesis.

The complete process is then similar to the one described above for the qTA method (D.1.1.2), with the differences in pitch target extraction (Figure D.1a) and contour synthesis (Figure D.1d) processes as described.

D.1.1.4 Contour template method

The notion of tonal complexes [67] and the stable three-syllable contours described in [77] suggests another method for F0 synthesis based on such "contour templates".

In this work we implemented a simple method for synthesising F0 contours based on this idea. The process is as follows:

1. Extract all phrase-internal three-syllable contours, normalise lengths by re-sampling using cubic-spline interpolation and determine the mean contour for each unique tone sequence as described in [77].

2. Synthesise a "standard length" utterance contour by overlapping and adding the first and last

thirds of consecutive mean contours stored in the previous step. A linear transition function is used to decrease and increase the weights of the respective overlapping regions resulting in a smooth transition (see Figure D.2).

3. Adjust the durations of syllables as required by re-sampling using cubic-spline interpolation.

We evaluated two variations of this method; the first simply as described above and the second adjusting the level of subsequent three-syllable contour templates by a fixed offset so that the mean F0 of the overlapping regions match. To set the absolute position of the resulting contour (for both methods) we shifted it by a fixed offset to match the *initial pitch* estimated as done in Section D.1.1.2.

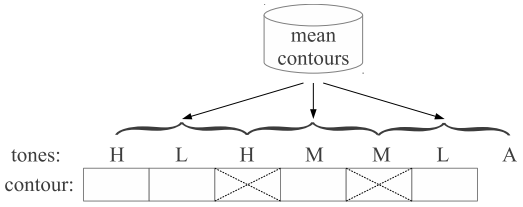

Figure D.2: *Example of the contour template synthesis process for a 6 syllable utterance. Diagonal lines illustrate the transition function applied and "A" refers to any syllable.*

D.1.1.5 HTS method

The standard method for modelling and generating F0 contours for HMM-based speech synthesis (HTS) employs multi-space probability distribution HMMs, that is HMMs that are able to handle discontinuities due to unvoiced regions. The single mixture HMM state densities represent log F0 and its first and second derivatives. HMMs are estimated using maximum likelihood with a large number of contextual features (specialised models) and states are tied using decision trees to deal with data sparsity. Contours are generated from these models using a parameter generation algorithm based on maximum likelihood and taking into account dynamic features and global variance of the static parameters. For the specifics of modelling and synthesis in HTS refer to [87].

In this paper we used the standard set of HTS (version 2.2) demonstration scripts[1] with all its associated default parameters and the HTS engine[2] (version 1.06). Two systems were tested:

1. A baseline system system without tonal information, only containing basic phone identity, syl-

[1] available at: http://hts.sp.nitech.ac.jp

[2] available at: http://hts-engine.sourceforge.net/

lable, word and phrase positional features.

2. A system including all of the features in the baseline system with the addition of *pre-previous syllable tone*, *previous syllable tone* and *current syllable tone* identities.

While this is by no means an exhaustive attempt to model F0 contours appropriately within the HTS framework, we consider the proposed systems a reasonable baseline in this context.

D.1.2 Experimental setup

We test the contour generation methods proposed on the manually verified speech corpus of 4 speakers described in Section 4.2.1. F0 contours were extracted as described in 3.2.2, however we did not in this instance interpolate F0 for unvoiced sections. The results in the following section are labelled according to the method and parameters used:

- `qta`: as described in Section D.1.1.2 with $\lambda \in \{20, 30, \ldots, 80\}$ and a coarse grid search of the SVM meta-parameters using 3-fold cross-validation in the training set ($C \in \{2^{-2,\ldots,4}\}$, $\gamma \in \{2^{-5,\ldots,1}\}$ and $\varepsilon \in \{0.01, 0.02, \ldots, 0.05\}$).
- `ppt`: as described in Section D.1.1.3 with SVM training as for `qta`.
- `ct1`: as described in Section D.1.1.4 without adjustment of contours.
- `ct2`: as described in Section D.1.1.4 with adjustment of contours.
- `htsbase`: HTS system without tone information.
- `htstone`: HTS system with tone information.

D.1.3 Results and discussion

To measure the effectiveness of the proposed methods we calculate the root mean squared errors (RMSE) and Pearson correlation coefficient between synthesised and actual F0 contours using ten-fold cross-validation. Experiments are repeated six times with different randomisations for splitting training and test sets. Results are presented in Figures D.3 and D.4.

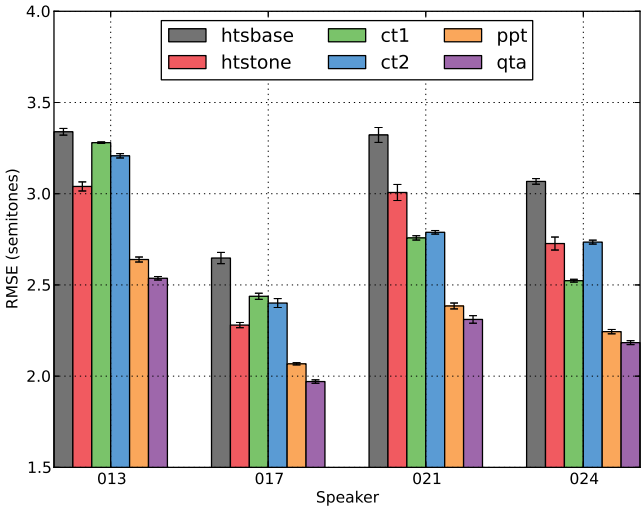

Figure D.3: *Mean RMSE values in semitones for each speaker and method for repeated cross-validation experiments. Error bars indicate the 95% confidence interval.*

It is clear that the HTS system lacking tone information results in the highest error rate and that by simply adding this information to this system as described, a significant improvement can be seen. The use of mean contours directly provide variable results, resulting in lower error rates than `htstone` in this context on our male speakers but higher for female speakers. This is understandable given the nature of the model and larger pitch ranges (variance) for female speakers. Further inspection of resulting utterance contours seemed to suggest that `ct1` underestimates overall declination while `ct2` fails to account for pitch resetting.

