ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULTY OF TECHNOLOGY

ELECTRICAL and COMPUTER ENGINEERING

DEPARTMENT

Prosodic Modeling for Amharic

By

Gebremichael Girmay

A thesis submitted to the school of Graduate studies of Addis Ababa University in partial fulfillment of the requirements for the degree of Masters of Science in Electrical and Computer Engineering

(Computer Engineering)

April, 2008

Addis Ababa, Ethiopia

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
FACULTY OF TECHNOLOGY
DEPARTMENT OF ELECTRICAL and COMPUTER ENGINEERING

PROSODIC MODELLING FOR AMHARIC

By

Gebremichael Girmay

Advisors
Dr. Manoj V.N.V
and
Mr. Molalgne Girmaw

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULTY OF TECHNOLOGY

PROSODIC MODELLING FOR AMHARIC

By

Gebremichael Girmay

APPROVAL BY BOARD OF EXAMINERS

_____	_____
Chairman Dept. of Graduate Committee	Signature
_____	_____
Advisor	Signature
_____	_____
Advisor	Signature
_____	_____
Internal Examiner	Signature
_____	_____
External Examiner	Signature

Declaration

I, the undersigned student, declare that this thesis work is my original work, has not been presented for a degree in this or any other universities, and all sources of materials used for the thesis work have been fully acknowledged.

Name: Gebremichael Girmay Signature: _____

Place: Addis Ababa

Date of submission: April, 2008

This thesis has been submitted for examination with my approval as university advisors.

_____ Signature: _____

Advisor

_____ Signature: _____

Advisor

ACKNOWLEDGEMENTS

I am deeply indebted to my advisor Mr. Molalgne Girmaw for his invaluable help and his caring attitude throughout my thesis project. He was a constant source of encouragement and hope in all the ups and downs of my project. In fact these few words of thanks are insufficient to express gratefulness to my advisor. His motivation and encouragement have always guided me through out my whole project process. I greatly, thanks to Dr. Manoj V.N.V for his willingness to become my advisor and for his kindness and readiness for any help I request.

I would also like to give great thanks to the department chairmen, for handling me in kindness. And my appreciation goes to teaching and other staff, for their totality effort.

I would like to especially thanks to those people who were giving me their sound and who participate in the test stage.

May God help me to repay the all type of support; I have had from my cousin w/r Aregay Tebeje, indeed I thanking you very much.

Table of Contents

List Of Abbreviations

AmhToBI	Amharic Tone and Break Indices
AM	Autosegmental Metrical theory of prosody.
ANN	Artificial Neural Networks
ASR	Automatic Speech Recognition
CART	Classification and Regression Tree
CV	Consonant Vowel
DRT	Diagnostic Rhyme Test
DSP	Digital Signal Processing
F0	Fundamental frequency
GAE	General American English
H	High tone
HMM	Hidden Markov Model
INTSINT	INternational Transcription System for INTonation
IP	Intonational Phrase
iP	Intermediate Intonational Phrase
L	Low tone
MART	Modified Rhyme Test
MOS	Mean Opinion Score
NLP	Natural Signal Processing
NP	Noun Phrase
POS	Part Of Speech
PRAAT	Speech analysis software tool
PROSPA	Prosodic Transcribing system
TILT	Prosodic Transcribing system
S.D	Standard Deviation
ToBI	Tones and Breaks Indices (a framework of describing prosody)
TTS	Text to Speech Synthesis System
VP	Verb Phrase
VODER	Voice Operating Demonstrator
WaveSurfer	Speech analysis software tool

List of Figures

List of Tables

ABSTRACT

During daily speech communications, we systematically employ variations in pitch, loudness, phrasing and duration of spoken segments to express our intentions, attitudes and assumptions. These cues are continuously picked by our listeners and our intentions are communicated. This phenomenon is what is termed as prosody. Various languages have been modeled using a prosodic modeling framework and have been used to varying degrees of success in applications like speech synthesis systems. The prosodic features of speech are largely dependent on the language, thus it is imperative that prosodic models be developed and tested for each language to apply them in (language specific) speech synthesis applications. In this study the prosodic nature of Amharic is investigated, and based on the investigation the prosodic cues of Amharic are identified and an Amharic prosodic model based on the ToBI framework is developed. This hypothesized model is then be subjected to a rigorous test to determine whether it results in improvement of Amharic synthetic speech in terms of naturalness and intelligibility.

INTRODUCTION

This chapter gives a brief description of the contents and the structure of this thesis as well as a discussion about problems associated with prosody modeling. The basic problem motivating this study was the lack of prosodic models for Amharic. Such models are necessary in many respects: first of all, they are an essential part of high quality speech synthesis systems and secondly, they provide a framework for the description of the phenomena that prosody comprises.

Prosody is the science of how speech is uttered as contrasted to what is spoken. The general problems with prosody modeling lie in the gray area between the discrete, symbolic representation of speech and its actual manifestation as a continuously varying signal. Basically, one needs to develop a methodology to associate a set of linguistic, paralinguistic and emotional instructions or representations with the prosodic parameters of synthetic or natural speech. The solution to the above problems presented in this thesis is based on the ToBI transcription system of conventions.

Perhaps the most important application of prosody is its importance in Text to Speech Synthesis (TTS) systems. This was, in fact, the main motivation behind the inception of this thesis. The lack of prosodic model has been a very significant bottleneck in the development of Amharic Text to speech Synthesis systems. Thus it is instructive to start with an overall description of TTS systems and the important role prosody plays in such systems.

Overview of Text-To-Speech Synthesis (TTS) Systems

A Text-To-Speech (TTS) synthesizer is a computer-based system that reads a given text aloud, or we can define Text-To-Speech as the automatic production of speech. There are several kinds of speech synthesizers that use various methods to convert textual input into sound output. Synthesizers are most commonly distinguished as being either rule-based or data-driven.

Rule-based synthesizers base their output on an acoustic model of speech production. This method can create a vowel by modeling how the signal is led through a filter, which mimics the human vocal tract, creating a set of particular resonance (formant) frequencies depending on the relative position of the different articulatory organs. Generalized rules are extracted from the filtered information. The phonetic input of the synthesis is tested on the rules; when a match is found, the synthesizer produces digital speech. Rule-based synthesizers are also referred to as formant synthesizers, because the speech information behind the generalized rules is related to formant and anti-formant frequencies and bandwidths.

As opposed to data-driven synthesizers, this method does not use human speech samples in order to create the output speech. These types of synthesizers are considered to have a greater degree of intelligibility, as well as to be resource-economic, since they do not depend on stored segments of human speech.

Data-driven or concatenative speech synthesis is a method for creating artificial speech by merging slices of pre-recorded human speech. Two subcategories can be distinguished within this method, diphone and unit-selection, differing mainly in the size of the units being concatenated. Both methods store the pre-recorded speech units in a database, from which the concatenation originates. Parts of utterances that have not been previously processed and stored in the database are "built up from smaller units." The diphone data-driven synthesis refers to storage of all possible diphones in a particular language, which are then merged in accordance with the phonetization of the input text. A diphone is a speech unit consisting of two half-phonemes, or of the phonetic transition in between.

The other method, unit selection, collects its speech data from a repository containing units of various lengths, including diphones as well as words and phrases.

Each pre-recording is stored in multiple occurrences, pronounced in different prosodic contexts. Hence, this type of synthesis requires an extensive storage facility, and has only recently become a popular method, since memories and performance of computers have

increased. The primary motivation for a large database is that with a large number of units available with varied prosodic and spectral characteristics it should be possible to synthesize more natural sounding speech than that can be produced with a small set of controlled units.

Since the diphone unit requires digital post-processing in order to incorporate prosodic information, the naturalness of the pre-recordings may be reduced. The unit selection synthesis, however, requires no digital post-processing; it simply concatenates the stored units as they are. Therefore, this type of synthesis usually gives the greatest naturalness.

Historical Background

Artificial speech has been a dream of the humankind for centuries. To understand how the present systems work and how they have developed to their present form, a historical review may be useful. The history of synthesized speech tells us that the effort of speech synthesis started from the first mechanical efforts to systems that form the basis for today's high-quality synthesizers. The earliest efforts to produce synthetic speech were made over two hundred years ago. In St. Petersburg 1779, a Russian Professor Christian Kratzenstein explained physiological differences between five long vowels (/a/, /e/, /i/, /o/, and /u/) and made apparatus to produce them artificially. He constructed acoustic resonators similar to the human vocal tract and activated the resonators with vibrating reeds like in music instruments.

A few years later, in Vienna 1791, Wolfgang von Kempelen introduced his "Acoustic-Mechanical Speech Machine", which was able to produce single sounds and some sound combinations.

In about mid 1800's Charles Wheatstone constructed his famous version of von Kempelen's speaking machine. It was a bit more complicated and was capable to produce vowels and most of the consonant sounds. Some sound combinations and even full words were also possible to produce.

The research and experiments with mechanical and semi-electrical analogs of vocal system were continued until the 1960's, but with no remarkable success. Stewart introduced the first full electrical synthesis device in 1922. The machine was able to

generate single static vowel sounds with two lowest formants, but not any of the consonants nor connected utterances.

The first device to be considered as a speech synthesizer was the VODER (Voice Operating Demonstrator) introduced by Homer Dudley at New York's World Fair in 1939. The speech quality and intelligibility were far from good but the potential for producing artificial speech were well demonstrated.

After the demonstration of the VODER, the scientific world became more and more interested in speech synthesis. It was finally shown that intelligible speech could be produced artificially.

The first full text-to-speech system for English was developed in the Electro technical Laboratory, Japan 1968 by Noriko Umeda and his companions. The speech was quite intelligible but monotonous and far from the quality of present systems. In late 1970's and early 1980's, considerably amount of commercial text-to-speech and speech synthesis products were introduced.

Modern speech synthesis technologies involve quite complicated and sophisticated methods and algorithms. One of the methods applied recently in speech synthesis is the Hidden Markov Models (HMM). HMMs have been applied to speech recognition from late 1970's. Neural networks have been applied in speech synthesis and results have been quite promising. However, the potential of using neural networks have not been sufficiently explored.

General Architecture of TTS Systems

The process of converting written text into speech can be considered into a two step process with two independent but interacting modules:

- Natural Language Processing (NLP) Module and
- Digital Signal Processing (DSP) Module.

The natural language process deals with extracting and organizing all possible information from the written text (and possibly additional extra help) that would aid the digital signal processing module generate the sound signals. The NLP module produces files with a phonetic transcription of the text, together with the desired intonation and rhythm. These serve as instructions for the digital signal processing module to actually generate the physical speech signals.

The DSP transforms the symbolic information it receives from the NLP module into speech. The process of converting text to speech requires a set of rules that translates

each grapheme of the text input into phonemes [4], [5]. This two step process is depicted graphically in Figure (1.1).

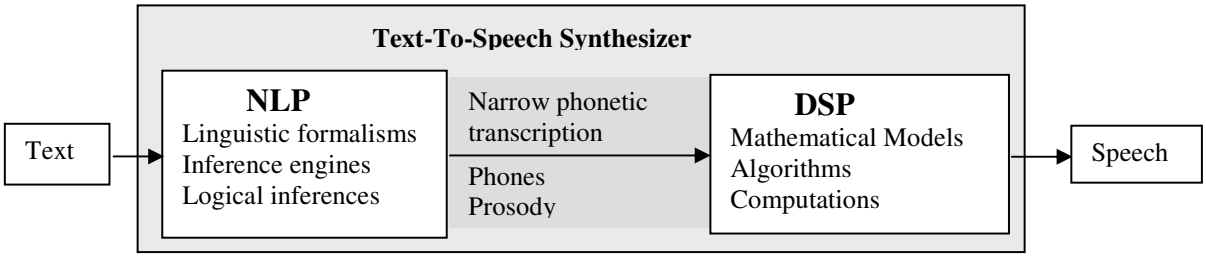


Figure 3.1 The two step process of speech synthesis [4]

The NLP module consists of the following main processing stages or components:

- Text Analysis
- Automatic Phonetization
- Prosody Generation

In the DSP module a transformation of the received symbolic information into actual speech waves is performed. There are three different categories of waveform generation (called types of speech synthesis):

- *Articulatory Synthesis:* a method of synthesizing speech by controlling the speech articulators and determines the characteristics of the vocal tract filter by means of a description of the vocal tract geometry,
- *Formant Synthesis:* specifies directly the formant frequencies and bandwidths as well as the source parameters. Formant Synthesizers are also referred to as Rule-Based Synthesizers, where generalized rules are extracted from the filtered information. The input is then tested on the rules.
- *Concatenative Synthesis:* This is the most commonly used technique, where segments of speech are tied together to form a complete speech chain. The speech output is produced by coupling segments from a database to create the sequence of segments.

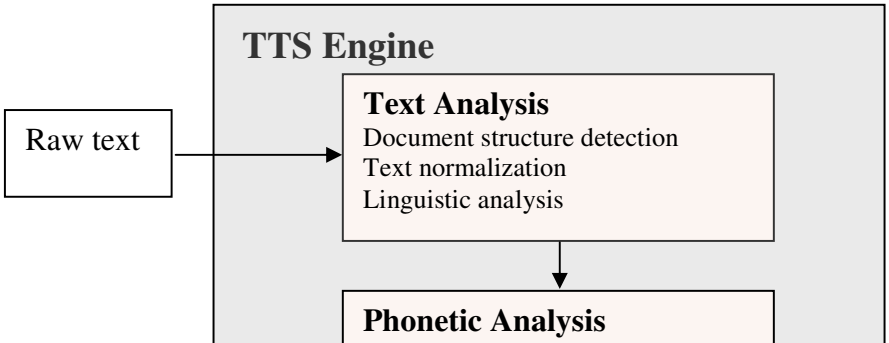


Figure 1.4 A more detailed architectural representation of speech synthesis [5]

The Text Analysis module/component normalizes the text to the appropriate form so that it becomes speakable. The input can be either raw or tagged text. These tags can be used to assist text, phonetic, and prosodic analysis. In text analysis, document structure detection, text normalization, and linguistic analysis are performed.

- *Document Structure Detection*: Provides context for all later processes, allows for prosodic classification. It separates the text into its document structures elements through sentence breaking and paragraph segmentation, allowing the TTS system to identify when to insert gaps during narration and may have direct implications for prosody. It also serves as the basis for syntactic and semantic analysis. The elements in a document structure are chapter and section headers, paragraphs, sentences and lists. These and other linguistic units are important in regulating prosody.
- *Text normalization:* Transformation of tokens (words, symbols) in a text, such as abbreviations, acronyms, idioms, numbers, mathematical formulas, currency, etc. into full text (common orthographic transcription) suitable for subsequent phonetic conversion.
- *Linguistic Analysis:* refers to syntactic and semantic parsing. It produces structural and semantic information about sentences. The fundamental types of information extracted in linguistic analysis module are:
  - Word part of speech – noun, verb, adjective etc.
  - Phrasal cohesion of words – idioms, clauses, phrases, sentences.

- Syntactic identification – questions, commands, quotes.
- Semantic focus – emphasis.
- Semantic type – requesting, informing, narrating

● *Phonetic Analysis Module:* it converts lexical orthographic symbols (the processed text) into phonemic representations (phonetic sequence), along with possible stress markers. Even though it is often referred to as grapheme-to-phoneme conversion, it is generally believed to have three sub modules: Homographic Disambiguation, Morphological Analysis, and Grapheme-to-Phoneme conversion.

- Grapheme-to-Phoneme (Letter-to-sound) Conversion: This step generates the phonetic representation of the text in question. Graphemes are any of a set of written symbols, letter, or combinations of letters that represent the same sound, for example, f in "fat," ph in "photo," and gh in "tough". Phonemes are the minimal units of speech sound in a language. Thus f, ph and gh in the above example all are represented by a single phoneme. In general, phoneme is the smallest unit of speech that can be used to make a word different from another that is same in every other way. E.g. the 'b' in 'big' and the 'p' in 'pig' represent two different phonemes. Phonemes are representation of sounds that are made in a language. They completely represent all the sounds made in a language and are limited in number for a given language. E.g. there are 42 or a little more phonemes in the English language. For example the word quick can be represented by the phonetic equivalent of /k w ih k/.
- Morphological Analysis: in this module words are decomposed into their morpheme constituents so that important cues are obtained to the pronunciation for inflected & derivational words. A morpheme is the smallest meaningful unit in a language, consisting of a word or part of a word, that cannot be divided without losing its meaning. These include prefixes, suffixes and stem words. E.g. 'Gun' is one morpheme; 'Gun-s' contains two morphemes; 'Gun-fight-er' contains three morphemes
- Homographic Disambiguation: the main purpose of this module is to identify which one of the possible pronunciations should be used in the context. Homograph refers to a word that has the same spelling as another, but is different

in meaning, origin, grammar, or pronunciation. E.g. the noun 'record' (reka:d) and the verb 'record'(rika:d) are homographs (of each other).

- Prosodic Analysis Module: - The prosodic analysis component attaches appropriate pitch and duration and other prosodic information to the phonetic sequence.

Amharic TTS Systems

Researchers had paid little attention to the study of Ethiopian linguistic technology until recent years. Few initial attempts as MSc thesis work on TTS have been done during the last five to ten years. These attempts, although encouraging, have not been generally followed up and used as a standard for educational purposes by universities nor applied in industry except, perhaps for a few telecommunication applications. Almost no attempt has been made to develop multilingual TTS systems.

As pointed out in [15] several computer fonts have been developed for the Amharic script, but for many years the languages had no standard computer representation. An international standard for the script was agreed on only in 1998 and later incorporated into the Unicode standard, but nationally there are still about 30 different standards regarding Amharic scripts. The thesis works done (on Amharic TTS) until now can be taken as initial simple prototypes, which with their limitation can be taken as positive starting tracks. These developed TTS systems are a good start to producing realistic speech from text, but there are several areas which can be improved up on.

Scope and Limitations of Previous Studies

In most of the papers, the following limitations and scope can be observed:

- A limited number of diphone speech data was stored in the diphone database. Thus it is an important issue to apply more experiment on the corpus data and hence to complete the diphone database. Here it is required to develop an intelligent algorithm that is able to separate a speech corpus data into diphone units automatically.
- Minimizing noise during recording (Selecting sound library), careful selection of corpus data, careful extraction of the required diphones and a speaker are most important.
- Text input could only extend up to paragraph level. A TTS system for discourse

type has not yet been implemented.

- The designed systems do not properly deal with numbers and abbreviations: the incorporation of number converter and word normalization are important.
- Sounds that may be acoustically different in different places of a word has not yet been thoroughly analyzed - this is important for speech quality.
- In some papers, the words within the sentences are assumed to be separated only by blanks…. Punctuation marks like slash, semicolons were not considered and these words could contain abbreviations, acronyms and the like.
- Dialect issues of the same language in different places were not addressed (a variety of language, spoken in one part of a country, which is different in some words or grammar from other forms of the same language.)
- Handling special characteristics and morphographemic structure of Amharic alphabets (like che(ቼ), gne(ኘ), lwa(ኗ), etc)
- Further attention is required in handling prosodic and intonation conditions
- Multilingual systems have not been attempted
- Naturalness (the sound quality and naturalness still remain a major problem)

Goals & Objectives of this Thesis

Goals of this thesis

The general goal of this research is the development of a prosodic model for Amharic. Amharic is the national language of Ethiopia and various attempts have been made to develop an automatic speech synthesis system for the language. However, the prosodic aspects of Amharic have been under-researched and no unified model for describing the prosodical features of Amharic exist. With the development of such a model it is expected that Amharic speech synthesizer programs can produce more intelligible and understandable synthetic speech. Furthermore, the prosodic models that are developed in this thesis can help speech researchers better understand the processes of Amharic speech production and perception.

The main outcome of the research is **AmhToBI**, which is a symbolic collection of pitch movements (tones), durations and pauses that apply specifically to Amharic. These collection of prosodic cues can be applied to any written Amharic text as an input to

speech synthesizers and are expected to help the synthesizer produce a more natural and intelligible synthetic speech.

Objectives of this thesis

The specific objectives of the research in studying and modeling the prosodic features of Amharic are:

- studying the prosodic features of a unit of speech, at a word, phrase, clause, or sentence level
- studying the intonational characteristics, syllable duration, accent, tone, and stress or pitch contour (F0)
- modeling the prosodic characteristics using a high-level symbolic prosody description system known as ToBI, based on the knowledge of category or structure of intonation, measurements of acoustic features (fundamental frequency (f0) for pitch, duration, and intensity for amplitude). In an effort of this work, the intonational phonology and ToBI (Tones and Break Indices, a transcription system of intonation and phrasing) model of Amharic will be proposed, this model is to be referred to as AmhToBI.
- Using the developed model to attach prosodic information to Amharic utterance segments, which will then be used to aid the synthesis of Amharic speech and then be subjected to a rigorous test to determine whether it results in improvement of Amharic synthetic speech in terms of naturalness and intelligibility.

Organization of the Thesis

Chapter1 presented an introduction into the problem, the motivation behind it, its role in speech synthesis and the objectives of the project work.

Chapter 2 presents a literature review of some of the fundamental concepts of prosody as well as the different methods used in prosodic modeling.

Chapter 3 presents the methodological concerns relevant for this thesis work. Specifically the data collection, data analysis, model selection and evaluation methods are discussed in this chapter.

Chapter 4 describes the experiments that were conducted and the results obtained from these experiments. This chapter describes the main work performed in this thesis to

come up with the Amharic prosodic model. It also indicates, as a result of the experiments, the model that is proposed in this thesis work.

Chapter 5 describes the experiments performed to evaluate whether or not the proposed model results in improvement in intelligibility and naturalness of synthesized Amharic speech. The results of the evaluation experiment is indicated and analysis is given on the results.

Chapter 6 gives concluding remarks summarizing the findings of the research, the limitations, the way forward and some of the possible applications of the research work.

