Enhancing Speech Segmentation with Prosodic Features and Neural Network Models

Anonymous ACL submission

Abstract

This study focuses on developing a sentence-level automatic speech segmentation system for Amharic. Two approaches were explored. The first approach utilized an automatic tool for segmenting and labeling Amharic speech data, creating an acoustic model through HMM modeling. The system's segmentation was refined using forced alignment AdaBoost techniques. In the second approach, prosodic features were extracted directly from the speech waveform, and statistical methods including AdaBoost were employed. Additionally, LSTM and Bi-LSTM models were utilized, achieving impressive accuracies of 94.62% and 95.23%, respectively. These approaches contribute to advancing automatic speech segmentation for Amharic, promising improved accuracy and efficiency.

1 Introduction

Current research in the field of speech technology aims to develop efficient speech systems that can be used for communication between peoples and devices for processing of information. Unfortunately, the ability of a computer to understand speech is still weak. Since human speech is continuously generated, the most difficult aspect of speech that challenges machines is its segmentation. Several speech processing systems require speech segmentation wave form into principal acoustic units (phonemes, syllables, sentence, and paragraph). In the field of speech technology, it is very primary phase. The primary purpose of this segmentation process is to use the outcome for other areas of speech research. In several speech research areas such as speech recognition, speech synthesis, language generation-based system, and language identification and speaker identification system, speech segmentation is an important preprocessing phase.

Therefore, to achieve this objective, well-organized, precise, and simple technique is required. While tending to the segmentation task, it must be considered that speech is not clearly organized as written text, especially spontaneous speech.

Amharic is one of the most spoken Semitic languages in Ethiopia with at least 27 million native speakers [1]. 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. Speech segmentation is a technique for discovering speech signal in different parts [2][3]. 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.

2 Experiment

In this study, we gathered 50,000 Amharic speech sentences from diverse sources like Amharic bibles, broadcast news, conversations, and fiction to build corresponding text corpora. These corpora reflected two distinct speaking styles: read speech and spontaneous speech. The read speech corpus combined existing broadcast news and Amharic bibles, while the spontaneous speech corpus merged broadcast conversations and Amharic fiction.

Acoustic models were developed using HMM modeling. In the second approach, an automatic speech segmentation system classified speech and non-speech segments, focusing on non-speech segments as potential sentence boundaries. Prosodic features such as pause duration, sentence duration, fundamental frequency (F0), and energy were extracted for each sentence candidate to capture contextual information. A statistical method, Ad-

aBoost, was then applied to all sentence boundary candidates using decision tree and support vector classifiers (SVC). Additionally, LSTM and Bi-LSTM neural network architectures were employed in our experiment for comparison and evaluation.

3 Result

The study evaluated a model for Amharic sentence boundary detection using various approaches and metrics. Initially, forced alignment was implemented with monosyllable, tied-State tri-syllable, and monophone acoustic models. The data sets were randomly split into 10% test and 90% training sets to assess model performance.

The results of the evaluation revealed that the decision tree classifier achieved impressive accuracy rates. Specifically, it obtained 91.3% accuracy for read aloud speech and 85% accuracy for spontaneous speech when applied to the monosyllable model. These results outperformed the tied-State and monophone models, showcasing the effectiveness of the decision tree classifier in this context.

In a different approach, a baseline system utilizing pause features initially achieved correct sentence boundary discrimination rates of 88.8% and 78.9%. As additional prosodic features were incorporated into the model, the accuracy significantly improved to 93.67% for read aloud speech and 84.3% for spontaneous speech. This enhancement highlighted the importance of considering various prosodic features in improving the accuracy of sentence boundary detection.

LSTM and Bi-LSTM models later achieved notable results, LSTM and Bi-LSTM models achieved 92.3% and 87.5%, and 93.36% and 89.78% accuracy for read aloud and spontaneous speech on the monosyllable model, respectively. Further enhancements with additional prosodic features resulted in 94.62% and 85.39%, and 95.23% and 87.89% accuracy using LSTM and Bi-LSTM models, respectively. These findings emphasize the importance of prosodic features and model complexity in accurate Amharic sentence boundary detection.

These comprehensive findings underscore the significance of prosodic features and model complexity in accurately detecting sentence boundaries in Amharic speech data. To further enhance model performance, consider integrating more linguistic features, and evaluate the potential benefits of data augmentation techniques.

4 Conclusion and Future Work

The primary objective of this research was to achieve sentence-level speech segmentation by accurately identifying boundaries in continuous speech signals. Future work will explore Advanced neural network based approaches to significantly enhance classification efficiency. Additionally, research will extend to automatic speech segmentation of other discrete units, including syllables, phonemes, and words, to further advance the field.

References

1. *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.*

2. *J. Kolář, "Automatic segmentation of speech into sentence-liked units," 2008. [3] E. Shriberg,*

3. *A. Stolcke, and D. Hakkani-t, "Prosody-based automatic segmentation of speech into sentences and topics," vol. 32, 2000*

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