RESEARCH ARTICLE OPEN ACCESS

# Automatic Speech Segmentation for Amharic Phonemes Using Hidden Markov Model Toolkit (HTK)

Eshete Derb Emiru [1], Walelign Tewabe Sewunetie [2]
Department of Information Technology
Debre Markos University
Debre Markos
Ethiopia

## ABSTRACT

Speech segmentation is the process of identifying the boundaries between meaningful units like phonemes in a continuous speech. A need exists for reliable, automatic determination of phonemes boundaries in different speech research areas such as automatic speech recognition (ASR) to improve the performance of the recognizer, to improve the quality speech synthesis system through segmented database, and to improve performance of language identification and speaker verification system. 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 in order to have parallel speech corpus. Both text and speech corpuses are split in to training (90%) and test (10%) data sets. Phoneme based speaker dependent Hidden Markov Model is preferred, being the one that is most widely used, and also due to the ready availability of the HTK software suite. 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. Letter and phoneme were used as a basic unit to model the HMM in context independent, context dependent with single Gaussian mixture and context dependent with Multiple Gaussian mixtures. The system is also evaluated in terms of percentage of boundary deviations with 5ms, 10ms, 15ms and 20ms tolerance values with reference to manual segmentation results. The evaluation of the experiments shows that best performance with Minimum percentage of time boundary deviations are achieved using phoneme based approach in context dependent environment with two Gaussian mixture.

***Keywords***:- Phonemes, unsupervised method, Automatic speech segmentation, Hidden Markov Model(HMM), Hidden Markov Model ToolKit(HTK), Mel Frequency Cepstral Coefficient (MFCC).

## I. INTRODUCTION

Segmentation of continuous speech into its corresponding phonemes is a very important issue in the area of speech like ASR, speech synthesis, speech database, and language identification and speaker verification. The most commonly proposed phoneme level speech segmentations are using either manual segmentation or automatic segmentation techniques. In manual speech segmentation expert/phonetician is required and its segmentation is based on listening and visual judgment on required boundaries. However, manual phonetic segmentation is tedious, expensive, inconsistent, prone to errors and time-consuming task[1]. There is also a disagreement between phoneticians and there is no clear, common and coherent strategies in order to segment speech waveforms [2]. Considering these and other disadvantages the development of automatic speech data is becoming increasingly important[3, 4].

Generally, automatic speech segmentation methods are divided into two types, namely supervised and unsupervised segmentation methods. Supervised methods require a priori knowledge about phoneme boundaries [5-7]. These boundaries of phonemes are existed in the form of their pre-segmented. It also requires pre-defined models of phoneme set of a specific language[8].

On the other side, unsupervised methods don’t require pre-defined model and knowledge about phoneme sets and their boundaries respectively. It is most commonly used in automatic speech segmentation through new modeling and training data sets[8]. Thus unsupervised method yields a desirable and more flexible framework for automatic segmentation of a speech at phoneme level[9].

Hidden Markov Model is the most commonly used model for automatic speech segmentation in unsupervised method[10-14]. This model able to handle new data robustly with in different working environments and enables to predict similar patterns efficiently[15]. It is also language independent and computationally efficient to develop and evaluate due to the existence of established algorithms[16-18]. Generally all the above benefits of both HMM model and supervised methods are the driving forces to conduct a research on supervised method of automatic speech segmentation using HMM model at phoneme level.

## II. DATA COLLECTION AND CORPUS PREPARATION

Optimal text selection technique is used to prepare Amharic text corpus from various Amharic document. These Amharic documents are used as data sources to get phonetically rich and balanced collections of sentences. Amharic Bibles, Health News, political News, sport News, Economy News, penal code, federal Negarit Gazeta and Amharic fictions named as “Fiker eskemekaber” are data sources used for text corpus preparation. This text corpus contains 1000 Amharic sentences.

