Chapter 8
Acoustic Modeling of Speech Signal using Artificial Neural Network: A Review of Techniques and Current Trends

Mousmita Sarma
Gauhati University, India

Kandarpa Kumar Sarma
Gauhati University, India

ABSTRACT

Acoustic modeling of the sound unit is a crucial component of Automatic Speech Recognition (ASR) system. This is the process of establishing statistical representations for the feature vector sequences for a particular sound unit so that a classifier for the entire sound unit used in the ASR system can be designed. Current ASR systems use Hidden Markov Model (HMM) to deal with temporal variability and Gaussian Mixture Model (GMM) for acoustic modeling. Recently machine learning paradigms have been explored for application in speech recognition domain. In this regard, Multi Layer Perception (MLP), Recurrent Neural Network (RNN) etc. are extensively used. Artificial Neural Network (ANN)s are trained by back propagating the error derivatives and therefore have the potential to learn much better models of nonlinear data. Recently, Deep Neural Network (DNN)s with many hidden layer have been up voted by the researchers and have been accepted to be suitable for speech signal modeling. In this chapter various techniques and works on the ANN based acoustic modeling are described.

INTRODUCTION

Acoustic and temporal modelings are the two major issues associated with an Automatic Speech Recognition (ASR) system. Speech is naturally dynamic in nature. The spectral and temporal variations of speech signals are due to speech production nature. Speech signal is produced by moving the articulators

DOI: 10.4018/978-1-5225-0159-6.ch008

to different position necessary for the target sound unit. Due to the variation in the articulator's motions, instead of producing a sequence of clean identical phonetic units, a sequence of trajectories or signature is obtained in the form of a speech signal. This makes it difficult to extract exact timing information as well as spectral information of the speech units from the speech signal. Therefore, modeling of speech signal needs to consider both these issues.

An ASR system uses acoustic models to extract information from the acoustic signal. In the pattern recognition based approach of speech recognition, basic recognition units are modeled acoustically based on some lexical description, which is essentially a mapping between acoustic measurement and phoneme. Such mappings are learned by a finite training set of utterances. The resulting speech units are called phone like unit (PLU) which is an acoustic description of that speech unit as present in the training set (Brown, 1987).

Thus handling temporal and spectral variability are the main challenges of ASR and currently the best known speech recognition technology prefers Hidden Markov Model (HMM), which provides solution to both these problems. Acoustic modeling is performed by discrete density models and temporal modeling is performed by state transitions (Xiong, 2009). HMM considers the speech signal as quasi- static for short durations and models these frames for recognition. It breaks the feature vector of the signal into a number of states and finds the probability of a signal to transit from one state to another (Rabiner & Juang 1993). Viterbi search, forward-backward and Baum-Welch algorithms are used for parameter estimation and optimization (Rabiner, 1989) (Juang & Rabiner, 1991). But in speech recognition HMM based acoustic modeling has a serious disadvantage. It suffers from quantization errors and poor parametric modeling. The standard Maximum Likelihood (ML) training criterion leads to poor discrimination between the acoustic models. Also the independence assumption makes it hard to exploit multiple input frames; and the first-order assumption makes it hard to model co- articulation and duration (Tebelskis, 1995)

Later after the introduction of Expectation Maximization (EM) algorithm (Rolf, 1974), GMM has been used for acoustic modeling. The probability distribution of the feature vectors associated with the HMM states can be modeled by GMM with higher accuracy. This facilitates the successful implementation of GMM-HMM systems for speech recognition as preferred by present day systems.

Despite of its outstanding performance in terms of accuracy, GMM has some disadvantages, like it requires huge amount of training data and processing speed. But the major drawback of GMM is that it requires a large number of diagonal Gaussians or a large number full covariance Gaussians to model data which lies near a non linear surface in the data space (Hinton, Deng, Yu, Dahl, Mohamed, Jaitly, Senior, Vanhoucke, Nguyen, Sainath & Kingsbury, 2012). Using large coefficients is not statistically efficient since underlying structure of speech signal is much lower dimensional.

This makes the researchers to think about some other methods for acoustic modeling. In this regard, Artificial Neural Network (ANN) appears to be an effective alternative, due to its inherent capability to learn data both linear and non-linear, show adaptive behavior, exhibit robustness to sudden variations like noise, parallel and discriminative nature of learning, retain and use it subsequently. In this respect, it closely resembles the processing nature of the human brain. Further, recent advances in very large scale integrated (VLSI) circuit technology system on chip (SoC) design of ANN have enabled implementation of bio-inspired attributes in ASR systems. ANN can model a diversity of speaking styles and background conditions with much less training data, because of the distributive representation of input. ANNs are constituted many layers of artificial neurons each of which individually can learn, retain the learning, use it subsequently and contribute to the cumulative processing capability of the network. Many neurons simultaneously process each of the stimulation received which maybe segments of a pattern and