THESIS

TOWARDS A TEXT-T0-SPEECH SYSTEM FOR MOORE: PROCEDURE AND ANALYSIS

Benemanda Medard Zoungrana

May, 2012

TOWARDS A TEXT-T0-SPEECH SYSTEM FOR MOORE: PROCEDURE AND ANALYSIS

by

Benemanda Medard Zoungrana

B.A., University of Ouagadougou, Ouagadougou, Burkina Faso, 2005

A Thesis

Submitted to the Graduate Faculty

of

St. Cloud State University

in Partial Fulfillment of the Requirements

for the Degree

Master of Arts

St. Cloud, Minnesota

May, 2012

This thesis submitted by Benemanda Medard Zoungrana in partial fulfillment of the requirements for the Degree of Master of Arts at St. Cloud State University is hereby approved by the final evaluation committee.

_____
Chairperson

_____

_____

_____
Dean
School of Graduate Studies

TOWARDS A TEXT-T0-SPEECH SYSTEM FOR MOORE: PROCEDURE AND ANALYSIS

Benemanda Medard Zoungrana

The goal of this study is the creation of a database of diphones that can be used as input to synthesize speech sounds in Moore.

> Diphones are speech units that begin in the middle of the stable state of a phone and end in the middle of the following one. Since diphones contain sections of neighboring phones, they contain the contextual information which determines the differences between allophones of a phoneme. When diphones are concatenated, this information is preserved. (…). The diphone represents a compromise between quality and effort in database construction. (Gibbon, 2010, p. 2)

Speech synthesized this way is known as text-to-speech and it is one of the speech synthesis technologies available today, the other one being speech recognition. Usually these services are designed for languages like English or French among others. This thesis focuses on the text-to-speech side of speech synthesis, that is, the ability for computers to read a written text in a loud and intelligible way. In order to do that, a computer needs a set of coded instructions or algorithm telling it how to convert these written words or sentences into intelligible voices. There are many algorithms, techniques or methods that are used in text-to-speech (TTS) synthesis.

The most common of them is the concatenative method which uses prerecorded speech sound of units of variable length and a set of units' database to produce synthesized speech. These units can be words, syllables, diphones or just phonemes. MBROLA is a speech synthesizer software that is based on the concatenative method. The MBROLA software uses as input a set of diphones, called diphone database, and a prerecorded sound voice, called voice, in order to produce synthesized speech of a particular language.

In addition to creating a diphone database, this study aims at creating a voice that can be used to produce synthesized speech in Moore, a West African

gur language which is widely spoken in Burkina Faso. The study being presented here uses the MBROLA software as the framework for the implementation part.

The present thesis offers a brief overview of the history of text-to-speech in general, and in Africa in particular, while discussing the Moore language in general with a linguistic perspective. It describes the phonetics, the phonology, and syllable structure of the language pointing out the transparency of its orthography. The study ends with an evaluation of the database and voice quality to determine the performance of the system in terms of intelligibility and naturalness. Lastly, some major issues encountered during this study are discussed as well as some suggestions for future work.

The result of this study shows that the sound of the synthesized speech of Moore language through MBROLA is quite intelligible and natural. Hence this thesis has led to the creation of a voice and a phoneme database, with which we can synthesize speech sounds from written words or sentences in Moore.

_____

Month Year

Approved by Research Committee:

_____

Ettien Koffi Chairperson

ACKNOWLEDGEMENTS

We would like to take this opportunity to thank Dr. Ettien Koffi, without whom the completion of this thesis, especially this type of thesis would have been impossible. Indeed, not only Dr. Koffi deeply committed himself to helping in the completion of this thesis but most importantly, he helped to open our mind to the possibility of using linguistics in conjunction with other disciplines like computer science. We would like also to thank Dr. John Madden and Dr. Bryant Julstrom for their time and feedback.

We also wish to express our sincere and deep gratitude to Dr. Dafydd Gibbon, professor of Linguistics at the University of Bielefeld in Germany, who graciously accepted to be our guide during this long journey, despite his busy schedule and the long distance. Dr. Gibbon's support went beyond what we could expect. As an expert in text-to-speech development, his guidance was extremely valuable to the completion of this thesis.

To the participants, thank you for your precious contribution. Last but not least, we would like to thank our entire family back in Burkina Faso for their constant and unshakable support, especially dad, Julien Issa Zoungrana, who has always done everything possible to help us reach our goals.

In memory of

Mom, Cecile Lamoussa Zoungrana.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ASCII: American Standard Code for Information Interchange

CE: Categorical Estimation

DRT: Diagnostic Rhyme Test

H: High tone symbolized by the accent (´)

Hz: Hertz

IPA: International Phonetic Association

L: Low tone symbolized by the accent (`)

LLSTI: Local Language Speech Initiative

MOS: Mean Opinion Score

MRT: Modified Rhyme Test

Ms: Millisecond

PC: Pair Comparison

PHO: Phonetic File Output

SAMPA: Speech Assessment Methods Phonetic Alphabet

SCSU: Saint Cloud State University

SUS: Semantically Unpredictable Sentence

TTS: Text-To-Speech

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

INTRODUCTION

Linguistics is a scientific field of study with multidisciplinary connections to the fields of sociology, neurology, speech therapy, and computer science among others. The merger of the latter and linguistics has produced computational linguistics, which interprets languages from a computational prospective. According to Fromkin (2000) "Computational linguistics is concerned with natural language computer applications, e.g. automatic parsing, (…), computational linguistics has the goal of modeling human language as a cognitive system" (p. 4). The modeling of natural languages, divides itself into two main areas, speech recognition and speech synthesis. The latter area, called text-to-speech (TTS), has made possible the development of computer programs, with the ability to transform a written text into an audio speech.

TTS system while opening new perspectives for many applications is a system whose development can follow many methods. The different methods used in speech synthesis can be classified in two groups: the parametric system and the concatenative system. Both of these systems require a preliminary linguistic work, for any language, before an eventual implementation or development of a TTS system.

In the present study, we did first an overview of both text-to-speech system and the linguistic system of Moore. The overview of the linguistic system of Moore includes the orthography and the tone system. Next, we follow with an implementation of these linguistic features using a multilingual prototype synthesizer (MBROLA) which is a concatenative system. The study ends with an evaluation of the output sound speech, a summary of the problems encountered during the implementation and possible solutions and suggestions for future studies.

Chapter I

BRIEF HISTORY OF TEXT-TO-SPEECH

The first attempts to develop a talking machine started about two hundred years ago. According to (Schroeder, 1993), Professor Christian Kratzenstein from Russia designed a mechanism capable of reproducing the vowels /a/, /e/, /i/, /o/, and /u/ in 1779. Later, scientists like Wolfgang von Kempelen in 1791, and Charles Wheatstone in 1800, among others led more experiments on the mechanical imitation of the human vocal tract until the 1960's.

In addition to these mechanical experiments, other scientists were working on an electrical speech synthesis system. In 1922, Stewart presented a speech synthesis device that was at the time the first electrical device of its kind (Klatt, 1987). The electrical speech synthesis system was capable of producing single constant vowels vibrations, but not any consonants. Homer Dudley presented the first speech system considered as a speech synthesizer at a fair in New York in 1939. This speech synthesizer, VODER, was not easy to manipulate, but helped prove that intelligible speech generated artificially is possible (Lemmetty, 1999).

Noriko Umeda and his colleagues in 1968 (Klatt, 1987) designed the first complete TTS system, in the Electrotechnical Laboratory in Japan. Ten years later

in 1979, Allen, Hunnicutt and Klatt introduced a new system called MITalk. They developed the system at the MITalk text-to-speech system laboratory. The first commercial speech synthesis systems appeared on the market between the 1970's and 1980's among which, the Votrax chip, followed by the LPC (linear prediction coding) just to mention these (Lemmetty, 1999).

TEXT-TO-SPEECH IN AFRICA

Computer automated speech, as a field of study is relatively young in Africa. It is a field that involves linguistic and specific technical expertise requiring considerable financial resources. There are initiatives such as the Local Language Speech Technology Initiative (LLSTI)[1], which provides the tools, expertise, support, and training necessary to enable the development of text-to-speech system for local languages in Africa. LLSTI made possible the realization of studies like (Gibbon, 2006) and (Ngugi, 2005) who, among many others, have investigated the main challenges and issues related to the development of speech synthesizers for African languages like Swahili and Ibibio.

(Ngugi, 2005) gives an overview of the development of a text-to-speech system for Swahili, one of the most widely spoken languages in Africa. The study discussed the development of the language text-to-speech system, based on the Festival Unit Selection Speech Synthesizer. It required a prior work on the

[1] LLSTI website consulted on 07-15-2011 http://www.llsti.org/vision.htm

linguistic features of the language, needed in the development of the system. These basic features first comprised the definition of the phone set, that is, the alphabet of the language, vowels and consonants including all clusters. In addition, the prosodic features of Swahili, lexical and rhythmic phones in particular, are part of the needed input. Lastly, according to (Ngugi, 2005), the pitch and intonation represent two features that are crucial in synthesizing a language. He defines the pitch as "the melodic height of a phone while intonation describes the patterns of pitch in a language" (p. 2).

Ngugi, in the same study, draws also attention to the syllable structure issue. According to him, "The function of the syllable is to regulate the structure of complex-segments. The syllable serves as a building block for higher-level phonological and morphological processes" (p. 84). He later stresses the importance of the syllable structure by specifying four points, among which the recording database of the possible syllables of the language. The study shows the necessary steps that one needs to follow in the development of a text-to-speech system for a language like Swahili using the Festival speech synthesizer.

In a 2006 study, Dafydd Gibbon investigated the problems and solutions related to one African language, the Nigerian tone language called Ibibio. He pointed out a certain number of linguistic issues that needed to be addressed as a prerequisite for an Ibibio text-to-speech system. One of the main linguistic issues that he encountered in his study related to the syllable phonotactics of the language:

> The complexity of syllable phonotactics is a major determinant of the size of the unit database resource: whether a language is fundamentally CCCVVCC, like many Indo-European languages, or CV, CCV, CCVC, like many West African languages, leads to significant differences in combinatory options and therefore in resource size. (Gibbon, 2006, p. 2)

The inflectional system, the morphotactic of words formation and the sentence structure are also among the issues raised in this study.

(Anberbir & Takara, 2009) present the speech synthesis developed for Amharic, the official language of Ethiopia. They explain how they molded the linguistic features of the language and developed the system using a formant or rule-based synthesis approach, which required the elaboration of complicated rules. The correct adaptation of these rules is a crucial factor that determines the degree of quality of the speech synthesizer of any language. (Anberbir & Takara 2009) note that "For any language, appropriate modeling of prosody is the most important issue for developing a high quality speech synthesizer" (p. 48). The study was able to produce an effective text-to-speech system for Amharic even though they pointed out the lack of naturalness of the system.

To the best of our knowledge, this thesis is the first in the area of text-to-speech system for Moore language. However, Moore is a language that has been the subject of many linguistic studies among which (Canu, 1976), (Nikiema, 1976), (Nikiema, 1982), (Malgoubri, 1985), (Malgoubri, 2000) just to mention these. Therefore, much of the pre-required linguistic work has been done. The orthography of Moore language has been established and standardized for over

thirty years now. There are multiple studies of Moore done by linguists from Burkina Faso as well as linguists from other countries.

SOURCES OF CHALLENGES

Today's computers are able to reproduce the human voice with remarkable quality, even though some challenges remain. Some of the main challenges relate to the text preprocessing or tones treatment. In the preprocessing step, all abbreviations, acronyms, and symbols have to be converted first into their full forms. Therefore, a TTS system must be able to assimilate, terms like Dr. Nikiema, Nikiema Dr., $ 40, or USA respectively as Doctor Nikiema, Nikiema Drive, forty dollars, and United States of America. The system must also be able to distinguish between numerals and ordinals, or if a given numeral indicates a date or a measure of some kind. The same distinction is applicable for Roman numerals and common abbreviations (Lemmetty, 1999). The system must also recognize abbreviations that are homonyms. Terms like "Dr." should be pronounced as <doctor> if it precedes a proper noun. If it follows a proper noun, it should be read as <drive>.

Another problem is the conversion or passage from text to spoken form accompanied with the correct stress and intonation. Speech synthesizers have yet to find a way to make this conversion with explicit emotions, i.e., the way a human would speak. This part seems more challenging for languages with irregular pronunciation like English, German, Danish, or French than for other

languages whose spelling systems are based on the phonemic principle, namely: one symbol, one sound and one sound one symbol. In this latter case, we can cite languages like Italian, Spanish, or Moore, which all have regular pronunciation, with a one-to-one correspondence between a symbol and a sound.

However, in either case there may be a need to deal with homographs, which are words with different meanings and different pronunciations, but having the same spelling. The word <desert>, in English, falls into this case. If pronounced /dəˈzɛrt/, <desert> is a verb meaning “to abandon a place, a person or a group”. If pronounced /ˈdɛzərt/, it is a noun meaning “dry area”. These are different meanings associated with different pronunciations of the same word that a TTS system must be taught to recognize. The pronunciation may also be tricky due to the context in which a particular word occurs, or because the word is from foreign origin.

The idiosyncrasy (characteristics of each language) plays also a role. As a result, it seems less challenging to make a TTS system for some languages while it is not so easy for others. One of the main characteristics of African languages, i.e. Moore, is that they are tonal languages, which turn out to be a source of difficulties in TTS system. The words <kiisi>, <kiibu>, <kibri>, and <kidbri> have each at least two meanings in Moore depending on how they are pronounced. The first word <kiisi>, when pronounced with a rising pitch on the first syllable and a falling pitch on the last one, i.e. <kíisì>, means “months”. However, when it

is pronounced with just a rising pitch on the first syllable, i.e. <kíisi>, it means “to extinguish”.

As we can see, tones languages are languages where the pitch plays a phonemic function. That is, two identical words can differ in their meanings just because they are pronounced with different pitch. (Fromkin et al., 2010) point out that “Languages that use the pitch of individual vowels or syllables to contrast meanings of words are called tone languages”. Thus, it is not surprising that tones languages’ lexicon are usually full of words spelled the same but pronounced differently. Being able to use the pitch appropriately on a tone language’s homographs is not an easy task for non-native speakers not to mention a machine.

This difficulty resides in the fact that suprasegmental features of languages are responsible for giving the speech its melody, rhythm, and emphasis. (Fromkin et al., 2010) define the suprasegmental features of tones as “features over and above the segmental values such as place or manner of articulation”. According to (Gibbon, 2006), the tone issue in developing a TTS system is part of a bigger issue that is the syllable phonotactics, which deals with the different allowable phonemes combinations that define the syllable structure of a language.

SPEECH QUALITY AND EVALUATION

The challenges we just mentioned above constitute some of the factors that influence, in one way or another, the quality of synthetic speech. The assessment of the quality of synthetic speech can be based on many criteria among which are intelligibility, naturalness and suitability to a given application (Klatt 1987, Mariniak 1993). Depending on the purpose of the speech synthesizer, a feature such as intelligibility would be more valuable than naturalness, for example. There are many methods of synthetic speech evaluation used to test the quality of a speech in general, but there are as well other methods designed to test some proprieties like the acoustic characteristics (Lemmetty, 1999). Below is a brief description of some of these evaluation methods.

Segmental Evaluation Methods

This set of methods evaluates the intelligibility of only a single segment or phoneme using many tests among which, rhymes test such as Diagnostic Rhyme Test (DRT) which tests the intelligibility of consonants in words' initial position (Goldstein 1995, Logan et al 1989). Another rhyme test used is the Modified Rhyme Test (MRT) which tests the intelligibility of consonants in words' final position (Goldstein 1995, Logan et al 1989).

Sentence Level Tests

A set of sentences is used to test the comprehension level of the synthesized speech. There are many types of this test among which the Harvard

Psychoacoustic Sentences, which evaluate words' intelligibility in a sentence context. The Haskins Sentences, which test the speech comprehension in a sentence or word level, and the Semantically Unpredictable Sentences, which test randomly selected words in sentences are also part of sentence level tests (Lemmetty, 1999).

Comprehension Test

As its name indicates, this test evaluates the comprehension of the synthesized speech. The participants respond to questions about the content of synthesized sentences to which they have been exposed (Allen et al. 1987).

Prosody Evaluation

This test evaluates the prosodic features of the speech, which are one of the most challenging tasks in synthetic speech. The evaluation may focus on the emotional feature of the speech with questions such as "Does the sentence sound like a question, statement or imperative" (Lemmetty, 1999).

Intelligibility of Proper Names

This method tests the intelligibility of proper names like Benemanda Zoungrana or Barack Obama, for example. The correct pronunciation of such names is not always easy in particular for someone who reads these names for the first time.

Overall Quality Evaluation

Among these methods, we have the Mean Opinion Score (MOS), which asks for the opinion of the listener about the speech naturalness and the Categorical Estimation (CE), which evaluates the speech based on several attributes or aspects. In addition, there is the Pair Comparison (PC) which tests the system overall acceptance (Kraft et al. 1995).

Since the evaluation is done through human ears, there is some level of subjectivity involved. Indeed, it could take time for one's ears to get accustomed to unnatural speech. In addition, there are some phones or combinations of phones that are more or less complicated to understand, through a speech synthesizer, especially nasalized consonants like /m/, /n/, and /ng/ considered difficult to the human ears (Carlson et al. 1990). Other consonants or combinations of consonants like /d/, /g/, /k/, /dr/, /gl/, /gr/, /pr/, and /spl/ are also difficult to perceive through a speech synthesizer (Lemmetty, 1999).

POSSIBLE APPLICATIONS

The range of applications of a TTS system is very wide. Telecommunication services, public announcements in settings like airports, educational institutions, and visually handicapped people, just to mention these, benefit considerably from speech synthesizer systems.

A TTS system can allow visually handicapped people to enjoy books they want to read, just like Peter Ladefoged's mother-in-law.

> My mother-in-law was nearly blind for many years at the end of her life. What she needed was a way of turning anything that could be written into good spoken English. She wanted to be able to put a book in front of a computer, sit back, and enjoy it. (Ladefoged, 2001, p. 68)

People with visual difficulties would certainly be thrilled to be able to enjoy books through a speech synthesizer device. Such device, according to (Lemmetty, 1999), would also be able to give in advance the length of the text to read and the approximate time it would take to read it.

Telecommunication and multimedia companies are already benefiting from the progress made in the field of automated speech. As consumers, we interact every day with automated speech systems when we call a phone or cable company customers' service. Today, smart cell phones users have the ability to place calls just by saying the name of the person they want to call. Moreover, it is now possible to have electronic mails, and short text messages be read to people. Today computers are manufactured with ready to use voice recognition and text-to-speech systems that allow the user to interact and perform some considerable tasks without using a keyboard.

The educational sector is a major area of application, the one that would benefit many countries like Burkina Faso. Indeed, unlike a human teacher, a computer equipped with a speech synthesizer can be used as a teaching device all the time. Therefore, it could be used as a helping device by teachers to boost the rate of literacy among adult speakers of local languages.

SUMMARY

In this chapter, we have done a literature review of the TTS system domain going from the first attempt to create a talking machine to its possible applications in a country like Burkina Faso. Many challenges face anyone who wants to develop a TTS system for African languages. The most considerable of them being how to design a system that can read tones accurately. Last, but not least, is the commercial factor. Most TTS systems being developed target widely spoken languages like English, French, and Spanish. Therefore, for languages like most African languages, the lack of potential users and markets has a negative impact. The following chapter lays out the approach followed in this thesis.

Chapter II

METHODOLOGY

Text-to-speech is "the production of speech by machines, by way of the automatic phonetization of the sentences it utters" (Dutoit, 1997, p. 13). This implies that the speech synthesizer automatically converts the input text into an output speech. There are many different ways to perform this conversion, which can be classified in two groups (Bachan, 2007). The first group is the parametric group, which is based on a model of the articulatory or acoustic properties of the human vocal tract. There are two methods of synthesis in the parametric group: the formant synthesis and the articulatory synthesis. The second group is the concatenative model. It is based on samples of recorded speech. This group contains two methods of synthesis: the diphone synthesis and the unit selection synthesis. Each group of speech synthesis has its own advantages and disadvantages. This study is based on the concatenative group, i.e. the diphone synthesis.