CHAPTER 2

LITERATURE REVIEW

In this chapter, a background review of what prosody is and what its constituents are will be made. In addition, concepts like intonation, stress, accent, prominence, pitch, pitch range, tone, prosodic phrases, timing, rhythm and tune will be introduced and their role in prosody described in detail. It will also focus on the description of a formal system for studying the prosodic structure of languages viz. Symbolic Prosody.

Symbolic prosody concerns itself with representation of the prosody of a spoken language using symbols to represent the elements of prosody, i.e. relative pitch, phone duration, loudness and pause insertion. Symbolic prosody deals with

- Breaking sentences into what are known as prosodic phrases, possibly separated by pauses and
- Assigning labels, such as emphasis, to different syllables or words within each prosodic phrases

Each language uses the basic elements of prosodic cues in its own way. Thus, symbolic prosody models are clearly language specific and it makes it necessary to study the prosodic structure of the language in question to identify what sorts of prosodic components manifest and their particular roles as prosodic cues. The result of this study must then be used to construct a symbolic system of prosodic modeling. This symbolic prosody model can then be applied for labeling text and uttered speech. The system presents a unified model that can be used in the aid of natural sounding automatic speech synthesis or as an analysis tool for linguistic researchers. This chapter then proceeds to describe the framework through which symbolic prosody models are constructed. It, thus, sets the theoretical background required for the development of prosodic models in general and specifically as applied to Amharic. The models developed for Amharic, which are described in later chapters, will be based on this framework.

What is Prosody?

Simply put, prosody refers to how speech is uttered. It can refer both to the melody of the speech as well as to the science that is dedicated to the study of how speech is uttered. Prosody, by its very nature, is concerned with features of speech that manifest over large

units of speech. Prosody doesn't concern itself with realizations of single phonetic elements but rather larger segments of speech including words, phrases, sentences or even larger utterances. For that reason, it is usually referred to as a supra-segmental feature of speech or simply supra-segmentals.

Prosody structures the flow of speech and is manifested as accentuation, loudness, relative duration and insertion of breaks between utterance segments. These manifestations are referred to as the melodic features of spoken language. Listeners depend on prosody to fully understand speech in addition to the juxtaposition of the words making up the speech utterance. Speech utterance without the systematic variations of the melodic features is known as segmental speech.

For example, the melody of speech will indicate to the listener: whether a question is being asked or a statement is being made: whether the speaker is happy, sad, or angry etc. The term prosody as pointed out in [1], [4], [5], is used to refer to aspects of sentence's pronunciation.

From the listener's point of view, prosody consists of the systematic perception and recovery of a speaker's intention based on

- *Pauses:* to indicate phrases and avoid running out of air
- *Pitch:* rate of vocal fold cycling
- *Rate/relative duration:* phoneme durations and timings
- *Loudness:* relative volume of speech parts.

In the general case the relevant factors and the behavior of the prosodic parameters range from the simplest, phonetically determined variation on the segmental level to the linguistically determined variation on the level of the utterance.

Prosodic events can be studied at various levels of representation including:

- *Acoustic level:* the acoustic manifestation of prosody (fundamental frequency (F0), amplitude, and duration).

- *Perceptual level:* represents the prosodic events as captured by listeners.
- *Linguistic level:* represents the prosody of an utterance as a sequence of linguistic units (phoneme, syllable, phrase, etc).

Acoustic	Perceptual	Linguistic
F0	Pitch	Tone, intonation, aspect of stress
Amplitude, energy, intensity	Loudness	Aspect of stress
Duration	Length	Aspect of stress
Amplitude dynamics	Strength	Aspect of stress

Table 2.4 Levels of Representation of Prosodic Phenomena

The variations in the fundamental frequency, amplitude and duration are perceived by listeners as the pitch, loudness and length of the speech. These variations depending on the language in question will have linguistic importance and the listeners can extract meanings from these variations. A good example will be the sentence "She ate the food". This sentence can be spoken as a simple declarative statement, as an exclamation (indicating surprise), or as a question. Although the sequence of words that are uttered is the same for all three of the aforementioned sentences, there are obvious differences in the meanings of the three realizations of the same sentence. This change in meaning is manifested through the variation of the four important components of prosody, namely pitch, loudness, duration and pauses.

The Role of Prosody in Text-To-Speech (TTS) Systems

The essence of text-to-speech synthesis is to automatically convert symbols into signals using a computer program. As the previous sections describe there are quite a large number of steps and interacting modules before the actual conversion is successful. The success or failure of the TTS system depends both on the quality of concept of each of the modules and on the quality of the data used to develop the rules in each module. The prosodic analysis module, being a part of the interacting TTS modules, plays a significant role in the synthetic speech production process. From an end user's perspective a synthetic speech with proper prosodic features will sound more natural and is generally

easier to understand. Thus, it might be stated that the main problem of adding naturalness and intelligibility to the systems can largely be solved through the incorporation of better prosodic models [8].

The example of the previous sentence makes it clear that a decent speech synthesis system needs to have a level of understanding of prosody, otherwise, the output will not distinguish between questions, exclamations and declarations and a similar other set of prosodic variations. Furthermore, the absence of the systematic variation of the prosodic cues will make the uttered speech monotonous, unnatural and hard to understand, even if there are no ambiguities that are similar in nature to the example sentence mentioned previously.

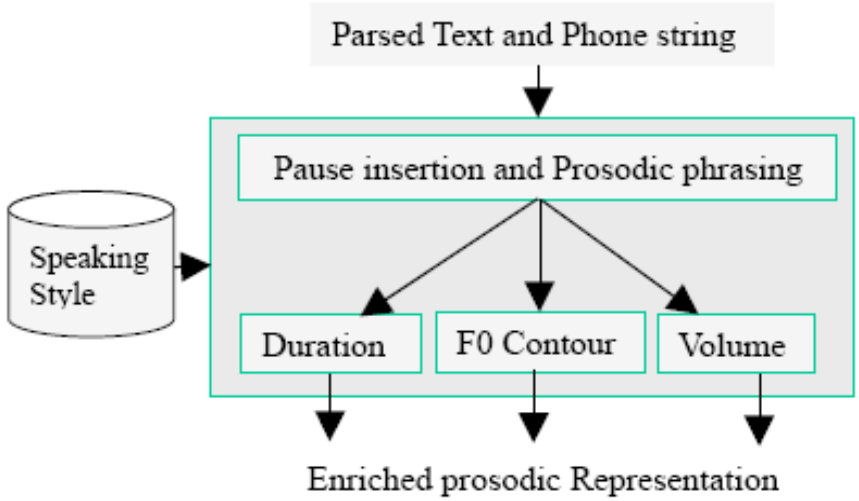


Fig 2.1: Architecture of a Prosodic Modeling System

Figure 2.1 presents the architecture of the prosody component. The input to the prosody module is normally parsed text with a phoneme string. The output specifies the duration of each phoneme and the pitch contour thus it generates an enriched prosodic representation for the speech to be synthesized. Depending on the speaking style different prosodic representations can be obtained. Speaking style can be for example, *character* (extra-linguistic properties of a speaker) and *emotion* (temporary emotional conditions such as joy, anger etc.). This shows that prosody depends not only on the linguistic

content of a sentence but also on extra-linguistic properties of the speaker and environmental effects.

The Autosegmental-Metrical (AM) Theory of Prosody

One of the most popular theories for prosodic analysis is the Autosegmental-Metrical(AM) model of intonational phonology. In this model prosody is described in terms of a prosodic structure and distinctive tonal categories (prominence relations) [14]. A prosodic structure is a hierarchical organization of prosodic units from the smallest prosodic unit to the largest. Prominence relations describe the relative prominence of the units of prosody within an uttered speech segment. Prominence can happen at the word or syllabic level. Within a phrase, some words are more prominent than others; and within a word, some syllables are more prominent than others. This prominence relation contributes to the perception of the uttered speech as having a certain connotational meaning. A prosodic structure and prominence relations of an uttered speech are realized by features such as variations in pitch, duration, and/or loudness as well as the realization of consonants and vowels.

In addition, the prosodic property of an utterance is a combination of prosody at the word level and prosody at the phrase level. The word level prosody is termed as lexical prosody and the phrase level prosody is termed as post-lexical prosody. The post-lexical prosody is constrained by the lexical prosody, and it also contains information about the lexical prosody.

The AM theory uses pitch accents and boundary tones as the fundamental components of prosodic modeling; hence it is instructive to continue with a brief discussion of what is meant by pitch accents and boundary tones.

Pitch accents and boundary tones

A fundamental notion within AM theory is that the variation of pitch on speech utterance segments which is referred to as *surface pitch contour* spread across an utterance arises from a linear sequence of *pitch events*. In some languages the variation of pitch changes the

lexical meaning of words and these languages are referred to as *tonal languages*. In tonal languages, the pitch contour will be specified both by lexical events as well as post-lexical events. But in intonational languages, like English, the variation in pitch are expected to be of two kinds only:

- *Pitch accents:* pitch movements which are manifested on stressed syllables; and,
- *Edge tones:* pitch movements which are manifested near the end of a phrase

Intonation has been defined as

> *"the use of suprasegmental phonetic features to convey post-lexical or sentence-level pragmatic meanings in a linguistically-structured way" [13].*

By this definition intonation is observed

> *"in all types of languages across the continuum of variation between an archetypal 'tone' language and an archetypal 'stress' language. In a tone language, such as Mandarin or Thai, pitch is part of the lexical specification of some if not all morphemes" [4].*

The main reason for choosing to analyze Amharic intonation within the context of autosegmental-metrical (AM) theory of intonation is its ability to capture the notion of *the unity of pitch phonology* [13]: A theory which can be used to analyze any language, regardless of whether pitch functions lexically and/or post-lexically. Thus the AM theory can be used to describe Amharic as good as it describes English, Korean or Swahili.

In AM theory all pitch events are defined using one of two level tones - either high (**H**) or low (**L**), with **H** being near the top of a speaker's pitch range, and **L** near the bottom.

The motivation in this project is that Amharic is assumed to be a purely stressed language, in which tone (variations in the pitch of uttered speech) is used exclusively post-lexically. In other words, it is assumed that in Amharic, changes in pitch do not signify lexical changes of meaning.

Transcription of intonation

AM theory provides a system for the transcription of intonational contours. Intonational tunes are seen as tonal sequences combining pitch accents and edge tones of various kinds; with all pitch events transcribed using **H** and **L** only. A set of standard notational devices used in AM theory is listed in [13].

Symbol	Meaning
H	high target
L	low target
*	pitch accent (associated with the main stressed syllable of some words)
-	phrase tone (associated with a phrase edge)
%	boundary tone (associated with a phrase edge)
!	Down step

Table 5.2 AM Symbols for representing pitch movements

Pitch accents may be composed of one target (monotonal) or at most two (bitonal) resulting in the set of possible pitch accents listed in the table below [4]. The star notation, when it appears in bitonal targes, indicates which of the two tones in a bitonal accent is associated primarily with the stressed syllable.

Stress	Type
H*	monotonal
L*	monotonal
H*+L	bitonal
H+L*	bitonal
L*+H	bitonal
L+H*	bitonal

Table 2.3 Possible Stress Combinations and Their Symbols

As regards, edge tones, in early AM work on English, notation was proposed for two types, showing affiliation to the edge of prosodic phrases at different levels [4]: boundary tones [**H%, L%**] align to the edge of a full prosodic phrase called intonational phrase (IP), and phrase tones [**H-, L-**] align to the edge of an intermediate phrase (iP), nested within the larger phrase. Since an IP is composed of one or more iPs, the right edge of an utterance was argued always to bear a sequence of a phrase tone and a boundary tone (the right edge of both iP and IP coincide at the right edge of the utterance [13].

The remaining symbol '**!**' is used to denote "*down stepping*" which refers to phonological lowering of the F0 target level of an **H** tone. In [7], down step was argued to be triggered phonologically by any bitonal pitch accent, whilst other authors have argued that down step is better analyzed as 'an independent linguistic choice' under the control of the speaker [13].

An influential AM theory transcription system for intonation is the Tones and Break Indices (ToBI) system, which was developed for General American English (GAE). English ToBI can be used to describe the intonation of GAE and possibly other dialects of English (British English). However, it cannot be directly used to describe the intonation of other languages since it is the result of a *phonological* analysis of a particular language, rather than a phonetic transcription system. The theoretical choices underpinning ToBI

have, however, been successfully adapted to other languages and as a result AM-style phonological models exist for many languages, using a similar notation system [4].

This thesis implements an AM-style transcription system for Amharic intonation in which the tonal sequence is analyzed using symbolic labels to represent pitch accents and boundary tones, and the correspondence between the tones and prosodic boundaries of different strengths is motivated in the text.

Pauses

In a long sentence, speakers naturally pause a number of times. These pauses have traditionally been thought to correlate with syntactic structure but might more properly be thought of as markers of information structure [4]. In a typical system, the most reliable indicator of pause location is punctuation. After resolution of abbreviations and special symbols relevant to text normalization, the remaining punctuation can be reclassified as essentially prosodic in nature. These include periods, commas, exclamation points, parentheses, ellipsis points, colons, dashes or their equivalent or similar symbols in languages other than English. Each of these can be taken to correspond to a prosodic phrase boundary and can be given a special pitch movement at its end-point.

In predicting pauses, although their occurrence and their duration have to be considered, the simple presence or absence of a silence (of greater than 30 ms) is the most significant decision [4]. The exact duration is only secondary, based partially on the current speaking rate and other factors.

There are many reasonable places to pause in a long sentence, but a few where it is critical not to pause. The goal is to make sure that the selection of the point where pause is inserted is not a place where ambiguity, misinterpretation or complete breakdown of understanding occurs [4].

Prosodic Phrases

Phrases, as speech processing and specially prosody is concerned, are places within speech that are signaled by silence, whether or not these are a result of semantic phrasing is of little concern to the speech processing system, although semantic parsing might help the system identify where these silences are to be inserted. Places where silence is inserted are referred to by the general name of *prosodic junctures*. Junctures that are clearly signaled by the presence of silence and usually by a characteristic pitch movement as well are called *intonational phrases* or IP for short. These junctures are required between utterances and usually at punctuation boundaries. Prosodic junctures that are not signaled by silence but rather by characteristic pitch movement only are called *phonological phrases* or *intermediate phrases* or iP for short.

The breaks or pauses that appear between utterance segments are not of equal durations. Some of the pauses are longer to indicate that the degree of juncture between the utterance segments either side of the pause is larger and some are shorter to indicate just the opposite. The break indices part of ToBI specifies an inventory of numbers expressing the strength of the prosodic juncture. For English, the prosodic association of words in an utterance is shown by labeling the end of each word for the subjective strength of its association with the next word on a scale from 0 (the strongest perceived conjoining) to 4 (most disjoint) [4]. Other languages also use a similar type of inventory but the list of inventories varies depending on the language. For example German has only 2 levels of pause inventories while Serbo-Croatian uses 6 [14].

Duration measurement

In order to produce natural sound, TTS systems should be able to provide segment and pause durations that do not significantly differ from those produced naturally. Durations can be modeled for different type of target units, for example durations of sub-phonetic units (such as vowel onset, steady-state part, and offset), phonemes, syllables, feet, words, or phrases. And appropriate relationships must be established between these units and the syntactic-prosodic information.

There are old and recent principles that can be followed in modeling duration units. For instance, there is an old principle called *isochrony*, which leads to analysis in terms of syllable or foot sized units. This is purely a rhythmic principle, and accordingly speakers would unconsciously use an initial internal clock to schedule speech segments (typically synchronized with syllables for fixed stress languages and feet for free stress ones).

Since many state-of-the-art TTS systems are based on phonetic elements as speech units, syllable-based duration systems incorporate an algorithm to derive segment durations from syllable ones. For this purpose elasticity principle may be used; in which all segmental durations in a syllable frame are obtained by one and the same factor k by the formula

$$Dur_i = \exp(\mu_i + k\sigma_i) \qquad (2.1)$$

in which Dur_i is the duration of the i^{th} phoneme of a particular syllable and μ_i and σ_i are the mean and standard deviations of its log-transformed durations in a large corpus. In equation 2.1, if the *syllable duration* and its segmental components μ_i and σ_i are given, then k is found and hence the actual phoneme duration is determined.

In general, though TTS systems focus on duration of phonetic segments, effects of syllable, word and phrase boundaries on segmental duration suggest that larger units should be considered.

Two main trends can be distinguished in duration modeling. In the first, and by far most common, one durations are computed by first assigning an intrinsic duration to segments (pauses being considered as particular segments), which is further modified by successively applying rules combining co-intrinsic and linguistic factors into additive or multiplicative factors. For instance several corpora can be analyzed to study speaker independent intrinsic durations and their modifications under the influence of a consonant upon the preceding vowel in a stressed syllable, of the type of word (content versus function word), of the location of the segment within a word, of the distance to major and minor phrase boundaries, and so on, as well as of the grouping of consonants into clusters

at various positions of the word/prosodic phrase/sentence. This leads us to establish a simple multiplicative rule:

$$\textit{Vowel duration} = ID\ V_i\ m_c \qquad (2.2)$$

$$\textit{Consonant duration} = ID\ C_{ij}$$

where *ID* refers to intrinsic durations, V_i and C_{ij} are shortening or lengthening coefficients dependent on the position i of a segment within a word and in a sentence, j refers to the phonetic class membership of consonants, and m_c reflects, in certain cases of closed syllables, the influence of a consonant or semivowel on the preceding vowel.

A second and more recent approach requires the availability of large speech corpora and of computational resources. A large number of possible control factors are simultaneously varied, this is a very general duration model proposed using CARTs or neural networks, and the parameters are automatically derived by the associated approximation algorithm (standard least squares methods, minimization of entropy, and error-back-propagation, respectively).

Phoneme duration

Numerous factors, including semantics and pragmatic conditions, might influence phoneme durations. There are several different duration-modeling mechanisms which can be used to model and determine durations. One example is Rule-based method [4], which involves table lookup of minimum and inherent durations for every phone type. In this mechanism, the duration of a phone is expressed as:

$$d = d_{min} + r(\bar{d} - d_{min}) \qquad (2.3)$$

where $\bar{d}$ is the average duration of the phoneme, d_{min} is the minimum duration of the phoneme, and the correction r is given by

$$r = \prod_{i=1}^{N} r_i \qquad (2.4)$$

for N rules being applied and each rule has a correction of r_i. Another commonly used duration modeling is CART-Based duration.[1]

1 Interested readers are referred to [4] for a detailed description of CART based duration modeling.

Syllable duration

Syllable duration can be determining by summing up the phonemes duration in a syllable as explained in the previous section. Several factors affect the duration of a syllable, such as it position in a word, whether it is stressed or unstressed, etc.

Several researchers studied the syllable structure of Amharic language and came up with different syllable templates. Some researchers concluded that there are six possible syllabic structures described as a juxtaposition of vowel and consonant phonemes. These are (V, VC, VCC, CVC, and CVCC), where V refers to vowels and C consonants. Other literature suggests that there are only two types (CV and CVC). Another view in literature argued that there are more than six syllable structures for Amharic (V, VC, VCC, CV, CVC, CVCC, CCV and CCVC) in which an initial cluster of two consonants could exist and when the second consonant in the cluster is liquid (l, r)).

In a limited experiment conducted on Amharic words structure in this research the number of possible syllable structure for Amharic is indeed more than six, although the reasoning (or justification) may vary.

In most cases, the duration of a syllable is in proportion to the number of phonemes contained within it, though greatly affected by other factors as well. For example: Amharic rhythm seems to exhibit isochrony, i.e., being equally spaced out in time, and is usually used in connection with the description of the rhythm of languages. In Amharic, stressed syllables tend to be longer than unstressed ones as is in other languages.

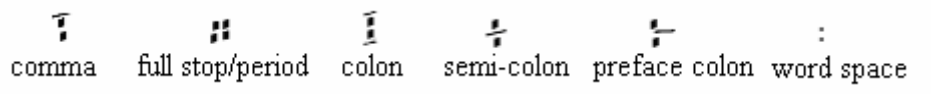
Pause Determination

Speakers introduce pauses for different reasons, such as, on arrival at punctuation marks, to give emphasis, on word junctures, to stop running out of breath, and other phenomena (such as a disfluency i.e. a repair or filled pause).

The most reliable indicator of pause location is, of course, punctuation. These include, periods, semicolon, commas, exclamation points, parentheses, ellipsis points, colons, dashes, etc. Each of these can be taken to correspond to a prosodic phrase boundary and can be given a special pitch movement at its end-point and a corresponding pause which duration depends on the type of punctuation.

The length of the pause is not as important as the simple presence of silence of equal length. Thus simple algorithms can implement pause with a fixed pause duration. However, techniques such as CART (Classification and Regression Tree) can be used for pause duration assignment. We can use POS categories of words, punctuation, and a few structural measures, such as overall length of a phrase relative to neighboring phrases to construct the CART.

In Amharic, as in other languages, punctuation can be taken as a main indicator for pause. Some of the commonly used sets of punctuations in Amharic that may involve signaling of pause are the following.



The word space punctuation (**:**) is nowadays becoming obsolete and replaced by space instead, and it has no significance for prosody modeling (it does not tell the degree of disjuncture between words).

Pauses can be introduced in between different size of linguistic or prosodic units, example between syllables (which is rare case), words, NP, VP, sentences, iPs, and IPs. And the duration of pause/silence for such category of units may vary depending, for example, on speaker's behavior or condition.

The following conditions may be useful in determining the location of pauses:

- When a sentence boundary is encountered (marked by punctuation e.g., **::**)
- If a comma or semicolon is occurred in some part of the utterance.
- Types of words (content or function word) --certain function words are more likely

to signal a break.

- How many content words occurred since the previous function word (if more than 4 or 5 words, a break is more likely)
- What is the length of current proposed major sentence?