Since parallel speech corpuses are required to continue the speech segmentation process, speech corpus is prepared for the corresponding text corpus. Speech corpus is indispensable part of speech segmentation because speech segmentation results vary as per speaker’s accent, age, gender and other invariants. By taking these invariants in to consideration, all sentences found in text corpus are recorded by in mono channel, *.wav format and 48 kHz sample frequency. It implies that speech corpus is prepared and this speech corpus is accessible in computer readable form, and have an annotation and documentation sufficient to allow re-use of data.

## III. AMHARIC LANGUAGE AND ITS PHONETICS

Amharic is a phonetic language and national language of Ethiopia. It has large number of speakers and large amount of published and unpublished documents in order to get phonetically balanced data sets. Humans can produce an infinite number of sounds; each language has a set of abstract linguistic units, called phonemes, to describe its sounds. A phoneme is defined as the smallest contrastive unit in the phonology of a language[19]. Amharic language is primarily comprised of 39 phonemes – 7 vowels and 31 consonants[20]. One additional consonant /ቭ/ [v] is inherited and included summing up to a total of 39 phonemes.

## IV. DESIGN OF AMHARIC AUTOMATIC SPEECH SEGMENTATION

### A. Design approaches

Two main approaches are proposed for design of automatic speech segmentation system and these approaches mainly differ in basic units of the pronunciation dictionary. These approaches are grapheme based and phoneme based where letters and phonemes are the basic units of pronunciation dictionary respectively.

Approach I, Grapheme based approach which considers the transliteration of every word in to Latin alphabet during its pronunciation dictionary preparation. It contains sequences of Latin alphabets or letters as pronunciation dictionary. This approach is bench mark for our system and it is the approach followed in many Amharic speech recognition systems.

Approach II, Phoneme based approach which integrates Grapheme to Phoneme (G2p) conversion developed for this purpose. This G2P conversion is achieved by applying epithetic vowel insertion rules in order to develop epithetic vowel insertion algorithm. The epithetic Amharic vowel is which is found in the speech utterances. This epithetic vowel is not found from directly translated words of the Latin representation but exists from acoustic signal of a speech. By considering this epithetic vowel during pronunciation dictionary preparation, approach II is believed by the researchers to enhance the performance compared to the previous approach. The G2P converter is adopted in[21].

### B. Automatic speech segmentation system

The general model of automatic speech segmentation system includes data preparation, manual labeling, language modeling, HMM model building, HMM segmenter and data verification sub-systems. The tasks of these sub-systems are accomplished in the order of what they have listed. Especially manual labeling is carried out after preprocessing section also known as data preparation sub-system. It implies that manually labeled phonemes with their time boundaries are found before completion of automatic speech segmentation in order to be free from biased during manual labeling. All sub- systems are included in both Grapheme and phoneme based approaches which can be implemented individually as phase1 and phase2 respectively during experimentation.

**Data preparation**

Data preparation is the main part of automatic speech segmentation system since it has high contribution to the performance of automatic speech segmenter. It includes core processes like corpus preparation which encompasses both text and speech corpuses, Lexicons preparation as pronunciation dictionary, sampling techniques used to split both text and speech corpuses in to test and training sets and feature extraction process to prepare speech corpus usable format to HMM acoustic modeling and HMM segmenter as input. The feature vector that represents the distinctive properties of the phoneme is designed to be of length 39, consisting of 12 mel-cepstrum coefficients and energy component, and additionally their delta and acceleration coefficients.

#### *1) Language Modeling*

Language Model is responsible for detecting connections between letters in a word and words in a sentence during Grapheme based approaches and for detecting connections between phonemes in a word and words in a sentence during phoneme based approach. Bigram and trigram language models can be used to handle multiple pronunciations of a word but in our text corpus the possibility of getting such kinds of words less. So unigram language modeling is enough to build the new automatic speech segmentation system.