THE CONCATENATIVE SYNTHESIS

The concatenative system uses small sets of recorded speech. First, a speech is recorded and then segmented in small units. These units can be phones, diphones, triphones, half syllables, syllables, words or other types of unit (Jurafsky & Martin, 2000, p. 274). The concatenative system produces a great quality of speech even though some glitches are often present in the speech output. The main two methods used in the concatenative synthesis system are the diphone synthesis and the unit selection synthesis.

The Unit Selection Synthesis

This method is similar to the diphone synthesis, but the speech database is much larger. In addition, the units instead of diphones can be simple phones such as [a], [b], or [k], selected according to criteria based on phonetic and prosodic environments so that the best possible units are extracted from the large database. The selected units are then concatenated to produce an output speech (Black & Taylor, 1997, p. 601). One of the most popular unit selection speech synthesizers is Festival (Taylor et al., 1998). The Bonn Open Synthesis System (BOSS) is another speech synthesis of the same type.

The Diphone Synthesis

A diphone (or a dyad) is a sound unit that begins in the middle, which is the most stable part of the phone, of a phone and ends in the middle of the next

phone (Gibbon, 2007). The following examples are some of the american[2] english voice diphones:

E.g. p as in drop or proxy.

p_h as in pod (aspirated allophone of p).

t as in plot or tromp.

t_h as in top (aspirated allophone of t).

k as in rock or crop.

k_h as in cot (aspirated allophone of k).

This type of synthesis is based on a recorded speech database that contains all the possible diphones existing in the target language. The convenience of this method comes from the fact that each diphone bears the necessary phonetic transitions and coarticulations and from the relatively small size of the speech corpus (Bachan, 2007). The concatenative system synthesis appears to have some particular problems because some glitches can be heard from the speech sound output. However, these anomalies can be minimized using diphones or some special methods to smooth the speech signal (Lemmetty, 1999, p. 34). There are many diphone synthesizers, among which the MBROLA software with a built-in clear language-to-speech interface called PHO file (phonetic output file) allowing the user to input text.

[2] Diphones extracted from US1 American Female Voice diphones database. It is one of the voices available on http://tcts.fpms.ac.be/synthesis/

THE MBROLA PROJECT

The MBROLA [3] project is an initiative from the "Faculte Polytechnique de Mon (Belgium)"[4] and is aimed at collecting a set of speech synthesizers for as many languages as possible, making them available for free and for non-commercial purpose. The project goal is to encourage academic research on speech synthesis, and particularly on prosody generation, known to be one of the biggest challenges in text-to-speech synthesis. The engine of the MBROLA project called MBROLA itself uses databases of concatenated diphones to run.

The MBROLA synthesizer is not really a speech synthesizer, since it does not accept raw text, but rather phonemes from a diphone database tailored for the MBROLA format in order for the synthesizer to run. TTS system can be thought of as a procedure consisting of two steps that result in the output of a speech sound waveform. Before the sound waveform is produced, MBROLA needs data or information about the phonetic and prosodic features of the input raw text. The input raw text is linguistically analyzed to extract the phonetic characteristics of phonemes composing the raw text.

The phonemes from the first step, along with their prosodic information are the only input accepted by MBROLA. These phonemes are called diphones in the MBROLA terminology because each one of them is made of a sound unit that

---

[3] MBROLA website, http://tcts.fpms.ac.be/synthesis/, consulted 08-05-2011

[4] Polytechnic faculty of Mon (Belgium)

starts in the most stable part of a phoneme, which is the middle part, and ends in the most stable part of any next phoneme. The MBROLA project has not only diphone databases for many languages but also many ready voices, with natural intonation in different languages from around the world. Each language voice, which sounds natural and respects the prosodic intonation of the language, was created based on a database of diphones unique to that language.

THE MBROLA TOOL PACKAGE

The MBROLA tool package contains:

Mbrola.dll	:	MBROLA engine
Mbrplay.dll	:	Easy to use interface to the engine
Mbroli	:	A pho player
PhoPlayer	:	A pho script player
Control Panel:		A control panel for managing the MBROLA databases installed in the computer
MbrEdit	:	Another pho file (written in VB, source included)
C and VB source codes documentation:		Interface to the DLLs

SPEECH SYNTHESIS BY MBROLA

The MBROLA speech synthesizer software can be downloaded from the MBROLA homepage[5]. Each voice is an output speech produced using a database of diphones created for the language to which the voice belongs. Each input phoneme of any word is accompanied by its duration in millisecond (Ms) and pitch height in Hertz (Hz). The information about the pitch is given in pairs, with the position of the pitch value in the phoneme (as a percentage of the duration of the phoneme), and the pitch value (in Hertz, cycles per second). In the following example, the phoneme /p/ is read with a span of 80 Ms and a pitch of 131 Hz realized at 25% of 80 Ms.

E.g. p 80 25 131

t 55 45 105 75 145

The pair of values "25 – 131" can be repeated more than once with different values, depending on how one wants the phoneme to sound. In the second example, /t/ is read by the system during 55 Ms with a first pitch of 105 Hz realized at 45% of 55 Ms, and a second pitch of 145 Hz realized at 75% of 55 Ms. By using PHO-Mbroli, an interactive window of MBROLA, one can input phonemes in order to synthesize words or sentences in any of the available languages' databases.

[5] http://tcts.fpms.ac.be/synthesis/

Synthesis Procedure

Once the software[6] is downloaded,

step 1: from the start menu,

step 2: select programs,

step 3: select MBROLA tools,

step 4: from MBROLA tools items, click on Mbroli.

A window that reads "Untitled-Mbroli" at the top will pop up.

Step 5: from this window click on the menu *file*, and on *open* to look for one of the MBROLA voices, previously downloaded, for example the voice en1, a British male voice. By selecting en1, a number of synthesized sentences are available. A double click on one of these voices pops up a window similar to the one below which is a PHO file of the famous sentence "To be or not to be, that is the question" of Shakespear.

[6] Like the MBROLA software, the voices can be also downloaded from http://tcts.fpms.ac.be/synthesis/

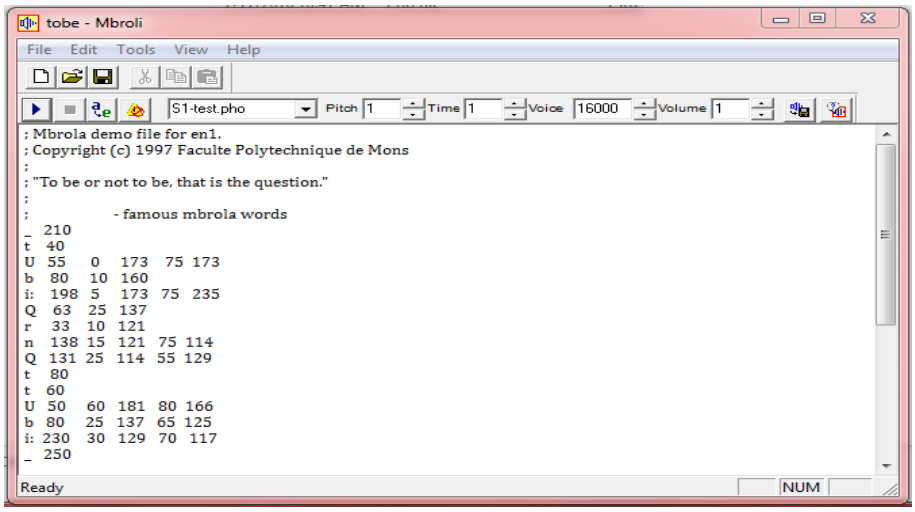


Figure 2-1. An English sentence PHO file

A click on the *play* button outputs a male voice of the above sentence. As explained above, the input line "b 80 10 160" means that the phoneme /b/ is synthesized with a span of 80 Ms and a pitch of 160 Hz realized at 10% of 80 Ms. The input line "U 50 60 181 80 160" is read by the system as /U/ with a span of 50 Ms and a first pitch of 181 Hz realized at 60% of the 50 Ms duration, and a second pitch of 160 Hz realized at 80% of the same 55 Ms. The input line "t 80" would be read in a period of 80 Ms with a monotone pitch.

The following window is a PHO file of the Moore sentence /bắng bõe n be tãngà̀ sɛ́ɛga/ which means, "guess what is at the bottom of the mountain". The phonemes in the bottom window illustrating the Moore sentence PHO file are read exactly the same way as in the above English sentence PHO file.

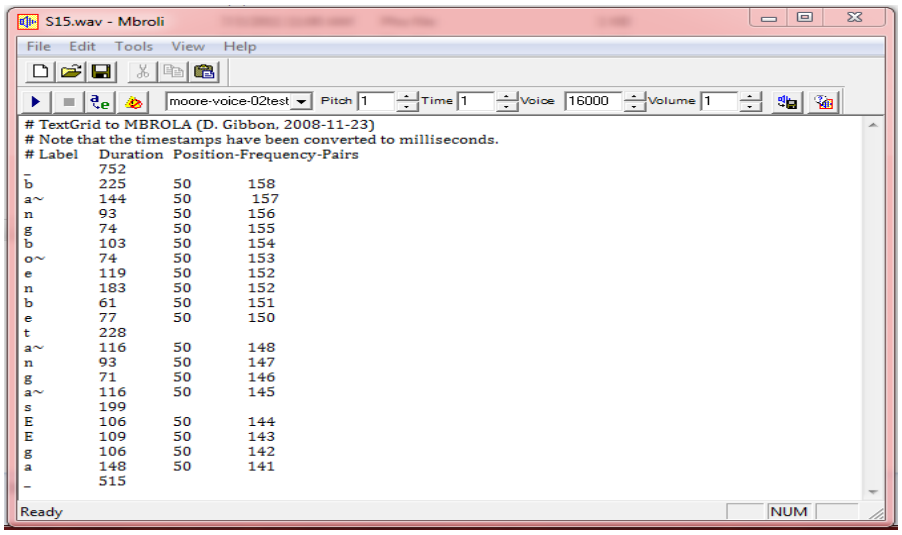


Figure 2-2. A Moore sentence PHO file

From the above window called PHO Mbroli, anyone can synthesize a word or sentence using the database of one of the provided voices such as en1, which can be accessed by clicking on the last button from the same line as the *play* button or by holding the control button while pressing the letter “i” from the keyboard. The en1 voice database, like all the voices available, contains ASCII characters based on the IPA symbols. It is also called SAMPA which stands for Speech Assessment Methods Phonetic Alphabet. Each language SAMPA is particular to that language and the phonemes it contains can be used to synthesize speech that respects the prosody of that language. The following figure is an example of diphone database that contains the phonemes or SAMPA of the voice en1. The voice was modeled from a British male voice, indicating a British accent.

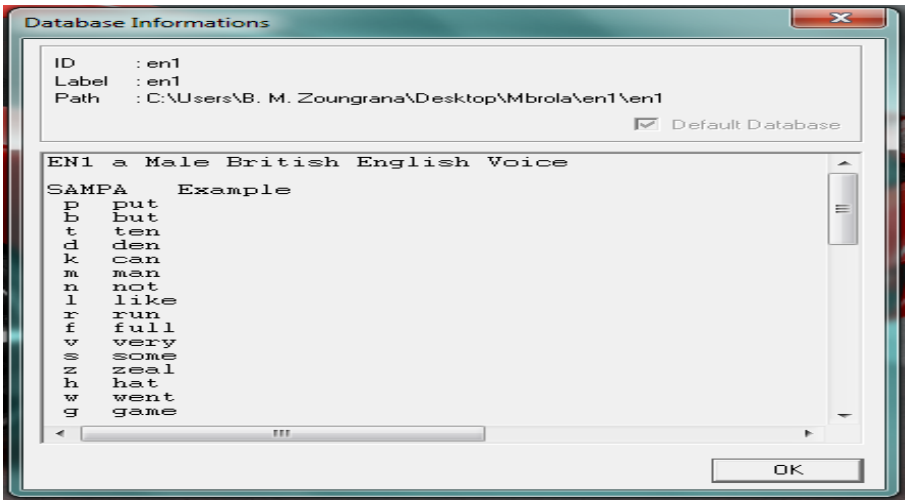


Figure 2-3. en1 SAMPA or diphone database

According to the word or sentence to be synthesized, the phonetic transcription uses these phonemes (diphones), which can be copied and pasted in the Mbroli window accompanied with the duration and pitch information in the subsequent columns.

Similarly, words or sentences of a different language, such as Moore can be synthesized using the diphone database of English or French, which have a similar orthography to Moore. The result will be a comprehensible Moore sentence, but with a British, an American or French voice. It would sound like a natural speech spoken by a native speaker of Moore if there were a database of Moore diphones or SAMPA available to be used as input. What we intend to do in the rest of this thesis is to create a voice and a diphone database that can be used to synthesize speech from written words or sentences in Moore.

The rest of this thesis is then about creating a database of Moore diphones. In doing so, these four major steps are followed:

1. First, the orthographic and phonetic system of Moore are introduced where the phonemic symbols are presented.
2. Second, the diphone database, which is the fundamental part of a diphone synthesizer, creation can be done in four steps as suggested by (Dutoit et al., 1996):
    a. First, creating the corpus:
    - A list of the phones of Moore is made according to the phonemic symbols of the language.
    - From the list of phones, a list of all digrams (sequences of two phonemes) is generated.
    - A list of keywords is made in such a way that every digram is included at least once.
    - Last, the corpus is ready when every keyword is included in a sentence called carrier sentence.
    b. Second, recording the corpus.
    c. Third, segmenting the corpus.
3. Lastly, an evaluation of the voice is performed.

A detailed account of Moore diphone database creation is presented in chapter V. The following PHO file explains how the synthesized speech of words or sentences in Moore or any other language without diphone database can be

generated using the British en1 voice database or any other database available. It is assumed that the software and one of the available voices were downloaded and can be accessed from a folder named MBROLA tools.

1- Formulate and transcribe the sentence(s) in the original language, Moore in this case, using the IPA symbols.

E.g. sentence in Moore.

Ned ka pẽgd a meng ye.

One not respect a self.

One should not have excessive self-respect.

In the IPA transcription, of the above sentence that follows, note that beside the high tones that were added, nothing changed. That confirms that the one symbol, one sound and one sound, one symbol rule is respected.

Néd ka pẽgd à méng yè.

2- Transcribe the sentences using the diphone set or SAMPA of the MBROLA voice to be used - Open 'control panel' from 'MBROLA tools' and choose the diphone database to be used - in this case, en1 voice (a British male voice).

E.g. transcription using the English voice en1 database from figure 2-3.

neD k{ pengD { meng je.

Using the en1 database to transcribe the Moore sentence, the phonemes /d/, /a/, and /y/ were replaced by their corresponding en1 database phonemes,

that is, /D/, /{/, and /j/. Since there is no nasalized /e/ in the database as in /ẽ/, it was replaced by the juxtaposition of /e/ and /n/.

3- From 'MBROLA tools', double click on 'Mbroli' which pops up an empty PHO file window in which the transcribed phonemes should be written down in the first column. The second column requires information about the duration of each phone. The third and fourth columns respectively require information about the percentage of the duration to which the pitch should be applied as in the following picture.

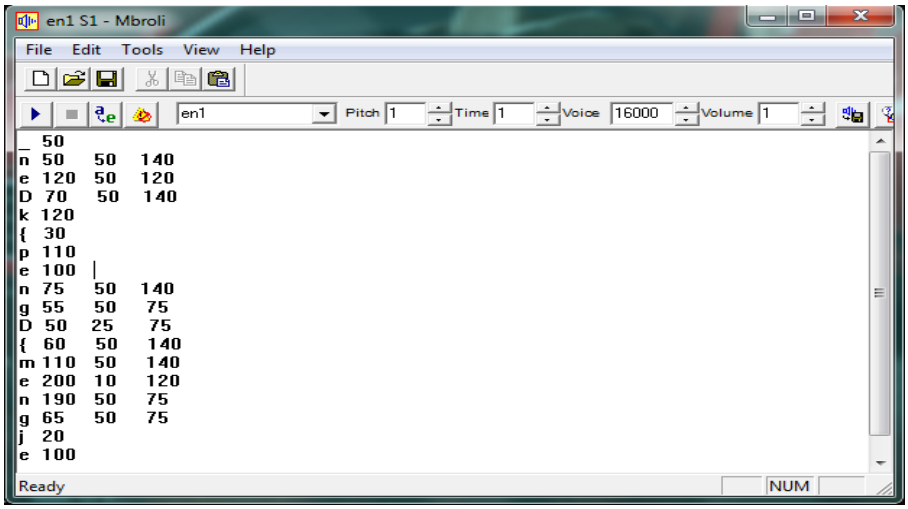


Figure 2-4. en1 PHO file of a Moore sentence

4- Run Mbroli by clicking the *play* button. The result is a relatively intelligible Moore synthesized speech minus the native accent.

The tones values (the third column of numeric values) of this sentence in Moore vary from 75 to 140 Hz. Each phone is realized in a period of 20 to 200 Ms.

These pieces of information give us an idea about the phonemes duration and their pitch height. The expectation at the end of this thesis is to be able to repeat the last four steps, but with a voice and a phoneme database proper to Moore.

SUMMARY

There are numerous systems of automatic speech synthesizer among which a few were presented in this section. These systems include the formant, the articulatory, the unit selection and the diphone synthesis. A particular accent was put on the diphone synthesis method because the framework of implementation retained in this thesis, MBROLA, is a diphone synthesizer. This choice is justified by the convenience and the flexibility of MBROLA that allows prosodic manipulation, through a PHO player interface, of speech segments (diphones or phonemes) pitch and duration.

Chapter III

SOCIOLINGUISTIC CONTEXT OF MOORE ORTHOGRAPHY

Moore is spoken predominantly in Burkina Faso, previously known under the name of Upper Volta before the year 1984. It is a country of about 274200 km^2 located right in the middle of the west region of the African continent. The country gained its independence from France on 08-05-1960 and has thirteen (13) regions, which are subdivided in forty five (45) provinces. The two major cities are Ouagadougou, the political capital, and Bobo Dioulasso, the economical capital. Burkina Faso is surrounded in the south, from west to east, by four countries that are Cote d'Ivoire, Ghana, Togo, and Benin, in the north- east by Niger, and in the north-west by Mali.

The name Burkina Faso [Burkinafaso], which means the country of people of integrity, is a compounded noun from two words of the two most spoken languages in the country. The term Burkina, which means integrity, honor, is from the Moore language, while the term Faso, which means territory or land is from the second most spoken language, Dioula. The citizens of Burkina Faso are called Burkinabe [Burkinabe], which is a blending of the word <Burkina> from

Moore and <-be>, which is a plural suffix from the third most spoken language in the country, Fulfulde. There are 68[7] listed locale languages in Burkina Faso.

THE MOORE LANGUAGE

Moore is the language of the Moosse [Moose] or Moossi [Moosi] people. They constitute the majority of the over 17 Millions (about 17, 2 75, 115)[8] burkinabe in the country. Moore is part of the Oti-Volta group, which is a sub-group of the Niger-Congo languages family. The language is mainly spoken in the central part of the country. It is also spoken in neighboring countries like Cote d'Ivoire, and Ghana. The term Moore refers to each one of the six official dialects, recognized by the "Sous-commision Nationale du Moore"[9]. These dialects are Saremde, Taolende, Yaadre, Ouagadogou, Yaande, and Zaoore. The orthography used is standardized and is valid for all dialects. In an effort to reduce the illiteracy rate in Burkina Faso, the government is using local languages like Moore to alphabetize the adults who have not had the chance to go to formal school.

Like many other languages in Africa, Moore is predominantly an oral language. The first written texts of Moore appeared around the time when the

---

[7] http://www.ethnologue.com/show_country.asp?name=BF consulted on 04/12/2012.

