
**ADDIS ABABA UNIVERSITY**

**COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES**

**SCHOOL OF INFORMATION SCIENCE**

# PROSODY BASED AUTHOMATIC SPEECH SEGMENTATION FOR AMHARIC

By

RAHEL MEKONEN TAMIRU

FEBRUARY, 2019
ADDIS ABABA, ETHIOPIA


**ADDIS ABABA UNIVERSITY**

**COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES**

**SCHOOL OF INFORMATION SCIENCE**

# PROSODY BASED AUTHOMATIC SPEECH SEGMENTATION FOR AMHARIC

A Thesis Submitted to the School of Information Science of Addis Ababa University in Partial Fulfillment of the Requirement for the Degree of Master of Science in Information Science and Systems (Language Technology)

By: RAHEL MEKONEN

Advisor: SOLOMON TEFERRA (PhD)

February, 2019

Addis Ababa, Ethiopia


**ADDIS ABABA UNIVERSITY**

**COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCE**

**SCHOOL OF INFORMATION SCIENCE**

# PROSODY BASED AUTHOMATIC SPEECH SEGMENTATION FOR AMHARIC

By: RAHEL MEKONEN

Name and signature of Members of the Examining Board

| | | |
|---|---|---|
| Solomon Teferra (PhD) | _____ | _____ |
| Advisor | Signature | Date |
| _____ (PhD) | _____ | _____ |
| Examiner | Signature | Date |
| _____ (PhD) | _____ | _____ |
| Examiner | Signature | Date |

## Declaration

This thesis has not previously been accepted for any degree and is not being concurrently submitted in candidature for any degree in any university.

I declare that the thesis is a result of my own investigation, except where otherwise stated. I have undertaken the study independently with the guidance and support of my research advisor. Other sources are acknowledged by citations giving explicit references. A list of references is appended.

Signature: _____

Rahel mekonen

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

Advisor's Signature: _____

Solomon Teferra (PhD)

## Acknowledgements

First and foremost, I would like to thank the Almighty God for His support and being with me in all directions of my life.

Next, Special thanks to my advisor Dr. Solomon Teffera, for his unreserved guidance, support, encouragement and constructive suggestion throughout this work. He was very caring and supportive like a father in the whole process conducting this research work. I thank you for encouragement from the beginning till the end of this research work.

I also thank my beloved family for their unconditional love and support. Last but not list, I would like to thank my friends for their valuable ideas and encouragements throughout this work.

# Table of contents

# Abstract

Many speech processing systems require segmentation of speech waveform into principal acoustic units. Speech segmentation is the process of identifying the boundaries between paragraph, sentence, words, syllables, and phonemes in spoken natural languages. It is the very primary step in the field of speech technologies. Automatic speech segmentation is a process segment any one of discrete units that occur in a continuous speech signal through algorithms developed for this purpose. Speech segmentation is a challenging task because the cues present for segmenting text are absent in a continuous speech.

The main goal of this work is to develop sentence level automatic speech segmentation system for Amharic. Sentence segmentation is a process of identifying the end of a sentence. In this study, sentence segmentation system is implemented in to two approaches. In the first approach, we used an automatic tool for segmenting and labeling of Amharic speech data. Acoustic model is created using speech and their text scripts and compiling them into a statistical representation of sounds which makeup words. This is done through HMM modeling. The approach one automatic speech segmentation system is done by forced alignment. In this approach we used rule-based and AdaBoost to discriminate the true boundaries from false. In the second approach, we extracted prosodic features directly from speech waveform and also statistical method, AdaBoost, is used.

The evaluation of the experiments shows that monosyllable acoustic model is the better model to get accurate forced alignment than monophone and tide state tri-syllable model. And also adaboost classifier showed consistently good results especially in decision tree classifier. In all experiment read-aloud speech perform higher accuracy than spontaneous speech. It also indicates that spontaneous speech is more difficult than read-aloud because, the spontaneous speech contains more noise and disfluencies. The evaluation in phase two indicates that pause feature is a basic discriminator for Amharic sentence boundary. And also when prosodic features are introduced, the performance is increased. The scope of the research work is narrowed down only to sentences level segmentation. It is also required to conduct a research on automatic speech segmentation of other discrete units.


Keywords: **Sentence segmentation, acoustic model, prosody**

# List of Tables

# List of Figures

# List of Acronyms

| | |
|---|---|
| HMM | Hidden Markov Model |
| HTK | Hidden Markov Toolkit |
| MFCC | Mel Frequency Cepstral Coefficients |
| MMF | Master Macro File |
| ANN | Artificial Neural Network |
| F0 | Fundamental Frequency |
| ZCR | Zero Crossing Rate |
| PSD | Power Spectral Density |
| BER | Boundary Error Rate |
| MAXENT | Maximum Entropy |
| CV | Consonant Vowel |

# CHAPTER ONE

# INTRODUCTION

This chapter describes the overall background for this thesis, statement of the problem, research questions and objectives, scope and limitation, the methodology we used and finally presents the organization of the thesis.

## 1.1. Background

Speech is the vocalized form of communication used by humans to exchange information. It is natural and fast means of communication for people. Modern research in the field of speech technologies aimed at creating strong speech systems that can be used for communication between human beings and information-processing devices. Unfortunately, a machine capability to interpret speech is still poor. The most difficult feature of speech that challenges machines is its segmentation because human speech is produced continuously.

Speech segmentation is a technique for discovering speech signal in different parts. In speech, a segment is any one of discrete units that occur in sequence of sounds, which can be broken down into sentences, words, syllables, and phonemes in a spoken language. The main objective of this segmentation process is to use the result for other area of speech processing. It is an essential preprocessing step in several speech research areas such as speech synthesis, speech recognition, language identification and speaker identification system. Language generation-based techniques such as question answering and summarization may also benefit from speech segmentation, as would speech play back in spoken document browsing application[1].

Consequently, an efficient, accurate and simple technique is needed to accomplish this objective. The most commonly proposed speech segmentations are using either manual segmentation or automatic segmentation techniques. In manual speech segmentation the segmentation is done by specialized experts and its segmentation is based on listening and visual judgment in order to detect the required boundaries. It is difficult for huge database. This makes it expensive, tedious, inconsistent, prone to errors and time consuming. Automatic speech systems segment any one of discrete units through algorithms.

Automatic speech segmentation can be classified into two types: namely blind segmentation and aided segmentation algorithms. In blind segmentation there is no use of preexisting or external knowledge of linguistic properties. On the other hand aided segmentation uses some sort of external linguistic knowledge of the speech[2].

When addressing the segmentation task, it must be taken into account that speech especially spontaneous speech is not as clearly structured as written text. In spoken language the timing and pitch patterns are lost. Such patterns are known as **speech prosody**.

In linguistics, the term prosody refers to supra segmental phenomena in speech, meaning that one has to look at entire utterances rather than just considering phonemes, syllables or words. From the acoustic viewpoint, prosody means the alteration of syllable length, pitch, rhythm or loudness. Speakers additionally use prosody to impress some sort of emotion and attitude[3]. In all languages, prosody is used to convey structural, semantic, and functional information. Prosodic cues by their nature are relatively unaffected by word identity. Prosodic feature extraction can be achieved with minimal additional computational load and no additional training data[4].

Automatic sentence segmentation of speech is a process of identifying the end of a sentence. It is a process of detecting the silences in the speech data and identifying whether the region of silence is a sentence end or a non-sentence end. A crucial step towards robust information extraction from speech is the automatic determination of sentence boundaries.

## 1.2. Statement of the problem

People use language more differently when they speak than when they write. Spoken language contains many interjections, filled pauses, etc. Speech segmentation is a challenging task because it is not like text segmentation, the cues present for segmenting text are absent in a continuous speech and the lack of other additional typographic indicators.

Nowadays, one of the hot topics in speech technology is how to develop a system that matches the characteristics of spontaneous speech. Information extraction from large audio database requires extracting the structure of audio file as well as their linguistic content. One of these processes is to add sentence boundaries to the automatic transcription of speech contents. Adding

this structure to the automatic transcript is a very challenging task when processing spontaneous speech because this kind of speech is characterized by disfluencies such as filled pauses, repairs, hesitations, repetitions, and partial words.

The identification of sentence break is confusing both for humans and computers. Jones demonstrated that sentence breaks are critical for legibility of speech transcripts [5]. Moreover, missing sentence segmentation makes meaning of some utterances ambiguous. Similarly, missing sentence boundaries cause significant problems to automatic downstream processes [6]. Sentence boundaries are important when analyzing the syntactic complexity of speech, which can be a strong indicator of potential impairment.

If it has to be done by humans, manual segmentation of speech is expensive, tedious, prone to errors, and time consuming. Considering these facts, automatic Amharic speech segmentation is important.

There is a research work done on speech segmentation of spontaneous Malay language[7]. Acoustic and prosodic features were used for identifying sentence boundary. Other works such as an Automatic segmentation of speech into sentences-like units[6], voice segmentation without voice recognition[6], prosody-based automatic segmentation of speech into sentences and topics [4] were done. However, there has been no previous work done on sentence level speech segmentation for Amharic.

The research works mentioned above used prosodic features directly from the speech for identifying sentence boundary. Even though there are similarities in form and function of prosody among diverse language like pause, natural tendency for fundamental frequency , there are considerable variations indicated by cross-linguistic comparison of prosodic features like different timing of essentially comparable phenomena, different relationships between or different mutual effects of fundamental frequency, duration and intensity[8]. We do not know which features or their combinations would result optimal speech segmentation for Amharic speech. Therefore this research investigates optimal method which performs segmentation using prosodic features.

Foreign researches done on sentence level speech segmentation did not use forced alignment method for sentence segmentation. For this study, we develop acoustic model and use these

model to make the segmentation process easier and to get segmented and labeled Amharic speech data.

## 1.3. Research questions

The research on automatic speech segmentation for Amharic tries to answer the following research questions:

1. What speech features or their combinations would result in optimal speech segmentation?
2. What model is optimal for automatic speech segmentation?

## 1.4. Objective of the Research

### 1.4.1. General Objective

The general objective of this research is to investigate prosody based sentences level automatic speech segmentation for Amharic read-aloud and spontaneous speech.

### 1.4.2. Specific Objective

- To investigate features of Amharic language for speech segmentation.
- To review related works focusing on automatic Amharic speech segmentation.
- To prepare training speech corpus for learning process.
- To explore and construct the appropriate model for sentences level speech segmentation.
- To develop prototype for Amharic speech segmenter.
- To evaluate the performance of Amharic speech segmentation.

## 1.5. Significance of the study

The experimental results of most speech research areas are highly dependent on segmentation performance. This research will have a significant contribution to show the way to develop Amharic speech segmenter and which features or their combinations would result in optimal speech segmentation for Amharic language. For other researchers this research can be considered as a ground to further study.

## 1.6. Scope and Limitation of the Study

Automatic speech segmentation is the process of dividing continuous speech signal into discrete units that range from phonemes to sentences even paragraphs. The scope of the research work is narrowed down only to sentences level segmentation.

## 1.7. Methodology

In this research design science research method has been used. Because it results an artifact that is implemented to address some problem identified using problem identification. Design science has six key steps. Those are problem identification and motivation, objectives of the solution, design and development, demonstration, evaluation and communication. This research takes these steps in to consideration.

**Problem identification and motivation**

In the first step, we define the specific research problem and justify the value of a solution. Justifying the value of a solution accomplishes two things: it motivates the researcher and the audience of the research to pursue the solution and to accept the results and it helps to understand the reasoning associated with the researcher's understanding of the problem. We did an observation and a literature review on the topic to identify the research problem.

**Objectives of the solution**

The objectives should be inferred rationally from the problem specification. Resources required for this include knowledge of the state of problems and current solutions and their efficacy if any. Therefore, after having the problems identified, the objectives of this study are defined well.

**Design and development**

After the objectives of this study are well defined, we continue to the design and development of the study. Resources required moving from objectives to design and development include knowledge of theory that can be brought to bear as a solution.

**Demonstration**

Demonstrate the efficacy of an artifact to solve the problem. This could involve its use in experimentation, simulation, a case study, proof, or other appropriate activity. We demonstrated the efficacy of an artifact using experiment.

**Evaluation**

Observe and measure how well the artifact supports a solution to the problem. We did this activity by comparing the objectives of a solution to actual observed results from use of the artifact in the demonstration. At the end of this activity, the researchers can decide whether to iterate back to step 3 to try to improve the effectiveness of the artifact or to continue on to communication and leave further improvement to subsequent researches.

**Communication**

Communicate the problem and its importance, the artifact, its utility and novelty, the rigor of its design, and its effectiveness to researchers and other relevant audiences, such as practicing professionals, when appropriate.

This research adopts novel method which performs segmentation through experiments and detailed studies. In order to achieve the stated objectives, the study used different methods such as literature review, data collection and preparation, modeling**,** tools and techniques and evaluation.

### 1.7.1. Literature Review

Literature review is done on related research works conducted on automatic speech segmentation for various languages and related research works on different approaches which can be used to investigate automatic speech segmentation. These literatures review conducted to give deeper understanding on the area and to have detailed knowledge on the various techniques that are essential for speech segmentation systems.

### 1.7.2. Data collection and preparation

A speech corpus is a database of speech audio files and text transcription. Amharic bibles, broadcast news, broadcast conversation and Amharic fictions are data sources used for speech corpus preparation. In this work 4000 Amharic speech sentence and the corresponding text corpus is collected for training. We created two corpora corresponding to two distinct speaking styles: a read speech corpus and a spontaneous speech corpus. The first corpus was created by collecting existing broadcast news corpus and Amharic bibles, whereas the second corpus was created by combining broadcast conversation and Amharic fictions (fikr skemekaber) as it is difficult to get spontaneous speech data easily. Data preparation includes Amharic text corpus

preparation, speech corpus preparation to the corresponding text corpus, data transcription to HTK usable format and parameterization of speech signals. Both text and speech corpuses are split into training and test data sets. Manual segmentation takes place on training and test data preparation. These manually segmented speeches will be used to evaluate the performance of automatic speech segmentation system.

### 1.7.3. Modeling approach

The most widely used approaches for speech segmentation are rule based, statistical and Artificial Neural Network (ANN) approaches. In this research rule based and the statistical approaches called Hidden Markov Models and Adaboost is used. HMM is currently the most commonly used and most successful approach for speech segmentation [9]. The main reason for this success is its ability to handle new data strongly, and it is computationally efficient to develop and evaluate, it is language independent and used in different working environments like speech segmentation under noise conditions. AdaBoost is an adaptive algorithm to boost a sequence of weak classifiers, where the weights are updated dynamically according to the errors in previous learning[12]. It uses decision tree classifier as default classifier[11]. In our experiment we have used decision tree classifier and support vector classifier (SVC) as a base estimator.

### 1.7.4. Tools

In this study open source tools such as Audacity, HTK and Python are used. HTK toolkit is a portable toolkit for building and manipulating hidden markov models. It is used for preprocessing of wave files. Python is used for text processing. Audacity is a free and open-source digital audio editor and recording application software.

### 1.7.5. Evaluation

The performance of speech segmentation systems is evaluated by the accuracy of the predicted segmentation based on well-known accuracy, recall and precision evaluation method. Recall and precision are chosen mainly because we are familiar with it and can easily measure the performance of the system. Accuracy refers to how close a measured value is to the actual (true) value. Precision refers to how close the measured values are to each other. Precision gives

information about how many of the detected boundaries are true, the measurement recall tells us how many true boundaries have been found.

## 1.8. Organization of the thesis

The rest of the thesis is organized as follows. Chapter Two presents literature review which includes the general overview of the possible approaches used in the last two decades and evaluation method for speech segmentation. It also gives an overview of the related works used in the thesis work. In the related work, the script of languages, the features used for parameterization and feature vectors, the corpus size used, models, techniques and the experimental results are presented in detail. And it also gives an overview of Amharic language and its phonology. Chapter Three covers the overall explanation and design of the new Amharic speech segmentation system. Chapter Four presents experimentation and results of the thesis work. Finally, conclusions and recommendations for future work are presented in the fifth Chapters.