#### *2) HMM Acoustic Modeling*

The acoustic models are statistical models which commonly uses the predominant HMM models. The Hidden Markov Models can be used to represent the sequence of sounds within a section of speech units like phonemes. Phoneme, an elemental speech sound, can be modeled by an individual left to right HMM. Phoneme based HMM model is used with three emitting states and two non-emitting states of the first and the last states without skipping. In acoustic modeling, individual phoneme based HMM model also takes place in context independent and context dependent with single Gaussian mixture and context dependent with multiple Gaussian mixtures environments.

#### *3) HMM Segmenter*

After getting Language modelling and HMM acoustic modelling resuts, the HMM segmenter assigns the corresponding letter or phoneme to acoustic signal as per the training and best selected pronunciation of a word. At the end, phonemes with their time boundaries are obtained as an output of automatic HMM based speech segmenter.

## V. EXPERIMENTAL RESULTS AND EVALUATION

### A. Automatic phoneme segmentation

Automatic speech segmentation using HTK toolkit is implemented in two phase with HTK toolkit. These phases differ in basic units of lexicon preparation which is used as pronunciation dictionary. The basic units are sequence of letters and phonemes in phase1 and phase2 respectively. In each phase, HMM modeling techniques and HTK commands are used to complete the task of automatic speech segmentation. Data preparation, HMM modeling and segmentation are the main stages of it.

Data preparation includes Amharic text corpus preparation, lexicon preparation, speech corpus preparation to the corresponding text corpus, data transcription to HTK usable format and parameterization of speech signals. In corpus based speech segmentation preparing training and testing data sets is required and for this purpose systematic random sampling technique is to split both text and speech corpuses in to training data sets (90%) and testing data sets (10%). Training data sets are used for language and acoustic modeling purpose where as testing data sets used for evaluation of automatic HMM segmenter. Since HTK doesn’t use speech data directly transcribing them into phone level and word level, and parameterization of them also required as part of data preparation. Parameterization of speech data takes place through feature extraction process.

HMM model with three emitting states and two non emitting states without skipping is used to model individual Amharic phonemes as shown in fig 1.

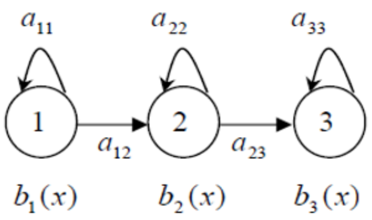

Fig. 1 Representation of Left-to-right HMM

Where: 1, 2 and 3 are states to represent a phoneme.

$a_{ij}$ is the transition probability from state i to state j and represented by the label on the edge from state i to state j

$a_{11}$, $a_{22}$ and $a_{33}$ are self transition probabilities.

$b_i(x)$ is probability of observation x on state i.

MFCC feature vectors together with their first and second, namely MFCCs + delta + delta-deltas are selected for individual HMM models. The delta and delta-delta coefficients are included to make the model sensitive to the dynamic behaviour of the signal[1]. Acoustic HMM modeling and HMM segmentation are taking place in context independent and in context dependent with single Gaussian mixture and context dependent with multiple Gaussian mixtures environments. Acoustic modeling with Multiple Gaussian mixtures values 2, 4, 8 and 16 are used to improve automatic speech segmentation results considerably, because they help avoid the problem resulting from the usage of the same type of probability density distribution for different models and states. At the end, phonemes with their time boundaries are obtained as an output of automatic HMM based speech segmenter.

### B. Manual labeling

Manual labeling is carried out via praat software. Praat shows the main characteristics of acoustic data both in its spectrogram and wave formats. The segmentation is also performed without any a priori information about test results fond during automatic speech segmentation by a single labeler but takes the split test text corpus during manual labeling. In manual segmentation, acoustic signal properties like pitch, formant, duration, intensity, energy, power spectral density and zero crossing rates are very important in order to carry out the segmentation.

Manual labeling also takes place on test data sets in addition to automatic speech segmentation applied on them. Manually segmented phonemes are its results and they are achieved by hand labeling of phonemes with their time boundaries. These manual segmented phonemes are used to evaluate the performance of automatic speech segmentation system.