[8] https://www.cia.gov/library/publications/the-world-factbook/geos/uv.html consulted on 04/12/2012

[9] Governmental institution in charge of regulating the language

country became a french colony, and Christian missionaries arrived in the 19th century. They created an alphabet of Moore for evangelical purpose. They were primarily interested in Bible translation. This orthography was based on the French language script. It can be noted in passing that the French orthography is opaque, whereas Moore, which is a tonal language, has a transparent orthography (Balima, 1997, p. 2).

(Alexandre, 1953) and (Hall, 1948) were the first to elaborate a system of transcription for Moore. Pastor Hall published Dictionary and Practical Notes Mossi-English languages in 1948. The orthography of Moore uses twenty-five graphemes. The standardization of Moore orthography gave birth to the realization of many projects that led to the translation of the Bible in 1983 and the publication of a dictionary of the language in 1997, among other projects. Moore as a language today is used in local radios like RNB (Radio National du Burkina)[10] and TVs stations like TNB (Television National du Burkina)[11].

MOORE ORTHOGRAPHY

The Alphabet

On September of the year 1976, the "Sous Commission Nationale du Moore", a governmental institution created in 1969, proposed an alphabet for

[10] National Radio of Burkina

[11] National Television of Burkina

Moore in a document titled: “Comment transcrire correctement le Moore”[12]. This document was an update of an earlier one and was intended for the public. The rules stated in this document regulate the alphabet and the grammar. Any document, written from then on, in Moore had to abide by these rules. The spelling of Moore reflects the pronunciation of the language as closely as possible. That means that it has a transparent orthography in which each symbol has a function (Balima, 1997, p. 17). In other words, the basic principle adopted is one symbol for one sound and one sound for one symbol. The orthography of Moore is composed of twenty (25) graphemes among which seventeen (17) consonants and eight (8) vowels. The graphemes are presented as following before a detailed illustration in Table 3-2 and Table 3-6:

Aa Bb Dd Ee Ɛɛ[13] Ff Gg Hh Ii ɩ Kk Ll Mm Nn Oo Pp Rr Ss Tt Uu Ʊʊ Vv Ww Yy Zz

Here are the main linguistic symbols used in this thesis and their meanings.

[12] Translate literally as “How to transcribe correctly Moore”

[13] /ɛ/ is not a not really a fundamental vowel but rather the result of these sounds: [ae], and [ea].

Table 3-1. Chart of symbols and conventions

Symbol	Interpretation
< >	Indicates words in graphemic notation
[ ]	Indicates words in phonetic notation
/ /	Indicates words in phonemic notation
[~]	Stands for nasal or nasalized vowels
[`]	Grave accent used to symbolize a falling or low pitch
[´]	Acute accent used to symbolize a rising or high pitch

Vowels

There are eight (8) oral vowels, among which only five can occur as nasal vowels, without contextual influences like assimilation, in the orthography of Moore. In addition, the same five vowels are nasalized automatically when they follow a nasal consonant /m or n/. These five vowels are / ã, ẽ, ĩ, õ, ũ/. Like oral vowels, nasal vowels can be short or long, with two, or even three vowels combined in one sound. When they are long, only the first vowel has the tilde or diacritic sign as a nasal sign. The vowel /ɛ/ is a combination of /a/ and /e/ and is the equivalent of the diphthong /ae/.

Table 3-2. The vowels of Moore

Phoneme		Grapheme		Example	Gloss
		Uppercase	Lowercase		
1	/a/	<A>	<a>	<zuga>	head
2	/e/	<E>	<e>	<zi-peelẽ>	white place
3	/ɛ/	<Ɛ>	<ɛ>	<gɛba>	onion
4	/i/	<I>	<i>	<mininzitã>	meningitis
5	/ɪ/	<ɪ>	<ɪ>	<tɪpa>	to heal
6	/o/	<O>	<o>	<yibeoogo>	morning
7	/u/	<U>	<u>	<miuugu>	red color
8	/ʊ/	<Ʊ>	<ʊ>	<pʊg-yaanga>	senior woman

Table 3-3. The nasal vowels of Moore

Phoneme		Grapheme		Example	Gloss
		Uppercase	Lowercase		
1	/ã/	<Ã>	<ã>	<ãdga>	star
2	/ẽ/	<Ẽ>	<ẽ>	<zi-peelẽ>	white place
3	/ĩ/	<Ĩ>	<ĩ>	<kĩnã>	pearls
4	/õ/	<Õ>	<õ>	<võre>	chunk
5	/ũ/	<Ũ>	<ũ>	<gũsi>	to sleep

Diphthongs and Triphthongs

Diphthongs are sequences or groups of two different vowels combined and pronounced in one vocal sound. Diphthongs can be oral or nasal. A triphthong can be described as a long diphthong, in which the last vowel has been lengthened, forming a sequence of three vowels. Not all sequences of vowels are called diphthongs or triphthongs. There are thirteen diphthongs and thirteen triphthongs in Moore.

Table 3-4. The diphthongs of Moore

Diphthong		Grapheme	Example	Gloss
1	/ae/	<ae>	<kae>	to boil
2	/ao/	<ao>	<sɪpaolgã>	summer
3	/ea/	<ea>	<keala>	left over
4	/eo/	<eo>	<peorko>	thick
5	/ɯ/	<ɯ>	<pɯlengo>	in a stealth way
6	/iu/	<iu>	<piuku>	large
7	/oa/	<oa>	<goama>	proper name
8	/oɛ/	<oɛ>	<logtoɛɛmba>	medical doctors
9	/oe/	<oe>	<koenoogo>	good news
10	/ui/	<ui>	<mui-zẽedo>	meal of rice
11	/ʊɪ/	<ʊɪ>	<lʊɪ>	to fall
12	/ʊɛ /	<ʊɛ>	<lambʊɛtga>	a tree specie
13	/ʊe/	<ʊe>	<bʊe-sablga>	black goat

Table 3-5. The triphthongs of Moore

Triphthong		Grapheme	Example	Gloss
1	/aee/	<aee>	<loaeega>	link
2	/aoo/	<aoo>	<taoore>	in front of
3	/eaa/	<eaa>	<tenteaaga>	large recipient
4	/eoo/	<eoo>	<kẽooge>	to fry
5	/ɯʊ/	<ɯʊ>	<tɯʊ>	to sew
6	/iuu/	<iuu>	<kiuugu>	moon or month
7	/oaa/	<oaa>	<soaala>	lord
8	/oɛɛ/	<oɛɛ>	<koɛɛga>	voice
9	/oee/	<oee>	<zoeese>	race
10	/ʊɪ/	<ʊɪ>	<pʊɪ>	to share
11	/ʊɛɛ/	<ʊɛɛ>	<kʊɛɛga>	short
12	/ʊee/	<ʊee>	<bʊeese>	goats
13	/uii/	<uii>	<tuiifu>	grain of a tree

Consonants

There are seventeen (17) consonants in the orthography of Moore.

Table 3-6. The consonants of Moore

Phoneme		Grapheme		Example	Gloss
		Uppercase	Lowercase		
1	/b/	<B>	<b>	<baaga>	dog
2	/d/	<D>	<d>	<daaga>	market
3	/f/	<F>	<f>	<laafi>	health
4	/g/	<G>	<g>	<piiga>	ten
5	/h/	<H>	<h>	<hato>	sunday
6	/k/	<K>	<k>	<kasenga>	elder
7	/l/	<L>	<l>	<peelle>	chovelle
8	/m/	<M>	<m>	<kasma>	older
9	/n/	<N>	<n>	<naasse>	four
10	/p/	<P>	<p>	<paase>	to add
11	/r/	<R>	<r>	<rulga>	column
12	/s/	<S>	<s>	<saaga>	rain
13	/t/	<T>	<t>	<tomaato>	tomato
14	/v/	<V>	<v>	<vele>	to swallow
15	/w/	<W>	<w>	<weefo>	bicycle
16	/y/	<Y>	<y>	<yeere>	cheek
17	/z/	<Z>	<z>	<zaala>	naked

In other documents written about the orthography of Moore, the symbol / '/ is added as a glottal sound. However, we have decided not to include this symbol in this study because it is hard to establish the phonemic distinction between / '/ and /h/ in spoken speech. The two sounds are so closely related that we have not been able to differentiate them during the recordings. Both sounds have a low frequency of occurrence. When they occur, it is usually in words borrowed from foreign language like Arabic. As a result, /h/, which has a higher frequency, will be used instead.

SUMMARY

This chapter gives a contextual overview of Moore itself and its orthography. It is a standardized orthography of twenty-five (25) graphemes with eight (8) vowels and seventeen (17) consonants. In addition, five of the vowels, which are all oral, can be nasalized with the tilde sign on top. The vowels can be short or long. A sequence of two simple or short vowels in the orthography that is pronounced as a single sound is called diphthong. Similarly, a sequence of three simple vowels, pronounced as a single sound, is called triphthongs. Diphthongs and triphthongs can also be nasalized. When they are nasalized, only the first vowel has the nasal sign. Diphthongs and triphthongs are phonemic sounds.

Chapter IV

PHONETIC FEATURES OF MOORE SOUNDS

THE VOWELS

The eight (8) oral vowel graphemes, the five (5) nasal vowel graphemes as well as the thirteen (13) diphthongs and (13) triphthongs are all based on seven (7) phonemic vowels presented in the following table.

Table 4-1. Phonetic features of Moore vowels

	Front	Central	Back
High	ɪ		u
	i		ʊ
Mid	e		o
Low		a	

THE CONSONANTS

The consonants can be classified according to the following table:

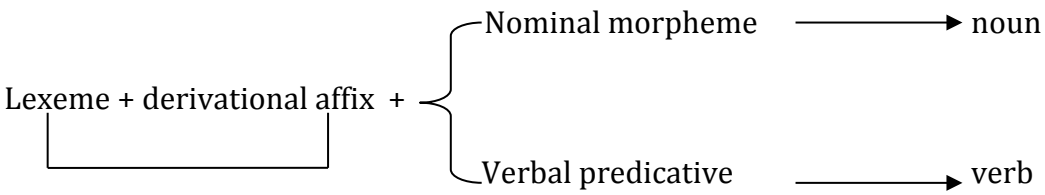
Table 4-2. Phonetic features of Moore consonants

Manner of articulation	Glottal state	Place of articulation					
		Bilabial	Labio-dental	Alveolar	Palatal	Velar	Glottal
Stops	- voice	p		t		k	
	+ voice	b		d		g	
Fricatives	- voice		f	s			h
	+ voice		v	z			
Nasals		m		n			
Liquids	Lateral			l			
	retroflex			r			
Glides					y	w	

WORD AND SYLLABLE STRUCTURE

Word Formation Process

Moore is a language with relatively short words. The dominant word formation process consists of the following morphemic structure: root + derivational suffix. Words that are created through this process are verbs or nouns for the most part (Houis, 1977, 1980, 1983). The syntactic constituent, the minimal unit capable of assuming a syntactic function in African languages has the following morphemic structure:



Moore words have the following properties:

- A free morpheme;
- A root and the affixes;
- A noun class marker, which indicates the part of speech of the words, as either a noun or verb. (Houis, 1980, p. 11)[14]

Below are some examples of the word formation process described above:

[14] Translated from French and adapted to the particular case of Moore.

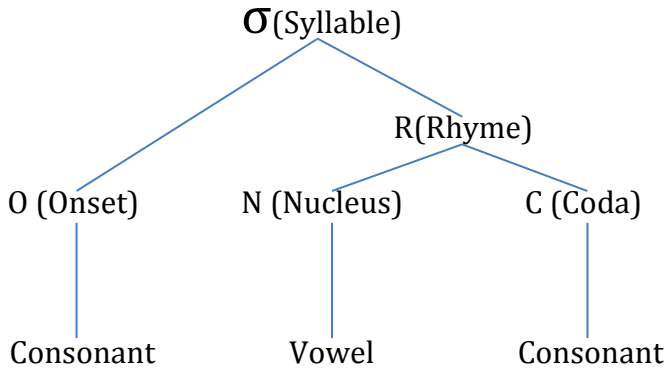
E.g.

ba	+	-g	+	a	⟶	baaga
/ root	+	derivational affix	+	nominal affix/		a dog
bʊ	+	-g	+	se	⟶	bʊgse
/ root	+	derivational affix	+	verbal predicative/		to guess

Syllable Structure

The words in Moore are often mono or disyllabic. The syllabic structure of Moore has been the subject of two interesting studies. (Kabore, 1980) proposed CV (consonant followed by a vowel) as the typical syllabic structure of Moore. (Nikiema, 1987) suggested that Moore has a CVC (consonant and a vowel followed by another consonant) structure in addition to having a CV structure. In contemporary linguistics, it is assumed that there are two components in a syllable: the onset, and the rhyme. The rhyme is further divided into a nucleus and coda. The nucleus, the heart of the syllable, is made of a vowel or more than one vowel.

E.g. architecture of the syllable



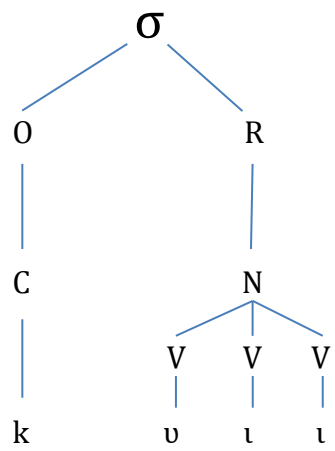
The vocalic system of Moore allows simple and long vowels as nucleus. In the case of diphthongs or triphthongs, the nucleus of the syllable can carry up to three vowels. The coda and the onset are either empty or occupied by one, two consonants or more. It is not uncommon to have in Moore, words with more than two consonants in the coda (which could be ideophones). So the syllabic structure of the language can be summarized by the formula CV (V) (V) (C). There are many types of syllables in Moore: open, closed, light and heavy syllables.

Open syllable. An open syllable is one that ends in a vowel (CV, CVV or CVVV) like /kʊɪ/ and /ma/ in <kʊɪma>.

Table 4-3. Open syllables

Syllable type	Syllable	Example	Gloss
CV	/ne/	<neda>	A person
CVV	/põo/	<põore>	A handicap
CVVV	/kʊɯ/	<kʊɯma>	A lazy

E.g. syllabic illustration of an open syllable

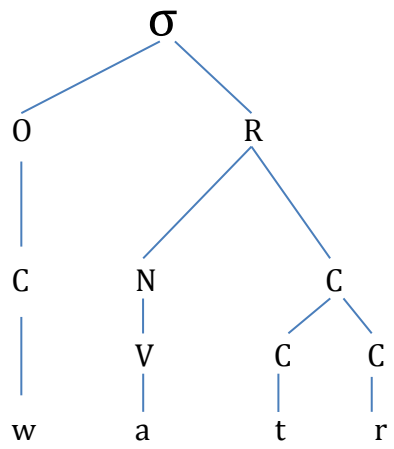


Closed syllable. A closed syllable is one that ends in a consonant (CVC or CVCC) like /watr/ in <watrwɛka>.

Table 4-4. Closed syllables

Syllable type	Syllable	Example	Gloss
CVC	/vig/	<vigsi>	Shake up
CVCC	/watr/	<watrwɛka>	Shea nut

E.g. syllabic illustration of a closed syllable

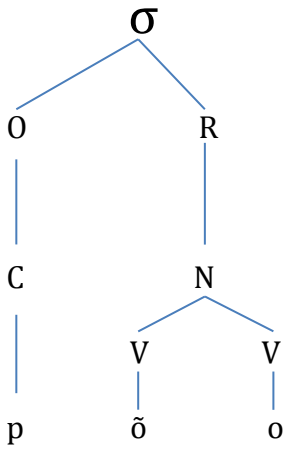


Heavy syllable. A heavy syllable is one with a branching rhyme which means that either the nucleus or codas (or both) respectively has more than one vowel such as /põo/ in <põore> or more than one consonant such as /watr/ in <watrwɛka>.

Table 4-5. Heavy syllables

Syllable type	Syllable	Example	Gloss
CVV	/põo/	<põore>	A handicap
CVVV	/kʊɯ/	<kʊɯma>	A lazy
CVCC	/watr/	<watrwɛka>	Shea nut

E.g. syllabic illustration of a heavy syllable

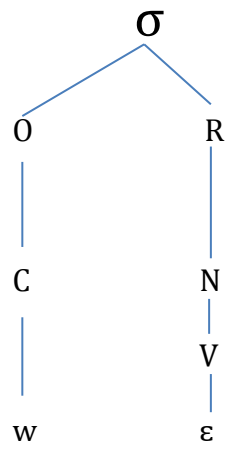


Light syllable. A light syllable is one with a rhyme that does not have branches or a node. The nucleus carries only one vowel such as /wɛ/ in <watrwɛka> or /ne/ in <neda>.

Table 4-6. Light syllables

Syllable type	Syllable	Example	Gloss
CV	/ne/	<neda>	A person
CVC	/vig/	<vigsi>	Shake up
CVCC	/watr/	<watrwɛka>	Shea nut

E.g. syllabic illustration of a light syllable



Importance of the Syllable Structure

The syllable structure plays a significant role in TTS system that uses the syllable as its basic speech unit of concatenation. Its structure constitutes the building block used in the development of the system. As stated in (Ngugi, 2005):

> (1) this structure is intended to help in recording the possible syllables and store them in the database. (2) It will also help in the design of the parsing algorithms, as these are the rules to be used to derive the syllables given a word. (3) It will help in the design of the Storage structure for easy and efficient retrieval of the audio files. (4) It will also help as a control to

> confirm the validity of a word with respect to its (contextual) structure. (p. 5)

MBROLA is a diphone concatenative system, that is, its basic unit of concatenation is the diphone. Therefore, instead of a syllable-based segmentation, all the speech sounds are segmented in diphones. As defined above, a diphone is a segment of speech unit that starts from the middle of one phoneme and ends in the middle of a second phoneme. The following is a picture showing an example of a segmented speech sound in diphones using PRAAT. The last two rows of the picture show thirty-eight (38) segmented diphones of the following sentence:

Ráwã yèelá maam tí ắadsằ yí wúsg zàamḗ zàabré.

The man told me that stars go out plenty of yesterday afternoon.

The man told me that yesterday evening sky was full of stars.

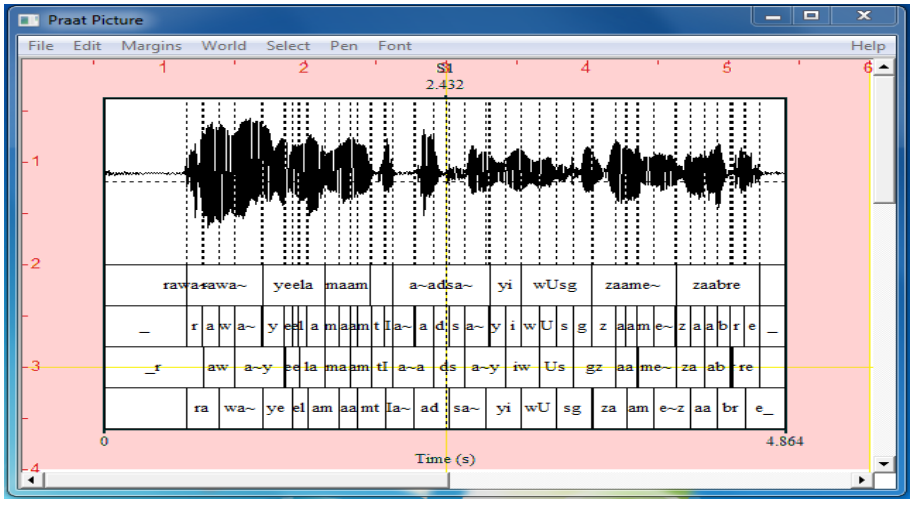


Figure 4-1. Speech sound segmented in diphones

The actual diphones are listed in the following table:

Table 4-7. Segmented diphones of a sentence

_r	aw	a~y	ee	la	ma	am	tI	a~a	ds	a~y	iw	Us	gz	aa	me~	za	ab	re	
	ra	wa~	ye	el	am	aa	mt	la~	ad	sa~	yi	wU	sg	za	am	e~z	aa	br	e_

The first diphone [_r] is the combination of MBROLA's pause sign (an underscore) and the first half part of the phoneme [r]. The third diphone [a~y] is the union of the last half part of the phoneme [a] and the first half part of the phoneme [y]. The tilde sign next to [a] indicates that [a] is nasalized. The segmentation details of a speech sound are explained later in chapter V.

