INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

WHETHER MFCC OR GFCC IS BETTER FOR RECOGNIZING EMOTION FROM SPEECH? A STUDY

Minu Babu[1]

[1]MTtech Scholar, Department of Computer Science and Engineering, Federal Institute of Science and Technology, Mahatma Gandhi University, Kottayam, Kerala
minubabu4@gmail.com

Abstract

A major challenge for automatic speech recognition (ASR) relates to significant performance reduction in noisy environments. Recently, the study of the emotional content of speech signals got more importance and hence, many systems have been proposed to identify the emotional content of a spoken utterance. The important aspects of the design of a speech emotion recognition system are pre-processing, feature extraction, training and classification, recognition. Typically, extracted speaker features are short-time cepstral coefficients such as Mel-frequency cepstral coefficients (MFCCs) and perceptual linear predictive (PLP) coefficients, or long-term features such as prosody. Such systems usually do not perform well under noisy conditions because extracted features are distorted by noise, causing mismatched likelihood calculation. By introducing a novel speaker feature, gammatone frequency cepstral coefficient (GFCC), based on an auditory periphery model, and show that this feature captures speaker characteristics and performs substantially better than conventional speaker features under noisy conditions. An important finding in the study is that GFCC features outperform conventional MFCC features under noisy conditions..

Keywords*:* Automatic speech recognition, Pre-processing, Feature extraction, Classification, Mel-frequency cepstral coefficient, Gammatone frequency cepstral coefficient.

1. Introduction

Speech is a complex signal which contains information about the message, speaker, language and emotions. Emotion on other side is an individual mental state that arises spontaneously rather than through conscious effort. The database for the speech emotion recognition system is the emotional speech samples. Features for emotion recognition are extracted from these speech samples. The features extracted from these speech samples [1] are, the energy, pitch, linear prediction cepstrum coefficient (LPCC), mel frequency cepstrum coefficient (MFCC) etc. Among them MFCC is widely used for speech related studies with a simple calculation and good ability of the distinction. But, it works poorly under noisy environments and it is necessary to obtain a new feature for efficient recognition system. Thus, the idea of GFCC came. Gammatone filters are realized purely in the time domain. Specifically, the filters are applied directly on time series of speech signals by simple operations such as delay, summation and multiplication. This is quite different from the widely adopted frequency-domain design, where signals are transformed to frequency spectra first and the gammatone filers then applied upon them. The time domain implementation avoids unnecessary approximation introduced by short-time spectral analysis, and saves a considerable proportion of computation involved in FFT.Then, classifiers are used to differentiate emotions such as anger, happiness, sadness, surprise, fear, neutral state, etc.

The classification performance is based on extracted features. Various applications of emotional recognition systems are dialog system for detecting angry users, tutoring system for detecting student's interest/certainty, lie detection systems for investigation purposes, social interaction system for detecting frustration, disappointment, surprise/amusement etc.

2. Emotion Recognition System

An emotion recognition system consists of various stages as shown in the Figure 1.

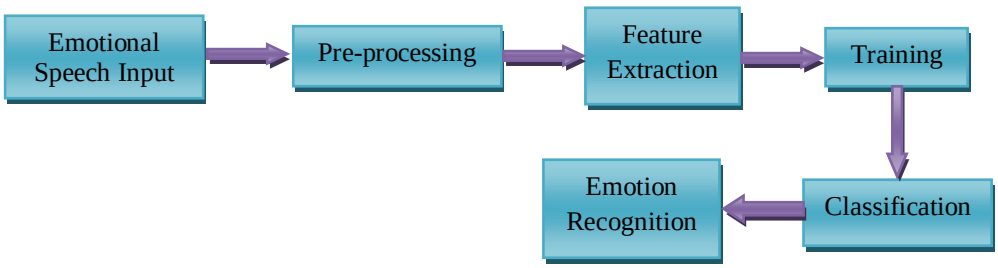

Figure 1: Emotion recognition system

2.1 Pre-processing

Pre-processing of speech signals [2] is considered a crucial step in the development of a robust and efficient speech or speaker recognition system. Pre-processing means segregating the voiced region from the silence/unvoiced portion of the captured signal is usually advocated as a crucial step in the development of a reliable speech or speaker recognition system. This is because most of the speech or speaker specific attributes are present in the voiced part of the speech signals moreover, extraction of the voiced part of the speech signal by marking and/or removing the silence and unvoiced region leads to substantial reduction in computational complexity. Other applications of classifying speech signals into silence/unvoiced region and voiced region are: Fundamental Frequency Estimation, Formant Extraction or Syllable Marking, End Point Detection etc.