The deviation of time boundaries of automatic segmented phonemes with reference to hand segmented phonemes since manual segmentation are considered as accurate results[22].

### C. Test results and evaluation

In order to measure the performance of automatic speech segmentation, phoneme mapping concept is used. These manually segmented phoneme boundaries are compared with each phoneme sequences found during automatic speech segmentation where as epithetic vowel is considered as part of letters in case of Grapheme based approach. The general formula used to calculate the deviation of initial time boundaries of a phoneme or a letter and the deviation of final time boundaries of a phoneme or a letter are indicated in (1) and (2) respectively.

$$t_1=|t_i-t'_i| \qquad (1)$$

$$t_2=|t_{i+1}-t'_{i+1}| \qquad (2)$$

where: $t_i$ is initial time boundary of manually segmented letters/phonemes.

$t'_i$ is initial time boundary of automatically segmented letters/phonemes.

$t_1$ is time difference between initial time boundaries of manually and automatically segmented letters/phonemes.

$t_{i+1}$ is final time boundary of manually segmented letters/phonemes.

$t'_{i+1}$ is final time boundary of automatically segmented letters/phonemes.

$t_2$ is time difference between final time boundaries of manually and automatically segmented letters/phoneme

The time differences found in (1) and (2) are deviation between manual segmented and automatically segmented phonemes or letters. The deviation of both initial and finale time boundaries of phonemes below 5ms are not considered as errors[23, 24]. Similarly, in our study boundary deviations are evaluated in four different tolerance values 5ms, 10ms, 15ms and 20ms[25]. It implies that the deviations exceeding these tolerances are considered as errors. Finally, the error is expressed in terms of percentage of deviation and it is calculated in quantitative method using (3).

$$\%\text{deviation} = \frac{\text{No of phonemes exceeding the tolerance value}}{\text{total number of boundaries tested}} \text{x}\,100 \qquad (3)$$

The phoneme boundary values beyond the tolerance values are considered as errors and these errors are expressed in terms of percentage through statistical technique. In this technique the number of phonemes occurred beyond the tolerance value and its percentage is given with reference to the total number of phonemes exist in each phases. The total numbers of phonemes exist in phase1 and phase2 are 5080 and 5404 respectively. Since epithetic vowels are inserted by plug in epithetic vowel insertion algorithm, the numbers of phonemes in phase2 are greater than phase1. These total phoneme sizes do not include short pauses exist in each speech utterances and this is expected to reduces the percentage of boundary deviation in each tolerance values. Having the performance evaluation technique, the test results of three automatic speech segmentation experiments in monophones, tied state and tied state with multiple Gaussian mixtures are presented.

The evaluation is taking place in terms of boundary deviations with in tolerances values of 5ms, 10ms, 15 ms and 20ms. In both phases, the evaluation of the experiments shows that the percentage of boundary deviation minimizes as we go from context independent to context dependent and from context dependent with single Gaussian mixtures to context dependent with multiple Gaussian mixtures due to considerations of phonemes context and different probability density functions per state respectively. Even the experiment conducted in Gaussian mixture values 2, 4, 8, and 16, best result is obtained at Gaussian mixture value 4 in both phases as shown in Table 1.

TABLE 1

Grapheme based and Phoneme based system with tied state at Gaussian Mixture four experimental results.

| Phoneme/ letter time boundaries | Phase1 | | | | Phase2 | | | |
|---|---|---|---|---|---|---|---|---|
| | % of boundary deviation in tolerances | | | | % of boundary deviation in tolerances | | | |
| | 5ms | 10ms | 15ms | 20ms | 5ms | 10ms | 15ms | 20ms |
| Initial (t1) | 5.71 | 1.58 | 1.08 | 0.95 | 3.77 | 0.50 | 0.07 | 0.06 |
| End (t2) | 5.18 | 1.61 | 1.12 | 1.02 | 3.29 | 0.46 | 0.07 | 0.06 |
| Both(t1+t2) | 12.83 | 3.60 | 1.93 | 1.32 | 8.29 | 2.15 | 0.65 | 0.26 |