TONES

Tones Marking

Moore is a tone language where the tones can be indicated in the lexical realization of words using two types of accent: an acute accent <ˊ> for a high tone (H) and a grave accent <ˋ> for a low tone (L). The pitch variations are used to distinguish words meaning in tonal languages. (Fromkin, 2000) notes this about tone languages: "A language is a 'tone language' if the pitch of the word can change the meaning of the word – not just its nuances, but its core meaning" (p. 229). She goes on later to give this precision: "A language with tone is one in which an indication of pitch enters into the lexical realization of at least some morphemes" (p. 231). In the orthography of Moore, the tones are not marked for

simplicity purpose, but in this study, the tones are marked wherever necessary to show the pitch variation of the speech sound. In noting the tones, only the changes are mentioned, that is, a syllable without tone is pronounced on the same register or level as the preceding syllable in the same word. A word with no tone is read monotonously.

Tonal Processes

Just like the vowels, the tones of Moore can undergo changes. The most typical of them are polarization, the raise of the radical tone, and the downstep (Nikiema, 1987). These tonal processes highlight the close relation existing between the syllabic structure and the tonal system of Moore.

Tone's polarization occurs in nominals (nouns) where the tone of the suffix is the opposite of the tone of the root.

E.g.[15]	zàk + ká	zàkká	House
	CVC+ CV	L H	L H
	pès + gó	pèsgó	Sheep
	CVC+ CV	L H	L H

The second change happens when a syllable is in a radical position in a word. The low tone of the root is raised when it interacts with a high tone of the syllable in the suffix.

[15] All examples in this section are adapted from Nikiema (1987)

E.g.	bìi+gá	bíigá	Child
	L H	H H	
	sòo+bó	sóobó	Bath
	L H	H H	

The last process, called downstep, is defined as the lowering of the tonal register that sometimes occurs between adjacent or identical tones. It is cumulative, successive and the phenomenon results in ever lower setting of the tonal register (Snider, 1998, p. 1). Moore downstep can be described by the following steps according to (Nikiema, 1987).

The first one is the lowering of a high tone such as the last one in a sequence like HLH. This type of downstep causes the last high tone (H) of a HLH word to be lowered one level down giving HLH_1 with H_1 still high, but one level lower than the first high tone (H).

E.g.	wénd + dòo + gó	wéndòogó	church
	H L H	H L H_1	

We will illustrate the downstep phenomenon by a pitch analysis using PRAAT. First, we show the realization of /wénd/, meaning God, and /dòogó/, meaning house, followed by the realization of /wéndòogó/, which means church, to see how the downstep phenomenon influences the tones' pitches.

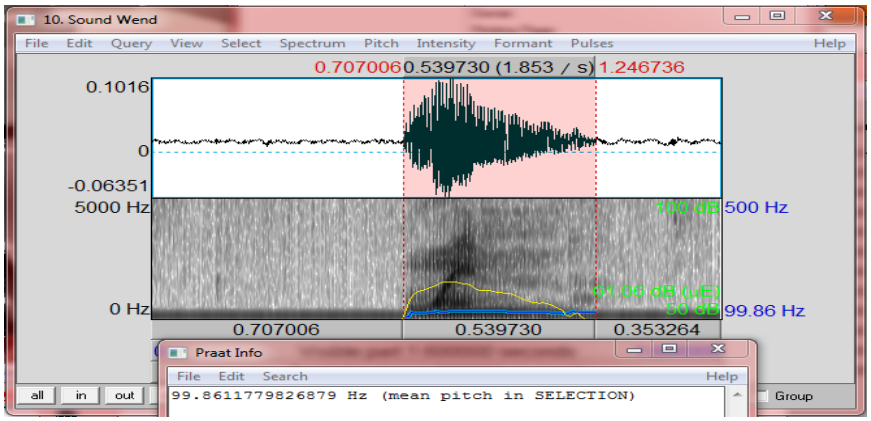


Figure 4-2[16]. Realization of /wénd/

In the above PRAAT picture, /wénd/ is realized at a pitch of 99.86 Hz as indicated at the bottom of the picture.

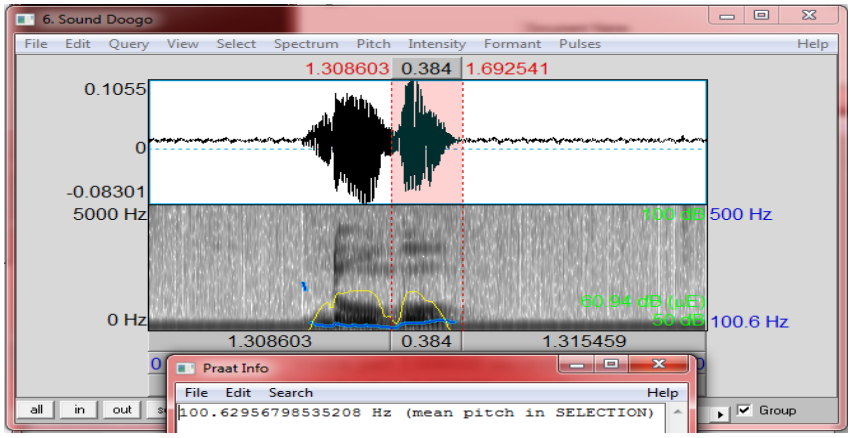


Figure 4-3. Realization of /dòogó/

[16] All PRAAT illustrations of Moore are based on my pronunciation.

In this picture, the last syllable <go> is realized at a pitch of 100.62 Hz. This syllable is the one that will be affected as /wénd/ and /dòogó/ are joined in one word /wéndòogó/.

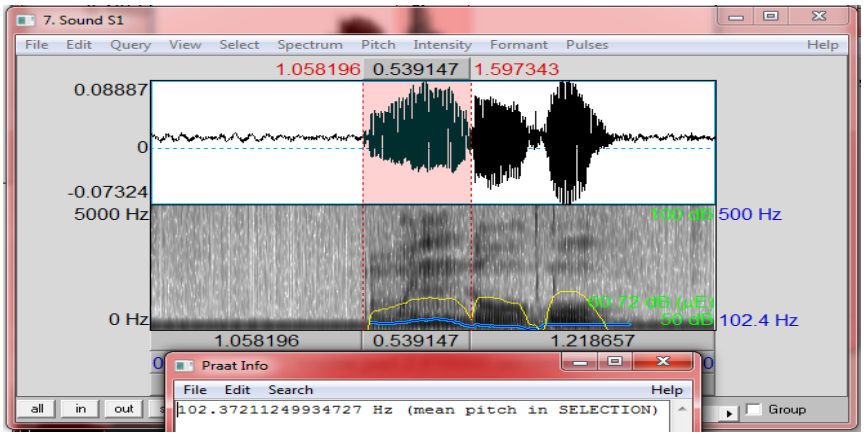


Figure 4-4. Pitch of /wénd/ in /wéndòogó/

Here, the first syllable of <wendoogo>, <wend> is realized at a pitch of 102.37 Hz, which is at least as high as in figure 4-2.

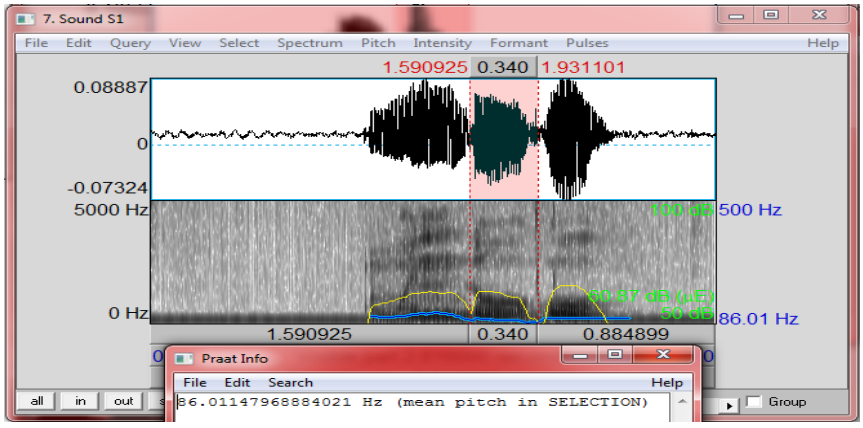


Figure 4-5. Pitch of /dòo/ in /wéndòogó/

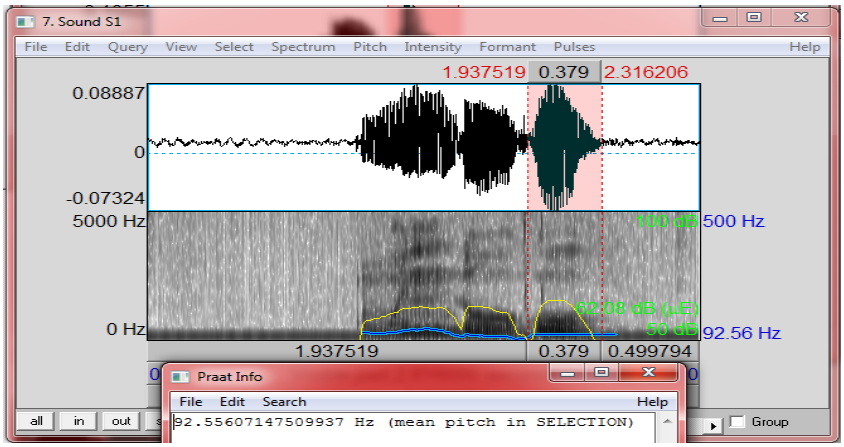


Figure 4-6. Pitch of /gó/ in /wéndòogó/

In this last picture, the pitch of <go> falls from 100.62 Hz in figure 4-3 to 92.55 Hz in the above picture. Compare to <doo> in the middle position of <wendoogo> whose pitch is about 86 Hz in figure 4-5, <go> is high but less high than <wend> pitch in figure 4-4, 102.37 Hz. This shows how the pitch of <go> is impacted by the downstep phenomenon.

The second step is the assimilation of a low tone meaning that a low tone between two high tones is raised to the same level as the preceding high tone.

E.g. ká + zɛ̀r + gá kázɛ́rgá cereal

H L H_1 H H H_1

In the orthography of Moore, <kázɛ́rgá> would be written as <kázɛrga> because of the rule mentioned earlier that says that a syllable without tone is pronounced on the same register as the previous syllable that has a tone.

Homographs

As mentioned in chapter I, homographs are words spelled the same but pronounced differently. In Moore, tones are used to distinguish homographs meaning. Without tones, reading a text passage in Moore would sound awkward. The following observation has been made in (Koffi, 2010):

> When reading homographs (out loud) in English, one must rely on contextual, morphological, and syntactical cues in order to assign the correct pronunciation. Sometimes, one may have to wait until much later in the sentence to find the cue. This causes the reading to lack fluidity. (p. 9)

That would lead to say that tones should be marked in the orthography of Moore, which has plenty of homographs. The following table lists just a few of the many homographs that exist in Moore.

Table 4-8. Homographs and their pitch height in Moore

Homograph	Pitch(Hz) (H or L)		Homograph	Pitch (Hz) (H or L)	
kìisì	kìi	Sì	kíisì	kíi	sì
	114.05 (L)	113.70 (L)		155.08(H)	95.93 (L)
kídbri	kíd	bri	kìdbrì	kìd	brì
	118.46 (H)	120.33(H)		105.73 (L)	103.34 (L)
kṹp	154.27 (H)		kṳ̃p	117.58 (L)	
mṍogo	mṍo	go	mõ̀ogó	mõ̀o	gó
	114.63(H)	114.53 (H)		96.53 (L)	115.68 (H)
píbi	pí	bi	pìbi	pì	bi
	133.77 (H)	133.55 (H)		108.82 (L)	107.34 (L)
sáagà	sáa	gà	sáaga	sáa	ga
	108.56(H)	86.80 (L)		100.78(H)	101.51 (H)
tấabò	tấa	bò	tã̀abó	tã̀a	bó
	108.27(H)	78.97 (L)		89.023 (L)	101.58 (H)
tàalé	tàa	lé	táalè	táa	lè
	90.54 (L)	105.26 (H)		110.40(H)	92.56 (L)
Average Pitch	H: 121.16	L: 96.81		H: 114.17	L: 101.87
Overall average pitch		H: 117.67		L: 99.34	

The above table contains homographs that can be differentiated by the tones or the pitches with which they are pronounced. The difference in meaning between these homographs is symbolized in the difference in pitch height. As noted earlier, tones are not marked in the orthography of Moore. The reason is that the orthography is intended to be as simple as possible. Another reason may be that one can distinguish homographs' meanings by relying only on contextual clues as if they were polysemic words. Polysemic words are words exactly written and pronounced the same way, because they have the same tones, but with different meanings. Therefore, the only way to differentiate them is to rely on the context in which they occur. Table 4-9 contains a few of these words.

We illustrate one case of the above homographs to show the pitch height role in differentiating the meaning of similar spelled words.

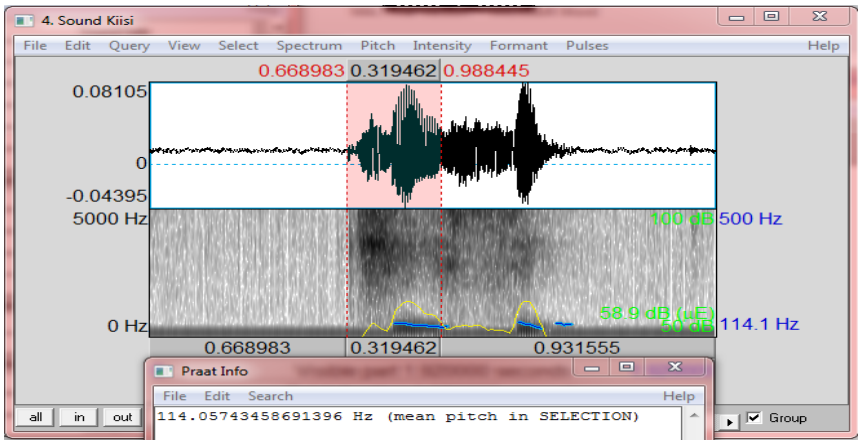


Figure 4-7. Realization of /kìi/ in the word /kìisì/

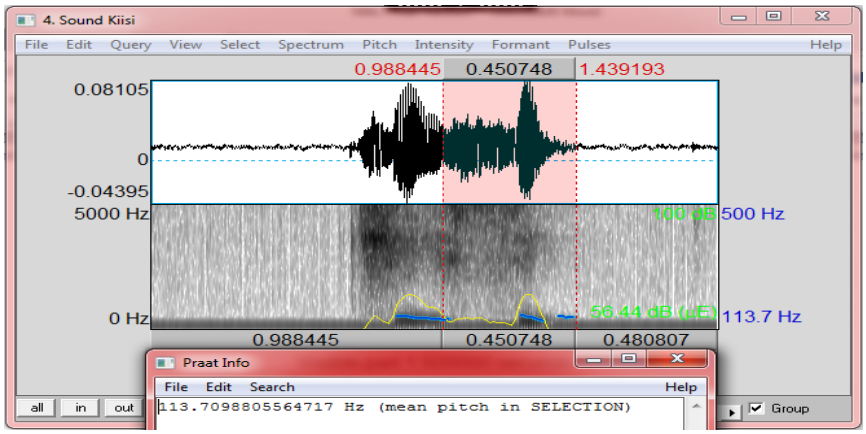


Figure 4-8. Realization of /sì/ in the word /kìisì/

In figure 4-7, /kìi/ has a 114.05 Hz pitch and /sì/ in figure 4-8 is realized with a pitch of 113.7 Hz. The two syllables of about the same low pitch concatenate to form the word /kìisì/, which means, “to extinguish”. In the following case, the same word has a different meaning because of the pitch height change.

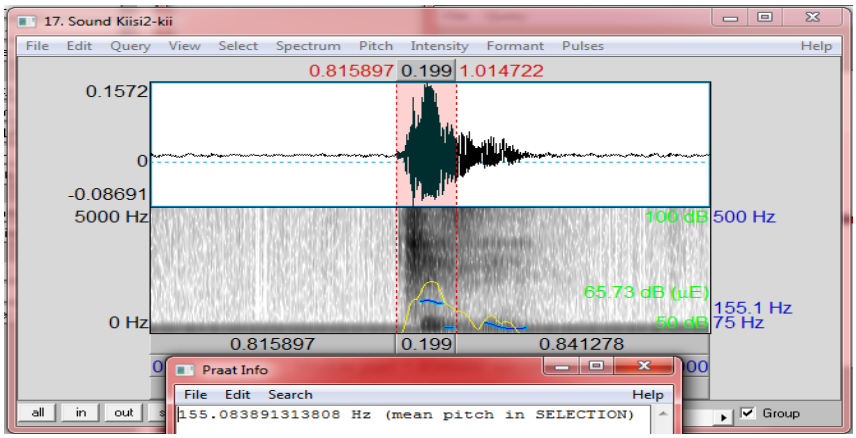


Figure 4-9. Realization of /kí/ in the word /kíisì/

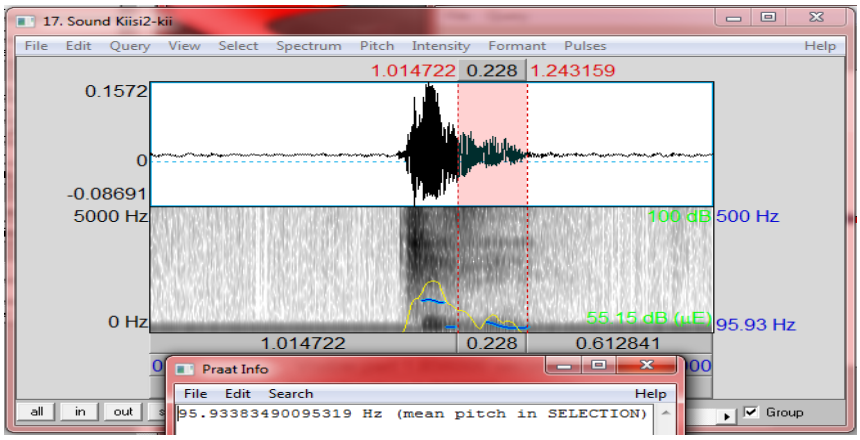


Figure 4-10. Realization of /sì/ in the word /kíisì/

In figure 4-10, /kí/ has a pitch height of 155.1 Hz, which is much higher than the 113.87 Hz average in figure 4-7, of /kìi/ and /sì/. The pitch of /sì/, 95.93 Hz, in figure 4-10 is much lower. These tonal changes also cause a change in the meaning of the word /kíisì/, which means, “to hate” or “taboo”. The following table contains a few examples of polysemic words in Moore.

Table 4-9. Polysemic words

Word	Gloss	Word	Gloss
gãagà	action of sleeping	gãagà	a fruit
píbi	surprise	píbi	struggle for survival or freedom
mɔ̀ogó	territory	mɔ̀ogó	world
mɔ̃ogo	grass	mɔ̃ogo	outside
à	determiner	à	pronoun
bág-bage	do something in an intense way	bág-bage	sharpness
hóbràgi	ostentation	hóbràgi	exlamation of admiration
kɛ̀egá	green color	kɛ̀egá	parrot
wãnà	how	wãnà	few
zɛ̃kè	lift	zɛ̃kè	dry

SUMMARY

This chapter is a brief overview of the phonetic features of the graphemes of Moore such as the vowels, the diphthongs, the triphthongs and the consonants. The word formation process, the syllable structure as well as the homographs

and polysemous words were introduced. As said above, the orthography of Moore avoids the use of tones.

Chapter V

IMPLEMENTATION THROUGH THE CONCATENATIVE SYSTEM

The previous chapter was dedicated to the linguistic and phonetic analyses of Moore. Now we are ready to use these linguistic and phonetic features in the creation of the database of diphones needed in order to synthesize sound speech through the Mbroli player. This section focusses on creating a database of diphones proper to Moore. The completion of this task requires a number of steps discussed previously. The present chapter presents in details the creation process of the database of Moore diphones.