# CHAPTER TWO

# LITERATURE REVIEW

This chapter covers the overall concepts of speech segmentation, sentence level speech segmentation, spontaneous Speech and the most common speech segmentation approaches. We present related works done on speech segmentation and a background review of what prosody is and what its constituents are will be made. In addition, concepts like intonation, Melody, Intonation Patterns, stress, accent, rhythm, pitch, pitch declination, pitch reset and tune will be introduced and their role in prosody described in detail. And finally an evaluation method of speech segmentation is going to be discussed.

## 2.1. Speech segmentation

Speech is a continuous acoustic signal. Segmenting of speech corpus is the essential task in the creation of annotated speech corpus. It involves defining a speech segment and labeling signal areas with symbolic information.

Due to a number of useful applications, speech segmentation has been a great deal of interest to researchers in the field of natural language processing. It can be defined as the process of the partitioning of the speech signal under study into manageable and will defined segments.[12] The main objective of this segmentation process is to use the result for other area of speech processing. It is an essential preprocessing step in several speech research areas such as speech synthesis, speech recognition, language identification, and speaker identification system.

Speech segmentation is the process of finding the boundaries in a certain spoken natural language between sentences, words, syllables and phonemes. **Phoneme** is a tiny unit of sound in speech. A phoneme doesn't have any inherent meaning by itself, but when phonemes are combined, they can make up words. A **syllable** is the smallest stretch of speech into which a speaker is able to divide his/her utterance [6]. People never hear phonemes one at a time but we hear two or more phonemes combined into a syllable. Therefore **syllable** is the smallest unit of speech perception. Syllables are often considered the phonological “building blocks” of words. The syllable consists of onset, nucleus or peak, and coda. The nucleus is the central part of the syllable and typically consists of a vowel. The onset is the sound or sounds found before the

nucleus, whereas the coda is sound or sounds that follow the nucleus. **Morpheme** is a smallest unit of speech perception that has a meaning. **Words** are made up of one or more morphemes. Words are combined to form phrases. **Sentence** is a set of words governed by a set of rules that expresses a complete idea. Naturally, it displays identifiable intonation patterns and is often marked by preceding and following pauses[6].

Automatic speech segmentation can be classified into two types: namely blind segmentation and aided segmentation algorithms. In blind segmentation there is no use of preexisting or external knowledge of linguistic properties. On the other hand in aided segmentation researchers use some sort of external linguistic knowledge of the speech[3].

## 2.2. Sentence level speech segmentation

Sentence segmentation of speech is a process of identifying the end of a sentence. It is a process of detecting the silences in the speech data and identifying whether the region of silence is a sentence end or a non-sentence end. As shown by a number of studies, the absence of sentence boundaries is confusing both for humans and computers. Jones et al [6] demonstrated that sentence breaks are critical for legibility of speech transcripts.

Sentence boundary detection also known as sentence segmentation decides where a sentence begins and ends. Previous method of sentence boundary detection is either done by linguistic approach or acoustic approach or combination of both approaches. Linguistic-based method used linguistic features in statistical language model to detect the sentence boundary and defined as syntactically and semantically coherent Linguistic units [6][7]. On the other hand, acoustic approach used prosodic features such as fundamental frequency (F0), energy, duration and pause in detecting the sentence boundary. However, combination of linguistic and acoustic methods always produced higher accuracy compared to linguistic and acoustic approach alone. Linguistic approach requires the need of a speech recognition component that includes the language context information and linguistic features for segmenting sentences. However, speech recognition often takes higher computational costs.

## 2.3. Spontaneous Speech

There is no general agreement on the spontaneous speech definition in the literature. Most typically, all speech that is not read-aloud is considered to be spontaneous[13]. The quality and fluency of speech is highly dependent on whether the utterance is prepared beforehand or not. The most fluent speech is typically produced when it is read-aloud from a paper or screen, while the least fluent utterances are usually produced when the speaker is completely surprised by an unexpected question or unforeseen situation and has to respond spontaneously[6].

There are differences between prosody of planned and spontaneous speech. The pauses are more frequent in spontaneous speaking and the pitch declination is less steep in spontaneous speech and pitch resets are less perceptible. And also there are difference in duration and the average of pitch values is usually lower in spontaneous speech.

## 2.4. What is prosody?

The term prosody is derived from the Greek word προσωδια. Which is a musical term meaning something like "song sung to music". In the area of automatic speech processing, prosody has largely been used in Text-to-Speech (TTS) synthesis. In recent years, speech scientists have also started to more often use prosody in speech recognition and understanding applications. The tasks for which use of prosodic features has been studied include linguistic segmentation of speech[14][15].

Prosody is the science of how speech is uttered. It is the study and rhythm of speech and how these features contribute to the meaning. In all languages, prosody is used to convey structural, semantic, and functional information [16]. Prosody doesn't concern itself with understanding of single phonetic elements but rather larger segments of speech including syllables, words, phrases, sentences or even larger utterances. Because of that reason, a prosodic feature is usually referred to as a supra-segmental feature of speech. Prosody may convey the locations of syllables, word, phrase, and sentence boundaries of information to the listener.

Processing of read or spontaneous speech in linguistics and related fields have shown that information units, such as syllables, word, phrase, and sentences, are often demarcated prosodically. In English and related languages, such prosodic indicators include pausing, changes

in pitch range and amplitude, global pitch declination, melody and boundary tone distribution, and speaking rate variation. For example, both sentence boundaries and paragraph or topic boundaries are often marked by some combination of a long pause, a preceding final low boundary tone, and a pitch range reset, among and other features. A reason to use prosodic information is that certain prosodic features can be computed even in the absence of availability of ASR, for example, for a new language where one may not have a dictionary available[16][14].

Listeners depend on prosody to fully understand speech in addition to the combination of the words making up the speech utterance. Suprasegmental features perceived by the listener include Pauses, pitch, loudness, rhythm, and speaking rate. It is not possible to measure these human-perceived features from a speech signal directly; we can only measure their correlates. Prosodic events can be studied at various levels of representation including: the acoustic (measurable) , psychoacoustic (perceived) and linguistic level of prosody[14][6].

| Acoustic | Perceptual | Linguistic |
|---|---|---|
| fundamental frequency (F0) | Pitch | Tone, intonation, aspect of stress |
| Amplitude, energy, intensity | Loudness | Aspect of stress |
| Duration | Length | Aspect of stress |
| Amplitude | dynamics Strength | Aspect of stress |

Table 2.1: Levels of Representation of Prosodic Phenomena

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

Prosodic cues are known to be relevant to discourse structure across languages[16] and expected to play an important role in various information extraction tasks. In [14], Grimay presented prosodic events can be studied at various levels of representation including: the acoustic (measurable) , psychoacoustic (perceived) and linguistic level of prosody. We focus on the acoustic (measurable) level of prosody used throughout this thesis work. These acoustic level prosodic features are based on fundamental frequency, amplitude and duration. The variations in the fundamental frequency, amplitude and duration are perceived by listeners as the pitch, loudness and length of the speech[14].

**Fundamental frequency (F0)**

The fundamental frequency (also called the ***fundamental***) is defined as the lowest frequency of a periodic waveform. Fundamental frequency of a periodic signal is the inverse (reciprocal) of the pitch period length[16][7]. The fundamental frequency is a measure of how high or low the frequency of a person's voice sounds. Pitch information is typically less robust and more difficult to model than other prosodic features, such as duration. F0 is calculated using

$$Fo = \frac{1}{T}$$

Where F0 is the fundamental frequency and T is the fundamental period.

**Energy**

Energy features aim to capture loudness patterns. It is expected that talkers tend to begin their utterances aloud and gradually taper off. It is very much related to the amplitude. It is a way of representing the amplitude changes in speech signal[17]. The amplitude of unvoiced segments is noticeably lower than that of the voiced segments. The short-time energy of speech signals reflects the amplitude variation. It is defined as:

$$\text{En} = \sum_{m=-\infty}^{\infty} x^2(m).h(n-m)$$

In order for $E_n$ to reflect the amplitude variations in time (for this a short window is necessary), and considering the need for a low pass filter to provide smoothing, h(n) was chosen to be a hamming window powered by 2. It has been shown to give good results in terms of reflecting amplitude variations.

**Pause**

Speakers introduce pauses for different reasons, such as, on arrival at punctuation marks, to give emphasis, on word junctures, to stop running out of breath, and other phenomena (such as a disfluency i.e. a repair or filled pause)[14].

In Amharic, as in other languages, punctuation can be taken as a main indicator for pause. Some of the commonly used sets of punctuations in Amharic that may involve signaling of pause are comma (፣), full stop (።), colon (፤), question mark (?), exclamation mark (!) and word space (፡).

In speech predicting the pauses occurrence and their duration have to be considered, the simple presence or absence of a silence is the most significant decision. A study by [16] related to pause behavior in spontaneous speech showed that fluent pause or sentence boundary's duration is observed to be statistically longer than disfluency pause. kolar [15], reported that the pauses marking sentence boundaries were longer than other pauses. Therefore, pause is used is the most basic feature for sentence boundary detection. Considering this, the pause features are extracted to represent characteristics of the sentence boundary candidate. We extracted the duration of the pause preceding and succeeding the boundary, to reflect whether speech before and after the boundary was just starting up or continuous from previous speech.

**Duration**

The other important prosodic cues to sentence boundaries in speech are changes in the speaking rate. Duration of a sentence is important in determining whether a speech segment is a sentence or non-sentence. Durations can be modeled for different type of target units, for example durations of phonemes, syllables, words, or phrases. And appropriate relationships must be established between these units and the syntactic-prosodic information. Speech segment with longer duration is classified as a sentence[7]. On the other hand, short segment is classified either

as a part of sentence or a short sentence. Therefore, speech segments with longer duration indicated a higher potential of occurrence of sentence boundary at the end of it[15][14].

## Prosodic terminology

The following subsections overview the most important terms of prosodic terminology.

### Focus

Focus is semantic accent. This accent is not linked with words but with sentences or clauses. It does not indicate prominence of a syllable but prominence of a word. The emphasized word usually shows a stronger stressed syllable with higher or lower pitch, or eventually it is longer or shorter[6].

### Melody, Intonation, Intonation Patterns

The term melody of speech refers to a pitch curve within an utterance. The term intonation is also often used as a synonym of melody. One way to formally describe a melody of an utterance is to disarticulate its pitch curve into abstract melodic patterns – intonation patterns. Linguists particularly focus on intonation patterns of sentence-final intonation phrases. The following patterns are usually distinguished: falling, rising, flat, falling-rising, and rising-falling. This classification system is phonemic – it is meant to express contrasts in a succinct manner rather than specify the realization[15].

### Pitch Declination, Pitch Reset

Running averages of F0 are higher on the beginning of an utterance unit than on the end. This feature of speech is usually denoted as a pitch declination. Pitch reset is often observed at utterance unit (sentence, paragraph) boundaries. The speaker returns to higher F0 values and the declination slope is usually changed. It was shown that stronger pitch resets usually correspond with more significant boundaries[6].

### Stress, Accent, Rhythm

In phonology, stress is a relative emphasis given to certain syllables in a word. We often distinguish two terms: stress and accent. The term stress usually denotes a “potential feature”, while the term accent stands for a realized emphasis. However, some other authors use these two

terms as synonyms. Overall distribution of stressed syllables within an utterance sets the rhythm (or timing ) of speech. The timing is determined by segmentation of the continuous stream of syllables into groups having similar length and acoustic characteristics[6][14].

**Tone**

It refers to an identifiable movement of pitch that is used in a linguistically contrastive way. In some languages (known as tone languages) the linguistic function of tone is to change the lexical meaning of a word. In other languages, tone forms the central part of intonation, and the difference between, for example, a rising and a falling tone on a particular word may trigger a different interpretation of the sentence in which it occurs. In the case of tone languages, it is usual to identify tones as being a property of individual syllables, whereas an intonational tone may be spread over many syllables and its purpose is to indicate prosodic features such as focus, contrast, exclamation [14].

**Pauses**

In a long sentence, speakers naturally pause a number of times. Important cues to boundaries between semantic units, such as words, sentences, topics, are breaks in prosodic continuity, including pauses. In a typical system, the most reliable indicator of pause location is punctuation. In speech predicting pauses, although their occurrence and their duration have to be considered, the simple presence or absence of a silence (of greater than 30 ms) is the most significant decision. There are many reasonable places to pause in a long sentence, but a few where it is critical not to pause[14][16].

## 2.5. Approaches of speech segmentation

In recent years, the development of an automatic speech segmentation has been a great deal of interest and is becoming a popular research area[6]. Speech segmentation systems developed so far, for different languages, are based on one of the two approaches Rule-based and machine learning.

### 2.5.1. Rule based approach

The rule-based approach has successfully been used in developing many natural language processing systems. Rule based systems are typically built from combined linguistic knowledge

of human experts in the problem domain. The linguistic knowledge acquired for one natural language processing system may be reused to build knowledge required for a similar task in another system[6]. A Rule-based system is built on two main components: a set of facts about situation and a set of rules for how to deal with those facts. The inference engine repeatedly selects a rule whose condition is satisfied and executes the rule [18].

The rule based speech segmentation can also based on acoustic characteristics of a speech like pitch, energy, zero crossing rate (ZCR), power spectral density (PSD), formant transitions, rhythm of consonants and vowels, fundamental frequency, duration, rate of speech, manner of articulation and place of articulation explicitly[18].

Rules are prearranged in rule-based sentence segmentation according to the acoustic features extracted from the speech segments. Each feature has its own threshold value and if a candidate features evaluated to true, the score is assigned to the boundary candidate indicating a sentence boundary. Meanwhile, if a candidate's feature evaluated to a false, a missed is assigned to the sentence boundary score. Boundary candidates that have a high score of boundary hits are classified as true sentence boundary[7]. Rule-based approach however, is not exhaustive and not robust to conflicts. Rule-based methods are very complex and hard to optimize their parameters efficiently; the performances reduce severely in real application, because learn-ability with being explicitly program is a not good method for classification.

### 2.5.2. Machine learning approach

Machine learning methods remained the focus of Speech segmentation work in recent research work. Unlike rule based methods, it is probabilistic and uses statistical model rather than deterministic rules. Machine learning approaches assume that the outputs can be described by a combination of input variables and other parameters. Machine Learning is the study of building computer systems that learn from experience.

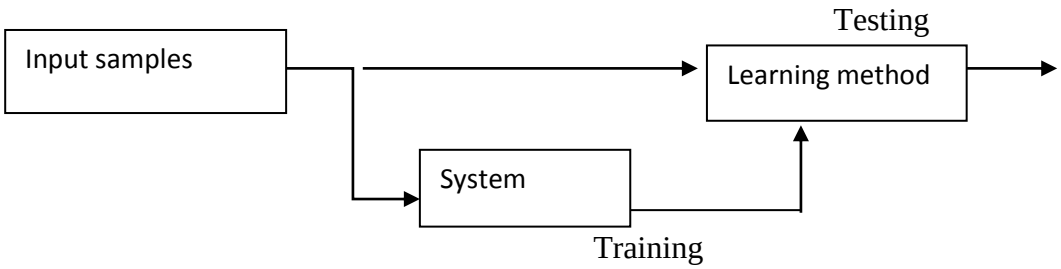



Figure 2.1: Learning system model

Steps in machine learning are:-

**Data collection**:-"training data", optionally with "labels" provided by a "teacher".

**Representation**:-how the data are encoded into "features" when presented to learning algorithm.

**Modeling**:-choose the class of models that the learning algorithm will choose from.

**Estimation/Selection**:-find the model that best explains the data: simple and fits well.

**Validation**:-evaluate the learned model and compare to solution found using other model classes. Apply learned model to new "test" data.

There are different statistical approaches to speech segmentation. This section covers the most common speech segmentation approaches used for sentence level speech segmentation such as hidden Markov model (HMM), a maximum entropy (MaxEnt) model, and a boosting-based model called BoosTexter.

#### 2.5.2.1. Hidden Markov Model

The Hidden Markov Model is one of the most important machine learning models in speech and language processing. HMM is a popular statistical tool for modeling a wide range of time series data. Andrei Markov gave his name to the mathematical theory of Markov processes in the early twentieth century, but it was Baum and his colleagues that developed the theory of HMMs in the 1960s [19].