As shown in Table 1, the percentage deviation is minimized in phase2 for both speakers since the pronunciation dictionary developed in phase2 is more phonetic than phase1. The percentage of deviation presented with initial time boundary, end time boundary, and both of them of phonemes. It is found that the percentage of boundary deviation is more in initial time boundaries of phonemes as compared to final time boundaries of them. On the other hand, the percentage of boundary deviation decreases rapidly when the boundary of deviation from 5ms to 20ms. The percentage of deviation beyond 20ms tolerance values are due to phoneme recognition errors and the result shows that all phoneme are almost within 20ms tolerance values[25].

## VI. CONCLUSION AND FUTURE WORKS

As phoneme based speaker dependent Hidden Markov Model is the most commonly used model for automatic speech segmentation[26], it is applied for our research. The HMM model with three emitting states and two non emitting states without skipping is used to model individual Amharic phonemes. MFCC feature vectors together with their first and second are selected for individual HMM models. HTK toolkit is used to implement the HMM model in two phases. These phases differ in basic units exist in the lexicon which is used as pronunciation dictionary; 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 are 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 taking place 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.

Further research is also required on the precision of phoneme boundaries and their consistency in different phoneme contexts and phonetic transitions between phonetic catagories like transition from nasal to semivowel, semivowel to vowel, semivowel to vowel, nasal to silence,stop to fricative, stop to silence, vowel to vowel and nasal to nasal using HMM model[27, 28]. It is also required we recommend to other researchers to continue the research with non uniform HMM topology for acoustic models since the duration of phonemes is variable [29].Automatic speech segmentation without speaker dependent is very essential. This speaker independent automatic speech segmentation expected to improve the performance of speech segmenter through speaker adaptation techniques[22].

## ACKNOWLEDGMENT

We would like to thank Dr. Sebsbie Hailmariam who is lecturer in Addis Abeba University, for initiating the research idea, giving us critical and constructive technical comments throughout this paper. Our thank goes to our friends Abraham Wobie and Nirayo Haylu for their support during corpus preparation and epithetic vowel insertion algorithm development respectively. Moreover, our heart bottom thanks goes to Mrs. Adina Berie for her endless support and patience during speech recording for experimentation. She has recorded properly all 1000 Amharic sentences collected from various sources.

## REFERENCES

[1] S. S. A. Archana Balyan1, Amita Dev3, "phonetic segmentation based on HMM of Hindi speech," 2010.

[2] D. F. a. M. O. plero Cosi, "A preliminary statistical evaluation of manual and automatic segmentation discrepancies," *Italy*, 2009.

[3] M. d. c. C. Dours, H. Kabr, J.M. pcatte,G.prennou and M.Vigoroux, "A multi-level automatic segmentation system:SAPHO and VERIPHONE," *preceedings of EUROSPEECH 89, Paris,france*, vol. 2, pp. 83-89, 1989.

[4] T. S. a. K. Vale, "Automatic alignment of phonetic labels with continous speech," *preceedings of ICSP-90, kobe,Japan*, vol. 2, pp. 997-1000, 1990.

[5] F. D. Brugnara F., and Omologo M., "Automatic segmentation and labeling of speech based on hidden Markov models," *Speech Commun.* , vol. 12, pp. 357-370, 1993.

[6] A. C. A. Kim Y.-J., "Automatic segmentation combining an HMM-based approach and spectral boundary correction," *in Proceedings of International Conference on Spoken Language Processing, Denver, CO*, pp. 145-148, 2002.

[7] B. L. Pellom, and Hansen, J. H. L., "Automatic segmentation of speech recorded in unknown noisy channel characteristics," *Speech Commun.25*, pp. 97-116, 1998.