MOORE VOICE CREATION

Components of MBROLA TTS System

A text-to-speech system based on MBROLA mainly consists of the following components:

- An NLP (Natural Language Processing) component, which converts sentences, through several steps, into PHO interface files with information about the duration and pitch of each phone.
- A DSP (Digital Signal Processing) component, i.e. in this case the

MBROLA runtime software, which has a diphone database and a PHO interface file, that produces audio files as output.

- The voice creation component, which is responsible for creating a diphone database, is called mbrolation.

The mbrolation component, using the *mbrolator* software, is licensed from the MBROLA team. In this study, we had the great opportunity to work with Professor Dafydd Gibbon. The mbrolator software converts a set of diphone files and a metadata file, with information about the diphones, into a diphone database.

DIPHONES CREATION PROCESS

First Phase: Corpus Creation

The corpus creation is the first step in the process that will lead us to the diphone database needed by the MBROLA TTS system. The corpus is a list of sentences in Moore from which the diphones are extracted. The corpus is created following these steps:

First step: list of Moore phones. We must use the phones of Moore identified earlier and presented in Table 3-2, 3-3, and 3-6, which is why the list of phonemes, composed of the 25 orthographic graphemes, of Moore[17] was used. To this list, the underscore "_" is added because it is used and recognized as a pause

[17] This list is the same as presented in chapter III.

sign by MBROLA. A total number of n+1 units (n representing the number of the language's phonemes and 1 representing the underscore sign) are therefore needed for any language. Allophones are eventually included in the reference list if the language contains some, which is not the case of Moore.

Second step: phoneme digrams.The reference list of Moore phonemes and the pause symbol were used to create a second reference list of phoneme digrams. According to (Gibbon, 2010), a phoneme digram is "a sequence of 2 phonemes (sometimes the word *diphone* is also used for this, but here the word diphone will be reserved for the actual items which are cut out of speech recordings)" (P. 2). This list matches every phoneme with every other phoneme, and the pause symbol "_". The pause symbol can be left out because the mbrolator software adds it automatically anyway. Therefore, a language with "n" number of phonemes would have $(n+1)^2$ -1 phoneme digrams according to (Gibbon, 2010).

In the particular case of Moore, the number of phonemes "n" is the sum of the twenty-five (25) orthographic graphemes plus the five (5) nasal vowels which give us n = 30. The five nasal vowels of Moore were added to the list, since these vowels are not conditioned by their contextual occurrence.

$$(30+1)^2 - 1 = 961 - 1 = 960 \text{ digrams}$$

This number includes all digrams or sequences of two phonemes in Moore. However, the list ended up being shortened because not all sequences of two phonemes or phoneme digrams can occur in Moore. Indeed, the phonotactics

of Moore, which tells us what sounds can be combined or put together to form parts of words, words or sequences of words in Moore, do not allow some phoneme digrams or phonemes combinations. So with respect to the phonotactics, a number of phoneme digrams like <kk>, <wb>, <zw>, which do not occur in a single word, among many others were left out from the phoneme digrams set. The following table lists the phoneme digrams identified in Moore. These phoneme diagrams can occur either in a single word or in a sequence of two words.

Table 5-1. Phoneme digrams

1	_a	20	ãm	39	az	58	bn	77	dd
2	aa	21	an	40	ãz	59	bo	78	de
3	ãa	22	ãn	41	a_	60	bõ	79	dɛ
4	ab	23	ao	42	_b	61	bp	80	df
5	ãb	24	ão	43	ba	62	br	81	dg
6	ad	25	ap	44	bã	63	bs	82	di
7	ãd	26	ãp	45	bb	64	bt	83	dɪ
8	ae	27	ar	46	bd	65	bu	84	dk
9	ãe	28	ãr	47	be	66	bũ	85	dl
10	af	29	as	48	bẽ	67	bʊ	86	dm
11	ag	30	ãs	49	bɛ	68	bv	87	dn
12	ãg	31	at	50	bf	69	bw	88	do
13	ah	32	ãt	51	bg	70	by	89	dõ
14	ai	33	av	52	bi	71	bz	90	dp
15	ak	34	ãv	53	bĩ	72	b_	91	dr
16	ãk	35	aw	54	bɪ	73	_d	92	ds
17	al	36	ãw	55	bk	74	da	93	dt
18	ãl	37	ay	56	bl	75	dã	94	du
19	am	38	ãy	57	bm	76	db	95	dw

96	do	117	ẽg	138	ey	159	fa	180	gã
97	dõ	118	ẽh	139	ẽy	160	fã	181	gb
98	dp	119	ek	140	ez	161	fd	182	gd
99	dr	120	ẽk	141	e_	162	fe	183	ge
100	ds	121	el	142	_ɛ	163	fẽ	184	gẽ
101	dt	122	ẽl	143	ɛb	164	fɛ	185	gɛ
102	du	123	em	144	ɛd	165	fg	186	gf
103	dw	124	ẽm	145	ɛɛ	166	fi	187	gg
104	dy	125	en	146	ɛg	167	fĩ	188	gi
105	dz	126	ẽn	147	ɛk	168	fɪ	189	gĩ
106	d_	127	eo	148	ɛl	169	fk	190	gɪ
107	_e	128	ẽo	149	ɛm	170	fo	191	gk
108	ea	129	ep	150	ɛn	171	fõ	192	gl
109	eb	130	ẽp	151	ɛo	172	fr	193	gm
110	ẽb	131	er	152	ɛp	173	fs	194	gn
111	ed	132	ẽr	153	ɛr	174	fu	195	go
112	ẽd	133	es	154	ɛs	175	fũ	196	gõ
113	ee	134	ẽs	155	ɛt	176	fʊ	197	gp
114	ẽe	135	et	156	ɛy	177	f_	198	gr
115	ef	136	ẽt	157	ɛ_	178	_g	199	gs
116	eg	137	ẽw	158	_f	179	ga	200	gt

201	gu	222	ib	243	iu	264	ʊ	285	kp
202	gũ	223	ĩb	244	ĩu	265	ʋ	286	kr
203	gʊ	224	id	245	iy	266	ʍ	287	ks
204	gv	225	ĩd	246	ĩy	267	ʏ	288	ku
205	gw	226	if	247	iz	268	ɪ_	289	kũ
206	gy	227	ig	248	i_	269	_k	290	kʊ
207	gz	228	ii	249	_ɪ	270	ka	291	kz
208	g_	229	iĩ	250	ɪb	271	kã	292	k_
209	_h	230	ik	251	ɪd	272	kd	293	_l
210	ha	231	il	252	ɪẽ	273	ke	294	la
211	hã	232	ĩl	253	ɪf	274	kẽ	295	lb
212	he	233	im	254	ɪg	275	kɛ	296	ld
213	hẽ	234	ĩm	255	ɪ	276	ki	297	le
214	hɛ	235	in	256	ɪk	277	kĩ	298	lẽ
215	hi	236	ĩn	257	ɪl	278	kɪ	299	lɛ
216	hĩ	237	ip	258	ɪm	279	kk	300	lf
217	ho	238	ĩp	259	ɪn	280	kl	301	lg
218	hõ	239	ir	260	ɪp	281	km	302	lh
219	h_	240	is	261	ɪr	282	kn	303	li
220	_i	241	it	262	ɪs	283	ko	304	lɪ
221	ia	242	ĩt	263	ɪt	284	kõ	305	lk

306	ll	327	mh	348	ne	369	õa	390	õo
307	lm	328	mi	349	nẽ	370	ob	391	op
308	ln	329	mk	350	nf	371	õb	392	õp
309	lo	330	ml	351	ng	372	od	393	or
310	lp	331	mm	352	ni	373	õd	394	õr
311	ls	332	mn	353	nɪ	374	oe	395	os
312	lu	333	mo	354	nk	375	õe	396	õs
313	lʊ	334	mp	355	nl	376	oɛ	397	ot
314	lv	335	mr	356	nm	377	of	398	õt
315	lw	336	ms	357	nn	378	og	399	ov
316	ly	337	mt	358	no	379	õg	400	õv
317	lz	338	mu	359	ns	380	oh	401	ow
318	l_	339	mv	360	nt	381	õh	402	oy
319	_m	340	mw	361	nu	382	ok	403	õy
320	ma	341	my	362	nv	383	õk	404	oz
321	mb	342	mz	363	nw	384	ol	405	o_
322	md	343	m_	364	ny	385	õl	406	_p
323	me	344	_n	365	nz	386	om	407	pa
324	mẽ	345	na	366	n_	387	õm	408	pã
325	mf	346	nb	367	_o	388	on	409	pd
326	mg	347	nd	368	oa	389	oo	410	pe

411	pẽ	432	rg	453	r_	474	sp	495	tĩ
412	pɛ	433	ri	454	_s	475	sr	496	tɪ
413	pi	434	rĩ	455	sa	476	ss	497	tk
414	pĩ	435	rɪ	456	sã	477	st	498	tl
415	pɪ	436	rk	457	sb	478	su	499	tm
416	po	437	rl	458	sd	479	sũ	500	to
417	põ	438	rm	459	se	480	sʊ	501	tõ
418	pr	439	rn	460	sẽ	481	sw	502	tp
419	pu	440	ro	461	sɛ	482	sy	503	tr
420	pũ	441	rõ	462	sf	483	sz	504	tt
421	pʊ	442	rp	463	sg	484	s_	505	tu
422	p_	443	rr	464	sh	485	_t	506	tũ
423	_r	444	rs	465	si	486	ta	507	tʊ
424	ra	445	rt	466	sĩ	487	tã	508	tw
425	rã	446	ru	467	sɪ	488	tb	509	tz
426	rb	447	rũ	468	sk	489	te	510	t_
427	rd	448	rʊ	469	sl	490	tẽ	511	_u
428	re	449	rv	470	sm	491	tɛ	512	ub
429	rẽ	450	rw	471	sn	492	tf	513	ud
430	rɛ	451	ry	472	so	493	tg	514	uf
431	rf	452	rz	473	sõ	494	ti	515	ug

516	ũg	537	uz	558	_v	579	wɪ	600	yy
517	ui	538	ũz	559	va	580	wo	601	y_
518	ũi	539	u_	560	vã	581	wõ	602	z_
519	uk	540	_ʊ	561	ve	582	wu	603	za
520	ũk	541	ʊb	562	vẽ	583	wũ	604	zã
521	ul	542	ʊd	563	vɛ	584	wʊ	605	ze
522	um	543	ʊg	564	vi	585	w_	606	zẽ
523	ũm	544	ʊi	565	vĩ	586	_y	607	zɛ
524	un	545	ʊɪ	566	vɪ	587	ya	608	zi
525	ũn	546	ʊk	567	vo	588	yã	609	zĩ
526	up	547	ʊl	568	võ	589	ye	610	zn
527	ũp	548	ʊm	569	vu	590	yẽ	611	wi
528	ur	549	ʊn	570	vũ	591	yi	612	zo
529	us	550	ʊr	571	v_	592	yĩ	613	zõ
530	ũs	551	ʊs	572	_w	593	yɪ	614	zu
531	ut	552	ʊt	573	wa	594	yn	615	zũ
532	ũt	553	ʊʊ	574	wã	595	yo	616	zʊ
533	uu	554	ʊw	575	we	596	yõ	617	z_
534	ũu	555	ʊy	576	wẽ	597	yu		
535	uw	556	ʊz	577	wɛ	598	yũ		
536	uy	557	ʊ_	578	wĩ	599	yʊ		

Third step: keywords formation. The next step consists of matching each one of the phoneme digrams, we just presented, with a word. If there is no one word containing the digram, a sequence of two words that contains it works. Such word is called keyword or sequence of keywords. In the case of Moore, matching digrams with keywords was done using our knowledge of Moore and a dictionary, the orthographic dictionary of Moore, going back and forth between the dictionary and the digrams list. What we would do is for each digram first letter, look up this first letter in the dictionary and go through the alphabetic order, write down all words containing that letter followed by any other letter. The digrams that we were not able to match with a keyword or a sequence of keywords were eliminated from our list of digrams, presented above. At the end, the digrams reference list was down from 960 digrams to 617. The following table is an example of digrams with their corresponding keywords.

Table 5-2. Digrams and their corresponding keywords

Digram	keyword		Digram	keyword
bo	bodgo		do	mui-zẽedo
bõ	bõe		dõ	rõsendõaaga
fd	nafda/nafdga		gd	lidgda

Fourth step: carrier sentences. Now that the keywords list is made, the next step consists of putting these words into sentences, called carrier sentences

or prompt sentences. (Gibbon, 2010) presents two methods that can be used to create the set of carrier sentences[18]:

- The first method consists of inserting each keyword or sequence of keywords into a sentence. The insertion can take place in the middle for any keyword, but at the beginning or the end for keywords containing digrams with the pause sign.

  E.g. _fú-yorgã yáa miuugù_

  Cloth the is red.

  The cloth is red.

A keyword like /_fú-yorgã/ contains digrams like /fú/, /yá/, /ug/. In the above example, the tones are marked just for illustration purpose. During the recording, the carrier sentences are read with a monotone voice.

- The second one is to select a set of phonetically rich sentences in which the words contain as many digrams as possible in as few sentences as possible. The more the sentence is long, the more digrams it contains. The following sentence is the same as the sentence illustrated in Figure 5-1 and contains 38 phoneme digrams if each phoneme is paired with every other phoneme.

  E.g. _ráwã yèelám tɪ ãadsã yi wʊsg zàamḗ zàabre_

  Man the said that stars the came out a lot yesterday night.

[18] The keywords and the carrier sentences are in the appendix I.

The man said that there were plenty of stars last night.

A corpus of one hundred (100) sentences was built following the two methods just described. The intention is that the corpus covers all the digrams allowed by the phonotactics of Moore. A few sentences contain only one keyword, but most of them contain at least two keywords, three or even more. Semantically speaking, the large majority of the sentences make sense in Moore, but a few of them do not, in that they were just designed to include the most possible digrams. However, these sentences still have an important characteristic needed here, in that they follow the phrasal structure (SVO) of Moore.

Second Phase: Recordings

The recordings were done, late during the night, at a time when there is much less background noise, especially cars passing, people whispering, etc. The reading had to be made in a monotonous tone from the beginning to the end of each sentence, keeping the pitch constant, with no emotions. The recordings were done accordingly to the required format of the mbrolation software:

Sampling frequency: 16 Khz

Resolution: 16 bit

Channels: Mono

The recordings were done using PRAAT software and the free talk audio software, which consist of a headphone with an integrated microphone and a USB key. To record with PRAAT, click on its icon on your desktop, assuming it is already installed on your desktop. If not, one can google it and download it. Then

click on *new* from PRAAT Objects window and choose *Record mono Sound*. It will bring up a small window *SoundRecorder*. From that window, check *Mono* under *Channels* and *16000Hz* under *Sampling frequency*. Then one can start to record, stop, play, save the recordings using the buttons *Record, Stop, Play*, and *Save*. Once recorded, the carrier sentences were saved in the format mentioned above. Each sentence sound file was saved, in the same folder as the others, using this format: S1.wav. The next phase is the post-recordings, which consists of the segmentation and annotation of the corpus.

Third Phase: Segmentation

The recorded corpus was segmented to extract all the diphones and annotate them accordingly. Each diphone annotation must match its corresponding phoneme digram presented earlier. The segmentation and annotation procedures were carried out manually using PRAAT. Each one of the corpus sentences was segmented in phonemes and diphones. One of the advantages of PRAAT is that the same speech wave file can be annotated using different types of labels, i.e., words, phonemes, syllables, etc, that can be stored in the same file. Therefore, each corpus sentence was segmented, labeled according to the phoneme digrams (diphones) it contains and saved. The segmentation process starts first by the annotation or labeling.

Labeling process. Click on *read* and then *read from file* from the PRAAT Objects window. A window opens from which one can open the folder that

contains all the sentences sound wave files and then select the speech sound file to be labeled, S1.wav for example, by double clicking on it. The selected item is copied and highlighted in blue, *as Sound S1*, in the window causing a new menu of commands to appear on the right side of the same window, as shown in Figure 5-1.

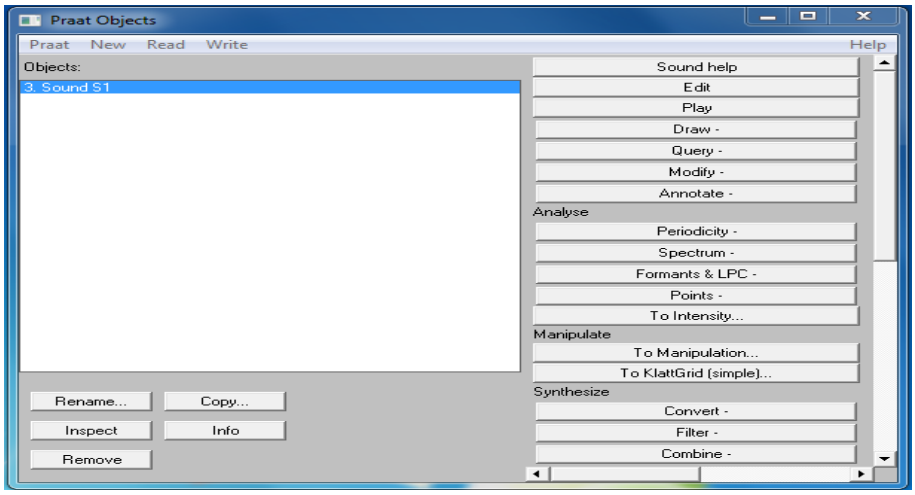


Figure 5-1. PRAAT Objects window

Next, still from the PRAAT Objects window, on the right side menu, click on *annotate* and then *toTextGrid*. This prompts the following window from which one chooses the desired labels. In this case, only the *tier names* box interests us. The words inside the first box are replaced by Phonemes[19]. The second box is left blank.

[19] The labeling process used here illustrates a segmentation in phonemes or phones.

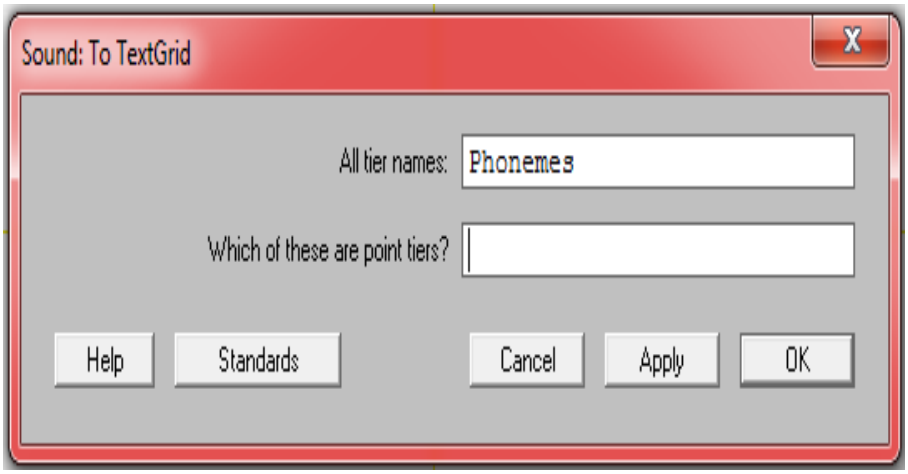


Figure 5-2. TextGrid annotation

A click on *Ok* prompts a highlighted new item that reads *S1TextGrid*, in the PRAAT Objects window just under the item *Sound S1*.