HMMs are very popular classifiers in many different tasks like in signal processing, and in particular speech processing, but have also been applied with success to low level NLP tasks. One of the reasons for the HMM popularity is due to the following advantages:

- there exist very efficient methods of HMM training and testing [19].
- able to handle new data robustly[18]
- computationally efficient to develop and evaluate due to the existence of established algorithms[18]
- used in different working environments
- language independent and does not assume complex linguistic knowledge

A hidden Markov model is a tool for representing probability distribution over sequence of observation. The general task of sequence classification is to assign a label to every element of the observed sequence. The HMM, given the sequence of elements, computes a probability distribution over all possible labels, and consequently finds the most probable label sequence [6]. The word "hidden" in the name of the model refers to the fact that we do not directly observe the state sequence that the model passes when generating the observation sequence, but only its probabilistic function [6][20].

The HMM model has the following components:

- O=o1,o2, . . . , oN :- an observation sequence (discrete or continuous valued);
- S = s1, s2, . . . , sN :- an underlying sequence of states;
- S = $s_0$, $s_{N+1}$ :- special start and end states not associated with observations
- A = a01, a02, . . . , $a_{nn}$ :- a transition probability matrix where $a_{ij}$ represents a probability of moving from state i to state j , satisfying the condition $\sum_{j+1}^{N} = 1$
- B = $b_i$ ($o_t$ ) — a set of observation likelihoods expressing the probability that an observation $o_t$ is generated by state i [6].

A first-order HMM makes two strong simplifying assumptions. First, the probability of the following state is only dependent on the directly preceding state

$$p(si|s1, s2, \ldots . si-1) \approx p(si|si-1)$$

Second, the probability of the output observation oi depends only on the particular state $s_i$ that generated the observation.

$$P(\text{oi}|\text{s1}, \text{s2}, \ldots, \text{si}, \ldots, \text{sn}, \text{o1}, \text{o2}, , \ldots, \text{oi}, \ldots, \text{on}) \approx \text{P}(\text{oi}|\text{si})$$

There are three fundamental methods for HMMs. First, we need a method to compute the likelihood of an observed sequence O given an HMM and its parameters (A, B). This likelihood can be efficiently computed using the Forward algorithm or Backward algorithm which is based on dynamic programming. A combination of the two algorithms, the Forward-Backward algorithm, can be used to estimate probabilities of the hidden state values given the output sequence [6][19].

Second, we need a method to find the best hidden state sequence S given an observation sequence O, and an HMM and its parameters (A, B). The best state sequence is usually decoded using the Viterbi algorithm[6][19].

Finally, we need a method to estimate HMM parameters (A, B) given an observation sequence O (i.e., training data). For this purpose, we can use the Baum-Welch algorithm, which is a special case of the Expectation-Maximization (EM) algorithm and enables training of HMM parameters in an unsupervised approach[6][19].

HMM approach to sentence segmentation, which was introduced by Shriberg and Stolcke, combines lexical and prosodic model within the hidden Markov model (HMM) framework. Lexical information is modeled by an N -gram language model, prosodic information is represented by posteriors output by an independent prosodic classifier. During testing, both knowledge sources are combined within an HMM [6].

#### 2.5.2.2. Boosting-based model

A popular adaptive boosting method known as AdaBoost combines weak-based classifiers to building up a strong classifier. The principle of boosting is to combine many weak learning algorithms to produce an accurate classifier. In boosting, each weak classifier is built based on the outputs of previous classifiers, focusing on the samples that were formerly classified incorrectly. At each iteration of the learning procedure, a new weak learner, $h_t$ is conjured through resampling and reweighting of the previous learners. A different weighting over the

training examples is used to give more emphasis to examples that are often misclassified by the preceding weak classifiers [19][6]. Finally, all the weak learners used in each iteration, t are linearly combined to form the classification function in equation

$$f(x, I) = \sum_{t=1}^{T} \propto_t h_t(x, I)$$

where $\alpha_t$ is the weight of the weak learner $h_t$ and T is the number of iterations.

**How does the AdaBoost algorithm work?**

It works in the following steps:

1. Initially, Adaboost selects a training subset randomly.
2. It iteratively trains the AdaBoost machine learning model by selecting the training set based on the accurate prediction of the last training.
3. It assigns the higher weight to wrong classified observations so that in the next iteration these observations will get the high probability for classification.
4. Also, It assigns the weight to the trained classifier in each iteration according to the accuracy of the classifier. The more accurate classifier will get high weight.
5. This process iterate until the complete training data fits without any error or until reached to the specified maximum number of estimators.
6. To classify, perform a "vote" across all of the learning algorithms you built.

AdaBoost uses many base classifiers. It uses decision tree classifier as default classifier[11]. In our experiment we have used decision tree classifier and support vector classifier (SVC) as a base estimator.

**Decision Tree Classifier**

Decision trees are one of the most popular algorithms used in machine learning, mostly for classification but also for regression problems. Decision Tree algorithm belongs to the family of supervised learning algorithms. The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by **learning decision rules**

inferred from training data. The decision tree algorithm tries to solve the problem, by using tree representation. Each **internal node** of the tree corresponds to an attribute, and each **leaf node** corresponds to a class label.

To build a decision tree, we need to make a first decision on the dataset to dictate which feature is used to split the data. To determine this, we try every feature and measure which split will give you the best results. After that, we will split the dataset into subsets. The subsets will then traverse down the branches of the first decision node. If the data on the branches is the same class, then you've properly classified it and don't need to continue splitting it. If the data is not the same, then you need to repeat the splitting process on this subset. The decision on how to split this subset is done the same way as the original dataset, and we repeat this process until we have classified all the data.

1. First test all attributes and select the one that would function as the best root
2. Break-up the training set into subsets based on the branches of the root node
3. Test the remaining attributes to see which ones fit best underneath the branches of the root node;
4. Continue this process for all other branches until
    - All examples have the same value
    - There are no more attributes
    - There are no more examples

Before we can measure the best split and start splitting our data, you need to know how to calculate the information gain. The split with the highest information gain is our best option. The measure of information of a set is known as the *Shannon entropy*, or just *entropy* for short. Entropy is defined as the expected value of the information. The Entropy measures the disorder of a set S.

- Calculation of entropy

$$Entropy(S) = \sum_{i=1}^{k} -\frac{|S_i|}{|S|} \log_2 (\frac{|S_i|}{|S|})$$

- S = set of examples
- $S_i$ = subset of S with value $v_i$ under the target attribute
- k = size of the range of the target attribute

We calculate the entropy for all values of an attribute as the weighted sum of subset entropies as follows:

$$\sum_{i=1}^{k} \frac{|S_i|}{|S|} Entropy(S_i)$$

- where k is the range of the attribute we are testing

Entropy is minimized when all values of the target attribute are the same. Entropy is maximized when there is an equal chance of all values for the target attribute (i.e. the result is random)

The Information Gain measures the expected reduction in entropy due to splitting on an attribute. We can also measure information gain (which is inversely proportional to entropy) as follows:

$$Entropy(S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} Entropy(S_i)$$

Choose the split that achieves most reduction (maximizes GAIN). For our classifier algorithm to work, you need to measure the entropy, split the dataset, measure the entropy on the split sets, and see if splitting it was the right thing to do. We will do this for all of our features to determine the best feature to split on.

**Support Vector Machines**

In machine learning, **support-vector machines** are supervised learning models. Support Vector Machine (SVM) is primarily a classier method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables.

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N—the number of features) that distinctly classifies the data points. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimentional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

SVM [21] was developed from the theory of Structural Risk Minimization. In a binary classification problem, SVM's decision function is

$$f(\mathbf{x}) = h\mathbf{w}, \varphi(\mathbf{x})i + b$$

where $\varphi(\mathbf{x})$ is a mapping of sample $\mathbf{x}$ from the input space to a high-dimensional feature space. $h\cdot,$ $\cdot i$ denotes the dot product in the feature space. The optimal values of $\mathbf{w}$ and $b$ can be obtained by solving the following optimization problem,

$$\text{Minimize: } g(\boldsymbol{w}, \xi) = 1/2\ k\ \boldsymbol{w}k\ 2 + C \sum_{i=1}^{N} \xi i$$

$$\text{Subject to: } yi(h\mathbf{w}, \varphi(\mathbf{x}i)i + b) \geq 1 - \xi i, \xi i \geq 0$$

Here, $\xi i$ is the $i$-th slack variable and $C$ is the regularization parameter.

## 2.6. Evaluation Metrics for speech Segmentation

Several evaluation metrics have been used for performance scoring of automatic speech sentence segmentation systems. There are various metrics that have been used in various researches but there is no measure that considered as a standard. Some of these metrics are discussed briefly bellow.

**Accuracy (A)**

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

$$\text{Total sentence boundary} = \frac{\text{Total correct sentence boundary}}{\text{Total sentence boundary}}$$

One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, you have to look at other parameters to evaluate the performance of your model.

**Precision (P)**

Let TP denote the number of true positives, TN true negatives, FP false positives, and FN false negatives. Then, precision, defined as a measure of the proportion of the detected event labels that the system got right, may be expressed as

$$P = \frac{TP}{TP+FP}$$

**Recall (R)**

Recall, defined as the proportion of the detected event labels that the system found, may be expressed as

$$R = \frac{TP}{TP+FN}$$

The higher precision and recall scores are, the better we consider the evaluated system.

## 2.7. Amharic language and its phonology

This section provides present description of the Amharic language. The general sound production system with especial reference to Amharic language is also discussed.

### 2.7.1. The Amharic Language

Amharic is the working language of the Federal Government of Ethiopia. It is the Semitic language that has the second large number of speakers in the world after Arabic. Thus, it is one of the most spoken Semitic languages in Ethiopia with at least 27 million native speakers. In contrast, it has also been identified as an under resourced language based on the following aspects, very limited research on acoustic features and spoken language technologies, lack of electronic resources for speech and language processing such as transcribed speech data, monolingual corpora and pronunciation dictionaries[22].

There are five different dialects of Amharic the Addis Ababa, Gojjam, Wollo, Gondar and Shewa dialects[23]. It is derived from the place where they are spoken. The speech of Addis Ababa has emerged as the standard dialect and has wide currency across all Amharic-speaking communities. The basic word order of the Amharic language is SOV. It is one of the languages that have their own writing system. It is written using fidel. Amharic has its own writing system, a semi-syllabic system. It has about 33 primary characters, each representing a consonant and 7 vowel which follow consonant, resulting in 231 CV syllables with 196 different pronunciations. Fidel is used across all Amharic dialects[14].

### 2.7.2. Amharic Phonology

Phonology is the study of the distribution and patterning of speech sounds in a language and of the tacit rules governing pronunciation[24]. In phonology, phoneme is the fundamental unit that describes how speech conveys linguistic meaning. The phoneme represents a class of sounds that convey the same meaning. The meaning of a word is dependant on the phoneme that it contains. But, all the letters of the Amharic script are not necessary for the pronunciation patterns of the spoken language; some were simply inherited from Ge'ez without having any semantic or phonetic distinction in modern Amharic[24].

Humans can produce an infinite number of sounds; each language has a set of abstract linguistic units, called phonemes, to describe its sounds. Amharic language is primarily comprised of 38 phonemes – 7 vowels and 31 consonants one additional consonant /v/ is inherited and included summing up to a total of 39 phonemes[23]. The following is the brief overview of each of the major categories of Amharic phonemes.

**The Amharic consonants**

Consonant phonemes are described in the following categories: voiced vs. unvoiced; manner of articulation. The place of articulation means where the constriction is located in the vocal tract. Based on manner of articulation, Consonants are generally classified as stops, fricatives, nasals, liquids and semi-vowels[22]. Table 3.1 shows the phonetic representation of Amharic consonants as to their manner of articulation, voicing and place of articulation.

| | Labials | Alveolar | Palatals | Velars | LabioVelar | Glottals |
|---|---|---|---|---|---|---|

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Stops | Voiceless | p | ፕ | T | ት | | | k | ከ | kwa | ኳ | Ax | ዕ |
| | Voiced | b | ብ | D | ድ | | | g | ግ | gwa | ጓ | | |
| | Glottalized | px | ጵ | Tx | ጥ | | | q | ቅ | qwa | ቋ | | |
| Fricatives | Voiceless | f | ፍ | S | ስ | sx | ሽ | | | | | H | ህ |
| | Voiced | | | z | ዝ | zx | ዥ | | | | | | |
| | Glottalized | | | xx | ጽ | | | | | | | Hwa | ኋ |
| Affricatives | Voiceless | | | | | c | ች | | | | | | |
| | Voiced | | | | | j | ጅ | | | | | | |
| | Glottalized | | | | | cx | ጭ | | | | | | |
| Nasals | Voiced | m | ም | N | ን | nx | ኝ | | | | | | |
| Liquids | Voiced | | | L | ል | | | | | | | | |
| | Voiced | | | R | ር | | | | | | | | |
| Glides | | w | ው | | | y | ይ | | | | | | |

Table 2.2: Categories of Amharic Consonants[22]

Some of the Amharic consonants have similar phonetic transcriptions like English. These include ብ [b], ድ [d], ፍ [f], ግ [g], ህ [h], ክ [k], ል [l], ም [m], ን [n], ፕ [p], ር [r], ስ [s], ት [t], ቭ ው[w], [y] and ዝ [z]. They correspond to English consonants b,d,f,g,h,k,l,m,n,p,r,s,t,v,w,y,and z. There are also sounds that have similar sound as English sounds, but are represented using different symbols. These include ች [ch], ኝ [nx], ሽ [sx] and ዥ[zx]. Sounds that are the characteristics of Amharic but not found in English are ጵ [px],ጥ [tx],ፅ [xx],ጭ[cx] and ቅ [q] [22].

Lengthening in time of the pronunciation of consonants is called ‘germination’. The local grammarians call this ‘tightening’ of consonants, showing that it is thought of as a form of emphasis or stress[23]. Unlike other languages like English, the rhythm of Amharic is marked mainly by longer and shorter syllables depending on the germination of consonants, and by certain features of phrasing. In Amharic, all consonants except ህ /h/ and ዕ /ax/ may occur in either a geminated or a non-geminated form[22].

**The Amharic Vowels**

Vowels have different categories based on the position and height of the tongue and their shapes during speech production. Based on the tongue position in the oral cavity, vowels are classified into three sub categories that are front, central and back. Based on the height of the tongue, these vowels are also classified into high, middle and low. Based on their shapes during speech production, vowels are classified into two sub classes that are rounded and unrounded[22][25].

Amharic has total of 7 vowels - five of the commonest vowels a, e, i, o, and u plus two additional central vowels E and I. their category is given in table 3.2.

| | Front | Central | Back |
|---|---|---|---|
| High | ኢ [i] | እ [I] | ኡ [u] |
| Mid | ኤ [e] | [E] | ኦ [o] |
| Low | | ኣ [a] | |

Table 2.3. Categories of Amharic vowels

## 2.8. Related Works

There are related works done on speech segmentation using Hidden Markov Model and other approaches in various languages. This section provides the study of related works which are done locally and internationally so far.

There is one research conducted on automatic speech segmentation for Amharic language. **Phoneme level Automatic Speech Segmentation for Amharic Language Using HMM approach** [18]. In this study unsupervised method of automatic speech segmentation is proposed as a solution. Text corpus with size of 1000 Amharic sentences is recorded by one female speaker. Both text and speech corpuses are split into training (90%) and test (10%) data sets. HMM approach is used to model Amharic phonemes in individual HMM with three emitting and two non-emitting states without skipping left to right HMM. MFCC feature vectors together with their first and second derivatives are selected for individual HMM models. For design of automatic speech segmentation system they used two main approaches these are grapheme based and phoneme based.

HTK toolkit is used to implement the HMM model in two phases. The second phase unlike the first phase includes epithetic vowels of Amharic language while the first phase is built with direct transliteration of Amharic words in to their corresponding Latin representations. In both phases three experiments were conducted; automatic speech segmentation in context independent, context dependent with single Gaussian mixture and context dependent with multiple Gaussian mixtures. The automatic speech segmentation system is evaluated with manual segmentation results by comparing automatic segmented phonemes to manually labeled phonemes with their time boundaries. The evaluation is done in terms of boundary deviations with in tolerances values of 5ms, 10ms, 15 ms and 20ms. Finally best performance with Minimum percentage of time boundary deviations are achieved at phoneme based speech segmentation in context dependent with Gaussian mixture value two.