2.2Feature Extraction

A speech signal composed of large number of parameters which indicates emotion contents of it. Changes in these parameters indicate changes in the emotions. Proper choice of feature vectors is one of the most important tasks in speech recognition. The feature vectors can be classified as long-time and short-time feature vectors. The long-time ones are estimated over the entire length of the utterance. The short-time ones are determined over window of usually less than 100 ms. The long-time approach identifies emotions more efficiently. Commonly used features [1] are energy and related features (the energy is the basic and most important feature in speech signal. The statistics of energy in the whole speech sample can be obtained by calculating the energy, such as mean value, max value, variance, variation range etc.), pitch and related features (the value of pitch frequency can be calculated in each speech frame), qualitative features (emotional contents of a utterance is strongly related with its voice quality. The acoustic parameters related to speech quality are voice level such as signal amplitude, energy and duration, voice pitch, phrase, word, feature boundaries and temporal structures), linear prediction cepstrum coefficients (LPCC) (LPCC embodies the characteristics of particular channel of speech. The Linear Predictive analysis is based on the assumption that the shape of the vocal tract governs the nature of the sound being produced. So we can extract these feature coefficients to identify the emotions contained in speech), mel-frequency cepstrum coefficients (MFCC) (MFCC is based on the characteristics of the human ear's hearing. It uses a nonlinear frequency unit to simulate the human auditory system. Mel frequency scale is the most widely used feature of the speech, with a simple calculation, good ability of the distinction, anti-noise and other advantages), perceptual linear predictive (PLP) coefficients etc. This study compares the differences between MFCC and GFCC for recognizing emotion from speech. The algorithm for MFCC and GFCC extraction [3] are given below.

MFCC Extraction (HTK version):

1. Pre-emphasize input signal
2. Perform short-time Fourier analysis to get magnitude spectrum
3. Wrap the magnitude spectrum into mel-spectrum using 26 tri- angular overlapping windows where center frequencies of the windows are equally distributed on the mel scale
4. Take the log operation on the power spectrum (i.e. square of mel-spectrum)
5. Apply the discrete cosine transform (DCT) on the log-melpower- spectrum to derive cepstral features
6. Perform cepstral liftering

GFCC Extraction:

1. Pass input signal through a 64-channel gammatone filterbank
2. At each channel, fully rectify the filter response (i.e. take absolute value) and decimate it to 100 Hz as a way of time windowing.
3. Then take absolute value afterwards. This creates a time frequency (T-F) representation that is a variant of cochleagram
4. Take cubic root on the T-F representation
5. Apply DCT to derive cepstral features

2.3 Training and Classification

Selection of classifier depends on the geometry of the input feature vector. Some classifiers are more efficient with certain type of class distributions. Various Classifiers used are Hidden Markov Model (HMM) [4], Gaussian Mixtures Model (GMM) [5], Support Vector Machine (SVM) [6], Artificial Neural Network (ANN) [7], K-Nearest Neighbours (KNN) [8], Decision Trees [9] etc. HMM has been used widely for speech emotion recognition due to its advantage on dynamic time warping capability. That is, its ability to estimate the similarity between two temporal sequences which may vary in time or speed. However, the classify property of HMM is not satisfactory. GMM are suitable for developing emotion recognition model when large number of feature vector is available. Gaussian Mixture Models (GMMs) are among the most statistically matured methods for clustering and for density estimation. The GMM and the HMM, are the most used ones for speech emotion recognition. Another common classifier, used for many pattern recognition applications is the artificial neural network (ANN). Their classification performance is usually better than HMM and GMM when the number of training examples is relatively low. The ANN based classifiers may achieve a correct classification rate of 51.19% in speaker dependent recognition, and that of 52.87% for speaker independent recognition. One of the important classifiers is the support vector machine (SVM). SVM classifiers are shown to outperform other well-known classifiers. The accuracy of the SVM for the speaker independent and dependent classification are 75% and above 80% respectively.