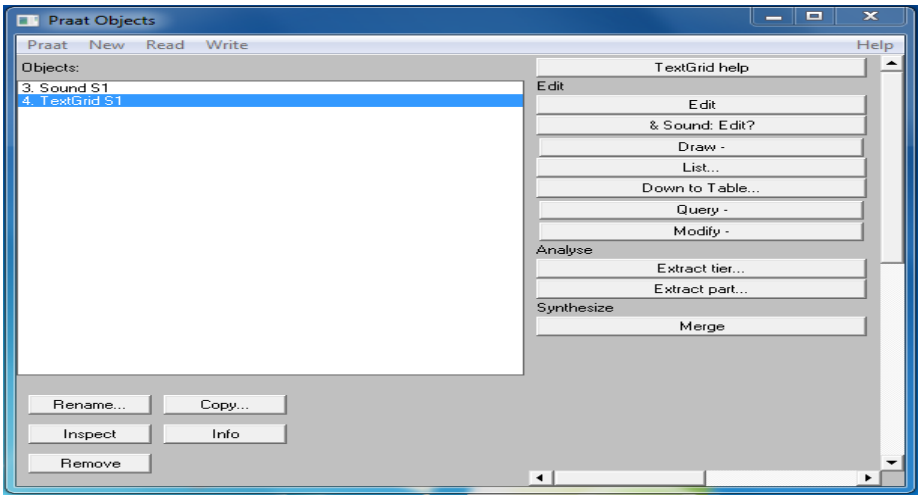


Figure 5-3. PRAAT Objects window with Sound S1 and TextGrid S1

Next by selecting both *Sound S1* and *TextGrid S1*, and clicking on *Edit* on the right side of the window, the following window pops up.

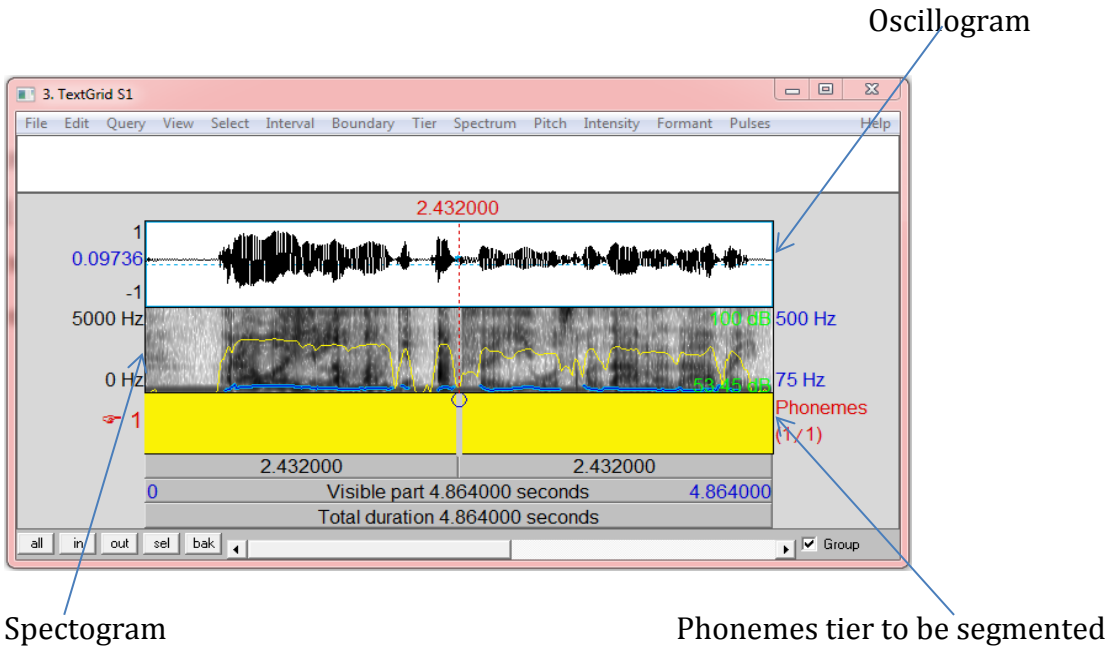


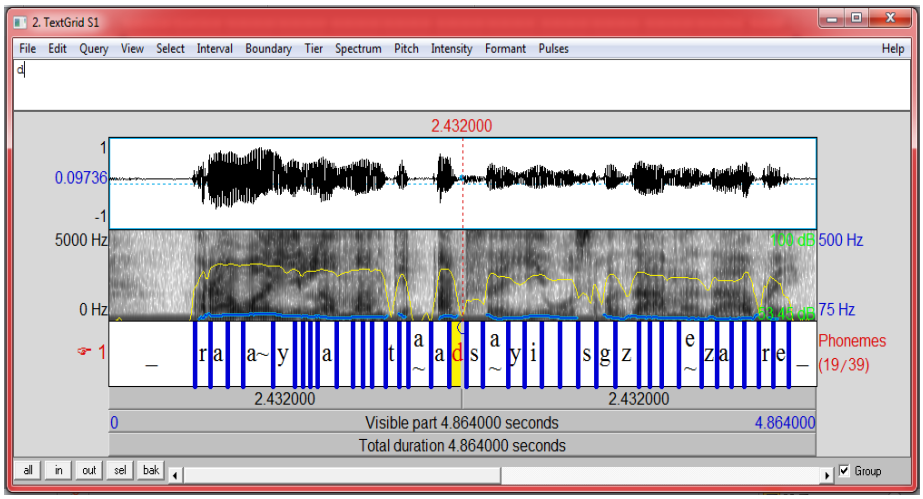
Figure 5-4. TextGrid S1 window with the annotation labels

The annotation is done from this window called *TextGridS1*, where the sound wave can be segmented according to labels such as Phonemes. This window is an interactive platform, where it is possible to play and listen to the entire speech *sound S1* as well as speech fragments or segments by clicking anywhere on the *Phonemes* tier row. This feature allows the segmentation of *Sound S1* into pieces of speech sounds corresponding to phonemes, phoneme diagrams, etc. In segmenting the sound wave files, the graphemes <ʊ, ɪ, and ɛ> were replaced by the following list of symbols because the MBROLA system does not accept special characters. The table also contains the symbol '~' as a nasal sign.

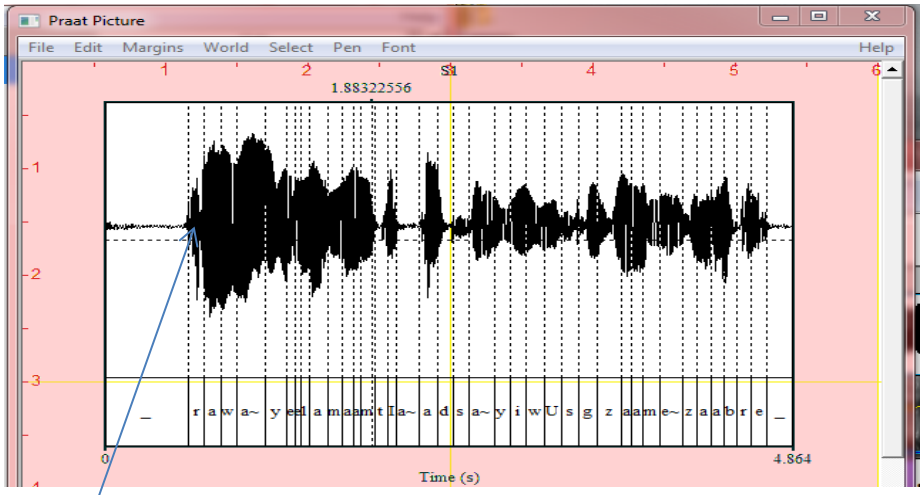
Table 5-3. Replacement symbols

Symbol	Interpretation
[_]	Silent pause sign
[~]	Nasal sign
[E]	<ɛ>
[I]	<ɪ>
[U]	<ʊ>

[E], [I], and [U] are capital letters. The following picture is the actual segmentation and annotation of the wave file of sentence one: *Sound S1*. All of the annotations use symbols from the keyboard. No special character is accepted.



Figure[20] 5-5. Segmented sound speech of sentence one



oscillogram

Figure 5-6. Segmented sound speech and oscillogram of sentence one

The above PRAAT picture shows a clearer segmentation of Sound S1 in phonemes, in correspondence with its oscillogram. The other ninety-nine (99)

[20] A complete tutorial on how to segment and annotate is available on PRAAT website.

sentences sound waves are segmented the same way as described in the third phase. After the segmentation is done, this *TextGrid S1* still highlighted in blue in the PRAAT Objects window, as every other sentence TextGrid, can be saved by clicking on *File* at the top of the window and then clicking on *Write TextGrid to text file*. All TextGrids were saved into the same folder. Once saved, a click on a TextGrid file gives us some technical information about the sound speech such as:

- the number of intervals or phonemes, 39 for Textgrid S1,
- the realization time of the speech sound, 4.864 ms for Sound S1,
- the timestamps of each phoneme. Phoneme "r" for example, is pronounced between 0.589 and 0.699 ms.

Following is a small part of sentence one TextGrid file:

```
File type = "ooTextFile"
Object class = "TextGrid"
xmin = 0
xmax = 4.864
tiers? <exists>
size = 1
item []:
   item [1]:
      class = "IntervalTier"
      name = "Phonemes"
      xmin = 0
```

xmax = 4.864

intervals: size = 39

intervals [1]:

xmin = 0

xmax = 0.5899586573950989

text = "_"

intervals [2]:

xmin = 0.5899586573950989

xmax = 0.6994285428984759

text = "r"

intervals [5]:

xmin = 0.9225786941168982

xmax = 1.128887324488647

text = "a~"

intervals [6]:

xmin = 1.128887324488647

xmax = 1.2804610121087077

text = "y"

intervals [7]:

xmin = 1.2804610121087077

xmax = 1.335195954860396

text = "e"

Fourth Phase: Diphones

The diphone files set consists of actual wave files each containing one single diphone cut out of the recorded corpus and annotated with the name of each diphone (or its corresponding digram) composing the file. When segmenting the recorded speech to cut out the diphones, 50 milliseconds of the speech signal are kept or added at the beginning and end of each diphone. This is a requirement for more accurate analysis by the MBROLA software. Each segmented diphone is saved as a separate file and annotated identically to the phoneme digram to which it corresponds in the reference list, e.g. the wave files like "_b. wav ", " af. wav " are named after the phoneme digrams /_b/ and /af/.

The procedure of cutting out the diphones from the segmented TextGrids previously saved is done from the PRAAT Objects window. For that, both speech sound wave and its corresponding TextGrid are selected by clicking on both items while maintaining the left button of the mouse pushed. Then, from the right side of the window, click on *Extract* and *Extract non-empty intervals*. The PRAAT system will extract all non-empty intervals or segments, which appear in the PRAAT Objects window. The diphones are saved individually as wave files in the same folder. The structure of a diphone file should therefore look like this according to (Gibbon, 2010):

> $|_1$ 50ms $|_2$ left half-phone $|_3$ right half-phone $|_4$ 50 ms $|_5$
> The vertical bars signify points in time:
> 1.$|_1$ beginning of file
> 2. $|_2$ beginning of diphone
> 3. $|_3$ middle of diphone (phone boundary)
> 4. $|_4$ end of diphone

5. $|_5$ end of file. (p. 5)

Fifth Phase: Metadata File

The diphone wave files are accompanied with a file of technical data about the diphones, called metadata file. The list of all the extracted and saved diphone files from the recorded speech along with information about the timestamps[21] for the beginning, middle and end of each diphone file are converted to a plain text file and saved as such. This plain text file is called metadata file or SEG file because of its extension ".seg" that stands for segmentation. A metadata file looks like this:

[21] The timestamps are given in samples with 16000 samples per second or 16 samples per millisecond.

Table 5-4: Metadata file sample of TextGrid S1

File name	Diphone	Start	End	Mid
_r.wav	_ r	800	1775	900
ra.wav	r a	800	2585	1675
aw.wav	a w	800	2585	1709
wa~.wav	w a~	800	3326	1675
a~y.wav	a~ y	800	3663	2450
ye.wav	y e	800	2450	2012
ee.wav	e e	800	1608	1237
el.wav	e l	800	1608	1170
la.wav	l a	800	2282	1237
am.wav	a m	800	2719	1844
ma.wav	m a	800	2248	1675
aa.wav	a a	800	1844	1372
am.wav	a m	800	2012	1271
mt.wav	m t	800	2315	1541
ti.wav	t i	800	2079	1574
ia~.wav	i a~	800	2517	1305
a~a.wav	a~ a	800	3056	2012
ad.wav	a d	800	2753	1844
ds.wav	d s	800	2618	1709
sa~.wav	s a~	800	3124	1709
a~y.wav	a~ y	800	3124	2214
yi.wav	y i	800	2618	1709
iw.wav	i w	800	2719	1709
wu.wav	w u	800	2787	1810
us.wav	u s	800	2787	1776
sg.wav	s g	800	2854	1810
gz.wav	g z	800	3157	1844
za.wav	z a	800	2719	2113
aa.wav	a a	800	2046	1406
am.wav	a m	800	2383	1439
me~.wav	m e~	800	3023	1743
e~z.wav	e~ z	800	2888	2079
za.wav	z a	800	2416	1608
aa.wav	a a	800	2315	1608
ab.wav	a b	800	2282	1507
br.wav	b r	800	2383	1574
re.wav	r e	800	2484	1608
e _.wav	e _	800	1775	1675

This metadata file was obtained automatically using the software TextGrid to Mbrolator voice creator format (seg file) converter. The software takes a TextGrid file on a word format like the TextGrid of sound 1, on page 82-83, as input. It contains technical data about the duration in milliseconds of the start, middle and end of each recording file or diphone file. Finally, the sentences wave files folder obtained from the second step (recording step); the sentence TextGrids folder obtained from the third step (segmentation process, Figure 5-5) and the metadata file (SEG file) are converted into the diphone database.

The following table is the actual diphone database and their illustration with MBROLA characters and Moore orthographic characters. The illustration with MBROLA characters shows the characters accepted by MBROLA. The orthographic characters illustrate the way they would be written in regular Moore script. As explained earlier only three orthographic characters, <ɛ>, <ɩ>, <ʋ>, and the tilde sign “˜” were replaced respectively by <E>, <I>, <U>, and “~” because MBROLA does not accept special characters.

Table 5-5: Moore diphone database

Moore diphones							
diphone		MBROLA character	Moore orthograhy	diphone		MBROLA character	Moore orthograhy
1	a	m*a*m	m*a*m	16	m	ru~*m*	rũ*m*
2	a~	raw*a*~	raw*ã*	17	n	*n*eda	*n*eda
3	b	*b*iiga	*b*iiga	18	o	f*o*	f*o*
4	d	*d*aaga	*d*aaga	19	o~	b*o*~e	b*õ*e
5	e	wa~b*e*	wãb*e*	20	p	le*p*re	le*p*re
6	e~	zaame~	zaam*ẽ*	21	r	*r*oogo	*r*oogo
7	E	p*E*ka	p*ɛ*ka	22	s	yE*s*e	y*ɛs*e
8	f	*f*uugu	*f*uugu	23	t	b*ɪt*o	b*ɪt*o
9	g	pa*g*a~	pa*g*ã	24	u	m*u*ka	m*u*ka
10	h	*h*ato	*h*ato	25	u~	y*u*~	y*ũ*
11	i	L*i*gd*i*	l*i*gd*i*	26	U	l*U*l	l*ʊ*ɪ
12	i~	s*i*~bga~	s*ĩ*bgã	27	v	*v*ima	*v*ima
13	I	t*I*pe	t*ɪ*pe	28	w	ra*w*a~	ra*w*ã
14	k	*k*e~ema~	*k*ẽemã	29	y	*y*aare	*y*aare
15	l	Yee*l*e	yee*l*e	30	z	*z*e~edo	*z*ẽedo

SUMMARY

In this chapter, the different steps required for the creation of Moore diphone database were listed and explained. The creation process includes the elaboration of three lists. The first list contains the phonemes of Moore (see Table 3-2, 3-3, and 3-6) and the second list contains the phoneme digrams (see Table 5-1). The third list contains the keywords that include the digrams and this list of keywords was used to create the corpus of carrier sentences to be recorded (see Appendix I). Lastly, the corpus is segmented and annotated, from which the diphones are extracted.

The speech sound wave files folder, the diphone set files folder, the TextGrid files folder, and the metadata file constitute the items used to create the voice and the database needed by MBROLA to read an input text of Moore in a loud speech. They were zipped and sent to Dr. Daffyd Gibbon, an mbrolator software licence holder, because we are not licensed by the software developer to carry out such operations. Table 5-5 contains the diphones that can be used to synthesize speech in Moore using MBROLA. The next section will evaluate how intelligible and natural, the synthesized speech is to the human ears.

Chapter VI

EVALUATION OF MOORE SYNTHESIZED SPEECH

The goal of this thesis was the creation of a database of diphones for Moore that can be used as an input to synthesize speech sounds. It is thus possible to assess how well the synthesized speech sounds to the human ears and how good is the diphone database. Assessing a synthesized speech is not an easy task because there are many aspects related to synthetic speech (Bachan, 2007, p. 60). Among these aspects is the fact that the evaluation is done through human listeners unaccustomed to other type of speech than the human voice. This means that "listeners must often expend more effort to understand and comprehend synthesized speech (…). Especially for users unaccustomed to synthesized speech, listening to a speech synthesizer for extended periods can be both tiring and unsatisfactory" (Lampert, 2004, p. 3). However, the most effective way to have a good idea on how natural and intelligible is a synthesized speech, would be to have humans listen to it and make judgments, no matter how subjective they could be. "One of the main system-level synthesis evaluation techniques is to have humans listen to the result and respond to specific questions or make subjective judgements " (Lampert, 2004, p. 3).

This chapter presents the results of the evaluation of Moore speech sound produced from Moore diphone database through MBROLA PHO player, Mbroli. Five listeners all students of SCSU and native speakers of Moore served as participants to this evaluation session. The speech sound of the words and sentences tested in this chapter were created following the methodology described in chapter III. The participants were exposed to synthesized speech and asked to respond to a number of questions depending on what aspect of the speech output is being tested. The two aspects tested here are the speech sound intelligibility and naturalness. This evaluation methodology is based mostly on (Gera, 2006).

INTELLIGIBILITY TEST

This test is concerned with how understandable is the speech sound to the listeners. It is composed of three tasks: a comprehension task, a phonetic task and a transcription task.

Comprehension Task

A short passage of five sentences is played to each one of the participants. After they listen to the passage, they are asked to repeat the sentences that they just heard. The results are presented in the table below and show how many sentences each subject understood from the passage. The sentences tested are the following:

(1) Tấpã pãrga rã-zãndå̀, wa nédẽ lʋi vãvãbdò.

Arrow the went through alcohol hangar the like someone fall noise.

The arrow went through the hangar of alcohol, like someone falling.

(2) Ráwã yèelám tí ấadså̀ yí wʋsg zàamḗ zàabre.

Man the said that stars the came out a lot yesterday night.

The man said that there were a lot of stars last night.

(3) Púg-peelèm néd ka mi widb yè.

Honesty man not curse.

An honest man does not curse.

(4) Ràkãagr n bé zak kãngà púgà.

Rich man is house this inside.

There is a rich man living in this house.

(5) Sʊ̀lg né yes-kõabnengà n bé yolg n wå̃.

Spider and insect is bag the.

There is a spider and an insect in the bag.

Table 6-1. Comprehension task scores

Subject	Number of correct sentences	Recognized sentences
1	4	(2), (3), (4), and (5)
2	3	(2), (3), and (5)
3	5	(1), (2), (3), (4), and (5)
4	3	(2), (3), and (5)
5	3	(2), (3), and (5)
Average	72%	

The five participants seem to have no difficulties with sentences (2), (3), and (5). Four subjects, due to its first two words Tấpã and pãrga, did not correctly recognize sentence (1). The reason they gave is that they could not hear the beginning of the sentence. Three subjects had trouble with sentences (1) and (4). The reason for sentence (4) is the same as for sentence (1), i.e., that they could not hear the beginning of sentence (4), which starts withRàkãagr. The overall success for this task is 72%.

Phonetic Task

The phonetic task consists of two tasks. Both are aimed at testing the ability of the participants to detect a specific sound segment, word-initial or word-final consonant.

Diagnostic rhyme test (DRT). The DRT tests how perceptible is the initial consonant of words in Moore synthesized speech. To perform this test, five pairs of words were played to the listeners who were asked to identify them. The two words of each pair only differ by the first phoneme and can be considered as minimal pairs. The five pairs of words tested are presented in the following table.