## Voice Segmentation without Voice Recognition

**Mulgi and Mantri** [26] presented a semantic voice segmentation approach without voice recognition. They present a novel algorithm based on vowel/consonant/pause classification result and some low-level feature sets. The sentence segmentation system is composed of three phases, in the first phase, basic features are extracted and the input audio is segmented into 20ms-long non overlapping frames, where frame features, including frame energy, Zero-Crossing Rate (ZCR) and pitch value, are calculated. In the second phase, three feature sets, including pause features, ROS and prosodic features are extracted and combined. In the third phase, a statistical method, AdaBoost, is used to detect the true sentence boundary. In this approach, an automatic pause threshold detection approach is proposed for adaptive V/C/P classification; features on pause, ROS and prosody are also considered to identify sentence boundary. For each sentence candidate, features mentioned above are extracted. Then, a statistical method, AdaBoost, is used to discriminate the true boundaries from false. They first implemented a baseline system which uses only pause features. About 79.8 boundaries are discriminated correctly. When prosodic features are introduced, about 82.3 boundaries are discriminated correctly.

**Prosody-Based Sentence Boundary Detection of Spontaneous Speech**

The work of [4] presented sentence boundary detection of spontaneous Malay language speech using acoustic and prosodic features. The Malay language is one of the world most spoken languages, being spoken by approximately 180 million people but due to lack of electronic resources for speech and language processing it has also been identified as under-resourced language.

Corpuses were collected from Malaysia Parliamentary Hansard debates which contain spontaneous and formal speeches surrounded with medium noise condition, disfluencies such as "um", repeat and self-repair, speaker's interruption, and different speaking styles like low, medium and high intonation. And also the data contains noises such as claps, laughter, whispers, and arguments. First 185 minutes of one parliamentary session document were selected as a dataset. The selected document consists of two sessions. The first session consists of formal speeches with read text prepared before the session. The second session is questions and answer (Q/A) session spontaneously answered during the parliamentary debate. Only the second session of the debate was used. The length of this session is 88 minutes. For faster processing this 88-minutes audio data is further segmented into 176 non overlapping segments of 30 seconds. But for sentence boundary detection experiments only 84 segments totaling to 42minutes question answer (Q/A) session of spontaneous speech comprising 12 adult male and 4 female were used. There are total of 227 sentence boundaries in 42-minutes dataset.

The experiments of sentence boundary detection is composed of four phase which are audio segmentation, feature extraction, speech/non-speech classification and boundary detection. In the first phase, the 42-minutes audio data speeches are further divided into 20 milliseconds (0.02 sec) non-overlapping 126,000 frames. In the second phase, there are two stages of feature extraction. The first stage is used for speech/non-speech classification while the second stage is for sentence boundary detection.

To categorize speech/non-speech fragments fundamental frequency (F0), energy and zero-crossing rate (ZCR) are extracted from each 126,000 frames. In the second stage of feature extraction, energy and F0 features are combined with seven prosodic features (rate-of-speech, volume change rate, pause, succeeding and preceding sentence duration, succeeding and preceding pause duration, and rate of-speech duration) to detect sentence boundaries.

In the third phase, first, by using Audacity 1.3.12-beta software a ground truth dataset is constructed by manually labeling the speech/non-speech segments of the speech datasets used to evaluate the sentence boundary detection's performance. A total of 6,413 segments are annotated from 84 segments consisting of 3,206 speech segments and 3,207 non-speech segments. Then frames that have high ZCR are categorized as speech segments and frames with low ZCR are categorized as non-speech segments. Frames that have very low value of F0 are categorized as non-speech segments and frames with high F0 are categorized as speech segment. A non-speech segment has much lower amplitude than the speech segment, resulting to non-speech segment to have lower energy.

Speech and non-speech classifications are done using the vowel/consonant/pause (V/C/P) classification rules adapted from [27]. However, they improved their methods by adding fundamental frequency and volume features at different stages of sentence boundary detection. Once all the frames are classified as vowel, consonants or pause, finally they merged vowel and consonant frames as speech segments and categorize pause frames as non-speech segments.

In the final phase, for closer analysis of the boundary candidates the pause duration for our speech dataset are shorter than 0.12 seconds are not considered as potential sentence boundaries. From a total of 3,207 boundary candidates, they removed 935 boundary candidates less than duration of 0.12 seconds. After omitting shorter non-speech segments, there are a total of 2,272 sentence boundaries in the speech dataset. Then a total of 10 audio features consisting of 7 prosodic features, 2 rate-of speeches (ROS) and a volume feature are extracted from the 2,272 boundary candidates.

Rule-based basic classification method was used for sentence boundary detection. Performance of the rule-based method is evaluated using accuracy rate and total error rate. Mean of each 10 features is calculated and used as a threshold for determining the sentence boundary. Each feature has its own threshold value and if a boundary candidate's feature evaluated to TRUE, a hit score is assigned to the boundary candidate indicating a sentence boundary. Meanwhile, if a boundary candidate's feature evaluated to a FALSE, a missed is assigned to the sentence boundary score. Boundary candidates that have a high score of boundary hits are classified as true sentence boundary.

Rule-based classification method classifies sentence boundary at 74.88% is achieved by candidates with at least 4 total hits. However, it produced the highest false alert that is 80.63%. All the results showed that as the detection rate of sentence boundary increases, the false alert rate also increases. Finally they conclude that acoustic/prosodic features are believed to be reliable properties for sentence boundary detection.

## Prosody-based automatic segmentation of speech into sentences and topics

A research on **Prosody-based automatic segmentation of speech into sentences and topics** [4] studied the use of prosodic information for sentence and topic segmentation, both of which are important tasks for information extraction and archival applications. Corpuses were collected from two speech corpora, Broadcast News and Switchboard. Switchboard data used in sentence segmentation was hand-labeled for sentence boundaries by the Linguistic Data Consortium (LDC). Broadcast News data for topic and sentence segmentation was extracted from the LDC's Broad cast News (BN) release. The two corpora are compared directly on the task of sentence segmentation, and the two tasks (sentence and topic segmentation) are compared for the Broadcast News data. For each task, they examine results from combining the prosodic information with language model information, using both transcribed and recognized words. They were extracted prosodic features reflecting pause durations, phone durations, pitch information, and voice quality information. Pause features were extracted at the inter-word boundaries. Duration, F0, and voice quality features were extracted mainly from the word or window preceding the boundary preceding the boundary. They also included pitch-related features reflecting the difference in pitch range across the boundary and non-prosodic features that are inherently related to the prosodic features.

To automatically reduce large initial candidate feature set to an optimal subset, they developed feature selection algorithm that involved running multiple decision trees in training. The algorithm proceeds in two phases. In the first phase, the large number of initial candidate features is reduced by a leave-one-out procedure. The second phase begins with the reduced number of features, and performs a beam search over all possible subsets of features.

The goal of language modeling for their segmentation tasks is to capture information about segment boundaries contained in the word sequences. Sentence segmentation performance for

true words was measured by boundary classification error. Topic segmentation was evaluated using the metric defined by NIST for the TDT-2 evaluation.

Prosodic features reacting pause durations, supra segmental durations, and pitch contours were automatically extracted, regularized, and normalized. The features were used as inputs to a decision tree model, which predicted the appropriate segment boundary type at each inter-word boundary. Two knowledge source integration approaches were investigated: one based on interpolating posterior probability estimators, and the other using a combined HMM that emitted both lexical and prosodic observations. Results showed that on Broadcast news the prosodic model alone performed better than purely word-based statistical language models, for both true and automatically recognized words. The integrated HMM worked best on transcribed words. They conclude that prosody provides rich and complementary information to lexical information for the detection of sentence and topic boundaries in different speech styles, and that it can therefore play an important role in the automatic segmentation spoken language.

**Automatic Segmentation of Continuous Speech on Word and Phrase Level based on Suprasegmental Features**

Vicsi & Szaszák[28] presented a cross-lingual study for Hungarian and Finnish about the segmentation of continuous speech on word and phrasal level by examination of supra-segmental parameters. In these languages stress is on the first syllable of the word. But not each word is stressed if stress occurs, then it is at the first syllable. In experiment they measured parameters like fundamental frequency, energy and time course. Each of these parameters is necessary for the perception of stress in Hungarian and Finnish language. Fundamental frequency and energy are mainly dominant.

Two methods were used and compared after the acoustical pre-processing of speech: one is a rule-based method and the other an HMM based statistical method. Hungarian BABEL and Finnish continuous read speech databases were used for the examination. The databases were segmented on phoneme level, and word, phrase and sentence boundaries were also marked. For Hungarian 1600 sentences from 32 speakers and for Finnish 250 sentences from 4 speakers were used. In rule based approach for the prediction of emphasized syllables they used peak-detection algorithms. This peak detection algorithm was performed on F0 and energy level data, but also

on data streams gained by computing absolute value of F0 and energy level differences between two neighboring.

In data-driven approach they used HTK to train and test a HMM recognizer to segment speech on prosodic level using F0, energy level parameters and their first and second order deltas. To train the HMM models prosodic segmentation was performed by an expert relying on fundamental frequency and energy cues. Finally a set of 5 different HMM models were trained to segment speech data on prosodic level. The 5 different models were trained for declarative, neutral, interrogative prosodic phrases and for silence.

Models using 11 states were found the best, with 1, 2, 4 or 8 Gaussian mixture components. Training was performed first with 14 persons speakers mixed, than with a small (1 or 4 persons) set of speakers, the results obtained were compared in both cases. To investigate the adaptively of the Hungarian word boundary detection system they carried out the same experiments on Finnish language. Again 5 different HMMs of 11 states were trained. Hence the Finnish database was collected only from 4 speakers; we involved all of them into the training corpus.

The rule-based algorithms and HMM-based methods are compared. In rule based algorithm the more accurate results were obtained by detecting stress on the basis of fundamental frequency and energy level changes (ΔF0+ΔE) from syllable to syllable. The overall best results can be achieved by calculating only with fundamental frequency differences (ΔF0. The overall best result was 77.4% accuracy with 57.2% effectiveness, obtained with HMMs trained on 4 speakers' F0 and Energy data for Hungarian language. On the other hand, segmenting Hungarian data with Finnish model gives 70.7%, accuracy which is a bit better than on Finnish data (67.1%).The best results were obtained by data-driven algorithms using the time series of fundamental frequency and energy together. By use of supra segmental features, word boundaries can be marked with high accuracy, even if they are unable to find all of them.

**Word boundary detection based on suprasegmental features: A case study on Bangla speech**

The work of [29] presented a method for detecting word boundaries in continuous speech signal for Bangla language. Bangla is a bound stress language with stress on the first syllable. In this study a total number of 225 sentences of different categories (simple, complex, compound,

passive, imperative and yes/no questions) spoken by 5 informants of both sexes and different age groups have been used for the analysis. These are directly recorded in wave files using standard multimedia cards available in PC. The signals are digitized at the sampling rate of 22050 Hz/16bits mono using Cool Edit software. This speech signal consists of three different kinds of waveforms which are quasi-periodic, the quasi-random and the silence.

A state-phase based Pitch Detection Algorithm (PDA) / Voice Detection Algorithm (VDA) is used to classify the speech signal, namely silence, sibilant and voiced as well as to detect F0 and the amplitude for each period in the voiced part of the sentences. This algorithm was automatically done syllabification of the recorded sentences by introducing a syllable marker at the beginning of a silent or sibilant zone. When it falls on a vocalic region the algorithm would fail to detect the syllable marker. In these cases syllabification was done manually by adding short silences at suitable position.

In this study the parameters related to the prosody of spoken sentences are used for marking the word boundary. These parameters are difference of the nucleus vowel duration across the syllable boundary, difference of the normalized nucleus vowel power across the syllable boundary, normalized F0 difference across the syllable boundary, difference of the average normalized F0 across the syllable boundary, difference of the normalized maximum periodic power of nucleus vowels across the syllable boundary and onset duration of the nucleus vowel. These parameters are extracted using the PDA/VDA algorithm. The normalized speech signal is broken up into three different types, voiced, silence and sibilant, using information provided by VDA.

They reported two classification experiments. In experiment one, the speaker independent word boundary detection. In the case of speaker independent classification there is only one group of two classes for the purpose of determining the class representatives. In experiment two, there are five groups corresponding to five speakers each with two classes. Word boundary recognition score for speaker dependent is 79.2% with only 7.6% false inclusion and speaker independent case 87.8% with only 11.1% false inclusion. The overall score of 87.8% with only 11.1% false inclusion for speaker independent case is quite encouraging.

**Automatic segmentation of speech into sentences-like units**

A research on **automatic segmentation of speech into sentences-like units [6]** studied the problem of automatic segmentation of speech recognition output into sentence-like units. The main goal of the work is to develop automatic systems for dialog act segmentation of English multiparty meetings and sentence unit segmentation of the two new Czech corpora. The work is focused on two languages – corpus of multiparty meetings in English and corpus of broadcast conversations in Czech. The English corpus contains multichannel conversational speech recorded by head worn microphones. The data were split into a training set (51 meetings, 539k words), a development set (11 meetings, 110k words), and a test set (11 meetings, 102k words). For Czech, the data included 159.1k words for training, 24.1k words for development, and 24.6k words for testing. The corpus is annotated based on LDC's Metadata Extraction (MDE) standard.

They examined three different modeling approaches relying on both textual and prosodic cues: HMM, maximum entropy, and a boosting-based model BoosTexter and evaluated them using both reference and automatic transcripts. The textual features describe lexical patterns associated with sentence-external and sentence-internal interword boundaries. They used features capturing word identities, parts of speech, and automatically induced word classes. The prosodic features for sentence segmentation of speech reflect breaks in temporal, intonational, and loudness contours in an utterance. Prosodic features for automatic classification are extracted directly from the speech signal based on time alignments from automatic speech recognition, without any need for hand-labeling of prosodic events. They measure SU segmentation performance using F - measure, which is the harmonic mean of Precision (P) and Recall(R).

The comparison of the three modeling techniques according to information sources used (textual and prosodic information). The system achieved for HMM(83.17%), MaxEnt (80.49%), BoosTexter(80.55%) and their combination(83.20%) in reference conditions and HMM(75.33%), MaxEnt (73.90%), BoosTexter(73.29%) and their combination(75.85%) in ASR conditions via linear interpolation [F %] for the English corpus. The system achieved for HMM(74.48%), MaxEnt (73.81%), BoosTexter(73.43%) and their combination(75.65%) in reference conditions and HMM(68.59%), MaxEnt (67.20%), BoosTexter(68.14%) and their combination(69.12%) in ASR conditions for the Czech corpus. The results indicate that the HMM model showed most consistently good results. It was the most successful approach for the

English corpus, but the difference among the three approaches is small. In contrast, BoosTexter was the best performing method for the Czech corpus, and the superiority of BoosTexter over the others (especially HMM) was greater than the differences in English corpus. This indicates that the discriminative models, BoosTexter and MaxEnt, are more robust to lexical irregularities frequent in conversational Czech. Overall, the best results for all our test sets were achieved by a model that combines HMM, MaxEnt, and BoosTexter via posterior probability interpolation.

In this section, we have reviewed works on automatic phoneme, word and sentence segmentation of speech. Many algorithms exist for the segmentation of speech into sentences. However, there has been no previous work on speech segmentation for Amharic. So we want to investigate optimal method for speech segmentation at sentences level for Amharic continuous speech.

# CHAPTER THREE
# DESIGN OF THE ARCHITECTURE FOR AMHARIC AUTOMATIC SPEECH SEGMENTATION

In this chapter we will discuss the overall design of our proposed system, automatic speech segmentation system for Amharic. First, we will illustrate the general overview of the proposed automatic speech segmentation system architecture from the perspective of the systems flow of operations, and then we will present the detailed explanation of the phases along with the subcomponents included in each approach.

## 3.1. Automatic Speech Segmentation Architecture

The proposed sentence segmentation system is composed of two approaches which can be implemented individually as approach one and approach two respectively. In the first approach we used an automatic tool for segmenting and labeling of Amharic speech data. The tool is based on Hidden Markov Model. In the second approach we used direct modeling approach in which prosodic features are extracted directly from the speech signal. The first approach of sentence segmentation system is composed of five stages, data preparation, feature extraction, HMM model building, forced alignment and sentence boundary detection. There are three stages involved in the second approach: audio segmentation, prosodic features extraction and sentence boundary detection. In each approach, statistical method, AdaBoost, is used.