The target-based methods evaluated here consistently have lower error rates than the baseline tone-aware HTS system in this context. One possible reason is the fact that we have reduced the number of parameters per syllable to only one value – making models more robust given the small amount of training data. An interesting result here is the fact that `qta` has a lower RMSE and `ppt` a higher correlation. Inspection of contours suggest that `qta` generally results in faster pitch change to flat targets (minimising error) while `ppt` is probably more suitable in HL and LH contexts where rising and falling contours are seen. These observations motivate further work on these models; allowing more parameters per syllable, λ and dynamic targets in the case of `qta` and temporal locations of maxima in `ppt` and optimising these parameters more appropriately (see Figure D.5 for synthesis examples). It must be noted that the HTS-based contours in taking into account detailed segmental structure of syllables generates more appropriate local patterns which are ignored by the other proposed methods

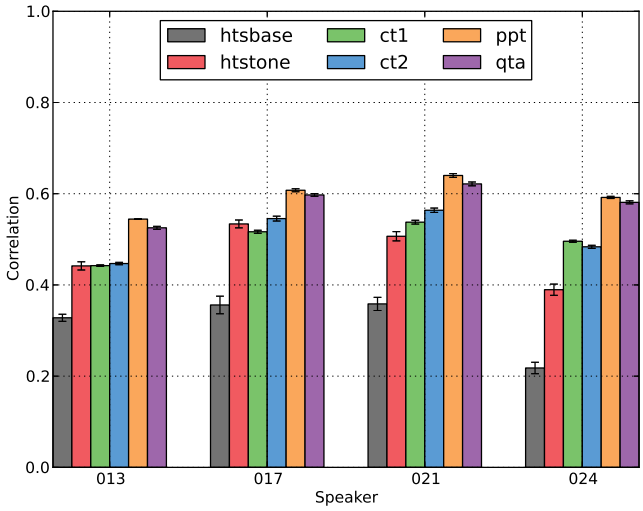

Figure D.4: *Mean correlation coefficients for each speaker and method for repeated cross-validation experiments. Error bars indicate the 95% confidence interval.*

and might be incorporated (along with voicing decisions) when building TTS systems.

D.1.4 Conclusions and future work

In this work we have proposed methods for modelling and generating utterance pitch contours in Yorùbá based on predicting syllable pitch targets [90]. The results presented here suggest that these methods will be useful for rapid development of TTS systems in this context; our work thus represents a step towards incorporating more accurate prosodic models for African tone languages.

While an HTS-based solution will certainly be more competitive with larger amounts of training data and further development effort with respect to synthesis for tone languages, the approaches followed here may also potentially allow more natural or efficient prosodic modification (e.g. to implement focus [21]) in systems built with larger amounts of data.

We have seen that significant gains in mean-squared error can be achieved with the methods that we have proposed. It would be interesting to see whether these gains can be improved further by more detailed modelling of the contours, and also to see whether these gains translate into perceptual improvements in the quality of synthesised utterances. We intend to investigate both of these matters in the near future.

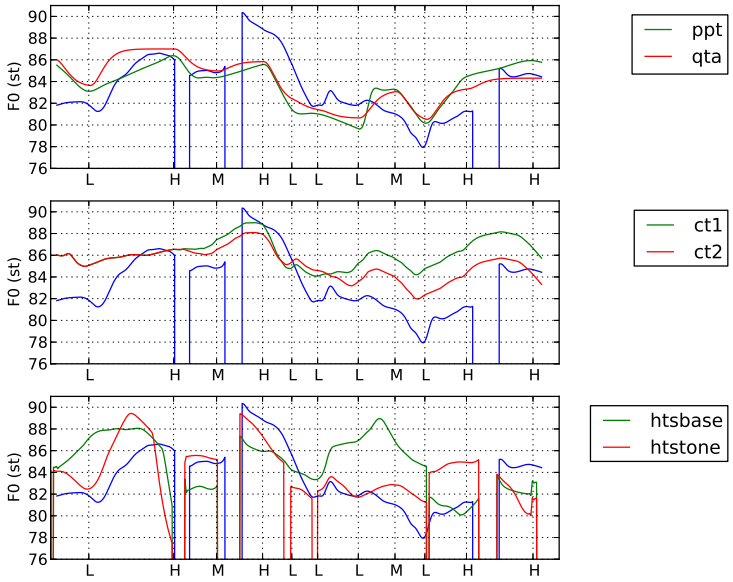

Figure D.5: *An example of synthesised contours for a specific utterance. Blue contours represent the reference extracted from the original speech sample. Grid lines indicate syllable boundaries with tones indicated on the x-axis.*