In Amharic writing system or script, a sentence can be terminated by any of the three punctuation marks (**!**, **?**, or **::**). This punctuation marks rarely, if at all, occur in any other context and thus ambiguity related to detecting sentence end punctuation is less difficult as compared to other languages, for example English sentences can be ended using either of (**.** , **?**, **!**, or **:**), of which only the question mark is almost unambiguous, since the other punctuation marks can be used in other contexts, e.g. the period can signify abbreviations which can occur in the middle of a sentence.

In the case of Amharic, exclamation mark can be ambiguous if it is used and refers to the mathematical symbol for factorial as is in English too. In Amharic alphabet characters there is no typeface options, such as small letter and capital letter varieties. So processing and detecting uppercase letters, which may help for detecting start of sentences, start of new line, acronyms, names etc., is not necessary.

Intonation Determination

The impression of naturalness of speech generated by a TTS system is a function of the richness of the melodic contours and the quality of the rhythmic patterns it applies to the speech it produces.

In its more restricted sense, “intonation” refers to the variations in the pitch of a speaker’s voice used to convey or alter meaning, but in its broader and more popular sense it is used to cover much the same meaning as ’prosody’, where variations in such things as voice quality, tempo and loudness are included.

The variations in human speech are mainly caused by alternation of the frequency of vibration of the vocal cord. The intonation variation can differ from language to

language. It is certainly possible to analyze pitch movements (or their acoustic counterpart, fundamental frequency or F0) and find regular patterns that can be described and tabulated.

Varying the intonation patterns in an utterance can indicate differences between words with otherwise identical phonological representation. This type of intonation is commonly referred to as tonal accent (or simply tone). Intonation can also cover an entire sentence, indicating whether the utterance is a question (interrogative) or a statement (predicative), as well as if the utterance is expected to be followed by more utterances of the same speaker, or if the speaker has completed speaking. This type of intonation is called clause intonation (or simply intonation).

Intonation is said to convey emotions and attitudes. Intonation includes things like the difference between statement and question. In conversational discourse it, involves aspects that indicate whether the particular utterance constitutes new information or old and the regulation of turn taking in conversation.

Generally two types of intonational phrases can be identified in a spoken language: Full intonational phrases (or simply Intonational phrases, IPs) and Intermediate intonational phrases (iPs). Intermediate intonational phrase are related to inter-sentence prosodic junctures while Intonational phrases are related to end-of sentence junctures.

Algorithms have been proposed which attempt to automatically break an input text sentence into intonational phrases. E.g. statistical models (incorporating probabilistic predictors such as the CART-style decision trees) for predicting intonational phrase boundaries based on such features as the part of speech of the surrounding words, the length of the utterance in words and seconds, the distance of the potential boundary from the beginning or ending of the utterance, and whether the surrounding words are accented.

Prosodic junctures that are clearly signaled by silence (and usually by characteristic pitch movement as well), also called intonational phrases, are required between utterances and usually at punctuation boundaries. Prosodic junctures that are not signaled by silence but rather by characteristic pitch movement only are called phonological phrases.

The primary phonetic means of signaling juncture are: silence (pause) insertion, characteristic pitch movements in the phrase-final syllable, lengthening of a few phones in the phrase final syllable, and irregular voice quality such as vocal fry. An end-of-sentence period may trigger an extreme lowering of pitch, a comma-terminated prosodic phrase may exhibit a small continuation rise at the end, signaling more to come, etc. Certain pitch-range effects over the entire clause or utterance can also be based on punctuation e.g. exclamations may have a heightened range, or at least higher accent targets throughout.

Accent

Accent may refer to prominence given to a syllable, usually by the use of pitch. In the broadest sense accent may mean stress, which is more often used to refer to all sorts of prominence (including prominence resulting from increased loudness, length or sound quality), or to refer to the effort made by the speaker in producing a stressed syllable. Accent also refers to a particular way of pronouncing words.

Tone

It refers to an identifiable movement of pitch that is used in a linguistically contrastive way. In some languages (known as tone languages) the linguistic function of tone is to change the lexical meaning of a word. In other languages, tone forms the central part of intonation, and the difference between, for example, a rising and a falling tone on a particular word may trigger a different interpretation of the sentence in which it occurs. In the case of tone languages, it is usual to identify tones as being a property of individual syllables, whereas an intonational tone may be spread over many syllables and its purpose is to indicate prosodic features such as focus, contrast, exclamation etc.

Tone determination

In practice, a number of rules are often used to determine pitch accent tones and phrasal tones, although it requires a complete analysis in the NLP module of a TTS system.

Abstract levels such as **H**(high) **L**(low) can be codified for a language to indicate a relatively higher or lower point in a speaker's pitch range. Using the **H/L** primitive distinctions we can form two types of pitch events:

(1) pitch accents, which signal prominence across syllables/word (E.g., **H***, **L***, **L+H***, etc.).

(2) Phrasal tones, which signal unit completion or delimitations. These Phrasal tones are further divided in to phrase accent tones (iP tones, E.g., **H-**, **L-**) and boundary tones (IP tones, E.g., **H%**, **L%**).

Evaluation

Once the models are developed, it is essential to evaluate them if they result in any marked improvement in terms of naturalness and intelligibility. As detailed in [5], there are different ways to evaluate the performance of a TTS system. These evaluation methods can be divided into two different classes: *subjective measure* and *objective measure*. With subjective measure, usually perceptual tests (a.k.a listening tests) are performed. Listeners are presented with the synthesized speech and are asked to judge for its intelligibility and overall quality.

There are two desirable qualities that a synthesized speech is expected to have. These are intelligibility and naturalness. These two qualities can be evaluated using either subjective or objective measures.

Intelligibility: the intelligibility of a synthesized speech measures the degree to which a human listener can understand the spoken output. Tests that measure the level of understandability of a synthesized speech are called *intelligibility tests*. Some examples of this sort of test methods are *Diagnostic Rhyme Test* (DRT) and *Modified Rhyme Test*

(MRT) on word level and the *Harvard Psychoacoustic Sentences and Haskins Syntactic Sentences test on sentence level.*[2]

Perceptual tests for intelligibility are designed to check whether listeners are able to identify the words that make up the sentence of the synthesized speech. They are made to listen to the uttered speech and asked to transcribe the sentence(s) they have just listened to. A scoring scheme that gives full points for sentences which have been correctly transcribed, half points for which some error in the transcription has occurred and no points for which the sentences were completely understandable is used rate the intelligibility of the uttered sentence(s). This test is conducted for multiple users and the average score for each utterance is used as an indicative value of the intelligibility of the utterance. Averaging all these values (for all utterances) will then give an overall score of the intelligibility of the output of the TTS system.

The ability of the perceptual test to properly measure (evaluate) the degree of intelligibility depends on two important factors:

- the number and variety of listeners used in the test and
- the number and variety of utterances used in the test.

The larger the number of listeners and sentences used, the better the ability of the test to measure intelligibility. Furthermore, care must be exercised to make sure that different groups of listeners representing all potential users of the system have been included in the system. Trained users might (e.g. speech researchers) might score high on intelligibility test. The sentences included in the test must include all types of sentences (i.e. declarative sentences, questions, exclamations, short sentences, long sentences as well as simple and complex sentences). This will assure that the score for the intelligibility of the system is not restricted to a small subset of sentence types but covers all possible utterance constructions.

Naturalness: the naturalness of a synthesized speech measures the degree to which the synthesized speech resembles natural speech uttered by a human speaker. The naturalness of a synthesized speech is evaluated using methods referred collectively as naturalness

2 For details on intelligibility test readers are referred to [4]

tests. Some examples of naturalness tests include the paired comparison, the MOS and Forced-Choice Ranking tests.[3]

Perceptual tests for naturalness are designed to measure how listeners rate the closeness of the synthesized speech to naturally produced speech.

In this test listeners are made to listen to the uttered speech and asked to rank, on a subjective measure, how closely the uttered speech mimics natural speech. The ranking varies from excellent to that of not at all. Each rank being associated with a point that varies from a maximum value (for excellent) to that of zero (for not at all). Like the intelligibility test, this test is conducted for multiple users and the average score for each utterance is used as an indicative value of the naturalness of the utterance. Averaging all these values (for all utterances) will then give an overall score of the naturalness of the output of the TTS system.

Much like the intelligibility test, the variety and number of the sentences and listeners used in the test affects the the naturalness test's ability to evaluate the level of naturalness of the synthetic speech.

Though several individual test methods for synthetic speech have been developed during the last few decades, there is still no single definitive test method that can be uniformly applied to all evaluation situations.. Therefore, the most suitable way to test a speech synthesizer is to mix several testing and evaluating methods. Depending on the required kind of evaluation, the methods can be applied on several levels of speech units including phonemes, words or sentences.

Prosody evaluation

As mentioned above evaluation for TTS systems can be done automatically or by using listening tests with human subjects, which holds too for prosody evaluation. In both the

3 For details on naturalness tests readers are referred to [5]

automated and listening tests, it is useful to start with natural recorded utterances with their associated text. It is required to start by replacing the natural prosody with the system's synthetic prosody. In the case of automatic evaluation, we can compare the enriched prosodic representations for both the natural recording and the synthetic prosody. The reference enriched prosodic representation can be obtained either manually or by using a pitch tracker and a speech recognizer.

Automated testing of prosody involves the following:

- **Duration**. It can be performed by measuring the average squared difference between each phone's actual duration in a real utterance and the duration predicted by the system.
- **Pitch contours**. It can be performed by using standard statistical measures over a system contour and a natural one. When this is done, duration and phoneme identity should be completely controlled. Measures such as root-mean-square error indicate the characteristic divergence between two contours, while correlation indicates the similarity in shape across different pitch ranges.

Listening test can be performed to evaluate a prosody module. The marked difference in this case to that of evaluating a general TTS system is the objective is to compare two synthetic speech utterances, or a synthetic speech and a natural utterance in terms of their naturalness and intelligibility. In the former case, one of the utterances is generated with prosodic models incorporated while the other is not while all other things are equal. The idea is to test whether the prosody model has improved the naturalness and/or intelligibility of the uttered speech. In the latter case, the synthetic speech (with prosody modeling) is compared with naturally produced utterance representing the same sentence and listeners are requested if the synthetic speech is a good approximation of the naturally produced utterance.

Subjects are made to listen to two different utterance realizations of the same sentence(s), either natural recording and synthetic speech, or to two synthetic speech utterances generated with and without prosody modeling. When the two synthetic speech utterances are used in the naturalness test, the listeners are asked to choose which one of the two

utterances sounds more natural or if there is no difference. The scoring is adjusted so that the utterance closer to the natural utterance gets full points while the other is given no points, in the case when both are equally close (or far) to the natural utterance, both are given equal points. This allows us to specifically measure the improvement (or degradation) attained because of the prosody modeling.

Description of the Tools and Systems used in the Project

A number of tools, systems and programs have been used in the development of the Amharic prosodic models. The following sections give a brief description of these tools and systems. The first of these is the ToBI transcription system, which provides the framework for developing the labeling inventories that are developed for Amharic in this project and described at the results section of this document. PRAAT and Wavesurfer are speech analysis toolkits that are used to analyze the properties of sample speech data in terms of variation in amplitude, pitch, pauses etc. The Eruxelf Amatets Amharic TTS system is used to generate Amharic synthetic speech to be used for testing the models developed in this thesis research.

Prosodic Transcription Systems

Prosodic transcription systems are the means that provide the framework for encoding prosodic phenomena. Encoding implies deciding which variations in the physical parameters of the speech wave carry out linguistic information and finding a way to describe them by means of a symbolic system. Several such transcription systems have been developed and used. The systems developed so far have been designed with different purposes in mind and within different traditions. Some examples are:

ToBI (Tones and Break Indices)

ToBI is a framework for developing community-wide conventions for transcribing the intonation and prosodic structure of spoken utterances in a language variety. A ToBI transcription of an utterance consists minimally of a recording of the speech, its fundamental frequency (F0) contour, and (in the transcription proper) symbolic labels for

prosodic events. The transcription proper is usually arranged in four time-aligned parallel horizontal panels or tiers, so that the symbolic labels can be easily matched with the corresponding F0 track and speech waveform. (Other tiers can be added for the needs of particular sites.) The four labelling tiers each appear in their own window:

- **orthographic tier –** used to transcribe the words or syllables
- **break-index tier –** marks the prosodic grouping of the words in an utterance by labeling the end of each word for the subjective strength of its association with the next word. For English this tier rates subjective strengths of associations on a scale from 0 (for the strongest perceived conjoining) to 4 (for the most disjoint).
- **tone tier –** consists of labels for distinctive pitch events, transcribed as a sequence of high (**H**) and low (**L**) tones marked with diacritics.
- **miscellaneous tier –** used to mark other events that are arguably not part of prosody.

Each tier consists of symbols representing prosodic events, associated to the time in which they occur in the utterance. Although primarily intended for English, work using the ToBI system has been extended for many other languages and as such has become the de-facto common framework for modeling prosody. Modeling the prosody of a language within the ToBI framework thus involves the following tasks

- determining the prosodic inventories for the language for the break-index tier and tone tier
- adding extra tiers if the language so requires and finding the inventories
- applying these inventories on speech segments

The following table taken partially from [14] [pp. 434] shows some of the inventories for a few selected languages.

Language	Types of Tiers – Extra only	Types of Break Indices	Type of Tones	Prosodic Units
English		0,1,2,3,4	L*,H*,L+H*,L*+H,H+!H*	
			L-,H-	iP
			L%,H%, !,<,>	IP
German		3,4	L*,H*,L+H*,L*+H,H+!H*,H+L*	
		2r (rhythm mismatch)	L-,H-	iP
			!H-%,L%,H%,^H%	IP
		2t (tone mismatch)	!, ^,<,>	
Italian		0,1,2,3,4	L*,H*,L+H*	
			L*n,L+H*n, L*+Hn, H+L*n	iP
			H(*)L-	iP
			L%	IP
			!(for *), ! (for nuclear pitch accent)	
Greek	Prosodic Word = (phonetic transcription)	0,1,2 (iP)	L*,H*,L+H*,L*+H,H*+L	
		3(IP)	L-,H-,!H-	iP
		s(sandihi)	L%,H%, !H%	IP
		m(mismatch)	!(for), <, >, w (for L* undershoot)	

Table 2.4 Prosodic models of some languages within the ToBI framework

WaveSurfer and **PRAAT** are good examples used as environments for doing ToBI labeling.

Speech Analysis Tools

Several software tools have been developed so far for speech recognition and speech synthesis analysis. Praat and WaveSurfer are two very good examples of software packages for the analysis of speech signals available free of charge via the Internet.

WaveSurfer

WaveSurfer is an Open Source tool for sound visualization and manipulation developed by the Sound and Music Research group of the Royal Institute of Technology, Sweden

available for download at http://www1.speech.kth.se/prod/wavesurfer. It can be used as a stand-alone tool suited for a wide range of tasks in speech research and education. Typical applications are speech/sound analysis and sound annotation or transcription. WaveSurfer can also serve as a platform for more advanced/specialized applications.

WaveSurfer has a simple but powerful interface. It works with sound files as the basic document to work on with ability to load a sound file from disk or record a new waveform using the tape-recorder like controls it provides. It allows users to manipulate files by opening multiple files, replaying them and other more advanced functions including combining multiple sounds, deleting segments of sound files, reordering, inverting, normalizing, reversing, amplifying and removing DC.

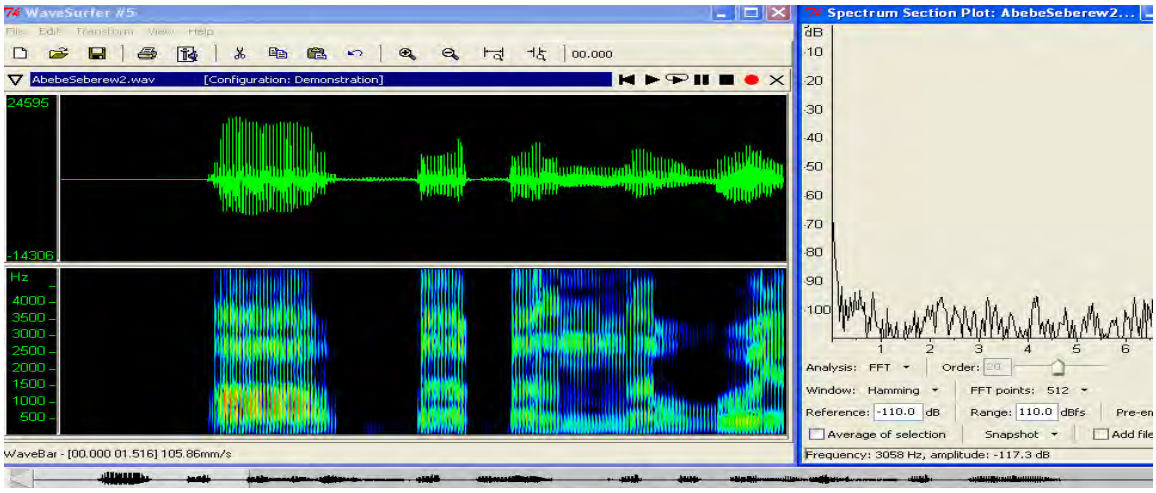
WaveSurfer can be used to visualize and analyze sound files in several ways. The standard analysis plug-in can display Waveform, Spectrogram, Pitch, Power or Formant panes. Many properties of these panes can be adjusted using the properties context sensitive menu. Special control windows are available for Waveforms and Spectrograms, which allows the user to make quick modifications. A special enlarged waveform window, which is cantered around the cursor, can be used for detailed inspection and adjustments.

WaveSurfer has many facilities for transcribing sound files. Transcription is handled by a dedicated plug-in and it's associated pane type. It allows users to create a special configuration for a certain combination of sound and transcription files, specifying file properties such as file name extension, label format, and character encoding. There are many options to control appearance and editing functionality. Depending on the transcription file format additional options might be available. Unicode characters are supported if using the source version of WaveSurfer, in order to keep the binary versions small. The transcription plug-in is used in combination with format handler plug-ins which handle the conversion between file formats and the internal format used by the transcription plug-in. Label editing is straightforward, achieved simply by click where it is wanted to insert a label and typing it in. The label fields are user-configurable and used

to insert a label directly at the cursor position. Time boundaries can be dragged using the mouse to right/left justify boundaries with the cursor.

Figure 2.2 Sample Window of Speech Analysis in WaveSurfer

The visualization of data related to a sound file is handled by the dataplot plug-in. Pitch, Power and Formants are examples of data that this plug-in can be used to plot. It can also be used to visualize other time aligned data, for example output from other programs. The plug-in plots tabulated numerical or text data. Optionally a spectrogram or a waveform can be drawn as a backdrop. The data can be plotted either as continuous curves or using



dots. The data values can be edited by simply dragging them with the mouse.

PRAAT

Praat is a computer program that can be used to analyze, synthesize, and manipulate speech. The program is created by Paul Boersma and David Weenink of the Institute of Phonetics Sciences of the University of Amsterdam. Praat is available for download from the web address http://www.praat.org or http://www.fon.hum.uva.nl/praat/. It has functions for speech analysis, speech synthesis, learning algorithms, labeling and segmentation, speech manipulation, listening experiments, and more.

The following are some of the functions that are available in Praat:

- ●View sound file as a waveform, pitch plot, spectrogram, various F1 vs. F2 displays, duration and intensity analysis.
- ●Playback with repetition with variable length delay between repetitions.
- ●Label intervals and time points on multiple tiers and transcribe speech files phonetically.
- ●Use sound files up to 2 gigabytes (3 hours)

The following two figures show how praat is useful to manipulate speech waveform. The first figure (Figure 2.3 bellow) shows the pitch movement (or variation) due to sentence final that can lead to recognize differences among sentences types. The dim F0 contour shows **L%** boundary tone for the Amharic declarative sentence "**የኢትዮጵያ መሬት ወጣ ገባ ይበዘዋል።**" But, the same sentence can be changed into yes/no question type sentence by varying the pitch movement of the sound utterance using the mouse to drag the anchor points of the F0 contour, so that it becomes "**የኢትዮጵያ መሬት ወጣ ገባ ይበዘዋል**?" converting the **L%** boundary to that of **H-H%** boundary tone sequence.

Similarly, Figure 2.4 below shows the Amharic sentence "**ገ'ና አልደረሰም።**" which was changed to "**ገና አልደረሰም።**" by modifying the relative duration for the **ን** phoneme. This example demonstrates how Praat can be used duration modification can be used disambiguate words' meanings by varying the relative duration of phonemes. Praat is specially very useful for prosodic modeling as it incorporates many functions that are specifically relevant to prosodic modeling. It runs in Macintosh, Windows, Linux, SPARC Solaris etc.

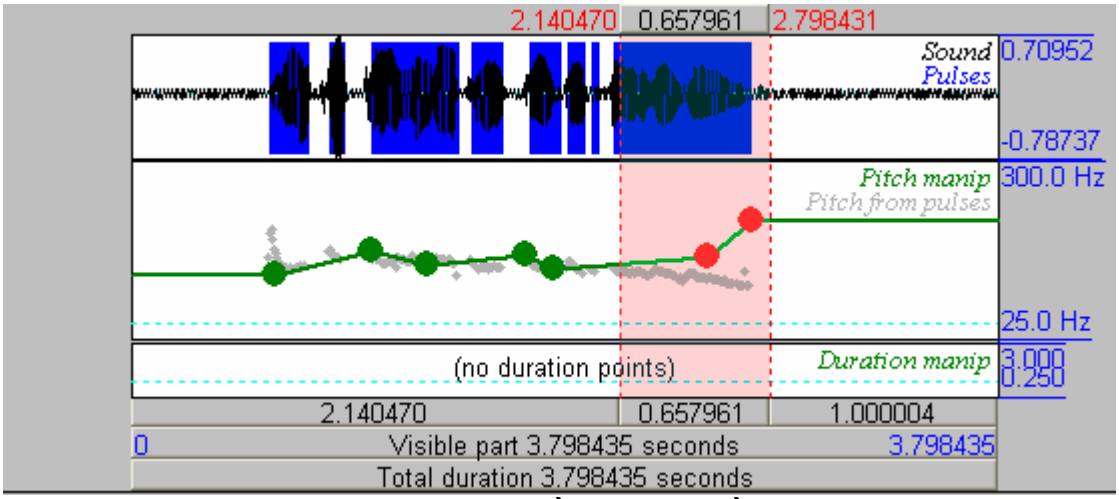


Figure 2.3 Pratt representation of "**የኢትዮጵያ መሬት ወጣ ገባ ይበዘዋል።**" to "**የኢትዮጵያ መሬት ወጣ ገባ ይበዘዋል**?" using pitch modification.