Table 6-2. Pairs of words used for the DRT

Pair number	Group A	Group B	Gloss A	Gloss B
(1)	nãmse	mãmse	suffer	try
(2)	daaga	raaga	market	market
(3)	kãoore	taoore	sauce ingredient	in front of
(4)	pɛka	tɛka	slap	all
(5)	ya	ta	3rd person pronoun	2nd person pronoun

The results obtained show how many pairs were identified correctly.

Table 6-3. DRT scores

Subject	Correct identifications			
	Group A		Group B	
	Number of words identified	Pair's word identified	Number of words identified	Pair's word identified
1	4	(2), (3), (4), (5)	4	(2), (3), (4), (5)
2	3	(3), (4), (5)	3	(1), (3), (5)
3	4	(1), (3), (4), (5)	4	(1), (3), (4), (5)
4	4	(1), (3), (4), (5)	4	(1), (3), (4), (5)
5	3	(3), (4), (5)	4	(1), (3), (4), (5)
Average	72%		76%	
Overall average	74%			

As the table shows, all the participants had no problem with group A words' initial phonemes of minimal pairs (3), (4), and (5). However, three subjects did not recognize the first word in minimal pair (1) and four subjects were unable to recognize minimal pair (2) first word. One subject did not identify minimal pair (1) group B word and four participants did not catch minimal pair (2) second word. The main reason why some of the words were not identified

and according to the participants is the quality of the audio sound. Nevertheless, the overall performance is 74% for both groups of words.

Modified rhyme test (MRT). The MRT is just another way of calling the DRT. The difference is that in MRT, the word final or last consonant, instead of the word-initial consonant, is being tested. As in the DRT test, the five subjects were asked to identify the last consonants of five pairs of words. The two words of each pair differing from each other by the last consonant. The following table contains the pairs of words that were tested.

Table 6-4. Pairs of words used for the MRT

Pair number	Group A	Group B	Gloss A	Gloss A
(1)	naare	naase	natural /fresh	four
(2)	lende	lenge	shallow	press
(3)	kãoore	kaodẽ	bitter ingredient	breaking
(4)	yoobe	yoole	six	excess of oil
(5)	wɪde	wʊbe	criticize	taboo

The results obtained show how many pairs were identified correctly.

Table 6-5. MRT scores

Subject	Number of correct identification out of 5 words			
	Group A		Group B	
	Number of words identified	Pair's word identified	Number of words identified	Pair's word identified
1	4	(2), (3), (4), (5)	1	(5)
2	5	(1), (2), (3), (4), (5)	5	(1), (2), (3), (4), (5)
3	3	(2), (4), (5)	4	(1), (3), (4), (5)
4	2	(2), (4)	4	(1), (3), (4), (5)
5	3	(2), (4), (5)	2	(3), (5)
Average	68%		64%	
Overall average	66%			

Four participants in this task were not able to identify the final consonant of pair (1) group A word and three of them could not identify pairs (1) and (3) first words final consonants. Only one participant succeeded in all five words from group A and B. In the second word or group B of each minimal pair, one subject recognized only one final consonant of pair (5) while two subjects failed in pair (2). Moreover, one subject did not recognize pairs (1), (2) and (4) last word consonants. The overall average of this task is 66%.

Transcription Task

In this task, three sentences of five words each were played to the listeners and they were asked to identify each one of the five words in each sentence. The idea was to have them write down or better transcribe what they have heard, but since the participants are illiterate in Moore, they were just asked to repeat what they heard. This is called Semantically Unpredictable Sentences test (SUS). The test is so named because the subject cannot guess in advance the words of a sentence, given that the sentences are meaningless and are made of a bunch of words put together. The three sentences used in this test are:

Sentence I

Kãngà - tɑ́pã - ràkãagr - zàabre - púgà.

That – arrow the – rich man – afternoon – inside

That – the arrow – rich man – afternoon - inside

Sentence II

Pʋgbi - mínisr – zɑ́ngà – zàabre - Pʋg-peelem

Little woman – minister – all – afternoon – stomach white

Fiancée – secretary – all - afternoon - honesty

Sentence[22] III

Kãngà - tɑ́pã - fú-yorgã - bùsɑ́angà - vãvãbdò

That – arrow the – shirt the – a member from the bùsɑ́angà ethny – noise of something falling down

That – the arrow – the shirt - a member from the bùsɑ́angà ethny – noise of something falling down

Table 6-6. Transcription test scores

Subject	Number of correct identification out of three sentences of five words each.		
	Sentence 1	Sentence 2	Sentence 3
1	4	4	3
2	4	4	3
3	3	4	3
4	3	3	3
5	4	4	4
Average	72%	76%	64%
Overall average	70.66%		

[22] The details (phoneme duration, and pitch heigth) of some synthesized words and sentences used in this section are in appendix II.

The table indicates the number of words each participant has been able to identify correctly. This task score reveals an average of 70%. Again, the quality of the sound was given as the main reason why some words were not identifiable.

NATURALNESS

This test uses a technique called MOS, which stands for Mean Opinion Score and is characterized as a subjective way to have an idea about the naturalness of the synthesized speech (Gera, 2006). The subject listeners were asked to rate the naturalness of all the speech sounds they have been exposed since the beginning of the evaluation. The rating is done on a scale of one to four, where one is very natural, two is natural, three is fairly natural, and four is not natural. The following table shows the results by participant and the average rating score out of maximum score of four.

Table 6-7. Naturalness test scores

Subject	Rating by subject
1	2
2	3
3	3
4	2
5	3
Average	2.6/4

RESULT ANALYSIS AND LIMITATIONS

The comprehension task table average score of 72% indicates that the participants were able to understand an important part of the passage played. The phonetic test shows a DRT of 74% and an MRT of 66% giving an average of 70%. It also demonstrates that the initial of the words are more perceptible than the finals. The transcription test score of 70.66% indicates that quite a considerable number of words can be heard and transcribed. Lastly, the naturalness of the sound score seems to indicate an average score of 2.6/4, which means that the synthesized speech can be ranked from fairly natural to natural. These different scores indicate that the participants performed relatively well at the comprehension task, which is why they also rated the naturalness of the speech at almost the same level.

The failures of the participants, in some of the different tasks, can be explained by limitations like the recording conditions. The corpus has been recorded late at night to avoid as much background noise as possible, but it turns out that there was some relatively perceptible background noise. Another limitation is the relative short length of the corpus. This corpus was intended to cover the maximum possible sequences of sounds that occur in the language. This is a challenging task to say the least, as it is virtually impossible to cover all the possible sounds occurring in a language in just one hundred sentences. We should also note that the testing was done on a voice made up with randomly selected diphones from the 617 segmented diphones. So obviously, there is no way to guarantee that the selection represents the best possible diphones, given the fact that the diphones were choosen randomly by the software. As a result, the quality of the voice depends on the quality of the selected diphones.

The performance of the participants could have been improved if the words and sentences tested were picked based on their actual frequency in the language. In addition, the participants are mostly inexperienced speakers of Moore, not to mention that they do not speak the same version (Moore has six dialects) of the language, which has certainly played a role in their performance. In addition each one of the participants is speaker of at least two languages, which may have caused some negative interferences in term of how they perceived the speech sound they were exposed in this test. Another aspect of the evaluation could have involved a task asking the participants to listen and

compare a natural recording and a synthesized version of the same passage. In addition, the sentences tested in this evaluation could have included more imperative and interrogative sentences as well as affirmative sentences.

SUMMARY

This chapter describes how intelligible and natural is the speech synthesizer of Moore. Given all the results, it is correct to say that the speech sounds synthesized are intelligible and natural, even though the participants had some difficulties with some synthesized speech words or sentences. The core of the MBROLA system is the diphone database of the language under investigation. In other words, the quality, intelligibility, and naturalness of the speech synthesized depend largely on the quality of the diphone database that was created. In light of the test result, it can be stated that the quality of Moore diphone database is acceptable.

FUTURE WORK ORIENTATION

In this study, what we did was to use a multilingual software, MBROLA, to create a system that can be used to read words, sentences in Moore. The results seem to indicate that the system works at least for a reasonable percentage of the tested speech. In order to make it work better, future studies would have to choose a much larger corpus and make sure the recording part is done in a soundproof environment, so that they can minimize background noises. The

MBROLA software was used here as a template, and the voice creation was done by a third party because we had no control over the software. An ambitious project would avoid such difficulties by trying to create a text-to-speech software proper to Moore.

CONCLUSION

This study can be viewed as an attempt for building and developing a text-to-speech system for Moore based on MBROLA framework. In doing so, we reviewed a fair amount of the previous works on the TTS of African languages. The previous studies and this one show that a phonetic, phonological, morphological, and orthographic study of the language is a prerequisite for the implementation of a good TTS system for the language under consideration. In the implementation section and in addition to MBROLA itself, we used a number of softwares such as PRAAT, Praat TextGrid to MBROLA synthesizer format (pho file) converter, and TextGrid to Mbrolator voice creator format (seg file) converter, among others. For languages having transparent orthographies, that is, they have a regular grapheme-to-phoneme correspondence, it seems less challenging to build a speech synthesizer than for languages with opaque orthographies. The most formidable challenge, for Moore and for other African languages that have lexical and grammatical tones, is how to design a robust TTS system that can read tones accurately. This is an area of investigation for future studies. The evaluation part of this study shows that the diphone database and the voice accompanying it can synthesize an intelligible and natural speech.

SOFTWARE

SOFTWARE

Boersma, P., & Weenink, D. (2001). Praat [Computer software]. Retrieved from www.fon.hum.uva.nl/praat

Dutoit, T. (1996). Mbrola [Computer software]. Retrieved from http://www.tcts.fpms.ac.be/synthesis/mbrola.html

Gibbon, D. (2008). Praat TextGrid to Mbrola synthesizer format (pho file) converter [Computer software]. Retrieved from wwwhomes.uni-bielefeld.de/.../PHONETICS/textgrid2mbrola.html

Gibbon, D. (2009). TextGrid to Mbrolator voice creator format (seg file) converter [Computer software]. Retrieved from http://wwwhomes.uni-bielefeld.de/gibbon/Forms/Python/PHONETICS/textgrid2mbrolator-example.html

REFERENCES

REFERENCES

Alexandre, G. (1953). La langue Moore. *IFAN*, *34*.

Anberbir, T., & Tomio, T. (2009). Development of an Amharic text-to-speech system using cepstral method. In *Proceedings of the first workshop on language technologies for African languages* (Vol. 1, pp. 46-52). Association for Computational Linguistics.

Bachan, J. (2007). *Close copy speech synthesis for perception testing and annotation validation* (Unpublished master's thesis). University of Bielefeld, Faculty of Linguistics and Litterature.

Balima, P. (1997). *Le Moore s'ecrit ou manuel de transcription de la langue Moore*. Ouagadougou, Burkina Faso: Promo-langues.

Black, A., & Taylor, P. (1997). *Festival speech synthesis system: System documentation* (Rep. No. Technical Report HCRC/TR-83). Edinburgh, Scotland: Human Communication Research Centre. doi: http://www.cstr.ed.ac.uk/projects/festival.html.

Canu, C. (1975). *A synchronic description of Moore (Dialect of Ouagadougou)*. Paris: SELAF.

Dutoit, T. (1997). *An introduction to text-to-speech synthesis*. Dordrecht: Kluwer Academic.

Faria, A. (2003). *Applied phonetics: Portuguese text-to-speech* (Unpublished master's thesis). University of California, Dept of Linguistics.

Fromkin, V. (2000). *Linguistics: An introduction to linguistic theory*. Massachusetts: Blackwell.

Fromkin, V., Rodman, R., & Nina, H. (2010). *An introduction to language*. Boston, Massachusetts: Wadsworth.

Gibbon, D. (2010, August 08). Text-to-speech system [E-mail to the author].

Gibbon, D., Eno-Abasi, U., & Ekpenyong, M. (2006). Problems and solutions in African tone language text-to-speech. In *Proceedings of the multiling conference*. Stellenbosch, South Africa: Justus Roux, ed.

Hall, J. F. (1948). *Dictionary and practical notes: Mossi-English languages*. Ouahigouya, Burkina Faso: Mission des Assemblees de Dieu.

Houis, M. (1977). Plan de description systematique des langues negro-africaines. *Afrique Et Langage*, *7*, 5-65.

Houis, M. (1980). Propositions pour une typologie des langues negro-africaines. *Afrique Et Langage*, *13*, 5-46.

Houis, M. (1983). De la derivation a travers quelques langues africaines. *Modeles Linguistiques*, 49-67.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing . An introduction to natural language processing, computational linguistics and speech recognition*. New Jersey: Prentice-Hall.

Kabore, R. (1980). *Essai d'analyse de la langue Moore (parler de Waogdogo: Ouagadougou)* (Unpublished doctoral dissertation). University Paris 7.

Klatt, D. H. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, *82*(3), 737-793.

Koffi, E. (2010). *Lexical tone, grammatical tone and orthography*. Lecture. Retrieved August 5, 2011, from http://www.orthographyclearinghouse.org/Lecture/Tone/Orthography.pdf

Lampert, A. (2004). *Evaluation of the Mu-talk speech synthesis system. Information and communication technologies* [PDF]. doi: http://www.ict.csiro.au/staff/andrew.lampert/writing/SynthesisEvaluation.pdf

Lemmetty, S. (1999). *Review of speech synthesis technology* (Unpublished master's thesis). Helsinki University of Technology, Department of Electrical and Communication Engineering.

Malgoubri, P. (1985). *Introduction a la morpho-syntaxe du moore* (Unpublished master's thesis). University of Nice, Faculte de lettres et Sciences humaines.

Malgoubri, P. (2000). Le Zaoore ou Jaoore: Donnees historiques et linguistiques. *Cahiers Du CERLESHS*, *2*, 26-53.

Malgoubri, P. (2000). Nasalisation en Moore: Elements de difference entre le Sare et le dialecte du centre. *Kuupole, D. D. (ed)*, 42-58.

Ngugi, K. W., Okelo-Odongo, W., & Wagacha, P. W. (2005). Swahili text-to-speech system. *African Journal of Science and Technology*, 80-89.

Nikiema, N. (1976). *On the linguistic bases of Moore orthography* (Unpublished doctoral dissertation). Indiana University, Department of Linguistics.

Nikiema, N. (1982). *Moor gulsg sebre: Manuel de transcription du Moore*. Ouagadougou: Imprimerie Presse Africaine.

Nikiema, N. (1987). Differences de comportement et rapports entre consonne finale de radical CVC et consonne initiale de suffixe en Moore. *Studies in African Linguistics*, 117-174.

Nikiema, N., & Kinda, J. (1997). *Dictionnaire orthographique du Moore*. Ouagadougou: SOGIF.

Schroeder, M. (1993). A brief history of synthetic speech. *Speech Communication*, 231-237.

Sous Commission National. (1976). *Comment traduire orthographiquement le Moore* [Brochure]. Ouagadougou, Burkina Faso: La Sous-commission du Moore.

Taylor, P. A., Black, A., & Caley, R. (1998). The architecture of the festival speech synthesis system. *The Third ESCA Workshop in Speech Synthesis*, 147-151.

APPENDICES

APPENDIX I

Carrier Sentences

Carrier Sentences

1. Ráwã yèelám tí ấadsằ yí wʊsg zàamế zàabre.

   Man the said that stars the came out a lot yesterday evening.

   The man said that there was a lot of stars in the sky last night.

2. Á watà zaabr la yíbeoogò.

   He comes afternoon and morning.

   He comes in during afternoons and mornings.

3. Víuugù zombấ ãdgà.

   Owl an sits tree.

   An owl is on a tree.

4. Rà-wếng kàngá nimbãan-zóeerằ yáa wúsgo.

   Man ugly this compassion is much.

   This ugly man has a lot of compassion.

5. Bì-ríblã wũga sàgba tí sấamề.

   Child boy the played food so that it is not eatable.

   The boy has played with the food so that it is not eatable anymore.

6. Mám pagằ núgbĩng paalằ n lùı ráb-yòáablẽ.

   My wife ring new fell six days ago.

   My wife new ring was lost six days ago.

7. Ràkãagr n bé zak kãngà púgà.

   Man rich is house this inside.

   A rich man is leaving in this house.

8. Mínisr kʊdgã yaa bʊ̀sáangà.

   Minister old the is bʊ̀sáangà (name used to designate members of the ethnic tribe bʊ̀sáangà).

   The former minister is of the bʊ̀sáangà tribe.

9. Fó nengẽ wã sɪd lèbgá nif-sãagré.

   Your attitude the became disrespectfull.

   You are becoming disrespecfull.

10. Lògtórã tʊ̀pá gúirenga bʊ̀igá.

    Physician the treated gúirenga (name used to designate members of the ethnic tribe gúirenga) child.

    The physician treated a gúirenga's child.

11. Kàmbá kʊ ràyúʊg n dát n wà̰be.

    Children killed rabbit want eat.

    Some children killed a rabbit and want to eat it.

12. Pʊg-yáangà réegà púʊsmà̰ n yètí ẽyyn.

    Lady old took salutation the saying ẽyyn (onomatopoeia).

    An old lady responded to someone greetings by saying ẽyyn.

13. Sípaolgã wàkáta là vũ̀nunvũug rũ̀m mam gɛ́y wã.

    Spring time insect bite my tigh.

    During the spring, I was bitten by a flying insect.

14. Fú-yorgã yáa miuugù.

    Cloth the is red.

The cloth is red.

15. Bắng bõe n be tãngằ́ sɛ́ɛga.

    Guess what is mountain the ass.

    Guess what is near the mountain.

16. Rõandã míninzitằ́ yốkà rắ-kõagdà yáo wã.

    Year meningitis the caught traditional beer brewer and saler brother the.

    This year meningitis epidemic disease caught the traditional beer brewer and saler's brother.

17. Tógs-d zĩiga ligd mèng hakíkà.

    Tell us land money itself real.

    Tell us the real value of this land.

18. Nín-kẽemằ́ yéelà yẽ tá yaa kìnkír-bagá.

    Old man the told him that he is phenomenon.

    An old man told him that he is a phenomenon.

19. Ḿ yaab yĩ̀ma yer-bắnga m pʋgdb yirằ́.

    My grandpa forgot steel tool my aunt house.

    Grandpa forgot the steel tool in my aunt house.

20. Kàren-saambá wề̀ karen-bíiga pɛ́kà.

    Teacher striked student slap.

    A teacher slapped a student.

21. Ấe loaeeg ǹ lʋɪ wắaga lahagì báõkằ́.

    Whose fishing rod fell slice lahagì shoulder the.

Whose fishing rod fell on lahagì's shoulder hurting him.

22. Tấpã pãrga rã-zãndầ wá nedẽ lʋı vãvãbdò.

Bow the tears veranda the like somebody falling.

The bow went through the veranda falling like somebody would.

23. Bìtó, tab-mòaagá, ne tab-kaoor bɛk ǹ bé roogề.

Veggie sauce, tobacco, and cigar peace are house the.

There are some veggie sauce, tobacco and a peace of cigar in the house.

24. Séb-bobdrấ ne kindfã ligdi ǹ lʋı bulgu wấ.

Schoolbag the and jewelry money the fell in well the.

The money for the schoolbag and the jewelry fell into the well.

25. Néd-kam fãa yẽd-nifầ yáa miuug yʋddd.

Everyone anus eye is red.

Everyone got the lesson.

26. Mám bẽdã ràyúʋgầ né bordì rõsendõaaǵa zugù.

I lure rabbit the with a banana tree head.

I lured the rabbit into a banana food trap.

27. Bódgò né libtùug yáa yembrè.

Bag and ignorant are same.

A bag and an ignorant are the same.

28. Á tɒpà a rab-nassẽ nodrã né tab-voak zẽedò.

He treated his four days wound with vegetable sauce.

He treated his wound of four days with a mixture of vegetables.

29. Ràb-wɛɛlẽ lekollẽ bɛɛbã sɛ́bgà mám weefã̀.

    Day nine school enemies the locked my bike.

    My school classmates locked my bike.