## 3.2. Automatic Speech Segmentation Architecture of approach one

Figure 3.1 illustrates the first approach architecture of the automatic speech segmentation system for Amharic. The architecture presented in Figure 3.1 shows the major elements of automatic Amharic speech segmentation system and the components that are required to address specific tasks.

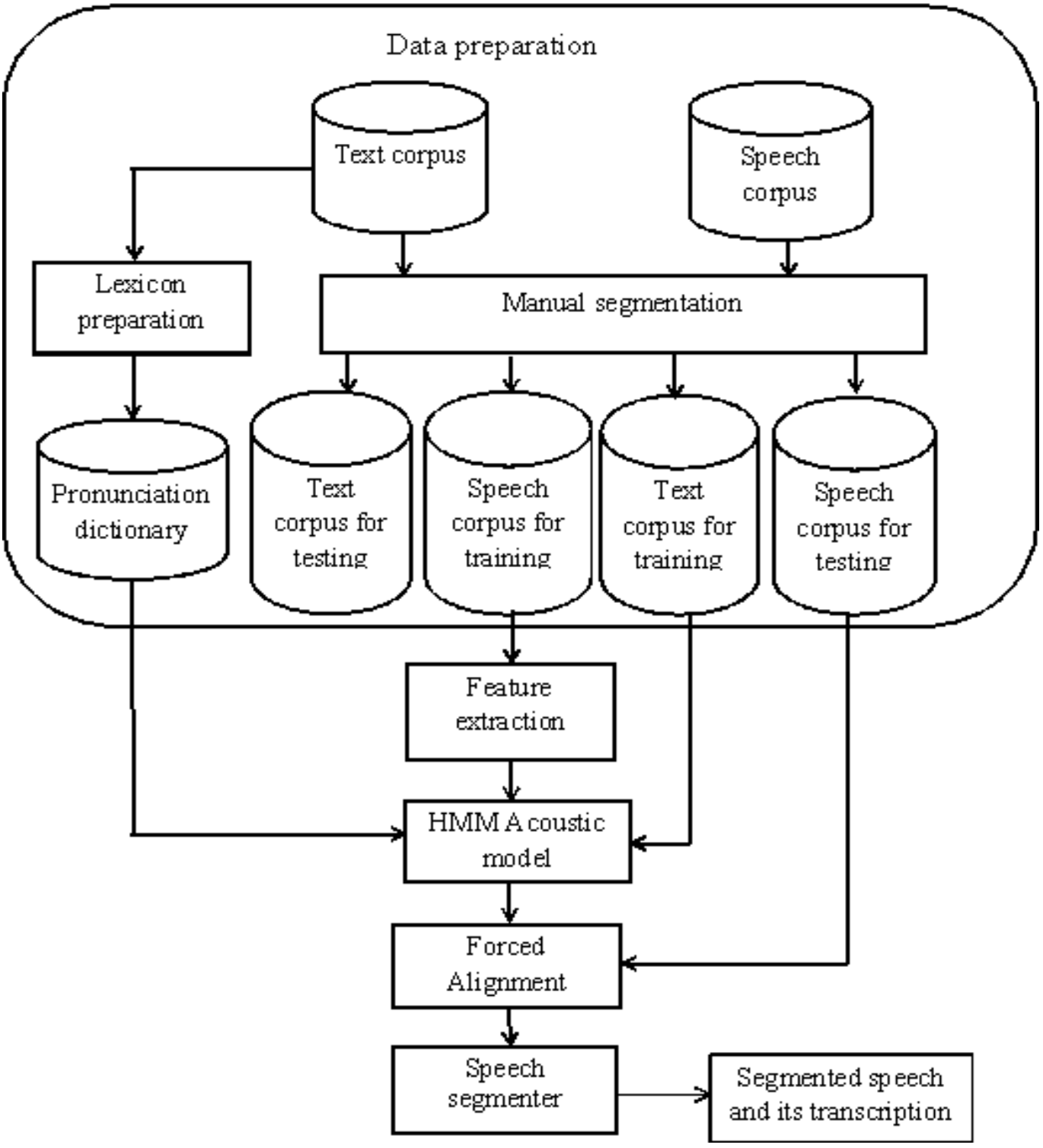



Figure 3.1: Architecture of the automatic speech segmentation system of approach one

### 3.2.1. Data preparation

Data preparation is an important step of solving machine learning problems. An accurate, properly prepared text and speech corpus is very important for speech research areas. It is the first stage of automatic speech segmentation system. It has high contribution to the performance of automatic speech segmenter.

In this paper the steps for getting data ready for speech segmentation system includes data collection, manual segmentation, lexicon preparation and pronunciation dictionary preparation.

**Data collection**

Data collection is the primary process of data preparation during automatic speech segmentation system. A speech corpus is one of the fundamental requirements for speech segmentation research. A speech corpus is a collection of speech recordings which is accessible in computer readable form, and its associated transcription. A standard speech corpus consists of a training set and evaluation test sets. The training set is intended to collect speech data for training and the evaluation test set is for the purpose of final evaluation of the segmenter.

In order to get optimal speech corpus, Amharic speeches and corresponding text data should be collected from Amharic bible, broadcast news, broadcast conversation and Amharic fictions. These Amharic speeches are used as data sources from different domain to get optimal data because there might be difference between these domain feature. The collected corpus contains over 4 hours of audio stored in 40 audio wave files along with their text corpus. On the text corpus, problems like spelling and grammar errors are corrected, abbreviations are expanded and numbers are textually transcribed.

**Manual speech segmentation**

Having the speech and corresponding text corpus, the next important step in data preparation is manual segmentation. One of the main goals of this thesis was to create a sentence level speech segmentation system for Amharic. Since there were no speech corpora with sentence break annotation such corpora had to be prepared as the very first part of our work. Both text and speech corpora are split into training and test sets. Manual segmentation should be done both for the training and test data. We have, therefore, segmented the collected speech at sentence level manually. This segmentation is carried out using Audacity software. Each of the segmented

sentence files are sampled at 44.10 kHz with 16-bit resolution, saved in the *.wav format. Due to time constraint, 2000 spontaneous speech sentences and 2000 read speech sentences are segmented for training while 400 speech sentences from both types of speech are segmented for testing. We have transliterated the text transcription of both the training and test sets into their corresponding ASCII representation so that it can used in the HTK settings.

**Lexicon preparation**

Lexicons are prepared as pronunciation dictionaries for our automatic speech segmentation system. Based on our system, lexicons are prepared with letter sequences. For example; Let us consider Amharic word "ፈጠረ". The Amharic word "ፈጠረ" has three orthographies which are ፈ, ጠ and ረ. Its corresponding ASCII transcription becomes " faTara". According to ASCII translation, "ፈ" is represented with "fa" , "ጠ" is represented with "Ta" and ረ" represented with "ra", and their letter sequences together becomes "faTara".

Usually, the first step in building the Pronunciation dictionary is to create a sorted list of the required words, one per line, with pronunciations. The word list is built from the sentences present in the training data. After we have sorted the word list, we constructed a lexicon dictionary.

The general formats used in our lexicons preparation are like this;

WORD [output] p1 p2 p3.... pn

This means that the word WORD is pronounced as the sequence of syllable/phone p1 p2 p3.....pn

**Pronunciation dictionary**

HTK toolkit is used to prepare pronunciation dictionary by taking lexicons as input. Since pronunciation dictionary for Amharic language is not built yet, preparing pronunciation dictionary plays a major role in the performance of automatic speech segmentation. The pronunciation dictionary maps the orthographic representation in the transcription file to its corresponding pronunciations.

### 3.2.2. Feature extraction

The final stage of data preparation is to parameterize the raw speech of the waveforms into sequences of feature vectors. This means that HTK is not as efficient in processing wav files as it is with its internal format. Therefore, you need to convert wav files to another format called **MFCC** format. Feature extraction is the process of transforming the speech waveform into a set of feature vectors. Mel Frequency Cepstral Coefficients is used to parameterize the speech signals into feature vectors with 39 MFCC coefficients.

### 3.2.3. HMM model building

An acoustic model is a file that contains a statistical representation of each distinct sound that make-up each word used in grammar. Acoustic models are statistical models which capture the correspondence between a short sequence of acoustic vectors and an elementary unit of speech. The elementary units of speech that are used in our research are syllable and phone. We use HMM to model the acoustic component in this research. We have created the acoustic model using Amharic speech and their text scripts and compiling them into a statistical representation of sounds which makeup words. Acoustic modeling process takes pronunciation dictionary, training text corpus, feature vectors of the training speech corpus as main inputs.

The first step in HMM training is to define a prototype model. This prototype defines the structure and the overall form of the set of HMMs. Here a 3-state left-right topology is used to model the HMMs. The second step is initializing HMMs. HMM initialization process involves the creation of HMM definition file, finding and storing of global mean and variance.

#### 3.2.3.1. Phone Based Acoustic Model

The phone models are context independent. It is a common practice to use context-independent HMMs for speech segmentation. In reality, this is not the case as two neighboring phonemes may influence each other in Amharic. To capture these effects, called co-articulations, models are needed that take into account the context of a phone[18][17].

Triphone model the context by taking into consideration the left and right neighboring phones. If two phones have the same identity but different left or right context, they are considered as different triphones. Having a set of monophone HMMs, the next stage of model building is to create context dependent triphone HMMS. Context-dependent triphones are made by simply

cloning monophones and then re-estimating using triphone transcriptions[3][9].With a triphone acoustic model, we are essentially looking for a monophone in the context of other monophones.

To generate a triphone (i.e. a group of 3 phones) declaration from mono phones, the "L" phone (i.e. the left-hand phone) precedes "X" phone and the "R" phone (i.e. the right-hand phone) follows it. The triphone is declared in the form "L-X+R"[6][25].

For example: - the conversion to a triphone declaration of the Amharic word ” faTara”. The first line shows the "monophone" declaration, and the second line shows the triphone declaration:

```
faTara [faTara]  f  a T a r a
faTara [faTara] f +a  a-T+a T-a+r  a-r+a  r-a
```

#### 3.2.3.2. Syllable Based Acoustic Model

The CV syllable is context independent. CV syllable is modeled with monosyllabic HMMs without considering its context.

For example: - to generate a tri-syllable declaration from monosyllable, the conversion to a tri-syllable declaration of the Amharic word “ፈጠረ”.The first line shows the "monosyllable" declaration, and the second line shows the " tri-syllable " declaration:

```
faTara [faTara]  faTa ra
faTara [faTara] fa+Ta fa-Ta+ra Ta-ra
```

The last step in the model building process is to tie similar states of a set of tri-syllable in order to share data and thus be able to make robust parameter estimates.

### 3.2.4. Forced Alignment

Forced Alignment is the process of taking audio file containing speech, and the corresponding transcript and determining for each fragment of the transcript, the **time interval** (in the audio

file) containing the spoken text of the fragment[30]. Forced Alignment process takes four main inputs; pronunciation dictionary, word level transcription, audio file and acoustic model to line up the written words in the words file with the spoken words.

### 3.2.5. Speech segmenter

The speech segmenter gives automatically segmented results of the test data set. In our initial experiment of sentence segmentation, a basic classification method rule-based is used. The segmenter performs the actual segmentation of the continuous speech into sentence level files and creates a corresponding transcription file. In our second experiment features from forced alignment are extracted. Then, a statistical method, AdaBoost, is used to discriminate the true boundaries from false.

## 3.3. Automatic Speech Segmentation Architecture of Phase Approach Two

The proposed approach two speech segmentation system is composed of three stages, as shown in Figure 4.4. These stages are: audio segmentation, features extraction and sentence boundary detection.

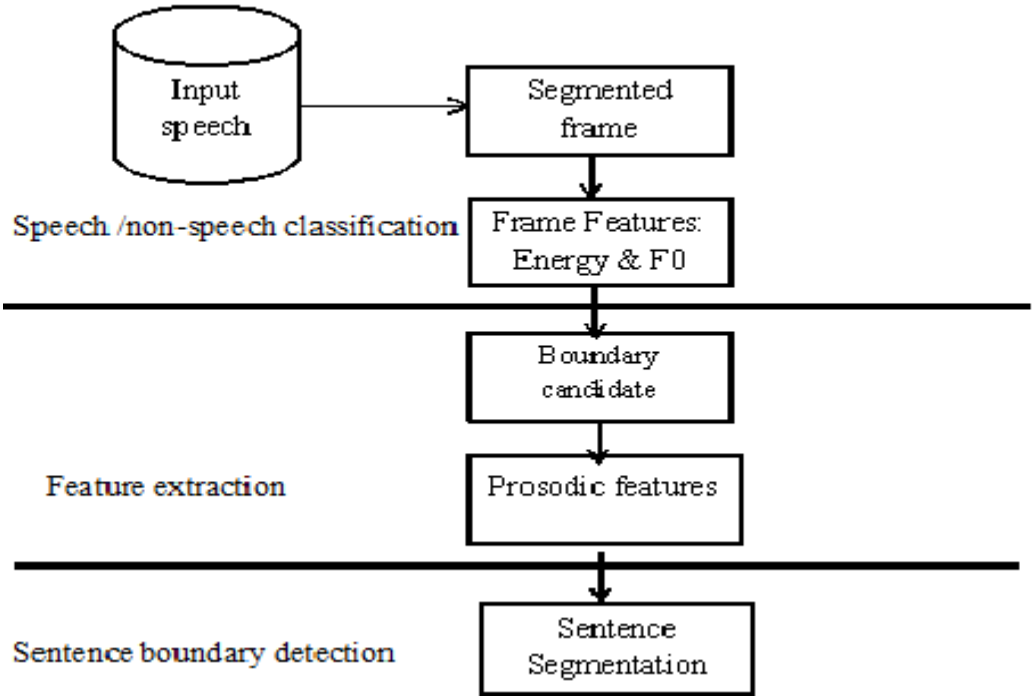



Figure 3.2. Automatic speech segmentation system of approach two

In the first stage, the input audio is segmented into non-overlapping frames. These smaller frames are used in feature extraction for classification of speech/non-speech segments. In the second stage, speech features are extracted. There are two stages of feature extraction in our experiment. In first stage, two acoustic features that are fundamental frequency (F0) and energy are extracted from each segmented frames individually to classify speech/non-speech segments. In the second stage, seven prosodic features are extracted to detect sentence boundaries. These features are pause, succeeding and preceding sentence duration, succeeding and preceding pause duration, fundamental frequency (F0) and energy. In the third stage, a statistical method, AdaBoost, is used to detect the true sentence boundary from the candidates based on its context feature

# CHAPTER FOUR

# EXPERIMENTAL RESULTS AND EVALUATION

## 4.1. Introduction

In this chapter, we describe the experimentation of sentence level automatic Amharic speech segmentation based on the design developed in Chapter Four and then the experiment results are presented for evaluation. This experimentation covers two techniques of speech segmentation. Firstly, automatic segmentation is implemented with HTK toolkit. HTK is used to compile speech audio and transcriptions into an Acoustic Model. Data preparation, HMM model building, forced alignment and sentence boundary detection are the main stages of it. Secondly, automatic segmentation is done based on prosodic features extracted directly from the speech signal. Audio segmentation, features extraction and sentence boundary detection are involved in stage two.

## 4.2. Experimental setup

The research used Audacity software for editing and segmenting sounds and also we have used Hidden Markov Model Toolkit (HTK) that is a portable toolkit for building and manipulating HMM. The system is developed and tested on a system with Intel Core i7 CPU of 2.5 GHZ speed, 8 GB RAM, 1 TB hard disk and Ubuntu 16.04 Operating system. We have used perl and Python 2.7 programming language and htk commands for developing pronunciation dictionary, hmm model building and forced alignment and speech segmenter. In addition, methods used for sentence boundary detection are done using python.

## 4.3. Implementation of automatic speech segmentation system of approach one

### 4.3.1. Data preparation

As mentioned in data collection, Amharic audio data gathered from various sources. We create two corpora read-aloud speech and spontaneous speech. A spontaneous speech contains disfluencies and noises such as um, repeats, claps, laughter and whispers and also different speaking styles.