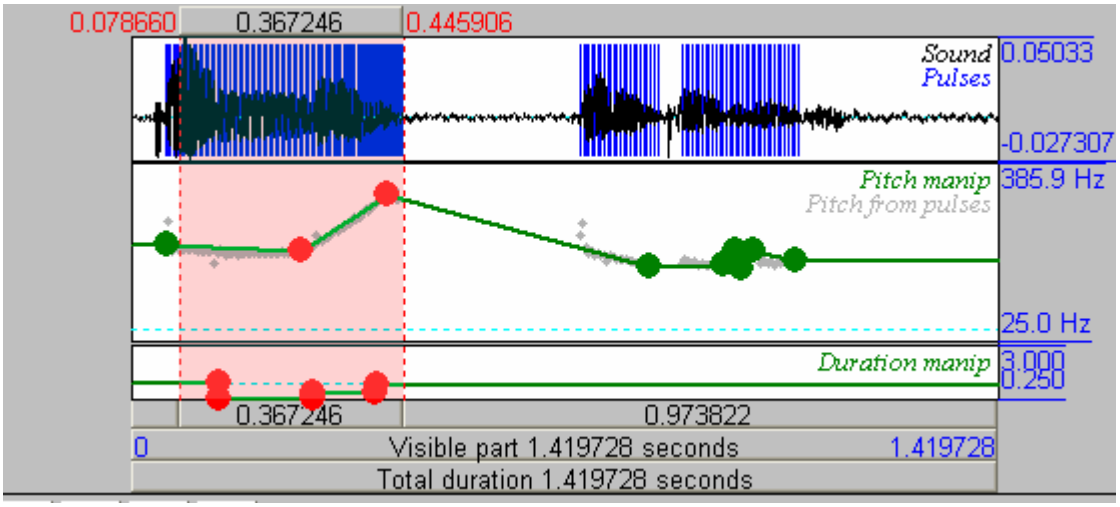


Figure 2.4 Pratt representation of " ገ'ና አልደረሰም::" to "ገና አልደረሰም::" using duration modification.

The Eruxelf Amatets Amharic TTS System

Eruxelf Amatets is an Amharic Speech Synthesis program primarily designed to help visually impaired Amharic computer users to create, modify and read Amharic documents. It is developed by Eruxelf Scientific Computing PLC., Addis Ababa. Eruxelf Amatets is a program that works in two modes – writer mode and reader mode. In writer mode users can create Amharic documents. Eruxelf helps users by echoing the Amharic and English characters as they are typed. In addition, Eruxelf Amatets can read out a saved Amharic file in reader mode.

Eruxelf Amatets (reader mode) is developed based on the Festival Speech Synthesis System. It uses the ked voice available in Festival to read out Amharic documents. Eruxelf Amatets takes the utterances of an English speaking person and tries to find a mapping for the corresponding Amharic utterances. In this manner the sounds (phones) are reassembled to imitate an Amharic utterance. Eruxelf Amatets is designed to generate intelligible Amharic utterances by mapping and assembling their equivalent English phones. Future versions are expected to use native speaker's voice to generate the Amharic utterances.

Eruxelf Amatets runs on Windows using the Cygwin Linux Emulator program. Both Festival and Cygwin are open source systems.

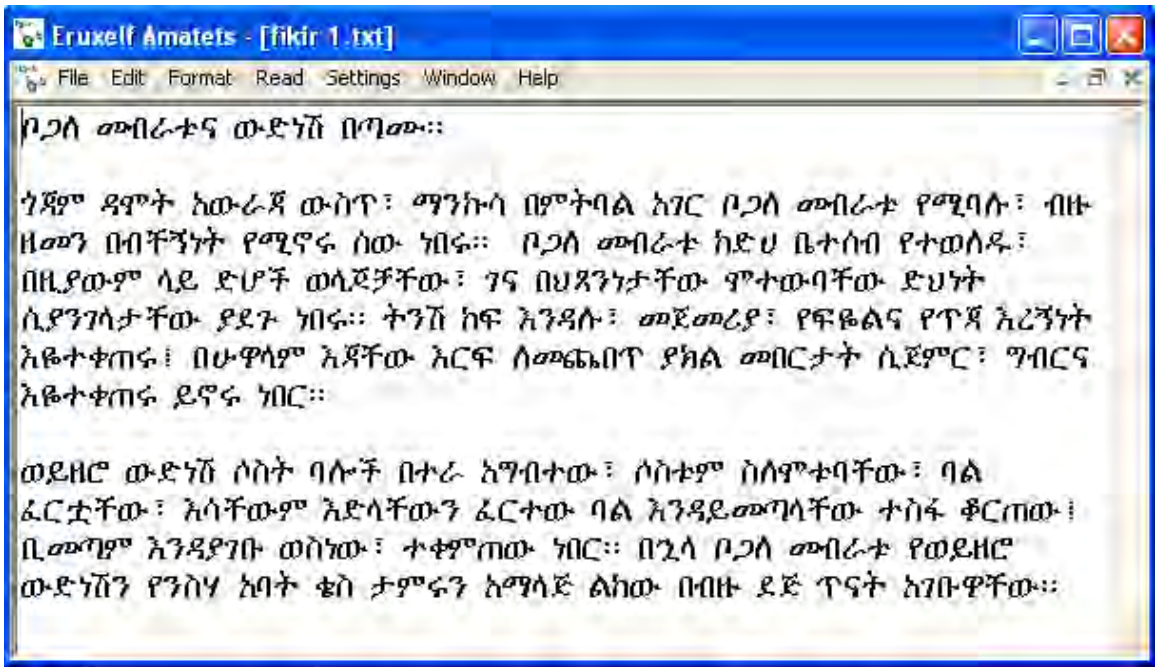


Figure 2.5 A screen shot of Eruxelf Amatets

The most common use of Eruxelf Amatets is to open an existing Amharic document and make it read the document by pressing ctrl+R. In this scenario, Eruxelf Amatets reads the current document out loud. Eruxelf Amatets can be made to save the sound output of the narrative by modifying its configuration file. Eruxelf Amatets does not use prosodic models specifically designed for Amharic. It applies only primitive prosody based on punctuation marks. Furthermore, these prosodic rules are based on English punctuation marks available in the Festival system, which it uses as its engine.

Although Eruxelf generates sounds, which are more or less phonetically correct for Amharic, it fails in the prosody part. This fact demonstrates the failure of Amharic synthesizers in terms of their prosodic sophistication. As the main goal of the thesis project is to improve the naturalness and intelligibility of Amharic synthetic speech by incorporating prosodic models, the output speech waveform of Eruxelf provides a very good test data to test whether the models of this project are successful in improving the naturalness and intelligibility of synthetic Amharic Speech. Eruxelf Amatets is, thus, used in this thesis project to generate synthetic Amharic speech to be used as test data for applying the prosodic modifications.

Summary

One method of describing the prosody of a language is by the autosegmental metrical (AM) theory which classifies prosody as composed of prosodic structure and prominence relationships. Prominences are generally indicated by variations in pitch. In the AM theory there are two levels of pitch (the high **H** and the low **L** pitch targets) which can be combined to give various prominence relationships. Each language defines its own set of prominence relationships. The other important prosodic component is the insertion of pauses. The main problem in insertion of pauses is locating where the pause should be inserted. This can be indicated either by punctuation marks or a set of other prosodic cues. The duration of the pause is of secondary nature compared to that of its existence. However, the degree of juncture between prosodic components is indicated by the duration of the pause. The set of available pause duration levels varies for each language. Specifically, this thesis wishes to add Amharic to the set of languages for which AM models based on ToBI are developed. The first step in this task is to demonstrate what is to be done in the methodology section of this thesis in chapter 3.

CHAPTER 3

METHODOLOGY

The purpose of this chapter is to describe in some detail the methodology to be used in the development of a prosodic model for Amharic. In section 3.1 the corpus collection methodology is described. In section 3.2 the procedure to be followed in determining the phonological inventory of Amharic will outlined. Issues that deal with duration measurement are described in 3.3 and how the various intonations are determined will be explained in section 3.4. Sections 3.5, 3.6 and 3.7 describe the procedures for labeling and transcription, analysis and testing methods and evaluation respectively.

Corpus Collection

Normally, as outlined in [1], NLP and then TTS systems require and have a well organized corpus database, which contains among other things: lexical dictionary, phonemically mapped words, phonemes, recorded sounds, diacritics indicating stress other similar data required by the different modules of the TTS system. Unfortunately these resources are not available for Amharic. The tagged textual information – the output of the NLP module – is a required input to the prosodic component. Since such data is not available, the initial task in the research is the development of a small text corpus specifically tailored to the research.

Text data collection

In this experiment, around 170 sentences were collected from different sources such as Amharic magazines, fictions, teaching books – designed to cover the different sentence types: declarative, yes/no question, Wh-question, interjection (exclamation), ambiguous statements and sentences having clear prosodic phrases (syntactic structures). Some of these are to be used as model text data – to be used in the development of the models, and the rest as test text data – to be used during the testing phase of the model. Manually normalized text data were used. This text corpus is listed in Annex 2.

Recording the data

All the 170 sentences were recorded by inviting 6 speakers (3 female, 3 male) between 25 and 43 years old of different mother tongue, gender and region. This is done with the

objective of incorporating the duration and pitch variability caused by gender and regional effects.

The model text data which are now in sound waveform (intended for modeling Amharic prosody) are investigated and analyzed using the speech analysis tool PRAAT for the following prosodic features

- ●Pitch characteristics movement: pitch accent, phrase accent and boundary tones.
- ●Relative duration of phonemes, syllables, juncture between words, Pause duration between sentences and iPs.

Duration Measurement

In this experiment durations measurements are performed. The durations that are measured are

- ●phoneme and syllable durations,
- ●pause durations
  - ○between words,
  - ○between sentences and,
  - ○between intermediate phrases.

The measurement is not absolute since the starting and end points of the target unit to be measured may have some variability due to manual marking. In reality we should not expect absolute duration for vowels, syllables and rhyme units, since the enunciation speed changes from one recording to another and from one speaker to another, the results are not directly comparable, especially in a small corpus.

There are mechanisms which can be used to apply the coefficients of correction of tempo of the enunciation, measured by syllables per second or by mean duration (phrase duration of utterance divided by the number of sounds or syllables) so that the variation in speed of each recording can be compensated. The only relevant item in a given rhythm is the proportion between units, not the absolute measurements in milliseconds.

3.3Intonational Inventory Determination

Lists of primitive intonational elements are to be determined so that intonational inventories for Amharic are to be proposed. These intonational inventories are classified into two categories:

- Pitch accent:- under witch **H*** and **L*** pitch accent tones are examined for their realization in Amharic accented and prominence syllables.
- Phrase accent (phrasal tone and boundary tone): tones realized due to phrase final characteristic pitch movements are to be modeled.

A number of Amharic utterances, from the data model, are to be considered for this particular purpose. These utterances will be deliberately selected so that they will have prominent words and clear intonational phrases (iPs and IPs). Now using PRAAT, the pitch contour (F0) is extracted from the waveform of the utterances and examined for its shape. The peaks and valleys in the fundamental frequency of each utterance are carefully investigated for their alignment in time with a particular syllable. This requires defining the appropriate pitch range setting for the speaker.

The pitch range is a quantity that defines the minimum, middle and maximum values of pitch between which the pitch of the utterance of a particular speaker might vary. Various age and sex groups show different pitch ranges; for example, the pitch range for adult males usually vary between 75Hz to 300Hz, for adult females this value ranges between 100hz to 500hz, and for children it is between 200 to 600hz, though these values greatly depend on the sampling frequency of the waveform and the software tool used. The pitch range is defined by three basic parameters: the top tone target, the low tone target and the middle target. Pitch values in the utterance that are above the middle target will be considered as High tone target and are indicated by the abstract level, **H,** and any tone target bellow the middle is considered as Low tone target and is indicated by abstract level, **L**. Then, the most significant peaks and valleys will be looked for. The significant high peaks will be denoted by **H*** and low targets (valleys) will be denoted by **L*.** If the peak and valley characteristic pitch movements are primarily due to prominent words or accented syllables in a word which is part of a prosodic phrase (iP or IP), it will be further examined to check whether or not there is another nearby pitch movement. If

such pair of targets are found, they are taken to constitute a single bitonal pitch accent. The pitches so identified can be stylized for better experiment.

The significant rise to high level and fall to low level of the speaker's range of the fundamental frequency in the intermediate phrase final position will be labeled By **H-** and **L-**, respectively. Similarly, the significant rise to high level and fall to low level of the speaker's range of the fundamental frequency in the Intonational phrase final position will be labeled By **H%** and **L%**, respectively. The following diagram shows how the High and Low tones will be obtained.

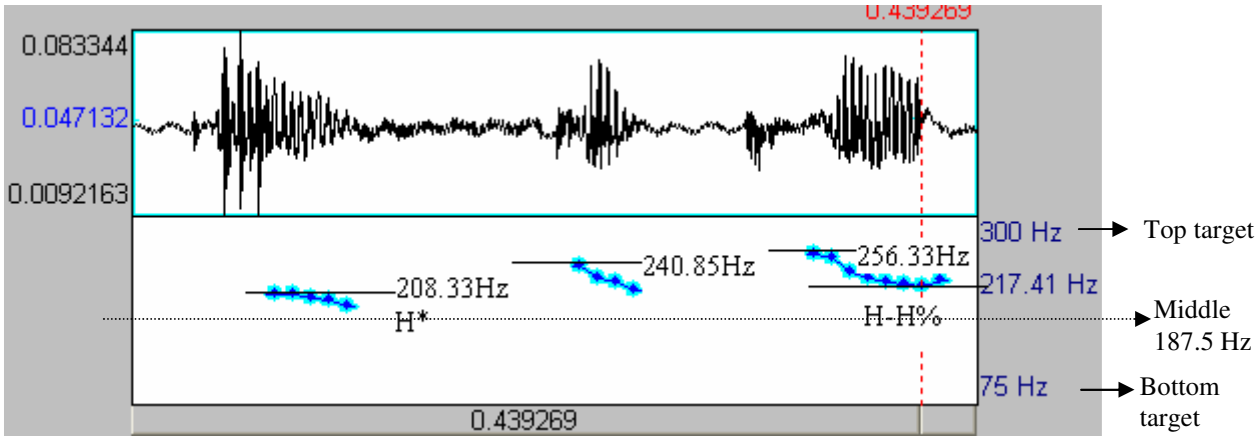


Figure 3.1: *Identification of Tonal Targets*

3.4 Labeling and Transcriptions

Transcription systems are helpful in labeling the prosodic elements of a particular language. They allow users to transcribe the most important aspects of prosody symbolically in a tiered structure aligned in time with the waveform of the utterance. The number of tiers can vary from system to system or from language to language. In this experiment, a ToBI like annotation system, provided with PRAAT is used.

There are four basic tiers in ToBI. The first tier is used to transcribe tones, and is called the *tone tier*, the second tier is used to label orthographic representation of the utterance and is called the *words or orthographic tier*, the third tier transcribes syllables and is referred to as the *syllable tier.* The fourth layer, called the *break index* tire, will be used for transcribing break indices. A *miscellaneous tier* can be included as a fifth tier for

storing miscellaneous information related to prosody, which may include comments and uncertainties in labeling. The following figure shows how the annotated layers looks when the utterance Africa is transcribed in the ToBI transcribing frame work using the Praat program.

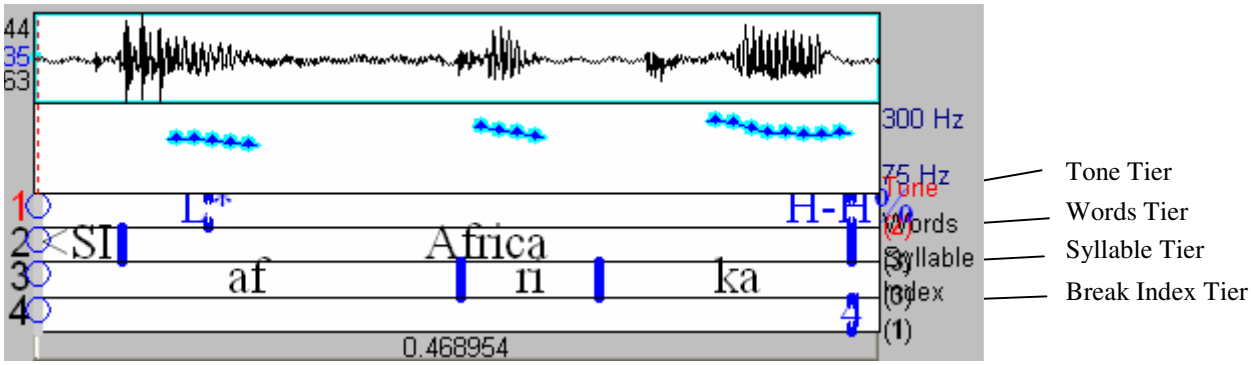


Figure 3.2: *The different layers of a ToBI transcription system shown for the utterance "Africa".*

The following table is another example of how the Amharic ToBI (AmhToBI) will look like.

	H*	L-L%			
ማን	ነው	የሰበረው			
ምኣን	ንአው	ይአ	ስአብ	ብአ	ርአው
0	1	4			

Table 3.1: *A transcription of an Amharic Utterance.*

The break indices will be placed at corresponding type of prosody junctures aligned with end of each word or phrase boundary.

3.5 Testing Methodology

3.5.1 Evaluation

In this thesis research, the evaluation of the prosodic models will be performed through listening tests. Once the Amharic prosodic model is developed, the objective of the test will be to test if there are any marked improvement in intelligibility and naturalness of

the uttered speech. For this purpose two synthesized speech utterances of the same sentence are required. The difference between the tow being one is made with prosodic modeling as developed in this thesis project and the other having no prosodic models. Apart from that it is desired that the two speech utterances be identical.

This can be obtained by synthesizing several Amharic written sentences using the Eruxelf Amatets program and saved in files on hard disk. These files are then modified using Praat to incorporate the prosodic model and the outputs are saved in another set of files.

To achieve this, several sentences that were developed in the test text corpus will be submitted to the Eruxelf Amatets Amharic TTS system, and synthesized. The synthesized speech will be manipulated using parameters for relative duration and pitch characteristic movement obtained from the Amharic prosody modeling step i.e. signal processing was performed.

Subjects will then be invited and listening tests will be made using both the speech synthesized with and without incorporating the prosodic models to identify if there was any improvement in understanding of the uttered speech (Intelligibility test) and overall quality of the uttered speech (Naturalness test).

Questions were presented to the subjects to rank the uttered speech in terms of intelligibility and naturalness. The synthesized speeches (with and without prosody) were played in random order so that a listener would not know which one incorporates the prosodic models and which one does not.

3.5.1.1 Intelligibility Test

- A subject is requested to repeat the speech that he/she has just heard. Rank giving was simple:
  - If a listener repeats the sentence exactly in the first try a score of 1 is given,
  - If the subject misses any word or could repeat the sentence after listening more

than once a score of 0.5 is given,

○If the listener cannot repeat the sentence a score o 0 is given.

●Listening tests were also performed to determine if a subject can correctly identify the type of sentence being uttered. In this experiment different types of sentences have been included: declarative sentence, yes/no question, simple type questions, sentences with clear phrasal structure, and sentences which need focus or disambiguating. The same scoring scheme as above was followed.

●Test for comparing which utterance correctly associated emphasis (focus) to specific words within a sentence and whether homographs were properly disambiguated. . Answer is in terms of yes/no or (1, 0, - (no change)).

3.5.1.2 Naturalness (or quality) test

To evaluate the naturalness or quality of synthesized speech, the widely used testing mechanism the Mean Opinion Score (MOS) has been used. This evaluating method is used to assess the relative closeness of the synthesized speech utterance to naturally produced speech. The scoring is based on a list of 5 possible values as shown in table 3.2.

Quality of speech (category)	Measure (Score)
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Table 3.2 Naturalness Test Scores

3.6 Summary

Corpus (corpora, plural) preparation is very important for prosody analysis. For prosody modeling, text corpus of different size, sentence of different type, etc., have to be recorded and studied using, for example speech tools such as PRAAT.

Phonologic inventory, which are the most expressive elements for prosody, should be determined. These include pitch events (pitch accent, phrasal tones), duration of pauses, phoneme and syllable duration.

The system should be evaluated for the intended goal. For these several evaluating approaches can be followed of which perceptual listening tests have been selected for this project.

CHAPTER 4

4 EXPERIMENTS AND RESULTS

In this chapter a description of the experiments done to develop the Amharic prosodic model is given. Three types of experiments have been performed to come up with the model.

The first of these experiments is dedicated to identifying the different pitch accents and boundary tones that exist in Amharic. The model speech samples are analyzed to identify the systematic pitch variations that appear in Amharic. The speech samples are fed to the PRAAT program and a contour of the pitch for the sample is made. The rises and falls in pitch are then observed to identify high and low targets. The high and low targets are the ones we are looking for in the pitch contour. These targets might appear alone or in combination with each other as well as together with phrase end prosodic elements. The objective is to find all those characteristic pitch movements that appear in the language and associating them with their prosodic roles (emphasis, contrast, sentence end etc.) Finally a ToBI like pitch accent and boundary tones classification is made to complete the pitch accent and boundary tone part of the prosodic model for Amharic.

