Chapter 9
Speech Feature Evaluation for Bangla Automatic Speech Recognition

Mohammed Rokibul Alam Kotwal
United International University, Bangladesh

Foyzul Hassan
United International University, Bangladesh

Mohammad Nurul Huda
United International University, Bangladesh

ABSTRACT

This chapter presents Bangla (widely known as Bengali) Automatic Speech Recognition (ASR) techniques by evaluating the different speech features, such as Mel Frequency Cepstral Coefficients (MFCCs), Local Features (LFs), phoneme probabilities extracted by time delay artificial neural networks of different architectures. Moreover, canonicalization of speech features is also performed for Gender-Independent (GI) ASR. In the canonicalization process, the authors have designed three classifiers by male, female, and GI speakers, and extracted the output probabilities from these classifiers for measuring the maximum. The maximization of output probabilities for each speech file provides higher correctness and accuracies for GI speech recognition. Besides, dynamic parameters (velocity and acceleration coefficients) are also used in the experiments for obtaining higher accuracy in phoneme recognition. From the experiments, it is also shown that dynamic parameters with hybrid features also increase the phoneme recognition performance in a certain extent. These parameters not only increase the accuracy of the ASR system, but also reduce the computation complexity of Hidden Markov Model (HMM)-based classifiers with fewer mixture components.

DOI: 10.4018/978-1-4666-3970-6.ch009

INTRODUCTION

Conventional Automatic Speech Recognition (ASR) systems use stochastic pattern matching techniques, where a word candidate is matched against word templates represented by Hidden Markov Models (HMMs) (Young, 2005). Although these techniques have a fair performance in limited applications, they suffer from huge computational cost at classifier stages, and also they always reject a new vocabulary or so-called Out-Of-Vocabulary (OOV) word. On the other hand, a traditional segmentation-based phone decoding technique can be used to solve these problems, but, until now, its recognition accuracy is far from sufficient performance.

These ASR systems could not be able to provide enough performance at anytime and everywhere. One of the reasons is that the Acoustic Models (AMs) of a Hidden Markov Model (HMM)-based classifier include many hidden factors such as speaker-specific characteristics that include gender types and speaking styles. It is difficult to recognize speech affected by these factors, especially when an ASR system contains only a single acoustic model. One solution is to employ multiple acoustic models, one model for each type of gender. Though the robustness of each acoustic model prevails to some extent, the whole ASR system can handle gender effects appropriately.

Most of these ASR systems use Mel Frequency Cepstral Coefficients (MFCCs) of 39 dimensions (12-MFCC, 12-ΔMFCC, 12-ΔΔMFCC, P, ΔP and ΔΔP, where P stands for raw energy of the input speech signal). Here, Hamming window of 25 ms is used for extracting the feature. The value of pre-emphasis factor is 0.97. Although these standard MFCCs are prevalent to current ASR system, but these features do not provide better performance because frequency domain information are not incorporated within the feature vector during the extraction process.

Recently, dynamic parameters such as velocity and acceleration coefficients of speech showed its necessity for embedding them as features to resolve the coarticulation effect due to widening the context window size. Though the coarticulation effects can be solved by incorporating the triphone models (Young, 2005), but a large-scale speech corpus is required to negotiate all the triphones. Besides, the training of triphone models incurs many complexities in HMM based classifiers. To eliminate these complexities at cost we need some parameters like dynamic parameters for solving the problem of left and right context.

Contemporary Bangla automatic speech recognition suffers from some difficulties: (1) lack of large scale speech corpus, (2) unavailability of labeled speech data, and (3) insufficient research opportunities though more than 220 million people speak in Bangla as their native language, which is ranked sixth based on the number of native speakers. These problems should be reduced immediately for constructing an ASR system for recognizing the voice.

The objective of this chapter is to design some ASR systems based on the above mentioned ground and to incorporate the some other speech features inside the ASR for improving the performance by eliminating the gender effects. The followings explicate the objectives of the chapter in details.

1. To construct a phoneme recognizer based on standard MFCC features to solve OOV problem.
2. To innovate a canonicalization method that resolves gender factor by incorporating both types of genders (male and female) in the process after selecting the maximum hypothesis.
3. To incorporate time and frequency domain information, new feature called local feature instead of standard MFCC is extracted from an input speech for an ASR system.
4. To extract phoneme probabilities based on time delay neural network by using MFCCs as input feature.