As mentioned in manual segmentation the collected speech segmented manually. After segmentation of speech and corresponding text corpus is completed transliteration of the Amharic texts into their corresponding ASCII representation is done. In order to do error free transcribed texts, a sed command which takes ASCII translation script as input is used as attached in figure 4.1. This command transcribes automatically Amharic texts to their corresponding Latin representation as per ASCII transliteration script.

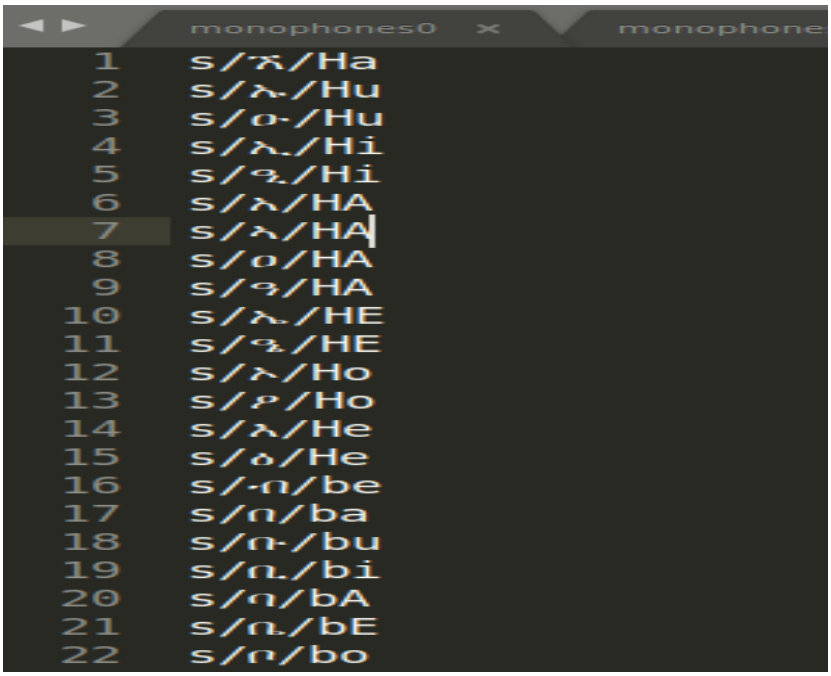



Figure 4.1 ASCII translation script

**Lexicon preparation**

First we created text file that includes list of sentence and the names of the audio files - one per line. The sample text file of the research is shown in appendix A.

The first column of the text file contains the name of the audio file saved in *.wav format and the next column contains the text transcriptions of audio file. The HTK Perl script **prompts2wlist** takes the text file created, and remove the file name in the first column and print each word in sorted order into a word list file "wlist". After we have sorted word list, the next step is to add pronunciation information to each of the words in the wlist file. We have constructed a lexicon dictionary by using python Script.

**Pronunciation dictionary**

We have used canonical pronunciation dictionary. For each word, a canonical pronunciation dictionary includes only the most probable pronunciation that is assumed to be pronounced in read speech[18][22]. We have prepared two Amharic pronunciation dictionaries. One is CV syllable based the other is phone based.

We have used HTK command **HDMan** to create pronunciation dictionary. HTK uses the **HDMan** command to go through the word wlist file shown in Appendix B, and look up the pronunciation for each word in a separate lexicon file, and output the result in a Pronunciation dictionary. The sample output of syllable based pronunciation dictionary is shown in appendix C.

**Creating the Transcription Files**

HTK toolkit cannot process text file directly. To train a set of HMMs every file of training data should have an associated syllable/phone level transcription. To make this task easier, creating a word level transcription before creating the syllable/phone level transcription is required. We use the HTK script **prompts2mlf** that outputs each word on a single line and each sentence terminated by a single period on its own. This script generates a **words.mlf** file as shown in appendix D.

Next we execute the HLEd command that expands the word level transcriptions to syllable and phone level transcriptions. This is done by replace each word with its syllable/phone, and put the result in a new Master Label File. HLEd edit script is used to insert the silence model “sil” at the beginning and end of each sentence. Next, we create a second Master Label File which will include short pauses “sp” after each word syllable/phone group.

**Coding the Data**

We use the HCopy tool to convert our wav files to MFCC format. To do this, a configuration file (config) which specifies all the needed conversion default parameters is created. Figure 4.2 shows "config" file.

```
SOURCEFORMAT = WAV
TARGETKIND = MFCC_0_D
TARGETRATE = 100000.0
SAVECOMPRESSED = T
SAVEWITHCRC = T
WINDOWSIZE = 250000.0
USEHAMMING = T
PREEMCOEF = 0.97
NUMCHANS = 26
CEPLIFTER = 22
NUMCEPS = 12
```

Figure 4.2: Configuration parameters for coding

The configuration file specifies that the source files of the speech are wav file, the target parameters are to be MFCC, the output should be saved in compressed format, and a cyclic redundancy check (CRC) checksum should be added. We create a file containing a list of each source audio file and its corresponding output MFCC file, and use that file as a parameter to the HCopy command. The result is the creation of a series of MFCC files corresponding to the files listed.

### 4.3.2. HMM model building

We use HMM to model the acoustic component in this research. The experiment is done to build acoustic model for new segmentation system. The work includes syllable and phone based acoustic model. The acoustic modeling system is carried out in three modeling techniques, namely syllable/phone based acoustic model and tied state acoustic model.

#### 4.3.2.1. Phone based acoustic model

**Creating Flat Start Monophones**

HMM based technique requires training before using it for the segmentation and labeling. The first step in HMM training is to create a prototype HMM. The prototype model used in this research is adopted from HTK book [20] attached in appendix E.

HMM initialization process includes the creation of HMM definition file, finding and storing of global mean and variance. There will be HMM definition for each phone in the phone list. Three state left-to-rights HMM with no skip is used to represent each phone. The HTK tool HCompV scans the data, and computes the global mean and variance for the whole corpus and outputs to a new prototype model.

Next, a Master Macro File called hmmdefs containing a copy for each of the required monophone HMMs is constructed by manually copying the prototype and relabeling it for each required monophone. Then we re-estimate the flat start monophones using the HTK command HERest. The purpose of this is to load all the models and re-estimate them using the MFCC files listed.

**Realigning the Training Data**

The dictionary may contain multiple pronunciations for some words. HVite command can consider all pronunciations for each word (in the case where a word has more than one pronunciations), and then output the pronunciation that best matches the acoustic data. We used HVite to get output with the most likely alignments. This command uses the HMMs to transform the input word level transcription to the new phone level transcription with time stamps using the pronunciations stored in the dictionary**.** Once the new phone alignments have been created, another 2 passes of HERest can be applied to re-estimate the HMM set parameters again. This is the final monophone HMM. The monophone models created in this step are used for forced alignment.

#### 4.3.2.2. Syllable based acoustic model

As it has been done on phone based model, we initialized the model with flat start methods and create monosyllable. Then, the monosyllable model is retrained and short pause model are added. In the next step tri-syllable model have been derived by cloning the respective monosyllable model and then re-estimating using tri-syllable transcription.

The last step in model building process is to tie states within tri- syllable sets in order to share data and thus be able to make robust parameter estimates[18]. HHED provides a mechanism that uses decision trees to tie the states within the tri- syllable sets. The final step of the acoustic modeling is the re-estimation of created tied state tri- syllables and this process is also same as

the earlier use of HERest command of HTK toolkit. This is also repeated for three times and the final output is trained acoustic model. The tied-State tri-syllables models created in this step are also used for forced alignment. The monosyllable and tied-State tri-syllables models created in this step are also used for forced alignment.

### 4.3.3. Forced Alignment

We used monosyllable, tied-State tri-syllable and monophone acoustic model to line up the written words with the spoken words. Different acoustic models may produce slightly different forced alignment results. The HVite command creates a file called **aligned.out** containing all the words from words file, with time alignments. The sample outputs from the HVite command is attached in Appendix F.

### 4.3.4. Speech segmenter

In our initial experiment, the Speech segmenter takes two main inputs: forced alignment and an audio file. Then, from forced alignment pause features (sp) are extracted. Pause duration is calculated and used as a threshold for determining the sentence boundary. In the first experiment we used rule-based for sentence segmentation. The rules for detecting sentence boundary are shown.

```
If feature candidate >= feature threshold
Candidate=1 # sentence end
else:
Candidate=0 # non sentence end
```

Rule-based sentence boundary detection

If a boundary candidate's feature evaluated to TRUE, the boundary candidate indicating a sentence boundary. Meanwhile, if a boundary candidate's feature evaluated to a FALSE, the boundary candidate is not sentence boundary. We tried different threshold value (10000, 500, 800, 1000 millisecond) to detect sentence boundary. In this experiment we decide minimum pause for sentence break is 1000 millisecond (1 sec).

In the second experiment pause features from forced alignment are extracted. Then, a statistical method, AdaBoost, is used to discriminate the true boundaries from false.

## 4.4. Experimental result and analysis

Experiments have been carried out to compare the performance of automatic speech segmentation. We used three acoustic models to build forced alignment. These are monosyllable, tied-State syllable and monophone acoustic model. We have also used two classification algorithms to detect sentence boundary. The first is rule based and the second is statistical method, AdaBoost. We evaluate our methods using two corpora – Amharic read-aloud speech and spontaneous speech. All experiments were evaluated using human-generated reference transcripts.

**Experiment I: Experimental result based on rule-based method**

The overall results of the automatic segmentation with two test set are shown in Table 4.1 and Table 4.2.

| | Accuracy(A) | Precision (P) | Recall (R) |
|---|---|---|---|
| monosyllable acoustic model | 69.1% | 53% | 56% |
| Tied-State syllable acoustic model | 61% | 67.6% | 63% |
| monophone acoustic model | 56% | 42% | 50% |

Table 4.1: Amharic read-aloud speech experimental results

| | Accuracy(A) | Precision (P) | Recall (R) |
|---|---|---|---|
| monosyllable acoustic model | 53% | 40.8% | 38.5% |
| Tied-State syllable acoustic model | 47% | 48.34% | 47% |
| monophone acoustic model | 45% | 40% | 43.3% |

Table 4.2: Amharic spontaneous speech experimental results

**Experiment II: Experimental Result based on statistical method, adaBoost**.

The overall results of the automatic segmentation with two test set are shown in Table 4.3 and Table 4.4.

| | | Accuracy(A) | Precision (P) | Recall (R) |
|---|---|---|---|---|
| monosyllable acoustic model | Decision tree classifier | 91.93% | 78.3% | 75.8% |
| | SVM classifier | 84.3% | 72.7% | 69.4% |
| Tied-State acoustic model | Decision tree classifier | 88.93% | 82.23% | 88.3% |
| | SVM classifier | 80.08% | 74.7% | 78% |
| monophone acoustic model | Decision tree classifier | 82.2% | 69.8% | 70.3% |
| | SVM classifier | 80.2% | 66% | 69% |

Table 4.3. Amharic read-aloud speech experimental results

| | | Accuracy(A) | Precision (P) | Recall (R) |
|---|---|---|---|---|
| monosyllable acoustic model | Decision tree classifier | 85% | 73.7% | 77.7% |
| | SVM classifier | 79% | 63.5% | 67.3% |
| Tied-State acoustic model | Decision tree classifier | 82.7% | 74.7% | 79.7% |
| | SVM classifier | 70% | 70.7% | 73.7% |
| monophone acoustic model | Decision tree classifier | 80.6% | 66% | 69.4% |
| | SVM classifier | 76% | 63.8% | 53% |

Table 4.4.Amharic spontaneous speech experimental results

**Experimental analysis**

As shown in experiment I table 4.1 and table 4.2 the percentage of accuracy is low in tied-State tri-syllable acoustic model for both speeches and also as shown in experiment II table 5.3 and table 5.4 the percentage of accuracy is low in tied-State acoustic model. The percentage of accuracy monophone acoustic model is low in both experiments. These indicate that different acoustic models produce slightly different forced alignment results. The better the acoustic model gives the more accurate the forced alignments. Therefore the researcher believes that monosyllable acoustic model is the better acoustic model to get accurate forced alignment. The overall performance of the system is affected by forced alignment.

The accuracy of tied-State acoustic model in experiment I table 4.1 and table 4.2 is lower than precision and recall. Accuracy refers to how close a measured value is to the actual (true) value.

Precision refers to how close the measured values are to each other. Precision gives information about how many of the detected boundaries are true, the measurement recall tells us how many true boundaries have been found. With high precision and low accuracy, each value will be off by a similar amount. With high accuracy and low precision, each value is closer to the true or expected value, but repeatability suffers. And also the percentage of accuracy in experiment I table 4.2 and in experiment II table 4.4 is low for all monophones, monosyllables and tied-State acoustic models in spontaneous speech. Spontaneous speech contains more noise and disfluencies. The researcher believes that the performance is lower because of this reason. Overall, the best results for all our test sets were achieved by experiment II statistical method, adaBoost. This indicates that machine learning method, are more robust than rule based method.

## 4.5. Implementation of automatic speech segmentation system of approach two

In this phase automatic sentence segmentation is done based on prosodic features. These features are extracted directly from the speech signal. Three stages are involved in this phase. These are audio segmentation, features extraction and sentence boundary detection.

### 4.5.1. Audio segmentation

In this stage the input audio is segmented into 20ms-long non overlapping frames. We used 30 minute Amharic read-aloud speech and 35 minute spontaneous speech. Both speeches are divided into 20 milliseconds (0.02 sec). From 30-minutes audio data we have got 143,384 segments of 0.02-seconds of read-aloud speech and from 35-minutes audio data we have got 196,834 segments of 0.02-seconds of spontaneous speech. These non-overlapping smaller frames are used in feature extraction for classification of speech or non-speech segments.

### 4.5.2. Feature Extraction

In this stage we extracted feature from small frames speech to classify speech/non-speech segments and to detect sentence boundaries. In our experiment there are two stages of feature extraction. The first extracted features are used for classification of speech or non-speech segments and the second features are used to detect sentence boundaries

In the first stage, two features are extracted from each frame individually. These are fundamental frequency (F0) and energy shown in figure 4.3.

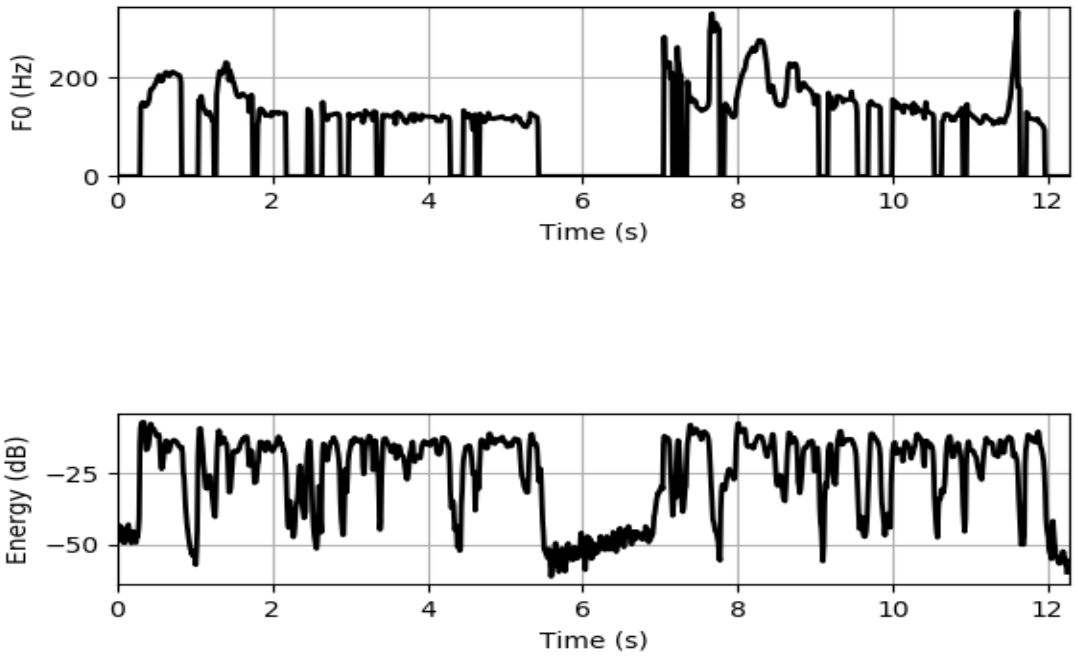



Figure 4.3 Amharic speech Sample of F0 and energy

We selected set of threshold to discriminate between speech and non-speech segments based on energy feature and frames that have very low value of F0 are categorized as non-speech segments and frames with high F0 are categorized as speech segment. The purpose of speech/non-speech classification is to categorize the speech audio dataset into speech and non-speech segments. The non-speech segments are used to detect sentence boundary and the speech segments are observed as non-boundaries

In the second stage, prosodic features are extracted from the speech boundary candidate. In our experiment we only consider non-speech segments for sentence boundary detection. These features are extracted and combine to represent the context of sentence boundary candidate. Prosodic features extracted from boundary candidate are shown in figure 4.4.