The second experiment deals with identifying the break indices of Amharic. In this experiment the model speech samples are analyzed to identify the various breaks that occur in the Amharic language. The breaks of interest are, breaks between sentences, breaks between intonational phrases, breaks between phonological phrases and breaks between words. The important aspect that we are going to look for in this experiment is the relative strength of each of the breaks and making a classification of the breaks following the break index nomenclature of ToBI. This will make up the pause (break index) modeling part of the research.

The third and last experiment deals with duration modeling. In this experiment the duration of phonemes, syllables and pauses is determined. The model data is once again used, this time, to find the average duration of the phonemes, syllables and pauses of

Amharic. The model speech data will be segmented into the phonemes, syllables and pauses and a measurement of the duration of each of these speech elements is made. Later the average of the measured durations for each speech element is calculated.

The following figure can help us in visualizing the prosodic units that are assumed to exist in Amharic language.

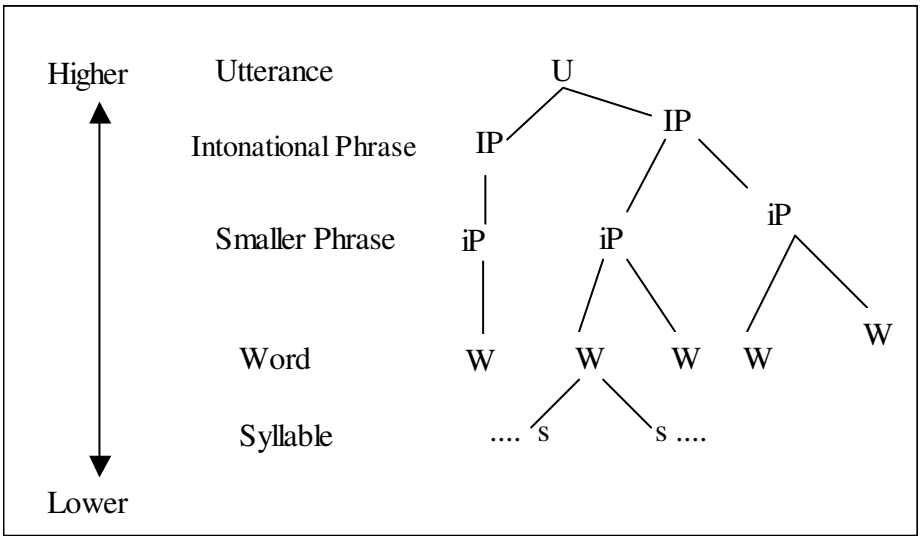


Figure 4.1 Amharic Prosodic Units

4.1 Pitch Accent and Boundary Tone inventory

Consider the following figure which shows the sound waveform and the pitch contour. for the Amharic sentence "**አበበ ተመርጠዋል::**" (for "Abebe has been elected."). This Amharic utterance when syllabified becomes as "**ኣብኣበአ ትኣምአርጥአውኣል::**" The most prominent syllable in this utterance is the second syllable (a**be**be) of abebe. The pitch contour shows a rise that peaks at the end of **be**- and then falls.

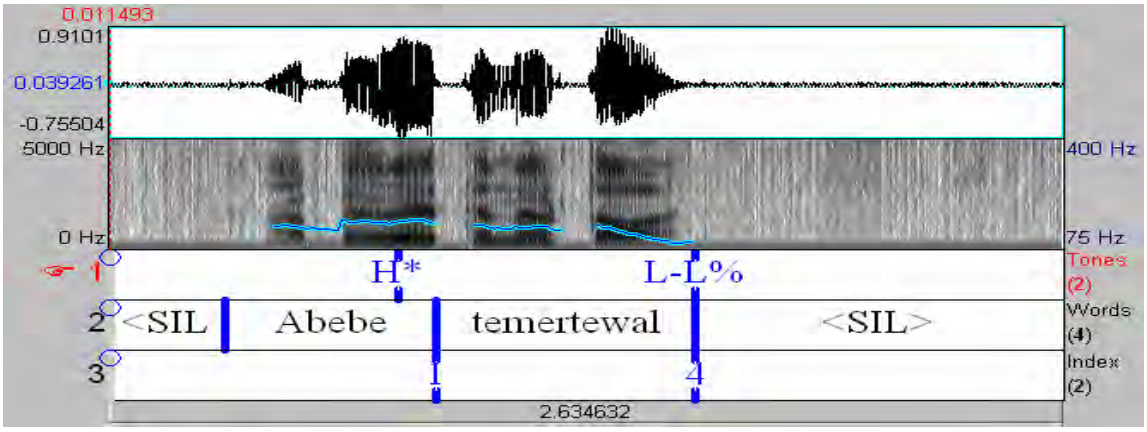


Figure 4.2 H* L-L% for "አበበ ተመርጠዋል::"

The **H** stands for a high tonal target and the * indicates that the **H** tone is associated with an accented syllable. Perceptually, this **H*** syllable is more salient than the other syllables around it. The **L-L%** sequence is often found at the end of spoken **declarative sentences** for Amharic as is for English, especially final sentences in a **turn** or **discourse**.

The following example has a falling F0 that reaches a minimum on the *–na*-syllable in *genana*. This is a Low pitch accent on *genana*, marked with **L***. Again, the **L** stands for a low tonal target and the * means that the low tone is associated with a prominent syllable. Perceptually, we can hear that the *– na*-in *genana* is more salient than other syllables in this utterance. Also we notice that the end of the utterance has a sharply rising F0. This intonational phrase has a high phrase accent (**H-**) and a following high boundary tone (**H%**), hence it is marked as **H-H%**. In Amharic, this intonational contour is one typical way of indicating that **an utterance is a question** for which the speaker expects a **yes/no** answer. Figure. 4.4 clearly shows that a question can be identified by its **H-H%** sequence of tonal patterns, that is high phrase accent followed by high boundary tone.

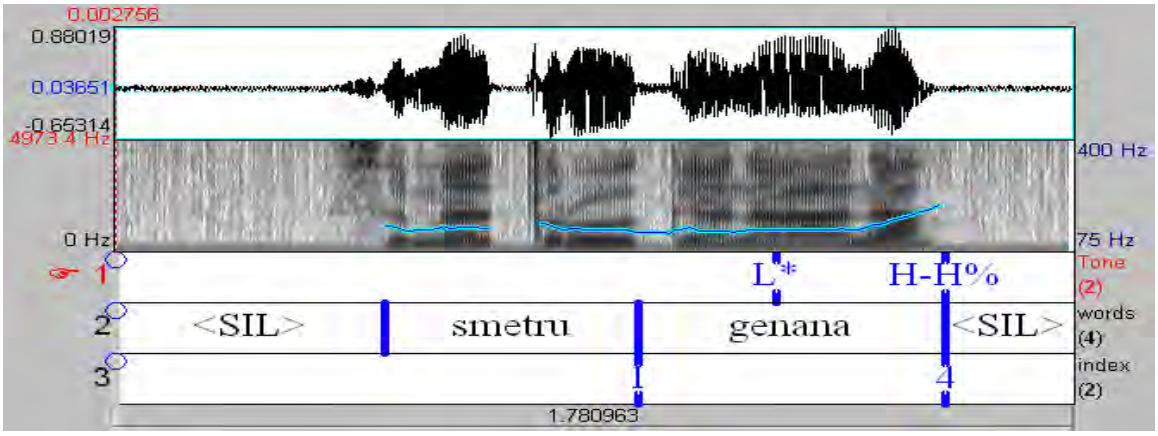


Figure 4.3 L* H-H* for **“ስመጥሩ ገናና”**

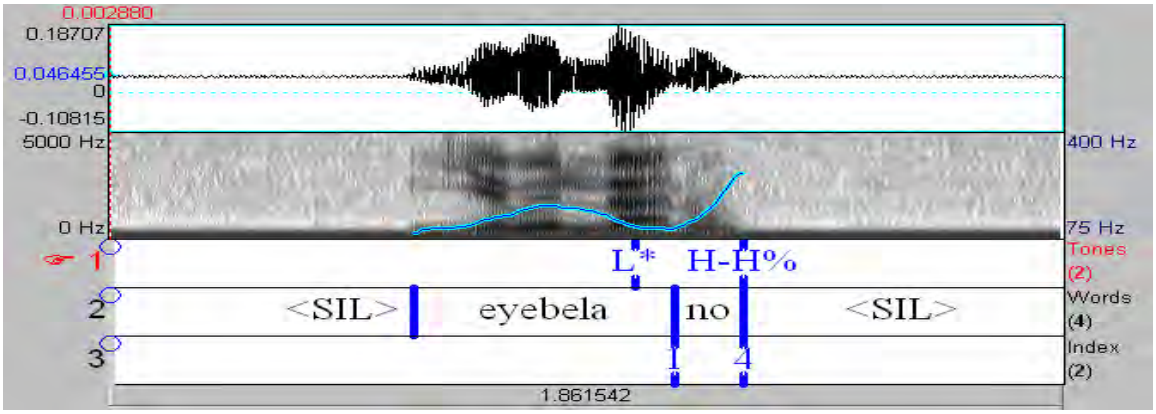


Figure 4.4 L* H-H% for “ **እየበላ ነው?** ”

In the next example, “Africa”, there is a chance placement of unvoiced segments /f/ and /k/ (sounds for which the vocal folds don’t vibrate, so there is no F0) separates the three tonal elements **L* H- H%**. This results in a step-like appearance. In other segmental contexts, the F0 for the successive tonal elements will flow smoothly from one to the next.

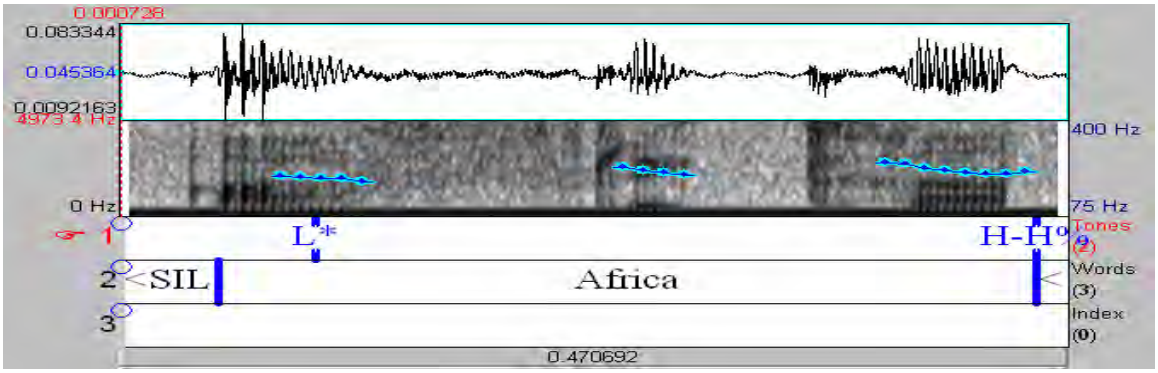


Figure 4.5 L* H-H% for ” **አፍሪካ**”

Speakers have choices, to convey the prominence (greater relative salience) of words and syllables in an utterance. There are times where two tones, both a High and a Low, are associated with the same prominence-marking event. Such events are called **bitonal pitch** accents. For example the bitonal pitch accents (**L+H*** and **L*+H**) have in common a Low tone followed by a High tone as part of the same pitch accent.

This difference in notation (**H*** and **L***) reflects: a perceptual difference in which of the two tones lends more prominence to the pitch accent and alignment characteristics of F0 movement in relation to the pitch accented syllable. The bitonal pitch accents **L+H*** and **L*+H** differ from the single-tone **H*** and **L*** accents by virtue of tone events that precede or follow the (starred) **H** or **L** target of the pitch accent. Specifically, the **L+H*** differs from the **H*** primarily by a more substantial rising pitch movement leading up to the **H*** target, i.e. the presence of a preceding **L** target. The **L*+H** differs from the **L*** primarily by a rising pitch movement that follows the **L*** target, i.e. the presence of a following **H** target.

Intonational Phrases frequently contain more than a single pitch accent. These pitch accents may combine in an Intonational Phrase: any pitch accent label (**H***, **L***, **L*+H**, **L+H***) may precede or follow any of these pitch accent labels. E.g. **H*** may be followed or preceded by another **H***, or by **L*** or **L+H*** or **L*+H**.

A down stepped pitch accent can be indicated by the **!H*** pitch accent label which may be used to indicate that there is a specific tonal relationship between the prominence labeled

!H* and the preceding pitch accent. Specifically, a down stepped **H** pitch accent indicates that the tone of the prominent syllable is realized by a perceptually lower F0 than that of an immediately preceding High tone: the tone has 'stepped down' from the preceding High. While the **!H*** is realized by a lower pitch than the preceding High, it is distinct from the **L*** pitch accent, which is characterized by a pitch excursion down towards the bottom of the speaker's pitch range (for that utterance).

The following pitch accents and phrasal boundaries inventories has been identified for Amharic prosody; we can call it as the tonal part of **AmhToBI** (Amharic Tone and Break Indices).

The **H*** and **L*** serve identical roles but are variants of each other. Similarly **L*+H** and **L+H*** serve identical (usually contrasting and exaggerated stressing) but are used as variants. The down step mark, **!H*** is similar in role to **H*** but is less prominent and invariably follows an accent containing another high accent. The **L*+!H** is similar to **L*+H** but the high accent is less prominent as in **!H*.**

Accent Symbol	Accent name	Accent use
H*	Peak accent	Focus (prominence) marker
L*	Low accent	Focus (prominence) marker
L*+H	Scooped accent	Focus (prominence) marker usually associated with contrast
L*+!H	Scooped downstep accent	Focus (prominence) marker usually associated with contrast following previous high tone
L+H*	Rising peak accent	Focus (prominence) marker usually associated with contrast
!H*	Downstep high tone	Focus (prominence) marker following a preceding H* accent

Table 4.1 Amharic ToBI Pitch accent tones

Note:

H stands for high tonal target

L stands for low tonal target

+ plus symbol used to indicate that the two tones are associated, and form a single unit: a complex (bitonal) pitch accent.

***** indicates that the H/L tone is associated with an accented syllable

Accent symbol	Accent name	Accent use
H-	Phrase accent, iP boundary for index 3	Accent to mark the end of an intermediate phrase
L-	Phrase accent, iP boundary for index 3	Accent to mark the end of an intermediate phrase

Table 4.2 Amharic ToBI intermediate Phrasal Tones -

H- and **L-** tones are used when the phrase end is not a full boundary.

Accent symbol	Accent name	Accent use
L-L%	Low Phrase accent, Low boundary tone	Accent to indicate the end of declarative sentence
L-H%	Low Phrase accent, High boundary tone	Accent to indicate the end of non yes/no questions
H-L%	High Phrase accent, Low boundary tone	Accent to indicate the end of a continuation phrase (commas, lists etc.)
H-H%	High Phrase accent, High boundary tone	Accent to indicate the end of yes/no question
H%	Disfluency marker	Accent to indicate the restart of a phrase after a disfluency break (as in a break due to hesitation)

Table 4.3 Amharic ToBI boundary Tones

Note:

Tables 4.2 and 4.3 list all the phrase-final tonal markers.

- Indicates the tone is associated with phrase.

% Indicates the tone is associated with boundary.

As explained and verified above, the **L-L%** sequence is often found at the end of spoken declarative sentences for Amharic as is for English, especially final sentences in a turn or discourse. Example in the, fig.4.2 above, **አበበ ተመርጠዋል** statement, the intonational phrase has a low phrase accent (L-) and following low boundary tone (**L%**); hence it is marked as **L-L%.**

The **H-H%** sequence is often found at the end of sentence that indicates an utterance is a question for which the speaker expects a yes/no answer. Example in the, figure 4.4 above, **እየበላ ነው?** statement, the intonational phrase has a high phrase accent (**H-**) and following high boundary tone (**H%**), hence it is marked as **H-H%**.

In the results we have seen so far, an F0 peak or valley occurs on the accented syllable. This is often the case, but not always. But whether there is a sharp peak/valley and whether it is aligned with the accented syllable or not, the perception is that the syllable is more prominent than other syllables (or than it would be if it were not accented), and that this prominence is associated with an **H*** or a **L*** tonal event.

Combinations of separate High and Low pitch accents can occur within the same Intonational Phrase as well. Example, **H*** followed by **H***, **L*** followed by **L***, **L*** followed by **H*** or **H*** followed by **L***. Thus we can have, for example:

H* H* L-L%

L* L* H-H%

L* H* L-L%

The **L-L%**, **H-H%**, **L-H%**, and **H-L%** intonational phrase boundary tone contours can occur with any final pitch accent (e.g. **L***, **H***), there is no constraint on what pitch accent can precede which of these boundary tones, i.e., **L*** and **H*** accents can be combined freely with various <phrase tone + boundary tone> sequences.

H* L-L%

H* H-H%

H* L-H%

H* H-L%

L* L-L%

L* H-H%

L* H-L%

L* L-H%

Often, mixed phrase tone/boundary tone combinations (**L-H%** or **H-L%**) do not end in the more extreme high or low levels that are typical of **H-H%** or **L-L%** respectively. Example, the **L-** makes it unlikely that the **H%** will rise to the top of the speaker's range.

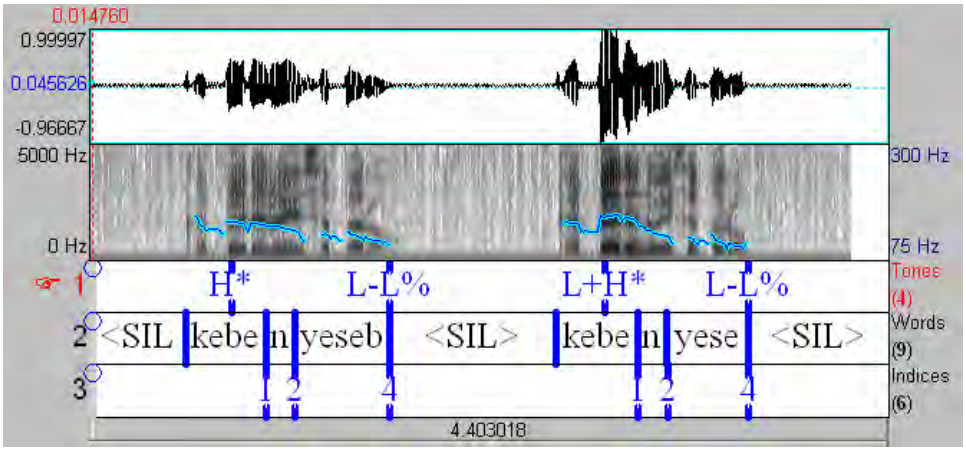


Figure 4.6 H* L-L% vs. **L+H* L-L%** for **ከበደ ነው የሰበረው::**

The contour with **L+H*** might be used in a context where the speaker is trying to make it clear that the person who broke the material was Kebede, as opposed to some other person:

Speaker A: **አበበ ነው የሰበረው::**

Speaker B: (**አይደለም**) **ከበደ ነው የሰበረው::** for the **L+H* L-L%** sequence.

In contrast, the **H*** pitch accent could be expected in response to a question about who broke the material:

Speaker A: **ማን ነው የሰበረው?**

Speaker B: **ከበደ ነው የሰበረው::** for the **H* L-L%** sequence.

The use of the label **L+H*** is constrained to places where a low F0 cannot be accounted for by some other tonal event. For example a rise from a Low tone to a High-toned

prominent syllable can sometimes be accounted for by a preceding **L-L%** phrase accent/boundary tone combination, or a preceding **L*** pitch accent.

4.2 Break Index Determination

Amharic words are normally spoken continuously, unless there are specific linguistic reasons to signal a discontinuity. In Amharic script, several punctuation symbols are used to indicate syntactic as well as semantic features of a group of words. These punctuation symbols can serve as means of indicating junctures between words, juncture between phrases, clauses or between sentences. Here, juncture refers to prosodic phrasing. Junctures cue where do words cohere, and where do prosodic breaks (pauses and/or special pitch movements) occur. Punctuations are not the only means to signal coherence and/or disjunction between the elements of an utterance.

In general, juncture effects that express the degree of coherency or discontinuity between adjacent words are determined by physiology (e.g., running out of breath), phonetics, syntax, semantics, and pragmatics.

In this experiment, an assessment of the punctuations and non-orthographic symbols used in Amharic script was done. And their function was analyzed to identify their role in specifying the strength of breaks in their corresponding position in the speech utterance. For example the punctuation mark **::** (which is a sequence of two colons) in Amharic indicates sentence termination, that is period; the punctuation mark ፡ **(**a single hyphened colon, equivalent to comma in English) is used for example to separate lists, or it is also used to introduce pauses in between the elements of utterance. To add more example, the single quotation mark symbol can be used to make two words to be merged as one by contracting the second word. For example the word sequence **ካው· እንዴ?** becomes **ካው'ንዴ?** Thus, as some of the punctuations in Amharic are prosodic by nature, they are the most reliable candidates for prosodic junctures as well as for pause locations determination.

Junctures can be signaled by phonetic means, such as silence insertion, characteristics pitch movements in the phrase-final syllable, lengthening of a few phonemes in the phrase-final syllable and irregular voice quality.

The junctures between words is not of the same strength. Thus, in this experiment, the following inventory of break indices was assumed based on the above basic concepts. Here, it is assumed that there are basically four distinguishable junctures strength between Amharic words.

0 weak disjunction (or no noticeable junction).

1 normal junction between words

3 semantically, pragmatically, introduced junctures which are basically related to intermediate intonational phrases (iPs).

4 junctures between utterances or at the end of utterance. The break index corresponds to Full intonational phrase.

