Extracting Unique Phonetic Features for Speech and Language Processing Applications

March 29, 2015

Feature extraction is a crucial step in developing any speech and language processing applications. Extracting features from speech can be viewed in two major perspectives, in the initial perspective a feature to represent speech signal is extracted by treating speech signal as any non-stationary signal. The later perspective is by considering speech as a special signal that has its own production mechanism and can only be produced by humans, this view of speech can also be viewed as a phoneticians view of a speech signal.

Traditionally a sequence of feature vectors is used to represent information in speech signal. Feature vectors are derived by block processing the speech signal using a segment of about 20-30 ms for a shift of every 5-10 ms. Typically the information in short time spectral envelope, which correlates to the shape of the vocal tract is represented in the feature vectors. Mel frequency cepstral coefficients (MFCCs) [1] and Perceptual linear predictive cepstral coefficients (PLPs) [2] and are the two widely used feature vector representations of speech. By hypothesizing that a sequence of feature vectors that share a common feature space can be used to represent the speech. Many speech systems such as speech recognition systems are designed, but these representations do not properly emphasize the information provided by the speech production mechanism. Though these features are easy to extract and computationally easy to deploy, as human is analog in nature a slight variation is imparted in the speech signal when ever he produces the same sound multiple times. But the features extracted by considering the initial view cannot provide tolerance to the variations that are possible and can not un-ambiguously represent the speech signal.

Speech is produced and perceived as a sequence of acoustic events, where each acoustic event can be distinctly represented by its production characteristics [3] [4]. Different sets of feature vectors may be needed to represent an acoustic event which may not share the common feature space, so a need of an alternate representation is a pre-requisite to develop a robust and highly practical speech system. A lot of scientific research is done in subjectively analyzing the acoustic events and phonetic features of various sounds in speech but quite a minimal amount of research is carried out to detect, model and quantify these features using signal processing techniques. The main objective of this work is to use the signal processing and advanced time frequency analysis techniques to develop signal processing measures to represent the presence of various acoustic events, phonetic features. These measures are used to detect the phonetic features from speech to provide an alternate way of developing speech and language processing applications. Though these features are hard to compute they provide robustness against the speech and speaker variability of speech.

Extracting the phonetically motivated features for representing various acoustic events of speech has wide range of applications in speech and language processing. Speech recogni-
tion and language identification are the two major systems where these features are highly influential. Present days automatic speech recognition systems (ASR) are mostly data driven approaches and building these systems need a lot of human effort in preparing the data to train the system. In spite of the human effort, these systems are not easily adaptable and are highly subjective to the speech variations. These systems need to be trained and all the possible variations of speech are to be covered in the training data. Despite of the efforts these methods use a language model for recognition along with the acoustic model. A language model is used as acoustic model derived form the conventional feature based representation is not reliable on its own. But the use of a language model customizes the operation of ASR system and in becomes operable only in a particular environment, which is not so readily acceptable. Phonetically motivated speech recognition is one possible alternative approach in which every sound is recognized by its unique phonetic characteristics [5]. This approach is easily generalizable, adaptable, more robust to speech variability and no training of the models is required [6]. Developing the signal processing measures to detect the presence of various phonetic features in speech plays a vital role in developing this system.

Phonetic feature extraction is also used in many other types of applications like language identification (LID) and speaker recognition systems. Due to the presence of multi-lingual culture in India a language identification (LID) system is highly demanded. Implicit language identification is used as a front end of a multi-lingual dialog systems. Owing to the decent from two or three major languages most of the sounds are same and automatically identifying them using conventional feature vector based representations may not yield acceptable level of performance. Some phonetic features are specific to a particular language [7]. Some of these unique phonetic features are presence of voiced retroflex approximants in Tamil and Malayalam [8], presence of breathy vowel in Gujarati [9] and presence of palatalized consonants in Kashmiri [7], presence of many aspirated sounds in north Indian languages compared to the south Indian languages etc. Detecting the presence of such similar phonetic features from speech can be used to improve the performance of existing LID systems for Indian Scenario.

Objectives

1. Analyzing the various phonetic features from speech using signal processing tools.
2. Developing signal processing equivalents for quantifying the phonetic features.
3. Detecting the presence of various phonetic features in continuous speech.
4. Developing various fusion techniques to use phonetically motivated features in combination with conventional feature vector based features to improve the performance of the existing speech systems.

References