30. Wéd-wĩirì n bé pẽkẽ wã wãbd bɛdensè.

    Bike rope is corner the eating worms.

    A snake is at the corner eating worms.

31. Búudù yaa bumb sẽn tog né pẽgbò.

    Family is something that deserves respect.

    A family is something that deserves respect.

32. Rík ningɛtgà né mankẽssã saglg lepre wã.

    Take eye glass and matche the put basket the.

    Put the eye glass and the matches box into the basket.

33. Á gẽmsà f bɛoolga bɛng rab-piilẽ.

    He harvested your field beans ten days ago.

    He harvested your beans crops ten days ago.

34. Á fadà fã́agà fika né fɛfɛ gilli.

    Fada riped away fan and ingredient all.

    Fada riped away the fan and all ingredient.

35. Wéd-yãoogà né a lemdã̀ mógà wúbsgù.

    Horse chest and chin eat dust.

    A horse chest and its chin are dirty of dust.

36. Bɛ̀lla rii bɛnga sẽn be lepr wá́.

Bɛlla (a member of the ethnic tribe of bɛlla) ate beans that was in plate the.

A bɛlla ate all the beans that was in the plate.

37. Wé-yũugù bɛsmá pɛpɛrgà né bɛrengà.

Non domestic cat broke tree with a hemp.

A non domestic cat broke a tree with a hemp.

38. Bɛ̀odgó ǹ lʋı wẽp febg fíkà.

Pitfall fell on fan of leaves.

A pitfall fell on the fan of leaves.

39. Mùnáanfɪglem ka nafd ned yè.

Dishonesty not benefit anyone.

Dishonesty does not benefit anyone.

40. Bògfʋllé fĩnfii n sã fẽgá

Larva small puped.

A small larva pupped.

41. Zʋ́lga zõondamè tá zũurà lʋı.

Zʋlga kneeled and tail fell.

A zʋlga (member of the ethnic clan of dioula) kneeled with his tail touching the ground.

42. Sèbgɑ́ zɛkà tíugɑ̃́ tí yìbrengá lùı a zelemdá zúgù.

Wind the lift tree so that scavenger fell tongue head.

The wind lifted the tree causing the scavenger to fall on its tongue.

43. Pʋg-zɛ́́egà zulgá vãadà kẽne wá yɛ.

Lady light skin carry leaves walk until tired.

The light skin lady carried the leaves walking until she gets tired.

44. Ḿ wobgà a kargá ta yik wõg kẽ weoogå̀.

They gave massage his foot and he got up to go in bush the.

They gave a massage to his foot and he went into the bush.

45. Vɛ́endà vő́omè wá vʊdga a nugà vuri.

Bullet came to penetrate his hand palm.

A bullet went through his hand palm.

46. Á vʊdgà vɛ́ennà yiis tí yaa tʊʊlg vimà

He extracted bullet and it was hot very.

He extracted the bullet and it was very hot.

47. Á vɛɛsẽ n ríkà sʊ-waorengá ne gutungúlunga ǹ vélè.

He diged in to take locust and bug to swallow.

He diged in to take a locust and a bug to eat.

48. Wídg bugmà la bas pʊ-yágbå̀.

Light fire and stop lying.

Put the light on and stop lying.

49. Á làmús sũ-sãamså̀ ne a pʊtɛ́erå̀ ya wúsgò.

Lamus problems and his memory are many.

Lamus has plenty of problems and a big heart.

50. Pʊg-péelèm ned ká mi widb yè.

Honest man do not know curse.

An honest man does not curse.

51. Súlg ne yɛ̃̀s-kɔ́abnenga n be yolg n wã.

Spider and insect are bag the.

There is a spider and an insect in the bag.

52. Sìlmíiga sɩ̃̀if né a nug-sũka zàabdámẽ.

Silmiiga (member of the ethnic clan of the silmiise) chest and wrist are painfull.

The chest and wrist of the silmiiga are painfull.

53. Gòetrngóes pis-tɛ̃̀ ság tenteaag n wã.

Ants thirty are in basket the.

There are thirty ants in the basket.

54. Á kònvɔ́lb vàaga fu-pókdã ǹ ning tɛ́ɛg wã zʊ̃̀zũnd-zũndi.

Konvolb took out clothes put them in basket in messy way.

Konvolbo took out the clothes and put them in the basket in a messy way.

55. Kʋ-bíiga kʋka zugà yɛs wa kúʋma.

Funeral kid has his hands over head as a lazy.

The orphan with the hand over his head is standing like a lazy.

56. Mùká fʋ̃̀ká rʋkɛ̃̀ lébg fʋ̃gná kʋ̃̀p.

Deaf uncovers pot and recovers it an a noisy way.

A deaf uncovered the pot and covered it with a lot of noise.

57. Kúmb bìt-zɛ́̃edà láag-wɛ̃̀ lá f tui pĩirà yʋngá.

Put veggie sauce plate the and get the mat out.

Put the veggie sauce in the plate and get the mat out.

58. Wʊk àlbáslã ne yams-nerɑ̀ tí yaa wubsgu.

    Gather onion the and salt because it is dusty.

    Gather the onions and the salt because of the dust.

59. Bìt-fáoogã pátbãkã yaa gagat gat bàla.

    Immature man the inextricable situation is irregular.

    The immature man inextricable situation is irregular.

60. Á yũ gàswɛɛgɑ́ ǹ yík ǹ látmedɛ̀.

    He drunk alcohol water the and got up with zig zag.

    He drunk the alcoholic water and got up walking in zig zag.

61. Wób-zãra vila ròogɑ́ wa gɪ̃sgá gìdg nóorà.

    Liana tree climb the house like a rope.

    The liana is climbing the house like a rope.

62. Á Pol gõdg n kell gʊ̃nuga bãg-víʊdɑ̀ né a yakã.

    Paul turned left clothes the with his neighbor.

    Paul suddenly left the clothes with his neighbor.

63. Yàafgó nafdà nédà.

    Grace benefit man.

    Grace is always benefic.

64. Fɔ̀ke rúkà ń dik sɪ̃bgɑ́, né kɛglɑ̃̀, páa ziũnfɑ̀.

    Uncover the pot take grapes the, fruits the, and fish the.

    Uncover the pot and take the grapes, the fruits, and the fish.

65. Ríẽ zɪ̃̀-líkrã la mam farans wàkɪ-vɑ́oogɑ̀ ménmɛ̀

Last year harvest time that my French one franc currency lost.

It is during last year harvest period that my one franc (french currency) was lost.

66. * Lìk f́ rìʊtgá ges kùkúrã n yũud ne sɩlsãka

Check out the pig drink with bottle.

Check out the pig drinking with a bottle.

67. Nó-tʊʊl-nedɛ̀ bas góamà la f́ rɩk sɩlpamdɛ̀ né neõedrɛ̀.

Talker man stop talking and take leveler the and shoe the.

Stop talking and take the leveler and the shoe.

68. Yùʊm-vɛ́kr lògtoɛ́ɛmba be yɛɛnɛ̀?

New year phisicians are where?

Where are the new year phisicians?

69. Fó nimbãal-zoeerã ka ta f yel-wẽnà né f sõm-zɪtlemã yè.

Your compassion does not measure up your sins and your ungratefulness.

Your compassion does not measure up to your sins and ungratefulness.

70. Nébɛ̀ gómdà fó maanì né f gõodemà yéllè.

People are talking about your acts and errancies.

People are talking about your acts and errancies.

71. Kò-vúɯgà n gõm be wanwarg n wã.

Water maggot stick there granular.

A water maggot is stuck there.

72. Ḿ rɛɛmb lʊɯm ta rĩirɛ̀ suk marmatig ko-rɛ́gdà.

My brother in law fell his forehead touch peanut water dirty.

My brother in law fell with his forehead touching the dirty water of peanut.

73. Ḿ lepdemm ne f yaar-yaarã̀ na rɪ ḿ lamè

Your flip floping and your lack of shame will eat you.

Your flip floping and lack of shame will kill you.

74. Rùruugā́ yẽ kàtrpṍaagã̀ né kortrko pugumdì.

Cobra the saw orange tree the and bird shake.

The cobra shaked itself after seeing the orange tree and the bird.

75. À alkamús tarà gõ-pṍõsga a yòpóe ne mo-fàoogá ye.

Alkamus has acacia flowers seven and herb.

Alkamus has seven flowers of acacia plant and herb.

76. Kámbà lɛmã né b làlldsā́ yaa lòm-lome.

Child cheeks the and testicles the are smooth.

The child cheeks and testicles are smooth.

77*[23]. Búgtellã̀ tűnugà á zɪ-wĩigà bélgè.

Glass bottle followed his vein infect.

The glass bottle infected his veins.

78. Níng sɛ̀t-tãoorã tí supĩimã̀ na tõbgá karsã̀.

Put thimble the pin the will hurt legs the.

Use the thimble to protect your legs from the pin.

[23] Sentences with * indicate that the sentence does not make sense. These situations happen when we could not construct a meaningful sentence that contains or combines particular digrams, words or sound (phonemes).

79. Pɔ̃orɑ́ kĩndfɑ̃ né a borfɑ̃ yaa árzɛkà.

Cripled man the necklace and his long sword are valuable.

The crippled man necklace and his sword are all he has.

80. Sɔ̃ng ý biigà né ko-yũud tà ra lʊı sobg ye.

Help child with water drink so that he doesn't pass out.

Give water to the child so that he doesn't pass out.

81. Fó mè mĩ n rúʊd pad-pade tì b yeel ti ẽhẽe.

You too pie pad-pade (onomatopoeia) so they can say yes.

You have to pie sometime like a man.

82. * Mòg-rõod-lá-f-kõ-m-zũuna hãrgà nóaagɑ̃ sĩd yítlì.

Idiot tear chicken insert a little knob.

An idiot is inserting a little knob into the chicken.

83. M̀hḿm fo ningà kíyà mui-zɛ́ed wã là?

Mhmm (onomatopoeia) you put cereals rice sauce the.

Mhmm did you put some cereals into the rice?

84*. Máanà lohórèm né a val valɑ̃ la á tulyã fɑ̃a.

Make him mercy with wind his and upside down all.

Pardon him with his windy attitude.

85. Ɔ̃hõo ko-zoetéma kẽdà võya tí pũndgdɛ̃.

Oh oh running water is siping into holes swelling them.

Oh oh a running water is siping into the holes swelling them.

86. Bi púufɑ̃ bas ǵ yes-yes goamà!

Child little fat the let your cheap talk!

Little fat child, stop saying anything!

87. Ḿ tĩngà bug-ráoogə̀ tá yi parrr.

I smashed rafle the and it fired parrr(onomapoeia).

I smashed the rafle on the ground and it fired several times.

88. Á pɩbendà bõn bõtrə́ n tar var-vare.

He struggles thing round with difficulty.

He caught the circular thing with struggle.

89. Ḿ nonga biigə̀ léoosgə̀ né sũm-yamsmà.

I like baby the purge with sweet peas.

I like purging the baby with sweet peas.

90. Hóbrgàgi, zĩ-peelə́ bɔ̀angá hĩrsàmé.

Hobrgagi (onomatopoeia), place white the donkey heehaws.

Hobrgagi, the white place donkey heehaws.

91. Bíigə̀ sã n y hə́e, m wẽed f là míh, hɛh wə̀

Baby the cry, I slap you mih (onomatopoeia), idiot the.

If the baby crys, I will slap you, idiot.

92. Héh maan gʊ̀usg né f hirp hirp wə̀.

Heh (onomatopoeia) make attention with your hic-cups.

Heh, be careful with your hic-cups.

93. Ɔ̀hõ ká yẽ da hũusdà wɛ́?

Ɔhõ (onomatopoeia) not him that was coughing?

Oh, wasn't he coughing?

94. Áh, kõn yellɛ̃̀-m bálà!

Ah (onomatopoeia), not say only!

Ah, that's the reason why!

95. Pʊg fáklgɑ̃̀ wéefɑ̃̀ tusb yáa nanà.

Lady flat the bike pushing is easy.

The flat lady's bike is easy to push.

96. Á yɛsãmè n lʊı kűi ǹ wẽneg n bé kũumɛ̃̀.

He scared fell kũi (onomatopoeia) turn into dead.

He got scared and died after falling.

97. Ráb-tãtẽ ãd-zʊʊrɑ̃̀ wɛ́ndà rõsendɔ́̃aagà.

Day three shooting star the like tree specie.

The shooting star I saw three days ago is like a certain kind of tree.

98. Sɯ̃̀-tó nedɛ̃̀ pɑ́̃b molf zɪ̃̀-peel wɑ́̃.

Self control someone strike place white the.

A self control man striked the white place.

99. Á tẽremdà fó halhaalà ḿengà.

He drag you essence itself.

He disrespected you.

100. * Ɛ̀i, rõ-víirɑ̃̀ ne bűmbɑ̃̀ yaa tʊʊlg vḿà.

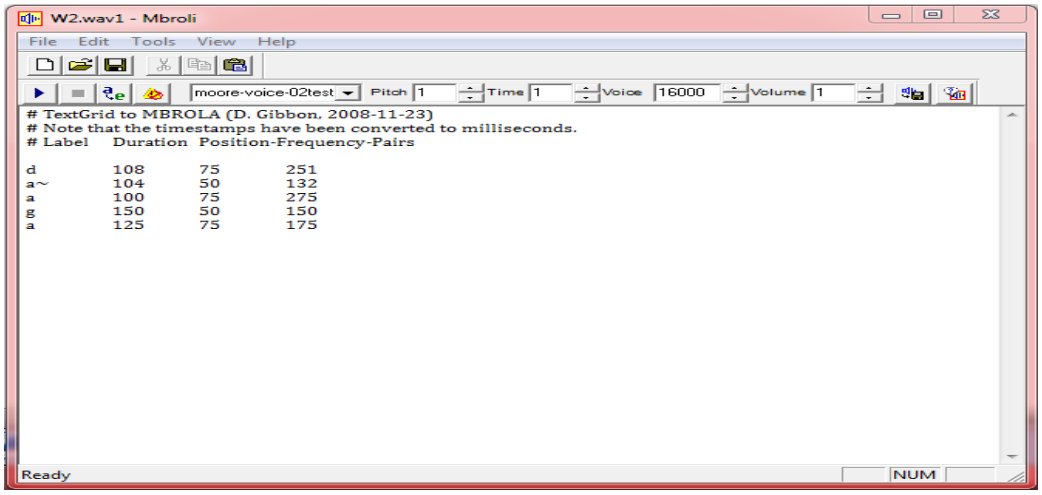
Ai(onomatopoeia), scare face and thing the are hot very.

Ai, the scares and the thing are very hot.

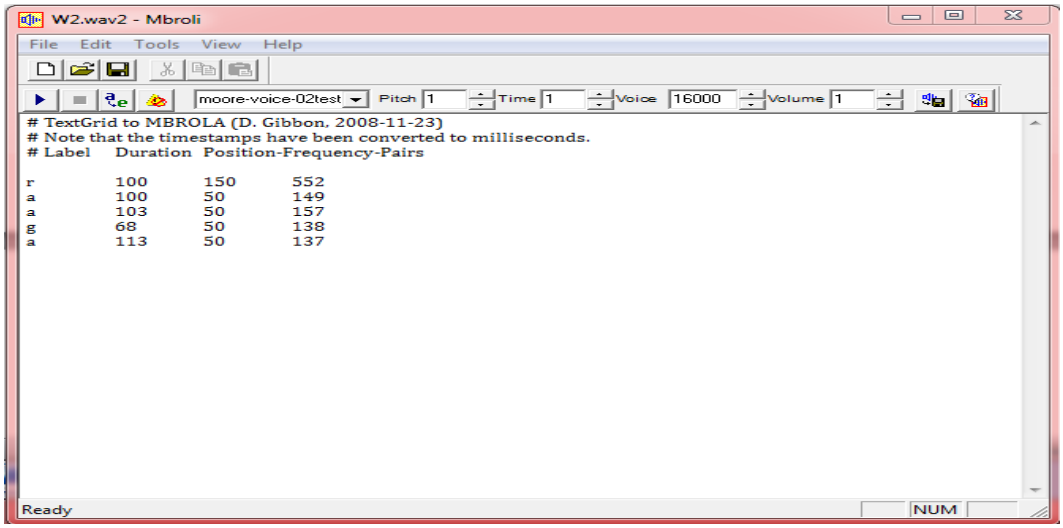
APPENDIX II

PHO Mbroli files

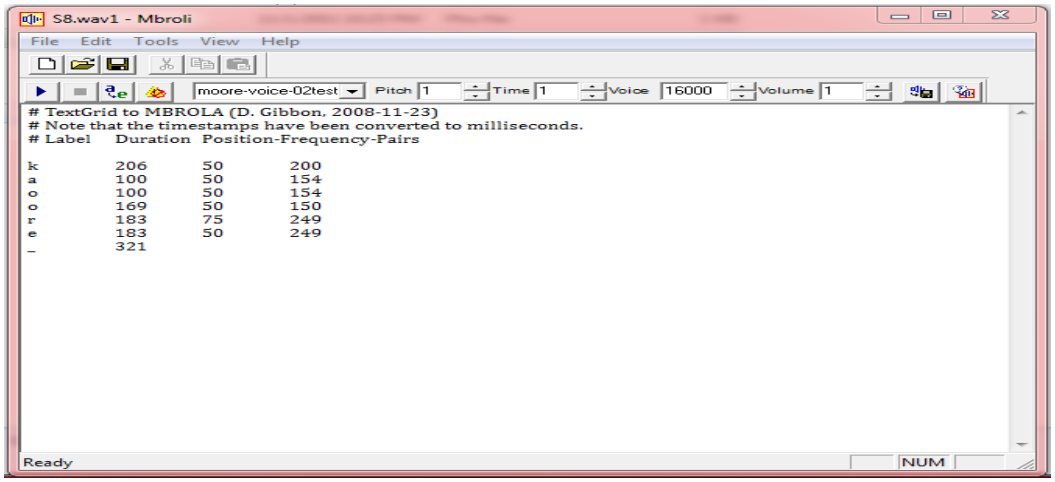
PHO Mbroli of the word <namse> for the DRT



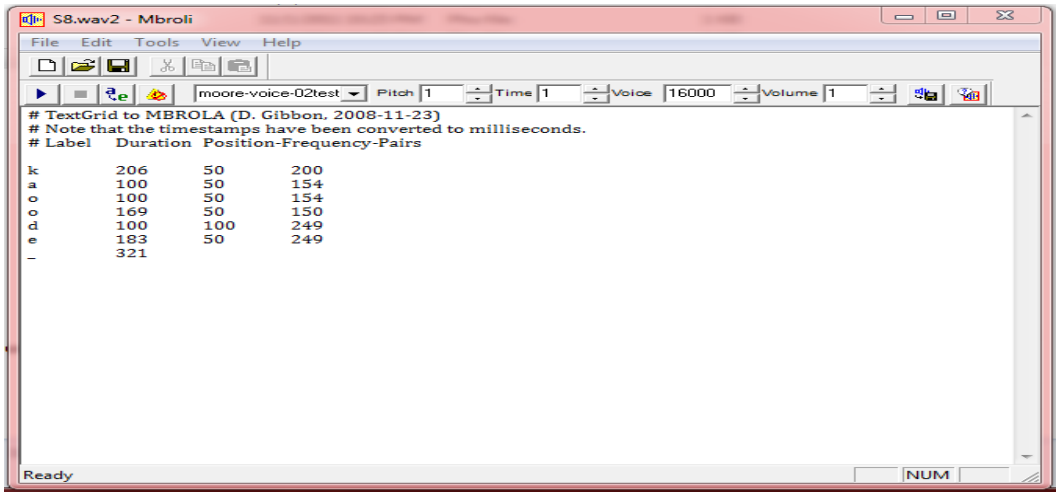
PHO Mbroli of the word <raaga> for the DRT



PHO Mbroli of the word <Kaoore> for the MRT



PHO Mbroli of the word <kaoode> for the MRT



PHO Mbroli of sentence I for transcription task 1.

<Kãngà - tãpã - ràkãagr - zàabre - púgà.>