Starting from this claim, investigation was made on some recorded text used for modeling Amharic prosody. First scan and determine the presence of any pause, location of the pause and duration of the pause. Here, pause can be predicted as presence or absence of silence of greater than 20 ms, though greatly influenced on the current rate setting and individual speakers and other extraneous factors. The following shows part of a recorded passage used to examine the prosodic cues signaled by punctuations as well as characteristics pitch movement.

Disjuncture between words, WS2 and WS3, seem essential pause insertion and WS4 seams less important. The pitch curve is also indicated to reflect the characteristic pitch movement with prosodic phrases. Here, in a standard rate of speaking (or reading a text), if the duration for pauses WS2 and WS3 avoided the intended message to be conveyed by this utterance becomes less emphasis. These junctures are not due to pauses only but also

due to characteristic pitch movements or lengthening of the prosodic phrase-final syllable. Example '**ው**' in '**የሚጠቅመው**' and '**ን**' in '**ሳይሆን**' seem somewhat lengthened.

The characteristic pitch movement, which is indicates by the pitch curve, above the juncture position, shows that for declarative sentences (sentence silence, SS2), goes down to the speaker's lower pitch range.

Thus, break index 3 corresponds to WS2 and WS3 while SS2 is related to break index 4. The other break indices, 0 and 1, are assigned for the word junctures other than prosodic junctures corresponding to break indices 3 and 4. The break index 0 is to be used, for several cases: one is to indicate two words that logically can be combined by contracting the second word, and seemingly form a single word; for example the word sequence **ነው እንዴ?** is frequently pronounced as **ነውንዴ?**, where the initial /e/ of **እንዴ** disappears and the disjuncture between the two words seems nonexisting. Second, if the joint between the two words is difficult to determine due to for example changing to other type of phone (if phoneme substitution occurs). The following are some more examples.

ነው እንዴ?
0 4

ማን ነው የሰበረው?
0 1 4

መጽሐፍ ማንበብ ደስ ይላታል።
1 1 1 4

Thus, prosodic junctures that are clearly signaled (or cued) by pause and/or by characteristic pitch movement can be determined in this manner.

Therefore the juncture types (that can be realized by break indices), proposed from this experiment for Amharic prosody are summarized in the Table bellow. These break indices can be frameworked as part of the AmhToBI (Amharic ToBI) transcription system and specifies an inventory of numbers expressing the strength of a prosodic juncture.

AmhToBI Index	**Description**
0	Word boundary apparently erased.

1	Typical b/n word disjuncture within a phrase
3	End of an intermediate phrase (intonational phrase)
4	End of an Intonational phrase

Table 4.4 Amharic ToBI Break Indices

4.3 Determination of Durations

4.3.1 Pause Duration

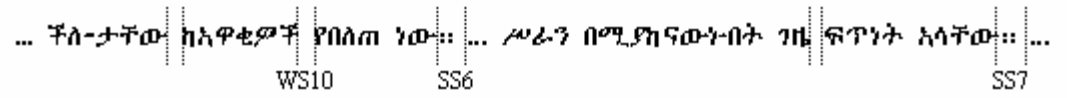
Most of the pauses introduced while making an utterance are significant and must be placed in their appropriate position in time. In this experiment pause durations was calculated based on the following facts:

• Is there a silence between words? This silence between words can be due to prosodic grouping of words or physical constraints or any other artifact pauses.

• Is there a pause silence between sequences of sentences?

• Is there a pause and pause silence with any punctuation marks position?

In this experiment silent pause determination and silent pause duration has been done using PRAAT as shown below.



Silences of utterances' final (SS, pause between two sentences separated by **::**), uttered by five speakers, and silences between words (WS), were randomly selected and measurement were taken to determine their duration. First the starting and ending point of the pause location is marked and automatically the duration of that assumed pause is obtained in units of second.

From this experiment, it was observed that (as can be seen from the tables below) duration assigned to pauses could vary from speaker to speaker for the same pause location.

Results

Table 4.5 gives the pause durations observed between several two adjacent word silences (WSi) taken from a recorded passage by five speakers (SPj). Average pause duration as

calculated from the table is thus, 174ms. Thus, we can conclude from this experiment that 150ms to 200ms is a sufficient duration for pauses between words.

Similarly, Table.4.6 lists the pauses between sentences, taking randomly for 5 speakers. Average pause duration as calculated from the table is thus, 476ms. The average value in the last row of table 4.6 shows how a speaker assigns duration to his/her rhythmic style and clearly shows every individual speaker/reader may produce different pause duration. Again the average value in the last column of table 4.6 indicates a given sentence silence (SSi) is recognized and assigned different pause duration by each speaker (SP_j).

Taking each silence average as $x_{i,}$ the variation and standard deviation is found to be respectively, as Vr = 0.0053, and S.D =0.073 for word prosody pauses and Vr=0.0132, and S.D=0.115 for pause duration between full intonation.

There may be also gap between syllables. E.g. 96ms, 135ms gap was observed from two speakers in the Amharic word **አል...ቀበልም።።** 50 to 150ms seems normal.

	SP1	SP2	SP3	SP4	SP5	**Avg.**
WS1	0.576	0.082	0.416	0.000	0.000	0.215
WS2	0.000	0.268	0.322	0.300	0.000	0.178
WS3	0.323	0.268	0.365	0.000	0.442	0.285
WS4	0.000	0.000	0.289	0.000	0.000	0.058
WS5	0.000	0.000	0.224	0.000	0.000	0.045
WS6	0.448	0.067	0.321	0.000	0.502	0.268
WS7	0.000	0.261	0.385	0.304	0.000	0.190
WS8	0.449	0.316	0.298	0.295	0.109	0.293
WS9	0.420	0.067	0.556	0.031	0.446	0.304
WS10	0.000	0.060	0.000	0.000	0.000	0.012
WS11	0.150	0.000	0.075	0.058	0.080	0.073
Avg.	0.215	0.126	0.295	0.09	0.144	**0.174**

Table 4.5. Silence duration between words in seconds.

	SP1	SP2	SP3	SP4	SP5	**Avg.**
SS1	0.599	0.426	0.442	0.424	0.518	0.482
SS 2	0.824	0.343	0.405	0.999	0.515	0.517
SS 3	0.651	0.284	0.496	0.404.	0.673	0.502
SS 4	0.669	0.282	0.419	0.436	0.524	0.466
SS 5	0.731	0.385	0.444	0.487	0.500	0.509
SS 6	0.705	0.342	0.470	0.406	0.390	0.443
SS 7	0.579	0.289	0.352	0.403	0.430	0.411
Avg.	0.680	0.333	0.433	0.437	0.502	**0.476**

Table 4.6. Silence duration between sentences in seconds

4.3.2 Phoneme Duration

Phoneme duration determination is not an easy task. First of all the number of phoneme in a particular language must be determined with their allophones type. But this is beyond the scope of this thesis project. In this experiment, a sample of phonemes durations was calculated. Here it is assumed that the grapheme to phoneme mapping is, i.e. before syllabification is as follows:

1:1 the six order Fidel (alphabet)

1:2 the CV form, or CC when a grapheme is stressed or lengthening

1:3 e.g. ሏ (lwa, i.e., lua)

1:0 this is rare case

Thus using PRAAT, a recorded word is segmented manually into its phoneme composition one by one, but first the word boundary is exactly specified.

Example:

ማን			ነው			የሰበረው									
ምላን			ንለው			ይለ		ስለብ			ብለ		ርለው		
ም	ላ	ን	ን	ለ	ው	ይ	ለ	ስ	ለ	ብ	ብ	ለ	ር	ለ	ው
102	86	111	47	31	30	75	43	84	66	75	45	86	31	47	75

Table 4.7: Example duration specification for the question "**ማን ነው የሰበረው?** "

According to this experiment on limited number of instance of phonemes duration, the result shows that phoneme duration can be vary from 30 ms to 11sms, for Amharic.

4.3.3 Syllable Durations

Syllable duration can be deduced from the number of phonemes it is composed of and other factors as discussed in chapter 2, once phoneme duration is determined. But in this experiment since durations of all phonemes was not determined, and for practical consideration, duration for most syllable (i.e., Amharic grapheme) was measured separately. And it was found that when Amharic grapheme (Fidel) is spoken separately, as shown in the waveform bellow (for **ሀ...ሆ**), the duration is in average 350ms to 450ms. It means, since Amharic grapheme is by it self a syllable, any Amharic syllable when pronounced separately at normal condition and rate is between 300ms and 400ms in average.

But due to many logical facts, syllable duration may vary from 100ms (even less) to 200ms(or more). For example, figure 4.7a shows the waveform for the Amharic word "**በአፍሪካ** (Africa)", taken from the Amharic sentence "**ኢትዮጵያ በአፍሪካ ቀንድ ትገኛ ለች።።**". In this word there are four syllables according to Amharic template classification (Amharic syllable templates are V, VC, VCC, CV, CVV, CVC, CCVC, and CVCC).

The boundary of the syllables for three different speakers was manually marked and measured accordingly using praat. As is observed from the measured value, syllable length can vary based on the speakers and/or speaking rate, based on the structure of the syllable in a particular language (number of phonemes in the syllable). Syllable length can also vary based on the position of the syllable in a word and/or utterance: fig 4.b shows two "**አፍሪካ**" words taken from the two utterances "**መጀመርያ አፍሪካ ማለት አለብን።።** " and "**አፍሪካ ሰፊ ገበያ ነች።።**" respectively for three different speakers. The variation in syllable length is also inherited from the variation of phonemes due to co articulation, reduction and assimilation.

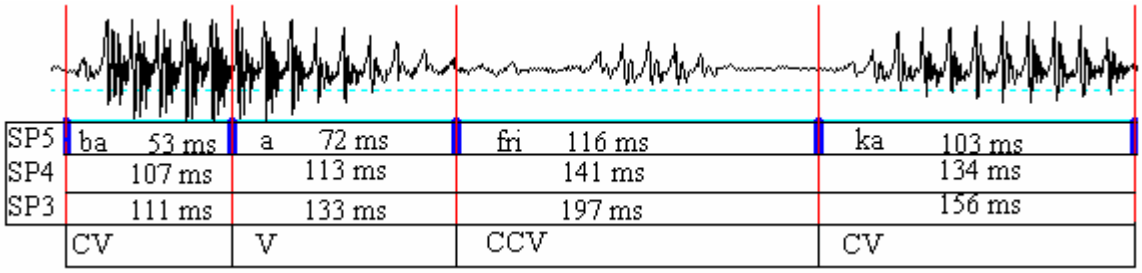
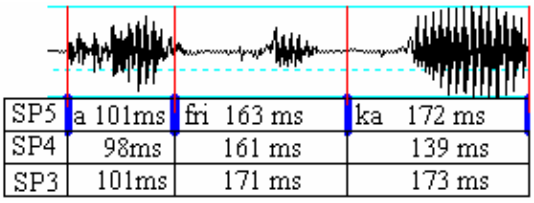


Figure 4.7a Syllables lengths of Africa in "**በአፍሪካ**"



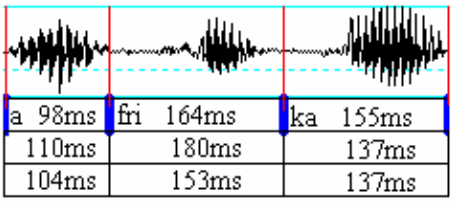


Figure 4.7b Syllables lengths of Africa in "**አፍሪካ**"

	አ			**ፍርኢ**			**ክኣ**		
SP3	133	101	104	197	171	153	156	173	137
SP4	113	98	110	141	161	180	134	139	137
SP5	72	101	98	116	163	164	103	172	155
Average	310/3=103.33ms			482/3=160.67ms			435.33/3=145.11ms		

Table 4.8 Syllables lengths in the Amharic word "**አፍሪካ**"

Following similar approach, the length of each syllable in the Amharic utterance "**ማን ነው የሰበረው?** " is shown bellow.

ማን	ነው	የሰበረው?			
ምኣን	ንኡው	ይኣ	ስኣብ	ብኣ	ርኡው
229ms	108ms	118ms	225ms	131ms	153ms

Of course it is quite true to conclude that most symbols (Fidel, ፊደል) in Amharic scripts corresponds (but not all) to two identifiable phones/phonemes in the form of consonant-vowel (CV).

Consider the Amharic word "ማር" which it's meaning is "honey". If we believe each symbol bears CV and each symbol is a syllable then the aforementioned word is a two syllable word.

4.4 Summary

This chapter described the experiments and results obtained in determining prosodic model of Amharic; the most important prosodic constituents have been experimented:

- Pitch accent – indicating prominence syllables/words.
- Phrasal tones – phrase accent and boundary tone.
- Break index – showing the disjuncture between words.
- Duration of pauses, phoneme and syllable.

CHAPTER 5

5 EVALUATION

Table (5.1) is the complete proposed prosodic model for Amharic. It is proposed that with a systematic application of pitch accents, boundary tones and durations following the model, a synthesized Amharic speech will perform better in terms of both intelligibility and naturalness. It is now time to test this claim. The purpose of this chapter is performing this evaluation.

Types of Break Indices	**Type of Tones**	**Prosodic Units**
0,1,3,4	H*,L*,L*+H,L*+!H,L+H*, !H*	
	L-,H-	iP
	L-L%,L-H%, H-L% H-H%, H%	IP

Table 5.1: Summary of the Prosodic Model for Amharic AmhToBI

There are two evaluations to perform. The first of these is to check the intelligibility test and the second is to perform the naturalness test. For both these tests the test text data is used. The text is synthesized using the Eruxelf Amatets Amharic speech synthesis system. This synthesized speech does not incorporate the prosodic model and will be stored for comparison. Next the synthesized speech is modified using PRAAT to incorporate the model that has just been proposed. Therefore, we will have two synthesized speech utterances that will be used in a listening test and compared. In the rest of the chapter the unmodified synthesized speech is referred to as the synthesized speech and the modified speech (which incorporates the prosodic model) will be referred to as resynthesized speech. Therefore a reference to the synthesized speech will mean a reference to an Amharic synthetic speech without incorporating the proposed prosodic model and to that of the resynthesized speech will be a reference to a synthetic speech that has been modified to put proper pitch accents, boundary tones, breaks and durations at the appropriate places within the synthesized speech utterance.

The listening tests are performed according to the procedure described in section (3.5). Listeners are asked to compare the synthesized and resynthesized speech utterances in random order and asked a set of questions as described in sections (3.5.2.1) and (3.5.2.2).

Now assuming the test text data "**ማን ነው የሰበረው**?" is synthesized without prosody first and a copy of it is modified and resynthesized using the prosody parameters as obtained from the prosody model; the following ToBI like annotation diagram (Fig.5.1) of several layers conceptually illustrates how the prosodic parameters can be fed to the synthesizer.

			H*			L-L%										
ማን			**ነው**			**የሰበረው**										
ምኣን			**ንአው**			**ይአ**		**ስአብ**			**ብአ**		**ርአው**			
ም	**ኣ**	**ን**	**ን**	**አ**	**ው**	**ይ**	**አ**	**ስ**	**አ**	**ብ**	**ብ**	**አ**	**ር**	**አ**	**ው**	
188		111	78		30	118		150		75	131		78		75	
0			1										4			

Table 5.2 Input data (prosody attributes) for: **ማን ነው የሰበረው?**

Here the basic prosodic parameters to be applied, as indicated in Figure.5.1 are the pitch accent **H***, phrase boundary tone **L-L%**, break indices 0, 1, 4 and corresponding durations between prosodic phrases. This can be applied to the synthesized speech to generate the resynthesized speech using PRAAT. Figures 5.2, 5.3, and 5.4, bellow, are presented for illustration purposes, and show the natural sound, the synthesized speech without prosody, and the modified speech with prosody, respectively.

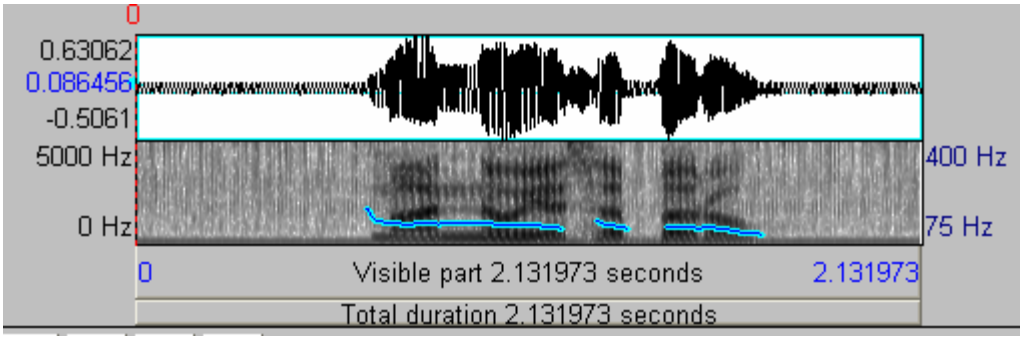


Figure 5.1 Original record data for: **ማን ነው የሰበረው?** "who had broken it?"

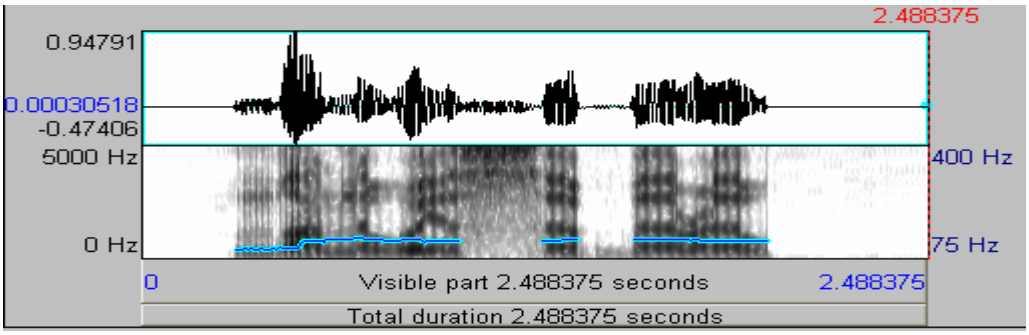


Figure 5.2 Synthesized speech (without prosody) for: **ማን ነው የሰበረው?**

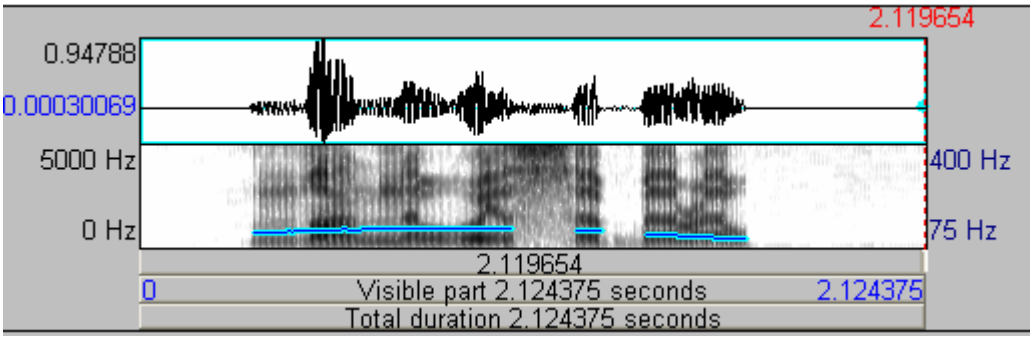


Figure 5.3 Prosodically manipulated sound for: **ማን ነው የሰበረው?**

This procedure is applied to all test data to complete the synthesized and resynthesized speeches to be used for the perceptual evaluation experiment. A total of twelve utterances were prosodically synthesized and provided for listening test. Out of these sentences six are question sentences (three yes/no question and three simple questions), and six sentences are deliberately selected to have emphasis/focus and homographs (words with the same spelling but with different pronunciations and different meanings).

5.1Perceptual evaluation

As explained in chapter 3, different approaches can be combined to evaluate TTS systems as well prosodic models. In this research the simple method for the intelligibility test and MOS for the naturalness test have been used. A total of ten subjects were selected to participate in the perceptual evaluation. Of these eight are male and two are female with seven of them native Amharic speakers and three of them with other mother tongues. The subjects were between 21 and 40 years of age.

The following questions were asked to the subjects to identify the relative qualities of the two speech utterances (synthesized and resynthesized) in terms of intelligibility and naturalness.

A. Intelligibility Test (scoring scheme)

In this test questions are asked to subjects as shown in the list below. The scores given for the subject, for the sentence in question is indicated in the second column of the tables.

1. Repeat the speech that has just been uttered. What?

What?	1	If subject repeats the sentence after listening only once.
	0.5	If subject repeats the sentence only after listening more than once or if he/she misses some words within the sentence.
	0	If subject fails to understand anything from the sentence

2. What type of sentence is it? Type?

Type?	1	If subject identifies correctly the intended sentence type.
	0.5	If subject is unsure about sentence type.
	0	If subject fails to identify anything.

3. Test for comparing which utterance correctly associated emphasis (focus) to specific words within a sentence and whether homographs were properly disambiguated. (Emphasis and Homographs).

Emphasis and Homographs	1	If subject identifies correctly identifies emphasized word or properly disambiguates homograph.
	0	If subject fails emphasized word or fails to properly disambiguate the homograph.

To help the analysis and simplify the tabulating of the results, let us denote the utterances as U, and the listeners as P. Now, under the intelligibility test: 12 utterances were taken,

out of these 6 utterances were purposely designed for the purpose of identifying their sentences type, and 6 utterances for recognizing emphasis and disambiguation.

Table.5.3 and Table.5.4 bellow show the result for intelligibility test for the twelve sentences, without and with prosody respectively. The method of scoring is as explained above. The last row in both tables indicates the average value for each sentence indicating identifiability by the ten subjects. The overall intelligibility test average for this test category is determined by averaging for all Ui values.