3. Result and Conclusion

Broadly speaking, there are two major differences between MFCC and GFCC. The obvious one is the frequency scale. GFCC, based on equivalent rectangular bandwidth (ERB) scale, has finer resolution at low frequencies than MFCC (mel scale). The other one is the nonlinear rectification step prior to the DCT. MFCC uses a log while GFCC uses a cubic root. Both have been used in the literature. In addition, the log operation transforms convolution between excitation source and vocal tract (filter) into addition in the spectral domain. Besides these two major differences, there are some other notable differences that are summarized in the Table 1. By carefully examining all the differences between MFCC and GFCC, it concludes that the nonlinear rectification mainly accounts for the noise robustness differences. In particular, the cubic root rectification provides more robustness to the features than the log. The cubic root operation better might be the case that some speaker information is embodied through different energy levels. In a noisy mixture, there are target dominant T-F units or segments indicative of this energy information. The cubic root operation makes features scale variant (i.e.

energy level dependent) and helps to preserve this information. The log operation, on the other hand, does not encode this information.

Table 1: Difference between MFCC and GFCC

Category	MFCC	GFCC
Pre-emphasis	Yes	No
# of Frequency Bands	26	64
Cepstral Liftering	Yes	No
Frequency Scale	Mel	ERB
Nonlinear Rectification	Logarithmic	Cubic Root
Scale-invariant (w/o 0^{th} coefficient)	Yes	No
Intermediate T-F Representation	Mel Spectrum	Variant of Cochleagram

REFERENCES

[1] Aastha Joshi1, Rajneet Kaur2, April 2013, "A Study of Speech Emotion Recognition Methods", *IJCSMC, Vol. 2, Issue. 4 pg.28 – 31.*

[2] Ayaz Keerio, Bhargav Kumar Mitra, Philip Birch, Rupert Young, and Chris Chatwin, 2009, "On Pre-processing of Speech Signals", *International Journal of Signal Processing 5;3.*

[3] Xiaojia Zhao and DeLiang Wang, 2013, "ANALYZING NOISE ROBUSTNESS OF MFCC AND GFCC FEATURES IN SPEAKER IDENTIFICATION" , *978-1-4799-0356 - 6/13/$31.00 ©2013 IEEE.*

[4] Phil Blunsom, pcbl@cs.mu.oz.au, August 19, 2004, "Hidden Markov Models".

[5] Douglas Reynolds, "Gaussian Mixture Models", MIT *Lincoln Laboratory, 244 Wood St., Lexington, MA 02140, USA.*

[6] CHRISTOPHER J.C. BURGES burges@lucent.com *Bell Laboratories, Lucent Technologies,* Kluwer Academic "A Tutorial on Support Vector Machines for Pattern Recognition", *Boston. Manufactured in The Netherlands.*

[7] http://www.cse.unr.edu/~bebis/MathMethods/NNs/lecture.pdf

[8] Gongde Guo, Hui Wang, David Bell, Yaxin B and Kieran Greer, "KNN Model-Based Approach in Classification".

[9] Kismat Maredia Fall, 2010, "A Study in Decision Analysis using Decision Trees and Game Theory".

Minu Babu – Currently doing her final year post graduation in Computer Science from Mahatma Gandhi University, Kottayam, Kerala. She took graduation from the same university in 2012 with distinction. This work is done as a part of her PG mini project. She is very much interested in Pattern Recognition, Image Processing, Machine Learning and Artificial Intelligence areas. Recently, she is interested in Emotion Recognition from speech and doing her main project in this area.