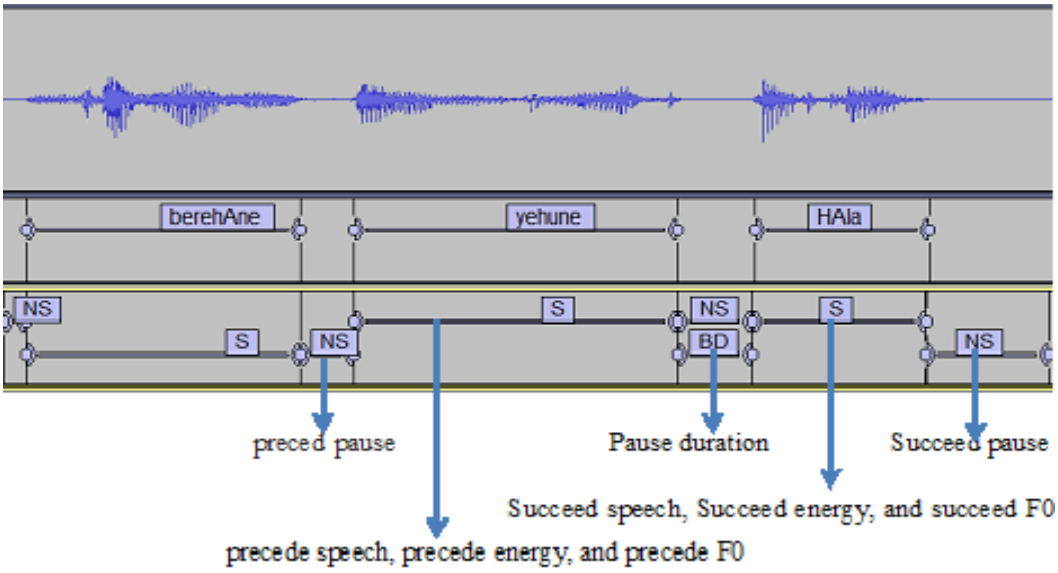



Figure 4.4 Extracted features from sentence boundary candidate

As shown in the above figure seven features are extracted to represent characteristics of the sentence boundary candidate,

1. Succeed and Precede speech: duration of the speech (non-pause) proceeding and succeeding boundary candidate.
2. Succeed and Precede pause: duration of the pause (non-speech) proceeding and succeeding boundary candidate.
3. Fundamental frequency: difference between preceding and succeeding fundamental frequency.
4. Energy: difference between preceding and succeeding energy.

### 4.5.3. Sentence boundary detection

In this stage, for each sentence candidate, features mentioned above are extracted. Sample features vectors extracted from boundary candidate are shown in figure 4.5.

Then, a statistical method AdaBoost, is applied on all sentence boundary candidates. It iteratively corrects the mistakes of the weak classifier and improves accuracy by combining weak learners[11]. AdaBoost uses many base classifiers. It uses decision tree classifier as default classifier[11]. In our experiment we have used decision tree classifier and support vector classifier (SVC) as a base estimator.

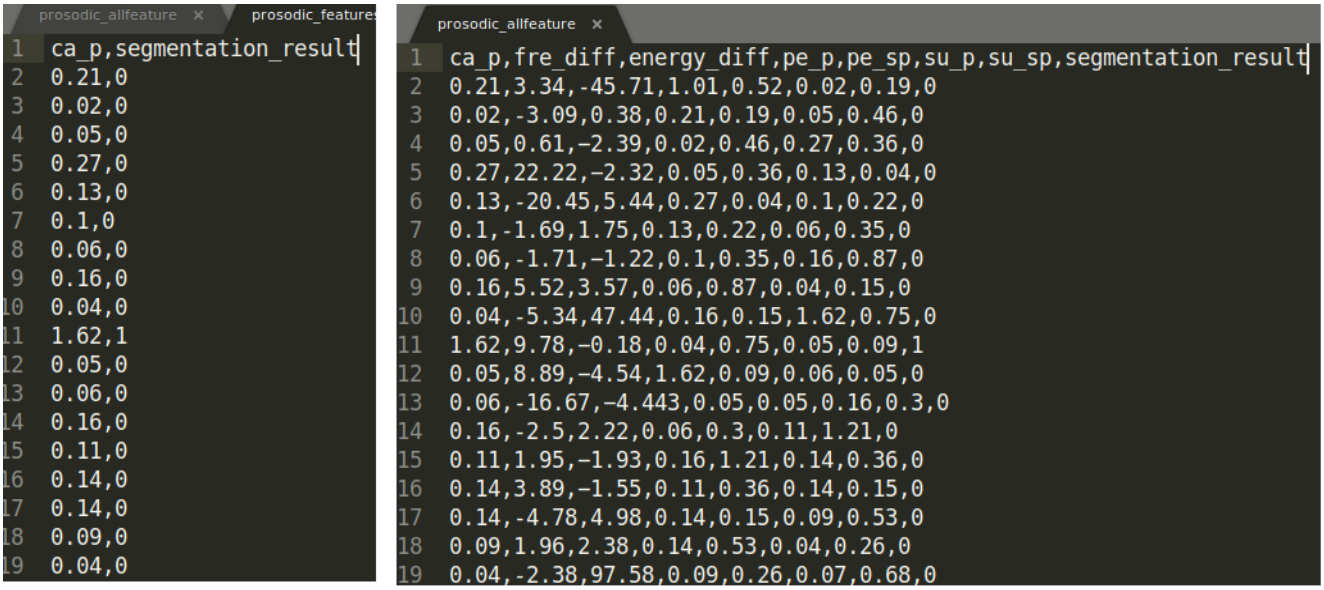



Figure 4.5. Features vectors

First we implemented a baseline system which uses only pause features. Then we implemented a system by using all extracted features.

## 4.6. Experimental result

We evaluate our methods using two corpora .We have also used pause features only and all seven features to detect sentence boundary.

**Experiment I: Experimental result based on pause features**

The overall results of the automatic segmentation with two test set are shown in Table 4.5 and Table 4.6.

| | Accuracy(A) | Precision (P) | Recall (R) |
|---|---|---|---|
| Decision tree classifier | 88.8% | 75.64% | 78.64% |
| SVM classifier | 85.47% | 73.82% | 74.64% |

Table 4.5 Amharic read-aloud speech experimental results with pause feature

| | Accuracy(A) | Precision (P) | Recall (R) |
|---|---|---|---|
| Decision tree classifier | 78.9% | 68.8% | 62.8% |
| SVM classifier | 75.9% | 65.8% | 60.8% |

Table 4.6 Amharic spontaneous speech experimental results with pause feature

**Experiment II: Experimental result based on all prosodic features**

| | | | |
|---|---|---|---|
| Decision tree classifier | 93.67% | 75.64% | 76.32% |
| SVM classifier | 85.47% | 88.8% | 82% |

Table 4.7 Amharic read-aloud speech experimental results with all features

| | Accuracy(A) | Precision (P) | Recall (R) |
|---|---|---|---|
| Decision tree classifier | 84.3% | 70.64% | 63% |
| SVM classifier | 80.7% | 62.8% | 60% |

Table 4.8Amharic spontaneous speech experimental results with all features

**Experimental analysis**

As shown in experiment I table 4.5 and table 4.6 pause features can perform well. When prosodic features are introduced, the performance increases, as the table 4.7 and table 4.8 shows. It indicates that pause feature is a basic discriminator for Amharic sentence boundary. And also the context of a sentence boundary candidate has significant contribution to detect true sentence boundary. Both experimental results showed that decision tree classifier is better than the algorithm support vector classifier. The resulting computing time by decision tree faster than support vector classifier.

The performance in experiment I table 4.6 and in experiment II table 4.8 is lower in spontaneous speech. The main reason why the performance of spontaneous speech is lower than read-aloud speech is that the pre-boundary lengthening is not that regular and expressive in spontaneous speaking as well as its relation with the strength of a boundary may differ. And also pauses are more frequent in spontaneous speaking and they also occur at locations where they do not have

syntactic or semantic motivation. Therefore the researcher believes that the performance is lower because of this reason.

Foreign researches done on sentence level speech segmentation did not use forced alignment method for sentence segmentation, which make the performance of speech segmenter better for segmenting and labeling of Amharic speech data. The performance of decision tree classifier for Amharic speech is better than rule based. It indicates that learn-ability without being explicitly programmed is a good method for classification. The research works done on sentence level speech segmentation of spontaneous Malay language[31] achieves 74% boundaries are discriminated correctly using rule based method. Acoustic and prosodic features were used for identifying sentence boundary. Other work voice segmentation without voice recognition have used only pause feature and 79.8% boundaries are discriminated correctly. And other features (pause, ROS and prosodic features) are introduced 82.3% boundaries are discriminated correctly using adaboost classifier. In is a research work we achieved better (93.67% and 84.3% accuracy) result using decision tree classifier for read aloud and spontaneous speech respectively.

# CHAPTER FIVE
# CONCLUSION AND RECOMMENDATION

This section discusses conclusions drawn from the findings of the experiment and possible future works to improve the work carried out by this research.

## 5.1. Conclusion

The primary objective of this research is speech segmentation at sentence level and that is achieved by identifying the boundaries between sentences in a continuous speech signal.

The work presented in the thesis can be divided into two major approaches based on the type of work. In the first approach, we used an automatic tool for segmenting and labeling of Amharic speech data at sentence level. In the second approach we used prosodic features directly from the speech to detect sentence boundary. We have evaluated automatic sentence segmentation of continuous speech using two corpora read-aloud and spontaneous speech. The automatic speech segmentation system is evaluated with manual segmentation results by comparing automatically segmented sentence to manually labeled sentences.

The evaluation of the experiments shows that speech segmentation based on monosyllable acoustic model perform higher than both from tied-State and monophone acoustic model. The results in the first approach show that adaboost classifier achieves higher accuracy than rule based method. As shown in the experiment the adaboost classifier showed consistently good results especially in decision tree classifier. It produced best results in the majority of our tests. Therefore, the answer for what model is optimal for automatic speech segmentation? , decision tree classifier is optimal classification method for Amharic speech segmentation.

The evaluations in approach two of the experiments shows that pause feature alone perform good accuracy. And also when other prosodic features are introduced, the performance increases a lot. Based on the findings of our experiment, for segmenting and labeling of Amharic speech data at sentence level, monosyllable acoustic model is the better model to get accurate forced alignment with regard to its sentences segmentation accuracy. And also the answer for what speech features or their combinations would result in optimal speech segmentation? , pause feature alone is an important indicator of sentence boundaries. When it is combined with other prosodic feature,

prosody provides complementary information to segmenter for the detection of sentence boundaries, and it can therefore play an important role in the automatic segmentation of Amharic spoken language.

## 5.2. Future work

The work presented in this thesis suggests extensions and future research directions. The following future works are made based on the findings and limitations of the current work:

- We have developed only 4000 sentence level speech corpus with their corresponding text from specific domain areas. A good corpus contains speech samples from significantly large number of speakers from different demographic and socio-linguistic groups. Thus, it is strongly recommended that the development of a reliable and sufficiently large corpus and make them available for researchers is good contribution.
- This thesis has only focused on automatic sentence boundary detection. It is also required to conduct a research on automatic speech segmentation of other discrete units (syllable, phoneme, and word level).
- This thesis has only focused on acoustic level prosodic features are based on fundamental frequency, amplitude and duration. It is also required to conduct a research on other prosodic features (perceptual and linguistic level).
- The automatic detection of disfluencies and filler words is important for spontaneous speech where these events are frequent.

# References


[1] M. Ostendorf *et al.*, "Speech segmentation and its impact on spoken document processing," *Signal Process. Mag.*, vol. 25, no. 3, pp. 59–69, 2008.

[2] M. Kalamani, S. Valarmathy, S. Anitha, and R. Mohan, "Review of Speech Segmentation Algorithms for Speech Recognition," vol. 3, no. 11, pp. 1572–1574, 2014.

[3] F. Pausch, "Automatic Segmentation of Speech into Sentences Using Prosodic Features," *Iem.Kug.Ac.At*, no. November, pp. 985–994, 2011.

[4] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tur, "Prosody-based automatic segmentation of speech into sentence and topics," *Speech Commun.*, vol. 32, no. 32, pp. 127–154, 2000.

[5] D. Jones *et al.*, "Measuring the Readability of Automatic Speech-to-Text Massachusetts Institute of Technology Information Systems Technology Group 2," pp. 1585–1588, 2003.

[6] J. Kolář, "Automatic segmentation of speech into sentence-like units," PHD *Dissertation*, Faculty of Applied Sciences, West Bohemia University , Pilsen 2008.

[7] N. Jamil, M. I. Ramli, Z. A. Bakar, and N. Seman, "Prosody-based sentence boundary detection of spontaneous speech," *Proc. - Int. Conf. Intell. Syst. Model. Simulation, ISMS*, no. July 2017, pp. 311–317, 2015.

[8] J. Vaissière, J. Vaissière, and P. Models, "Language-independent prosodic features To cite this version : HAL Id : halshs-00703571," pp. 53–65, 2012.

[9] S. C. Pammi and V. Keri, "HTKTrain : A Package for Automatic Segmentation."

[10] X. Li, L. Wang, and E. Sung, "AdaBoost with SVM-based Component Classifiers," pp. 1–26.

[11] K. Arras, C. Stachniss, M. Bennewitz, and W. Burgard, "Robotics 2 AdaBoost for People and Place Detection  Machine Learning : A Survey."

[12] M. T. Uliniansyah *et al.*, "A Tool to Solve Sentence Segmentation Problem on Preparing Speech Database for Indonesian Text-to-speech System," *Procedia Comput. Sci.*, vol. 81, no. May, pp. 188–193, 2016.

[13] R. Dufour, R. Dufour, V. Jousse, and Y. Est, "Spontaneous Speech Characterization and Detection in Large Audio Database," no. January, 2009.

[14] G. Girmay, "Prosodic Modeling for Amharic," M.S thesis, ELECTRICAL and COMPUTER ENG DEPARTMENT,ADDIS ABABA UNIVERSITY, ADDIS ABABA, 2008.

[15] W. Bohemia, P. Faculty, A. Sciences, C. M. Advisor, and J. Psutka, "AUTOMATIC SEGMENTATION OF SPEECH INTO SENTENCE-LIKE UNITS Ing . Jáchym Kolář," 2008.

[16] E. Shriberg, A. Stolcke, and D. Hakkani-t, "Prosody-based automatic segmentation of speech into sentences and topics," vol. 32, 2000.

[17] J. Kolář and Y. Liu, "AUTOMATIC SENTENCE BOUNDARY DETECTION IN CONVERSATIONAL SPEECH : A CROSS-LINGUAL EVALUATION ON ENGLISH AND CZECH J ´ achym Kola r ˇ Faculty of Applied Sciences , Dept . of Cybernetics , Univ . of West Bohemia in Pilsen , Czech Republic Department of Co," pp. 1–4.

[18] E. D. Emiru and D. Markos, "Automatic Speech Segmentation for Amharic Phonemes Using Hidden Markov Model Toolkit ( HTK )," vol. 4, no. 4, pp. 1–7, 2016.

[19] P. Blunsom, "Hidden Markov Models," pp. 1–7, 2004.

[20] S. Young, M. Gales, X. A. Liu, P. Woodland, T. Htk, and H. T. K. Version, *HTK Book*, no. December 2001. 2009.

[21] P. Harrington, *Machine Learning in Action*. 2012.

[22] S. T. Abate and W. Menzel, "Syllable-Based Speech Recognition for Amharic," *Work. Comput. Approaches to Semit. Lang. Common Issues Resour.*, no. June, pp. 33–40, 2007.