Subjects	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12
P1	0	1	1	1	0.5	1	1	0	1	1	1	1
P2	0	1	1	0.5	1	1	1	1	1	1	1	1
P3	0	0	1	1	1	0	1	1	1	1	1	1
P4	1	0	1	1	1	1	1	1	1	1	1	1
P5	0	0.5	1	0.5	1	1	1	1	1	1	1	1
P6	1	0.5	1	1	1	1	1	1	1	1	1	1
P7	1	1	1	0.5	1	1	1	1	1	1	0	1
P8	0	0	1	0.5	1	0.5	0.5	1	1	0.5	1	1
P9	0	0.5	1	0.5	0.5	1	1	0	1	1	1	1
P10	1	0	0.5	0	0.5	1	0.5	0	1	1	1	1
Average	0.40	0.45	0.95	0.65	0.85	0.85	0.90	0.70	1.00	0.95	0.90	1.00

Table 5.3 Intelligibility test for repeating the speech utterance (for utterance without prosody model). (Total Average: **80**%)

Subjects	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12
P1	0.5	1	1	1	0.5	1	1	0	1	1	1	1
P2	0	1	1	1	1	1	1	1	1	1	1	1
P3	0	1	1	1	1	0.5	1	1	1	1	1	1
P4	1	0	1	1	1	1	1	1	1	1	1	1
P5	0	0.5	1	0.5	1	1	1	1	1	1	1	1
P6	1	1	1	1	1	1	1	1	1	1	1	1
P7	1	1	1	1	1	1	1	1	1	1	0	1
P8	0	0	1	0.5	1	0.5	0.5	1	1	1	1	1
P9	0	0.5	1	0.5	1	1	1	0	1	1	1	1
P10	1	0	1	0	1	1	1	0	1	1	1	1
Average	**0.45**	**0.60**	**1.00**	**0.75**	**0.95**	**0.90**	**0.95**	**0.70**	**1.00**	**1.00**	**0.90**	**1.00**

Table 5.4 Intelligibility test for repeating the speech utterance (for utterance with prosody model). (Total Average: **85**%)

Table. 5.5 summarize the test result for identifying sentence type, that is, is an utterance U_i, a declarative sentence, yes/no question or other type of sentence? The value is entered in pair wise: the first value in each pair corresponds to the speech without applying prosody, and the second one for the speech after applying prosody. The minus (-) score indicates the synthesized speech could not be identified for its intended sentence type and score 1 is given otherwise. Score 0 is given if the listener is not able to identify the sentence type. The score 1 indicates, if prosodic features are appropriately incorporated with TTS systems, then, obviously as expected prosody plays important role in synthesizing different types of sentences as required. The last row of Table 5.5 shows the average value for each utterance as evaluated by ten subjects.

Subjects	**U2**	**U5**	**U6**	**U7**	**U8**	**U10**
P1	1, 1	-, 1	-, 1	-, 1	-, 1	-, 1
P2	-, 1	-, 1	-, 1	-, 1	-, 1	1, 1
P3	-, 1	-, 1	-, 1	-, 1	-, 1	-, 1
P4	0, 0	-, 1	-, 1	1, 1	-, 1	-, 1
P5	-, 1	1, 1	-, 1	-, 1	-, 1	-, 1
P6	-, 1	1, 1	-, 1	-, 1	-, 1	-, 1
P7	-, 1	-, 1	-, 1	-, 1	-, 1	-, 1
P8	-, 1	-, 1	-, 1	-, 1	1, 1	1, 1
P9	1, 1	-, 1	-, 1	-, 1	1, 1	-, 1
P10	-, -	-, 1	-, 1	-, 1	0, 0	1, 1
Average	**0.2/0.8**	**0.2/1.0**	**0.0/1.0**	**0.1/1.0**	**0.2/0.9**	**0.3/1.0**

Table 5.5 Intelligibility test for identifying sentence type. (Average: **15%, 95%**)

So the overall intelligibility test average for this test category has been found to be (by averaging for all Ui) 15% for the synthetic speech and 95% for the resynthesized. Improvement gained from this test is 80%.

Table. 5.6 summarize result for emphasis and homograph disambiguation in similar scoring way as above. For example, consider the sentence U2 " **ሰውዮው ሽፍታ' ነው ወይስ ሽፍታ የገደለው?**". In this sentence the word **ሽፍታ** appears twice bearing different meanings. In one hand it can be for the meaning 'bandit' in other hand it can mean a kind

of skin rash. Here, the pronunciation for each case is somewhat different although the spelling is the same. The aim is to determine which word (the first or the second, ሽፍታ) corresponds to bandit or skin rash. This cannot be disambiguated contextually. The only solution is, by determining the stress and duration required to distinguish such differences. Again focuses or emphasis can be realized by changing intensity (loudness) of speech, by changing pitch movement, etc., this kind of phenomena was explained in previous chapter.

Subjects	U1	U2	U3	U4	U9	U11
P1	0, 0	-, 1	-, 1	-, 1	-, 1	-, 1
P2	0, 0	-, 1	-, 1	1, 1	-, 1	-, 1
P3	0, 0	-, 1	-, 1	-, 1	-, 1	-, 1
P4	-, 1	-, 1	1, 1	-, 1	-, 1	-, 1
P5	1, 1	-, 1	-, 1	1, 1	-, 1	1, 1
P6	-, 1	-, 1	-, 1	1, 1	1, 1	1, 1
P7	-, 1	-, 1	-, 1	-, 1	-, 1	1, 1
P8	1, 1	-, 1	-, 1	-, 1	1, 1	1, 1
P9	1, 1	1, 1	-, 1	1, 1	1, 1	1, 1
P10	1, 1	1, 1	-, 1	1, 1	1, 1	1, 1
Average	**0.4/0.7**	**0.2/1.0**	**0.1/1.0**	**0.5/1.0**	**0.4/1.0**	**0.6/1.0**

Table 5.6 Focus or homograph disambiguation (Average: 36.7%, 95%)

In this thesis, six sentences are taken for focus/disambiguating test. The subjects were asked if they could identify where emphasis lies and properly disambiguate words. The meaning of score value (0, -, and 1) is as used in table.5.5. The overall intelligibility test average for this test category has been found to be (by averaging for all Ui) **58.3%** better than the speech without prosody.

b. Naturalness test

This test was made by letting the subjects to select which one of the synthesized speech (with and without prosody) is better and grade it based on MOS. The general question is how close is the synthesized speech to human speech. For this purpose, 12 utterances are taken for Table 5.7 shows the result for naturalness test before applying prosody, and Table 5.8 with prosody. The overall average improvement for this test type has been found to be (by averaging for all Ui) **7**%.

Subjects	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12
P1	1	2	3.5	3	2	3	3	1	2	3.5	3	3
P2	1	1	3.5	1	3	3	3.5	3.5	3	3	2	3
P3	1	1	2	4	3.5	1	3	4.5	3	4.5	2.5	5
P4	2.5	1	4	3.5	2.5	2.5	3	3.5	2.5	3	3	3
P5	1	2	4.5	1	4	3.5	3.5	4.5	3.5	3.5	3.5	4.5
P6	2.5	3.5	4.5	4	3	3.5	3.5	4.5	4	4.5	4	4
P7	1.5	2.5	2.5	1	2.5	3	3.5	3.5	2	3	1	5
P8	1	1	3	2.5	3	2	1	3.5	3	2.5	3	4
P9	2	1	4	1	1	3	3.5	1	1	2.5	3.5	4.5
P10	1	1	1	1	1	3.5	2	1	4	3	3.5	4.5
Average	**0.29**	**0.32**	**0.65**	**0.44**	**0.51**	**0.56**	**0.59**	**0.61**	**0.56**	**0.66**	**0.58**	**0.81**

Table 5.7 Naturalness (MOS) score (for utterance without prosody). (Average: **55%**)

Subjects	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12
P1	1	2.5	3.5	4	2	3.5	4	1	3	4	2	2
P2	1	2	4	1	2	4	4	4	3	3	2	3
P3	1	2	2	4	4	1	4	5	3	5	3	5
P4	3	1	3	4	3	3	3	4	3	3.5	3.5	2
P5	1	3	5	1	3.5	4	4	5	4	4	4	5
P6	3	4	5	3	3	4	4	5	3	5	4	4
P7	2	3	3	3	3	3.5	4	4	2.5	3.5	1	5
P8	1	1	4	2	4	3	1	4	4	4	4	5
P9	2	1	4	1	2	4	4	1	1.5	3	4	5
P10	1	1	3	1	2	4	4	1	3	3	4	5
Average	**0.32**	**0.41**	**0.73**	**0.48**	**0.57**	**0.68**	**0.72**	**0.68**	**0.6**	**0.76**	**0.63**	**0.82**

Table 5.8 Naturalness (MOS) score (for utterance with prosody). (Average: **62%**)

The results of the perceptual tests for limited number of sentences are statistically analyzed. The test results and improvements in average values are summarized in Table.5.9.

Target	Synthesized	Resynthesized	Improvement
What?	80%	85%	**5%**
Type?	15%	95%	**80%**
Emphasis and Homograph Disambiguation	36.7%	95%	**58.3%**
Naturalness	55%	62%	**7%**

Table 5.9 Summary of the Relative Performance of the two Speech Utterances.

5.2 Summary

The perceptual tests conducted to test the improvement in naturalness and intelligibility of a synthesized speech by incorporating prosodic models have shown that there are indeed significant improvements in both tests. The evaluation approach followed in this paper is subjective, meaning perceptual test. Specifically, perceptual intelligibility and naturalness tests (using MOS method) have been performed.

This shows that, the models that have been developed are in the right track and result in significant improvements.

CHAPTER 6

6 CONCLUDING REMARKS

The objective of this thesis was, in short, to develop a prosodic model for Amharic using the Autosegmental – Metrical theory of prosody. The end result expected from this modeling was a better understanding of the prosodical nature of Amharic and an improvement in the intelligibility and naturalness of automatically synthesized Amharic speech. The research proposes six pitch accents, two intermediate phrasal tones and five boundary tones for Amharic. Furthermore, four types of breaks occur in the language. The hypothesis of the thesis work was that applying this model systematically in synthetic Amharic speech shall result in significant improvement in intelligibility and naturalness. The perceptual tests conducted to test this idea have confirmed the hypothesis. Therefore, it is reasonable to claim the proposed model is a reasonable model for Amharic prosody.

This thesis is an original research work on prosodic modeling of Amharic that presented a description of the postlexical prosodic features of the language including pitch accent tones, boundary tones, intonational phrasal tones, breaks and durations in the autosegmental metrical framework of prosodic modeling.

The model was developed by investigating what important cues are used in Amharic prosody. The analysis was performed mainly through the Pratt speech analysis software. It follows the popular prosodic modeling framework ToBI and hence following the custom of speech researchers the model is named AmhToBI. The proposed model has been tested by investigating the improvement in the intelligibility and naturalness of Amharic speech synthesized using the Eruxelf Amatets Amharic speech synthesis program.

The evaluation results clearly show that the model results in significant improvements in all test questions for intelligibility and naturalness. Prosody is an important component of TTS systems and the model proposed by this thesis work can be incorporated into Amharic TTS systems to improve the intelligibility and naturalness of their output. The results can directly be applied for automatic labeling of written text based on punctuation marks and simple pause insertion rules for longer utterances (every fourth word for

example) with significant improvement on understanding and naturalness. As is the case in all other current TTS systems, manual labeling can be used based on the findings of this thesis to obtain good quality output. The results of this system can be a great benchmark to identify the mistakes new learners of Amharic (e.g. foreigners) make when speaking the language by recording pre-labeled text based on the proposed models and analyzing their output against the model.

6.1 Summary of Findings

In this thesis project great effort has been directed towards assessing and modeling the prosodic features of Amharic. The work mainly focused in addressing the basic prosody components in relation to their role in Amharic TTS systems. That is to achieve good quality synthesized speech output incorporating prosody. The research focused on the three important aspects of prosody, which are

- Pitch accents,
- Breaks and
- Duration

6.1.1 Pitch Accents

In Amharic, variations in pitch are used uniformly as suprasegmental quantities, which means that variations in pitch do not result in a change of the (lexical) meaning of any given word. The variations in pitch are used to indicate focus (prominence) and mark phrase boundaries. Thus Amharic is purely a pitch accent language.

Like all pitch accent models, the Amharic prosodic model (AmhToBI) has two levels of pitch – the high (H) tone and the low (L) tone. The high and low tones are relative measures of pitch which are indicated by a rise or fall in the pitch of an uttered speech. The high tone occurs when the pitch rises to a high frequency value close to the speaker's top pitch range and the low occurs when the pitch falls close to the speaker's lower range. These tones can appear either alone or in combination with each other. The high and low tonal targets can be followed by the (*) symbol. This symbol indicates that the target appears on the syllable on which stress falls. Thus, when an (H*) symbol is indicated, it signifies that the pitch rises to the top range of the speaker's pitch range on the stressed

syllable. This has a special significance when the pitch accent contains two tonal targets (which are referred to as bitonal pitch accents). Thus the pitch accent L+H* indicates that the pitch first falls to the speakers low range and then rises again. The (*) indicates that the high tone is placed on the stressed syllable. In other words, the pitch falls immediately before the stressed syllable and then rises when it reaches the stressed syllable. With this understanding the following pitch accents were identified.

Accent Symbol	Accent name	Accent use
H*	Peak accent	Focus (prominence) marker
L*	Low accent	Focus (prominence) marker
L*+H	Scooped accent	Focus (prominence) marker usually associated with contrast
L*+!H	Scooped downstep accent	Focus (prominence) marker usually associated with contrast following previous high tone
L+H*	Rising peak accent	Focus (prominence) marker usually associated with contrast
!H*	Downstep high tone	Focus (prominence) marker following a preceding H* accent

Table 6.1 Amharic ToBI Pitch accent tones

Note that although there are other possible combinations of pitch accent targets, only six appear in Amharic. These pitch accents invariably occur in the middle of a prosodic phrase. The first two symbols are simple focus markers. They are used during speech to attract attention to specific words, such type of focus is referred to as *informative focus* [16]. The last pitch accent is also used in informative focus. However, its pitch target is usually lower than that of a simple **H***. This accent occurs when two or more high tonal targets appear within the same phrase. In this case the first of these targets take the full (H*) accent and the subsequent ones take the **!H*** target. The **!H*** target still shows a rise in pitch but it does not rise as high as that of the **H*** target.

The bitonal pitch accents, **L*+H**, **L*+!H** and **L+H*** are again used to attract attention to specific words within a phrase. They differ from the monotonal targets because they are mainly used in contexts when a speaker is contrasting an idea. This type of focus is referred to as *contrastive focus* [16].

The distinction between informative and contrastive focus can be clarified by looking at this example. Here there are two excerpts of a conversation. In the first one, the first speaker asks the second who ate the bread?. The answer is "Kebede ate bread" and the stress falls on Kebede and the accent type used is **H***, the stress falls on Kebede. This is an example of an informative focus. On the other hand the second conversation starts with the first person claiming that Abebe ate bread, but the second speaker clarifies the mistake by stating that Kebede ate bread. Again the focus is on Kebede, however, the pitch accent that is used is **L+H*** with the purpose of contrasting Kebede and Abebe.

(**H***)
ማን ዳቦ በላ? (Who ate bread?)
ከበደ ዳቦ በላ። (Kebede ate bread.)

(**L+H***)
አበበ ዳቦ በላ። (Abebe ate bread.)
(አይደለም) ከበደ ዳቦ በላ። ((No) Kebede ate bread.)

The list of intermediate phrasal tones is given in table 6.2. These tones occur at the end of an intermediate intonational phrase. An intermediate phrases are characterized by containing at least one Pitch Accent followed by a phrase accent, which is either **L-** or **H-**. The final syllable of this constituent is lengthened more than at a typical phrase-medial word boundary and less than at a Full Intonational Phrase boundary. They are assigned when a listener perceives there is a phrasal end because of syllable lengthening and a distinct rise or fall in pitch but an inspection of the waveform shows that this phrasal end is not large enough to assign it as a full intonational phrase. Furthermore, the gap between the end of an intermediate phrase and the next phrase is very small. These tones are always followed by a boundary index 3.

Accent symbol	**Accent name**	**Accent use**
H-	Phrase accent, iP boundary for index 3	Accent to mark the end of an

		intermediate phrase
L-	Phrase accent, iP boundary for index 3	Accent to mark the end of an intermediate phrase

Table 6.2 Amharic ToBI intermediate Phrasal Tones

Table 6.3 lists the (full) boundary tones in the AmhToBI. These occur at the end of a full phrase boundary. The end of a full prosodical phrase is indicated by a characteristic pitch movement and a large break, which is given the level 4 break index. There are four possible characteristic pitch movements that indicate the end of such a full intonational phrase **L-L%**, **L-H%**, **H-L%** and **H-H%**. The selection of the particular accent indicates the phrase type. The **L-L%** accent is used at the end of a declarative sentence, the **L-H%** indicates the phrase is the last phrase of a question which answer is not Yes/No, the **H-L%** indicates that there is more to come as in lists or as a result of commas and semicolons, the **H-H%** accent is used at the end of a yes/no question.

Accent symbol	Accent name	Accent use
L-L%	Low Phrase accent, Low boundary tone	Accent to indicate the end of declarative sentence
L-H%	Low Phrase accent, High boundary tone	Accent to indicate the end of non yes/no questions
H-L%	High Phrase accent, Low boundary tone	Accent to indicate the end of a continuation phrase (commas, lists etc.)
H-H%	High Phrase accent, High boundary tone	Accent to indicate the end of yes/no question
H%	Disfluency marker	Accent to indicate the restart of a phrase after a disfluency break (as in a break due to hesitation)

Table 6.3 Amharic ToBI boundary Tones

The only remaining accent is the **H%** accent. Unlike the former three, this accent is used at the beginning of a full phrase. Its purpose is to indicate a continuation of a previously interrupted phrase. This interruption can occur because of hesitation, coughs etc. Thus this accent is called a disfluency marker.

6.1.2 Breaks

Breaks are an important feature of prosody. They mark the distinction between words, phrases, sentences, etc. In the ToBI framework the possible breaks are indicated by numbers starting from 0 up to the desired level. The numbers indicate the relative strength of the breaks. The total number of breaks varies from one language to another. In the AmhToBI model a total of four breaks have been identified as shown in table 6.4. Each of these breaks appear at particular points within an utterance and together with boundary tones are used to group utterance segments. It appears that in Amharic the breaks are used to separate words, intermediate phrases and intonational phrases. The break index 0 occurs between two words which are contracted together into an apparently single word, as in "ማን ነው?" into "ማነው?". The break index 1 occurs between words within a phrase. The break index 3 occurs at the end of an intermediate phrase. It occurs together with either the **H-** or **L-** intermediate phrasal tones. The last break index is 4 and it occurs at the end of full intonational phrases.

ToBI Index	Description
0	Word boundary apparently erased.
1	Typical b/n word disjuncture within a phrase
3	End of an intermediate phrase (intonational phrase)
4	End of an Intonational phrase

Table 6.4 Amharic ToBI Break Indices

The duration of each of these pauses is not an absolute quantity, rather depends on the speaker, mood of the speaker and environment, but with in the context of a single utterance they maintain the relative strength relationships according to the assigned numbers.

6.1.3 Duration

The final aspect of prosody that was investigated in this research was an indicative value for the duration of speech units. The research considered the durations for pauses, phonemes and syllables. The results indicate that the average duration of pauses between sentences (index 4) was 476ms, at the end of intermediate phrases (index 3) was also 476ms, between words (index 1) 174ms and between contracted words (index 0) 0ms.

The average syllable duration was found to vary between 100 to 200ms and phoneme duration was found to be between 30 to 115ms.

6.2 Applications

The findings of the thesis research has been summarized in the previous section. Simply put, it is a more or less complete model for Amharic prosody that lists the possible variations in pitch, the various types of breaks and indicative durations for the pauses, syllables and phonemes.

This model, as stated briefly in previous sections, can be used for two distinct purposes

- As a framework to study the process of producing Amharic utterances
- As a standard to tag the prosody of Amharic into written text so that it can be used in Amharic speech synthesizers.

In the first application scenario, researchers and teachers can use the model to explain the process of Amharic prosody. In other words it presents a framework for researchers and teachers to identify and mark Amharic prosodic cues using the proposed framework to explain and teach Amharic prosody. It can also be used as a tool for developing software for teaching proper Amharic prosody to (for example) foreigners and analyzing the uttered speech of learners in terms of prosody.

The initial proposed model can also serve as a starting point for studying Amharic prosody and to develop further applications based on the model. The model can be used to study the distribution of pitch accents among differing groups of Amharic speakers, study the role of prosody in Amharic understanding, study the role of prosody in learning Amharic, study the difficulty among differing type of students (based on ethnic background and native language) in producing proper Amharic prosody, study the role of prosody in speech recognition systems etc.

Some of the applications that can be developed based on the model include computer algorithms and programs to perform automatic generation of prosodic labels for Amharic text (pre-synthesis labeling), automatic prosody label generation for uttered speech (post-

synthesis labeling), performance comparison of Amharic speech synthesizers based on prosody accuracy etc.

In the second scenario, the model can be incorporated into Amharic speech synthesizers by manually or automatically (requires an automatic labeling program) Amharic text and modifying the pitch, inserting pauses and adjusting durations according to the rules of the labels. This second application of the research is expected to produce better synthetic speech in terms of both intelligibility and naturalness as verified through the evaluation results in chapter five of this research.

6.3 Limitations of the Thesis Work

Although every effort has been made to make this project work as complete as possible, nevertheless, time and financial constraints have made it impossible to include some of the possible interesting research questions from being asked. In this section, some of the limitations of the work have been indicated in the hope that other researchers will take up proceed with an all inclusive prosodic study for Amharic.

The thesis work focused on the post-lexical prosody features of Amharic, simply because post lexical features are more important in terms of their effect in the intelligibility and naturalness of a speech. However, lexical prosody is also important and hence must be properly studied.

Even within the context of post-lexical prosody, the research did not consider all possible prosodical tools, namely intensity. Literature [4] suggests that intensity plays a minor role in prosody as compared to pitch and durations. However, intensity does play a role and its study is important in the full specification of the prosody of a language.