[8] Y. C. a. Q. Wang, "A Speaker Based Unsupervised Speech Segmentation Algorithm Used in Conversational Speech," *Springer-Verlag Berlin Heidelberg 2007*, pp. 396–402, 2007.

[9] V. W. a. M. E. Odette Scharenborga, "Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries," pp. 1084–1095, 2009.

[10] T. H. Jianhua, H.U., "Syllable Boundaries based Speech Segmentation in Demi-Syllable Level for Mandarin with HTK," *Department of Computer Science and Technology,Tsinghua University, Beijing, China*, 2002.

[11] T. J. a. H. U. Hain, "Syllable Boundaries based Speech Segmentation in Demi-Syllable Level for Mandarin with HTK," 2002.

[12] E. A. b.-R. Maged Nofal, Hadia El Henawy and Nemat S. Abdel Kader, "Arabic Automatic Segmentation System and its application for Arabic Speech Recognition System," 2004.

[13] K. D. a. T. Laureys. (2010). *A Comparison of Different Approaches to Automatic Speech Segmentation*.Available: http://www.esat.kuleuven.ac.be/~spch.

[14] S. Nefti and O. Boëffard, "Acoustical and topological experiments for an HMM-based speech segmentation system," 2001.

[15] M. R. H. a. B. Nath, " StockMarket Forecasting Using Hidden Markov Model: A New Approach," 2005.

[16] R. B. S. Cox, and P. Jackson, "Techniques for accurate automatic annotation of speech waveforms," *in Proceedings of the International Conference on Spoken Language Processing, Vol V., Sydney, NSW*, pp. 1947-1950, 1998.

[17] J. K. a. A. Black, "CMU ARCTIC databases for speech synthesis," *Tech Rep. CMU-LTI-03-177, CMU Language technologies Institute*, 2003.

[18] C. B. John Kominek, and Alan W Black, "Evaluating and correcting phoneme segmentation for unit selection synthesis," *in Proceedings of Eurospeech, Geneva, Switerzland*, 2003.

[19] D. O'Shaugnessy., "Speech Communication," *Series in Electrical Engineering.Addison-Wessley*, 1987.

[20] Y. Baye, "የአማርኛ ሰዋሰው," *Addis Ababa. ት.መ.ማ.ማ.ድ.*, 1997.

[21] H. Nurayo, " Modeling improved Amharic syllbification algorithm (in press) , " *Addis Ababa University, computer science department*, 2011,.

[22] D. T. Toledano, Gomez, L.A.H. and Grande, L.V., "Automatic phonetic segmentation," *in IEEE Transactions on Speech and Audio Processing*, pp. 617-625, 2003.

[23] T. N. a. K. K. Svenden, "Automatic alignment of phonemic labels in continous speech," *Telenor Research COST 249, Nacy, March 6-7* 1995.

[24] B. P. Sarah Hoffmann, "Fully Automatic Segmentation for Prosodic Speech Corpora," *Interspeech 2010,Speech Processing Group, ETH Z¨urich, Switzerland,* 2010.

[25] M. I. A. a. M. M. A.-G. Mohammed A. Al-Manie, "Arabic speech segmentation: Automatic verses manual method and zero crossing measurements," *Indian Journal of Science and Technology,* vol. Vol. 3 No. 12, Dec 2010.

[26] A. L. Iosif Mporas, Todor Ganchev and Nikos Fakotakis, "Using Hybrid HMM-based Speech Segmentation to Improve Synthetic Speech Quality," 2009.

[27] A. E. a. G. Aversano, "Text Independent Methods for Speech Segmentation " 2005.

[28] G. K. Ladon Baghai-Ravary, John Coleman, "Precision of phonemes Boundaries Derrived using Hidden Markov Models," *INTERSPEECH 2009 BRIGTON,* 2009.

[29] J. C.-B. Kalu U. Ogbureke, "Improving Initial Boundary Estimation for HMM-based Automatic Phonetic Segmentation," *School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland,* 2009.