[23] E. Studies, "Institute of Ethiopian Studies Review Reviewed Work ( s ): Yamariñña Säwasäw by Bye Yimam and Baye Yimam Review by : Getahun Amare Source : Journal of Ethiopian Studies , Vol . 28 , No . 2 ( December 1995 ), pp . 55-60 Published by : Institute of Ethiopian Studies Stable URL : http://www.jstor.org/stable/41966049," vol. 28, no. 2, pp. 55–60, 2018.

[24] N. T. Mergia, "SCHOOL OF GRADUATE STUDIES FORMANT BASED SPEECH SYNTHESIS FOR Addis Ababa University School of Graduate Studies Faculty of Informatics Department of Computer Science," 2008.

[25] O. Jokisch and O. Jokisch, "Syllable-based prosodic analysis of Amharic read speech," no. December 2012, 2018.

[26] M. Mulgi, P. V. Mantri, and M. G. M, "Voice Segmentation without Voice Recognition," vol. 2, no. 1, pp. 2–6, 2013.

[27] D. Wang, L. Lu, and H.-J. Zhang, "Speech segmentation without speech recognition," *IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP '03).*, vol. 1, pp. 468–471, 2003.

[28] K. Vicsi and G. Szaszák, "Automatic Segmentation of Continuous Speech on Word Level Based on Supra-segmental Features," *Int. J. Speech Technol.*, vol. 8, no. 1, pp. 363–370, 2005.

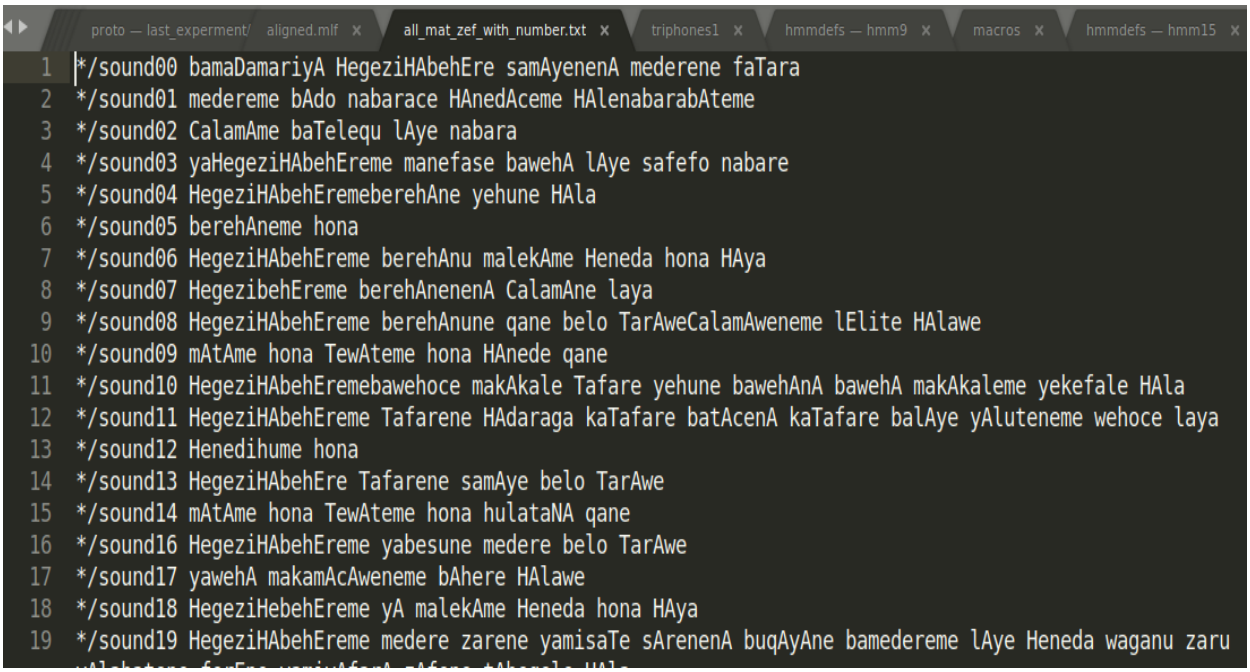
[29] S. Kr, D. Mandal, and B. Gupta, "Word boundary detection based on suprasegmental features : A case study on Bangla speech," pp. 17–28, 2007.

[30] J. Kolář, "Automatic segmentation of speech into sentence-like units," PHD *Dissertation*, Faculty of Applied Sciences, West Bohemia University , Pilsen 2008.

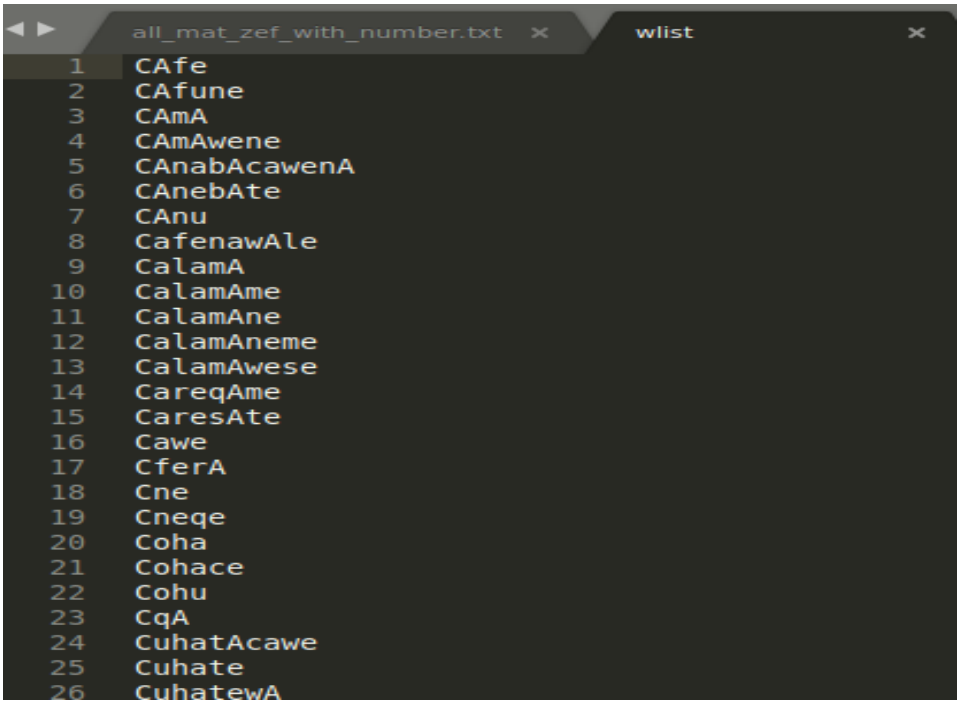
[31] N. Jamil, M. I. Ramli, Z. A. Bakar, and N. Seman, "Prosody-based sentence boundary detection of spontaneous speech," *Proc. - Int. Conf. Intell. Syst. Model. Simulation, ISMS*, no. July 2017, pp. 311–317, 2015.

# Appendixes

Appendix A: Texts file with audio name




Appendix B: wilst file



Appendix C: pronunciation dictionary based on syllable

aligned.mlf | all_mat_zef_with_number.txt | triphones1 | dict | hmmd

```
CAfe            [CAfe]          CA fe sp
CAfune          [CAfune]        CA fu ne sp
CAmA            [CAmA]          CA mA sp
CAmAwene        [CAmAwene]      CA mA we ne sp
CAnabAcawenA    [CAnabAcawenA]  CA na bA ca we nA sp
CAnebAte        [CAnebAte]      CA ne bA te sp
CAnu            [CAnu]          CA nu sp
CafenawAle      [CafenawAle]]   Ca fe na wA le sp
CalamA          [CalamA]        Ca la mA sp
CalamAme        [CalamAme]      Ca la mA me sp
CalamAne        [CalamAne]      Ca la mA ne sp
CalamAneme      [CalamAneme]    Ca la mA ne me sp
CalamAwese      [CalamAwese]    Ca la mA we se sp
CareqAme        [CareqAme]      Ca re qA me sp
CaresAte        [CaresAte]      Ca re sA te sp
Cawe            [Cawe]          Ca we sp
CferA           [CferA]         C fe rA sp
Cne             [Cne]           C ne sp
Cneqe           [Cneqe]         C ne qe sp
Coba            [Coba]          Co ba sp
```

Appendix D: word level transcription

all_mat_zef_with_number.txt

```
#!MLF!#
"*/sound00.lab"
bamaDamariyA
HegeziHAbehEre
samAyenenA
mederene
faTara
.
"*/sound01.lab"
medereme
bAdo
nabarace
HAnedAceme
HAlenabarabAteme
.
"*/sound02.lab"
CalamAme
baTelequ
lAye
nabara
.
"*/sound03.lab"
yaHegeziHAbehEreme
manefase
bawehA
lAye
safefo
nabare
```

Appendix E: prototype model

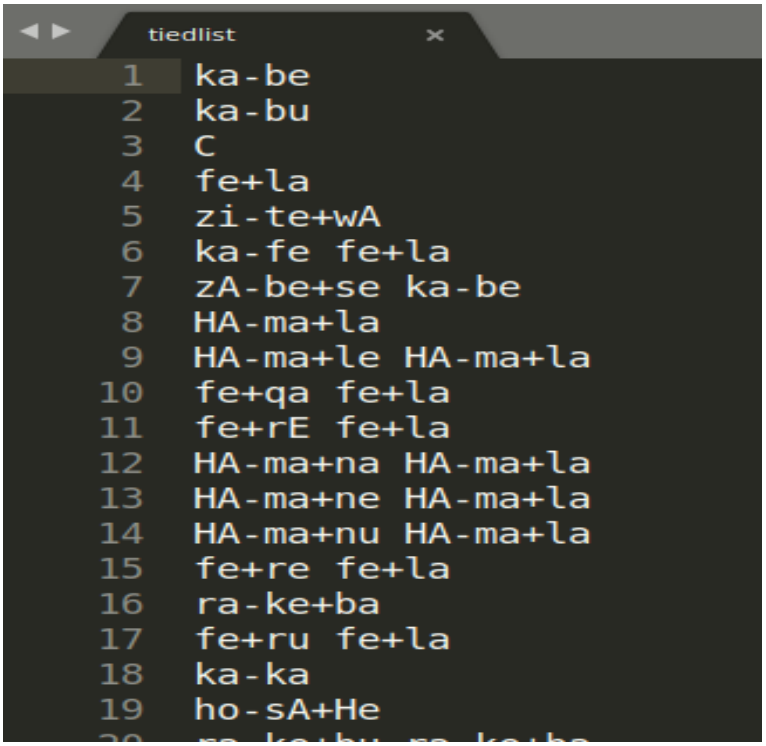
```
~o <VecSize> 25 <MFCC_0_D_N_Z>
~h "proto"
<BeginHMM>
  <NumStates> 5
  <State> 2
    <Mean> 25
      0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
    <Variance> 25
      1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
  <State> 3
    <Mean> 25
      0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
    <Variance> 25
      1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
  <State> 4
    <Mean> 25
      0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
    <Variance> 25
      1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
  <TransP> 5
   0.0 1.0 0.0 0.0 0.0
   0.0 0.6 0.4 0.0 0.0
   0.0 0.0 0.6 0.4 0.0
```

Appendix F: forced aligned file

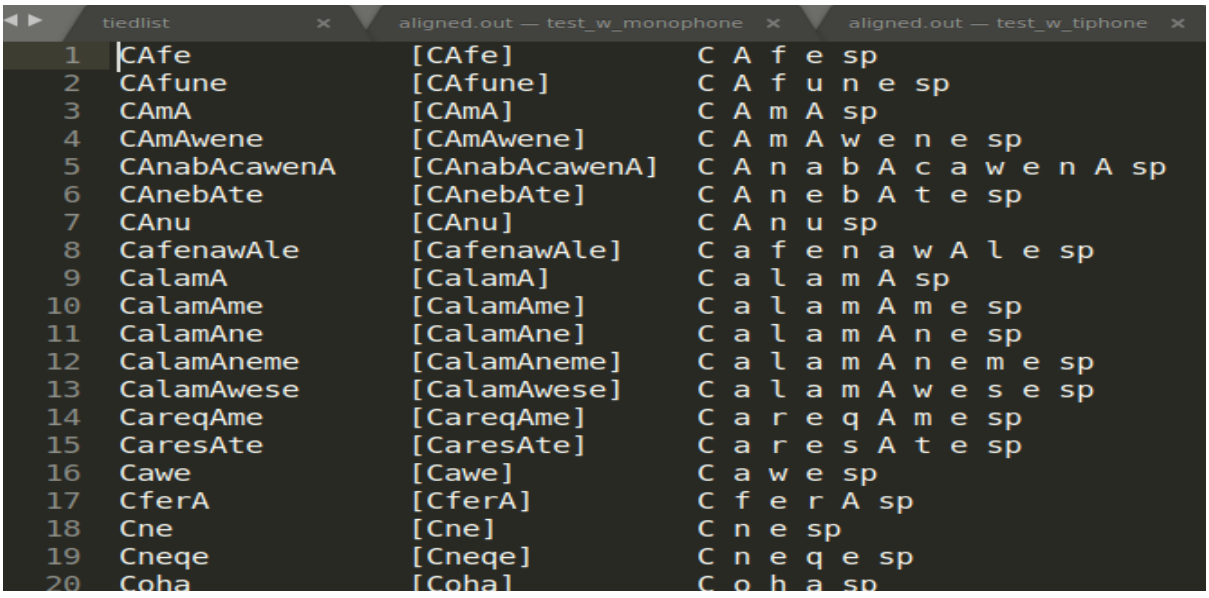
macros | hmmdefs — hmm15 | aligned.out — test_w_monophone

```
#!MLF!#
"*/tst2.rec"
0 9900000 sil -5765.594727 SENT-END
9900000 11100000 He -889.883667 HebAbeme
11100000 12900000 bA -1203.343872
12900000 13500000 be -445.061188
13500000 17300000 me -2863.603271
17300000 17300000 sp -0.310092
17300000 17300000 sp -0.310092
17300000 18500000 He -959.823303 HegeziHAbehEre
18500000 18900000 ge -288.399811
18900000 20900000 zi -1513.900391
20900000 21600000 HA -530.550232
21600000 21900000 be -240.588211
21900000 24200000 hE -1705.331665
24200000 25100000 re -641.611816
25100000 25100000 sp -0.310092
25100000 25100000 sp -0.310092
25100000 26400000 HA -947.264038 HAmelAke
26400000 26900000 me -338.530823
```

Appendix G: tidelist syllable




Appendix H: monophone pronunciation dictionaries



AppendixI:

codetrain.scp

```
./audio/all_wav/sound00.wav ./audio/mfcc/sound00.mfc
./audio/all_wav/sound01.wav ./audio/mfcc/sound01.mfc
./audio/all_wav/sound02.wav ./audio/mfcc/sound02.mfc
./audio/all_wav/sound03.wav ./audio/mfcc/sound03.mfc
./audio/all_wav/sound04.wav ./audio/mfcc/sound04.mfc
./audio/all_wav/sound05.wav ./audio/mfcc/sound05.mfc
./audio/all_wav/sound06.wav ./audio/mfcc/sound06.mfc
./audio/all_wav/sound07.wav ./audio/mfcc/sound07.mfc
./audio/all_wav/sound08.wav ./audio/mfcc/sound08.mfc
./audio/all_wav/sound09.wav ./audio/mfcc/sound09.mfc
./audio/all_wav/sound10.wav ./audio/mfcc/sound10.mfc
./audio/all_wav/sound11.wav ./audio/mfcc/sound11.mfc
./audio/all_wav/sound12.wav ./audio/mfcc/sound12.mfc
./audio/all_wav/sound13.wav ./audio/mfcc/sound13.mfc
./audio/all_wav/sound14.wav ./audio/mfcc/sound14.mfc
./audio/all_wav/sound15.wav ./audio/mfcc/sound15.mfc
./audio/all_wav/sound16.wav ./audio/mfcc/sound16.mfc
./audio/all_wav/sound17.wav ./audio/mfcc/sound17.mfc
./audio/all_wav/sound18.wav ./audio/mfcc/sound18.mfc
./audio/all_wav/sound19.wav ./audio/mfcc/sound19.mfc
./audio/all_wav/sound20.wav ./audio/mfcc/sound20.mfc
```