The model development process also suffers from data limitations. The model and test data used was only from read text data and no spontaneous utterance was used. This may have prevented in the omission of some prosodic cues which may occur during spontaneous speech. Furthermore, the text data corpus size was relatively small. Once again some important Amharic prosodic cues may have not occurred in the data. Additionally, the number of speakers used for the speech corpus was small, which may have limited the number of observed prosodic cues.

6.4 Future work

Considering that this work on prosody model for Amharic is a first attempt for its kind and that prosody covers broad concepts and involves complex techniques for its solution, the results only cover a small portion of the total problem. Furthermore, the usability of the models can be improved to a great extent with further development, therefore, I now put my recommendations on what can be done to further the initial prosodic model proposed in this research.

- **Syllabification of words:** will greatly improve prosodic modeling that is the syllabic structure of Amharic along with the segmental prosody of Amharic should appropriately studied and modeled. The proper identification of Amharic syllables is the first step in determining where the stress point should fall within a word. In addition, homographs in Amharic are identified through syllabification as has been exemplified in chapter 5.
- **Automatic prosodic labeling:** can be attempted based on the models presented together with automatic semantic and syntactic understanding and further studies on where post-lexical stresses fall on words,
- **Intensity Modeling:** should be incorporated into the prosodic model
- **Lexical (word level) prosody:** needs to be studied
- **More accurate phone duration models:** can be developed by considering their position within a word and the co-articulation effect of preceding and following phones. For this, a statistical measurement of duration for each phone considering the preceding and following phone (a tri-phone) can be made on a very large speech corpus. The average duration of each tri-phone can be computed from this large corpus together with the standard deviation and this can be incorporated into TTS systems for a more realistic duration modeling.
- **Models for other Amharic Dialects:** need to be developed. The model that has been developed in this thesis is for Amharic as is spoken in Addis Ababa. However, there is no guarantee that this model will work equally well for other dialects, e.g., Wollo Amharic, Gondar Amharic.
- **Amharic speech prosody analysis software:** should be developed to analyze the prosody of spoken utterances. This application can be used to analyze whether a

speaker is generating speech in line with the standard Amharic prosody, which might be useful to identify weaknesses in the prosody of a speaker (e.g. Foreigners that are learning Amharic).

ANNEX

Annex I. REFERENCES

1 Daniel Jurafsky & James H. Martin: **Speech and Language Processing**: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, *Pearson Education, inc., 2000*

2 Thomas F. Quatieri, **Discrete-time Speech Signal Processing: principles and practice**, *Pearson Education, inc., 2002.*

3 B. Gold, N. Morgan, **Speech and Audio Signal Processing**: Processing and Perception of Speech and Music*, John Wiley and Sons Inc., 1999*

4 Huang, Acero, Hon, **Spoken Language Processing**: A Guide to Theory, Algorithm, and System Development, *Prentice-Hall, inc., 2001*

5 Dutoit, T., **An Introduction to Text_To_ Speech Synthesis**, 1997, Kluwer Academic Publishers

6 Ronald W. Schafer, John D. Markel, **Speech analysis**, *IEEE, inc., 1979.*

7 L.R. Rabiner /R.W.Schafer, **Digital processing of Speech signals,** *Prentice-Hall, inc.,1978.*

8 Beckman, M.E. and G.M. Ayers, **Guidelines for ToBI labeling**, 1997, http://www.ling.ohio-state.edu/phonetics/ToBI/main.html.

9 Fujisaki, H., "**Prosody, Models, and Spontaneous Speech" in Computing Prosody**, *Y. Sagisaka, N. Campbell, N. Higuchi, Editors, 1997, New York, Springer.*

10 Kompe, R., **Prosody in Speech Understanding Systems**, *1997, Berlin, Springer.*

11 Liberman, M., **The Intonation System of English**, *PhD Thesis in Linguistics and Philosophy, 1975, MIT, Cambridge.*

12 Sonntag, G., T. Portele, and B. Heuft, "**Prosody Generation with a Neural Network: Weighing the importance of Input Parameters**," *Proc. Int. Conf. On Acoustics, Speech and Signal Processing, 1997, pp. 930-934.*

13 Ladd, D.R.. **Intonational phonology. Cambridge***, 1996, CUP.*

14 Jun, S.-A. **Prosodic Typology: The Phonology of Intonation and Phrasing. Oxford,** *2005. OUP.*

15. Bjorn Gamback, Gunnar Eriksson, **Natural Language Processing at the School of Information Studies for Africa.**

16. Samantha J. Hellmuth., **Intonational pitch accent distribution in Egyptian Arabic,** Doctoral Thesis, *School of Oriental and African Studies, University of London, March 2006*

Annex II. THE MODEL TEXT CORPUS

ds1
መጥበቅ ያለባቸው ቃላት የማጥበቅያ ምልክት ካልተደረገባቸው በስተቀር በጽሑፍ ወቅት የሚሰጡት የሓሳብ ፍች የተድበሰበሰ ሊሆን ይችላል። በተለይም አማርኛ ቋንቋ ሁለተኛ ቋንቋቸው ለሆኑ ያገራቱ ዜጎች ሊቸገሩ ይችላሉ።

ds2
ልጆች ከህጻንነት ጀምሮ በቤት ውስጥ ሥራቸውን በሓላፊነት □የወሰዱ መሥራት መለማመድ ይኖርባቸዋል። ይህም የሚጠቅመው ቤተሰቡን ለመርዳት ብቻ ሳይሆን ለወደፊት ህይወታቸውና ለሕብረተሰቡ ነው። በአገራችን አብዛኛው ግዜ ለልጆች ሓላፊነት አይሰጥም። ሁል ግዜ እንደማያውቁ ይቆጠራሉ። ነገር ግን ልጆች ለቤተሰብ ኑሮ መሻሻል ብዙ አስተዋፅኦ ሊያደርጉ ይችላሉ። ልጆች ለመስራት ዕድል ሲያጋጥማቸውና በአቅማቸው መጠን ሥራወችን እንዲሰሩ ሲደረጉ ከአዋቂወች ባላነሰ መጠን ይሠራሉ። ልጆች አንድን ነገር በቀላሉ የመቅዳትና የመረዳት ችሎታቸው ከአዋቂወች የበለጠ ነው። ስራን በሚያከናውኑበት ግዜ ፍጥነት አላቸው። ሰርተውም ስላልጠገቡ የመሰልቸት ዝንባሌ አይታይባቸውም። ስለዚህ በሚሠሩት ሥራ መናቅና መጠላት የለባቸውም። ከሌሎች እኩል መታየት አለባቸው። ለሚሰሯቸው መልካም ሥራወችና ለሚሰጡአቸው አስተያየቶች ሁሉ መወደድ፣ መከበር ማበራታቻ መሰጠት አለባቸው። እንደዚህ ከሆነና ልጆች ሰሜታቸው ተጠብቆ ካደጉ የቤተሰብ ኑሮ ከማማሩም በላይ ያ ቤተሰብ ጥሩ ዜጐችን ያፈራል ማለት ነው።

ds3
የሰው ስብእና ማለት አንድ ሰው በውስጡ የያዘውና ከሌሎች ሰወች ልዩ የሚያደርገው ጠቅላላ ፀባይ ነው። የሰው ስብእና ከህፃንነት ወራት ጀምሮ ያለና እስከ አዋቂነት እድሜ ድረስ የሚቀጥል የማያቋርጥ ሂደት ነው። የአንድን ሰው ጥሩ ስብእና የሚፈጥሩ ብዙ ጤናማ የሆኑ ስሜቶች አሉ። እነዚህም ለምሳሌ ያህል ትክክለኛነት፣ መንፈስ ጠንካራነት፣ ብርቱ ስሜት፣ በራስ የመተማመን መንፈስ የመሳሰሉ ይጠቀሳሉ።

ds4
ውድ ወንድሜ ዝናው እንደምን ከርመሃል?
እኔ ካንተ ከጓደኛዬ ናፍቆት በስተቀር ለጤናዬ በጣም ደህና ነኝ።

ds5 **አንተ ተነስ እየመጡ ናቸው!**
አንተ ተነስ እየመጡ ናቸው።

ds6 **መፅሓፍ ማንበብ እፈልጋለሁ።**
መፅሓፍ ማንበብ ደስ ይለኛል።
መፅሓፍ ማንበብ ደስ ይላታል።
መፅሓፍ ማንበብ ደስ ይለዋል።
መፅሓፍ ማንበብ ደስ ይላቸዋል።

ds7. **ኢትዮጵያ በአፍሪካ ቀንድ ትገኛለች። የህዝብዋ ብዛት ከ70 ሚልዮን በላይ ነው። ትክክለኛ ቁጥሩ አሁን በሚደረገው የህዝብ ቆጠራ ይታወቃል።**

ds8. **የመማርና የማስተማር ሂደቱ መገምገም አለበት፣ ምክንያቱም የሥራ ግምገማ ጠቃሚ ስለሆነና ገምጋሚና ተገምጋሚ ስለሚማማሩበት ነው።**

ds9
መጀመርያ አፍሪካ ማለት አለብን። መጀመርያ አንድ ነች ማለት አለብን። ያ ብቻ ነው ጥንካሬያችን። ያ ብቻ ነው ብርታታችን። ያን ግዜ አፍሪካ አንድ ስትሆን እንደዛሬው የአምሳ አገሮች ሹክሹክታ ሳይሆን ያንድ አፍሪካ ድምፅ ያስገመግማል። ያን ግዜ አለም ያዳምጠናል። ያን ግዜ ብርታታችንን ይገነዘቡታል። አፍሪካችን ገና የምትለማ ነች። ይህንን ገንቡ፣ ያንን አፍርሱ፣ስላሉን ሳይሆን የሚበጀንን እንገነባለን። የማይጠቅመንን እናፈርሳለን። አፍሪካ ሰፊ ገበያ ነች። ይህን በዚህ ሽጡ ስላሉን ሳይሆን በሚያዋጣን እንሸጣለን፣ ይህን በዚያ ግዙ ስላሉን ሳይሆን በሚስማማን እንሸምታለን። አፍሪካችን ድንግል ናት። ጥሬ ሃብት የኛ ነው፣ ከፋን ቢሉ እንበላልጣቸዋለን፣ አቃረን ቢሉ እናማርጣቸዋለን።

ያን ጊዜ ዛሬ የተሳለቁብን፣ ዛሬ የተዘባበቱብን ቀለማችንን አይተው እድፍ፣ ኋላቀርነታችንን ገላምጠው ዝንጀሮ ያሉን ሁሉ ወዳጅነታችንን ይመኙተል፣ ወንድማማችነታችንን ያከብሩታል።

ds10
"ታሪካችን ይኸው ነው። እየጣሉ መውደቅ ነው። የክብርና የመስዋትነት ሞት ማለትም ይህ ነው። በክብር ቀደመኝ እንኳን በጀርባው በኩል ተመትቶ አላገኘሁት።" ብለው በግል አሽከሮቻቸው አሸክመው ወደትውልድ መንደራቸው አስወስደው ቀበሩ።

Questions (q)
Wh-question:- What did you give him?
Alternative questions:- Did you give him a book, a pen, or a pistol?
Tag questions:-You give it to him, didn't you?
Yes/no questions:- Did you give him a book?

q1 **የቃል ክፍል ማለት ምን ማለት ይመስላችኋል?**

q2 **ከበደ ከሄደበት አልተመለሰምን?**

q3 **ምን አይነት ምግብ ነው የምትመገቡት?**
ምን አይነት ምግብ ነው የምታዘወትሩት?
ምን አይነት ምግብ ታዘወትራላችሁ?

ምን አይነት ምግብ ነው የምትመገብ?
ምን አይነት ምግብ ነው የምታዘወትር?
ምን አይነት ምግብ ታዘወትራለህ?

q4 **ምን አይነት ምግብ ነው የምትመገቢ?**
ምን አይነት ምግብ ነው የምታዘወትሪ?
ምን አይነት ምግብ ታዘወትሪያለሽ?
ምን አይነት ምግብ ትበያለሽ?
ምን አይነት ምግብ ትመገብያለሽ?

Yes/No question (ynq)

ynq1
እየበላህ ነው።

እየበላህ ነው? (በቁጣ)

እየበላ ነው።
እየበላ ነው? (የቁጣ ጥያቄ)
እየበላ ነው? (ልማዳዊ/ቀጥታዊ ጥያቄ)
ምን እየበላ ነው? (ቁጣ ጥያቄ)---የምግብ አይነት/እየበላ አይደለም
ምን እየበላ ነው? (ቀጥታዊ ጥያቄ)---የምግብ አይነት/እየበላ አይደለም

ynq2. Is that so!
Are you sure?

ynq3. እዉነትህን ነው?! (የመደንገጥ ጥያቄ)
እርግጠኛ ነህ? (የመደንገጥ ጥያቄ)
እርግጠኛ ነህ? (ልማዳዊ ጥያቄ)

ynq4. ነው እንዴ?! (የመደንገጥ ጥያቄ)
ነው እንዴ... (መገረም)

ynq5. መፅሐፍ ማንበብ ትወዳለህ?
መፅሀፍ ማንበብ ትወዳለህ እንዴ?
መፅሀፍ ማንበብ ደስ ይለሃል?
መፅሀፍ ማንበብ ደስ ይለሃል እንዴ?

መፅሀፍ ማንበብ ትወጅያለሽ?
መፅሀፍ ማንበብ ትወጅያለሽ እንዴ?
መፅሀፍ ማንበብ ደስ ይልሻል?
መፅሀፍ ማንበብ ደስ ይልሻል እንዴ?

መፅሀፍ ማንበብ ደስ ይላችኋል?
መፅሀፍ ማንበብ ደስ ይላችኋል እንዴ?
መፅሀፍ ማንበብ ትወዳላችሁ?
መፅሀፍ ማንበብ ትወዳላችሁ እንዴ?

Homograph statements (hgs)

hg1
ገና ገና ነው አልደረሰም።
ገና ገና አልደረሰም።
ገና አልደረሰም ገና ነው።
ገና አልደረሰም።
hg2 ገና ነው አልደረሰም ገና።
ገና ነው አልደረሰም ገና።
ገና ገና አልደረሰም።
ገና አልደረሰም።

hg3 በድምቀት የሚከበረው በዓል ገና ነው።
በድምቀት የሚከበረው በዓል ገና ይባላል።
በጥምቀት ዋዜማ የሚከበረው በዓል ገና ነው።

በጥምቀት ዋዜማ የሚከበረው በዓል ገና ይባላል።

hg4 አበበ በጠና ታሞዋል።
አበበ ሕመሙ ጠና ብሎበታል። አበበ ሕመሙ ጠና።
ጠና ያለው ሽማግሌ በጠና ታመዋል።

hg5 እናንተ አንድ ነገር በሉ እንጂ።
እንግዶቹ ምግቸውን በሉ።

hg6 ዘውዴ ነገ እመጣለሁ አለ።
ዘውዴ ቤት አለ።

hg7 ደራሲው ቆንጆና የተዋጣለት ድርሰት ደረሰ።
የፈተና ግዜ ቶሎ ደረሰ።
የፈተና ግዜ መቼ ደረሰ?

hg8 በፍጥነት ሲሽከረከር የነበረው መኪና ሰው ገዳ።
አበበች በፍጥነት ሲሽከረከር የነበረ መኪና ገዳ አርጓታል።

hg9 በቀለ ኪስዋን ቆንጆ አድርጎ ለጋ።
የበቆሎው እሸት አልደረሰም ገና ለጋ ነው።

hg10 ወተቱ ቶሎ ረጋ። ዉሃው ግን ረጋ ብሎ እየሞቀ ነው።

hg11 የኢትዮጵያ መሬት ወጣገባ ይበዛዋል።
ልጁ ወጣገባ ያበዛል። ሴትየዋም ወጣገባ አበዛች።

hg12 ቆሻሻውን ወጣ ብለሽ ጣዩ።
ውሻው ከግቢው ወጣ።

hg13 የአቶ ደሳለኝ ቤት ወደ ውስጥ ገባ ያለ ነው።
የሆነ ሰው ወደ ቤት ገባ።

hg14 ስፖርተኛው ዋና መዋኘት ይወዳል።
ስፖርተኛው ዋና ተጫዋች ነው።

hg15 እግርሽን ሰብስቢው አላት።
አስካለ ብዙ መፅሀፍት አላት።
hg16 እግርህን ሰበስበው አለው።
ተስፋይ ብዙ መፅሐፍት አለው።

hg17 መሬቱ ለኔ ይገባኛል። ችገሩ ምን እንደሆነ ይገባኛል። (stress on two syllables)
ለምለም እቤት አለች፣ ሆኖም ግን አልመጣም አለች።

hg18 ማር ሲበላ ይጣፍጣል፣ ሆኖም ከበደ ማር ሲበላ ያመዋል።
ጠጅ ሲጠጣ ይጣፍጣል፣ ነገር ግን ከበደ ጠጅ ሲጠጣ ያቅለሸልሸዋል።

hg19 አንተ በላ ተናገር፣ ምን አፍህን ይለጉመዋል?
ከበደ ምሳውን በላ።

hg20 ፈረሱ በመኪና ተገጭቶ ሞተ። ህገወጥ ቤቶቹ ፈረሱ።
ልጆች በሚሠሩት ሥራ መናቅ የለባቸውም። ወጣቶች ሥራ መናቅ የለባቸውም።
hg21 አበበ መፅሀፍ ደረሰ። ሆኖም ሳይጨርስ ግዜው ደረሰ።
አበበ ራሱን ገዳ። መፅሀፉን በግዜው ስላልደረሰ ኑሮውን ገዳ አርገታል።
አበበ ዕድሜው ለጋ ነው። ሆኖም ግን ትልቅ ኪስ ለጋ።
አበበ አይኖቹ ቀላ ያሉ ናቸው። ሆኖም አንደኛው አይኑ በጣም ቀላ።

hg22 ሰውዮው ሽፍታ ነው ወይስ ሸፍታ የገደለው።
መሰረት አንድ እንጀራ በወጥ ግጥም አርጋ በላች።
መሰረት ግጥም መግጠም ትችላለች።
hg23 እንግደቹን እቤት ግቡ በላቸው።
በገቹን ጅብ በላቸው።
በሬው መላ አካላቱ ተቃጥሏል። ህይወቱ ሳያልፍ መላ ይፈለግለት።

ቃል	ሲጠብቅ	ሲላላ	ቃል	ሲጠብቅ	ሲላላ
መላ			ደረሰ		
በላ			ገዳ		
ግጥም			ለጋ		
ሸፍ□			ቀላ		
መናቅ			ረጋ		
ገና			ወጣገባ		
ጠና			ወጣ		
በሉ			ገባ		
አለ			ዋና		
አላት			ይገባኛል		
አለው			አለች		
ሲበላ			መናቅ		
ሲጠጣ			ፈረሱ		
በራ					
ለማ					
በማይረሳው					

Phrasal statement (ps)

ps1
ደራሲ ሀዲስ አለማየሁ፣ እውቁ ኢትየጵያዊ ደራሲ፣ በማይረሳው ሥራቸው፣ ዘወትር ሲወሱ ይኖራሉ።

ps2
ደራሲ ሀዲስ አለማየሁ፣ እውቁ ኢትዮጵያዊ፣ ደራሲ በማይረሳው ሥራቸው፣ ዘወትር ሲወሱ ይኖራሉ።

ps3

ለግቢው፣ የብረት መዝጊያና ቁልፍ፣ ለሳሎኑ፣ ሶፋ፣ የምግብ ጠረጴዛና ቢሮ፣ ለመታጠቢያ ቤቱ፣ ሳሙናና ፎጣ አሟላ።

ps4
በቀደምት የሰዋሰው ስራወች የአማርኛ ቃላት በስምንት ተከፍለው ይገኛሉ። እነርሱም፡- ስም፣ ግስ፣ ቅፅል፣ ተውሳከ - ግስ፣ መስተፃምር፣ ተውላጠ - ስም መስተዋድድና ቃለ አጋኖ ናቸው።
ይህም አመዳደብ ከሌሎቹ ቋንቋወች የሰዋሰው ስራወች የተወረሰ እንጂ የአማርኛ ቋንቋ ባህሪ ተጠንቶ አለመሆኑ ግልፅ ነው።

ps5
እስከአሁን ያየነው ሁኔታ የአማርኛ ቃላት በአምስት **ዋና ዋና** ክፍሎች ሊመደቡ ይችላሉ ወደሚል መደምደሚያ የሚያደርስ ነው። ክፍሎቹም፡- ስም፣ ቅፅል፣ ግስ፣ ተውሳከ - ግስ፣ መስተዋድድ ናቸው።

ps6 & ps66
ከአንድ በላይ ፍች ያለው ቃል የቱ ነው?

ሀ. ተረጋገጠ	or	1. ተረጋገጠ
ለ. ጥፍር		2. ጥፍር
መ. ሰንበር		3. ሰንበር
ሠ. ሁሉም		4. ሁሉም

ps7 ልዩነታችን ውበታችን፣ ውበተታችን አንድነታችን ነው።

Annex III. THE TEST TEXT CORPUS

U1 ገና አልደረሰም።

U2 ሰውዮው ሽፍታ ነው ወይስ ሽፍታ የገደለው።

U3 የፈተና ግዜ ተሉ ደረሰ።

U4 የኢትዮጵያ መሬት ወጣ ገባ ይበዘዋል።

U5 ምን ዓይነት ምግብ ነው የምትመገብ?

U6 መጽሓፍ ማንበብ ደስ ይልሃል?

U7 መጽሓፍ ማንበብ ደስ ይልሻል?

U8 እየበላ ነው?

U9 አበበ ነው የሰበረው።
ከበደ ነው የሰበረው።

U10 ማን ነው የሰበረው?
ከበደ ነው የሰበረው።

U11 መገምገም

U12 አፍሪካ