Implementation of Emotion Detection by Cepstral Analysis and Noise Filtration in MATLAB

P V S Siva Tarun[1] and Humaira Nishat[2]

[1] CVR College of Engineering/ECE Department, Hyderabad, India
Email: tarunrd.95@gmail.com
[2] CVR College of Engineering/ECE Department, Hyderabad, India
Email: huma_nisha@yahoo.com

Abstract— **The main objective of the paper is to determine the emotional state of a person, through his speech. The two most important features that are used to identify the emotional states of a person are i. Pitch and ii. Formant frequencies. To understand the three emotional states of a person, these two pitch and formant frequencies are first extracted from the speech signal and, later their analysis is carried out. The three emotions that are considered are Anger, Neutrality and Happiness. Simulations are carried out in MATLAB and the TU-Berlin database is used for the analysis. For the extraction of pitch frequency Cepstral analysis method is used and the formant frequencies are estimated to detect happy emotions. The objective of cepstral analysis is to separate the speech into its source and system components without any prior knowledge about source and / or system. This separation is done due to difference in occupancy of frequency ranges. Following it, noise filtration is implemented to remove inaudible ambient noises.**

Index Terms— **cepstral analysis, emotion, filtering, formant frequencies, noise**

I. INTRODUCTION

In the present era of modernization, comes the need of automation. Many of the things around us tend to be controlled automatically. Once the word ‘automatic’ is being used, a person’s speech/voice plays an important role. Processing of one’s speech signal and extracting some features of the speech is known as speech signal processing. Speech signal processing [1] has its applications in various areas like speech coding, speech recognition, speaker recognition, speech enhancement, emotion detection etc. Study on speech signals has developed very much in the recent past due to various sophisticated algorithms being developed like vector quantization, hidden Markov model, cepstral analysis etc. In this paper, cepstral analysis is used to extract the pitch and formant frequencies from the speech signal to identify the emotional states of a person.

The paper is organized as follows. Section II covers the basic principles of cepstral analysis. The proposed work is described in section III. Simulations are carried out using MATLAB. Section IV gives the results and section V gives conclusions.

II. CEPSTRAL ANALYSIS

Considering the source-filter theory, the production of speech can be categorized into two stages. The first stage involves generation of a sound source which has its own spectral shape and spectral fine structure. In the second stage the sound source is then shaped or filtered by the resonant properties of the vocal tract. According to signal processing, speech can be considered as the output coming from a system (i.e., entire vocal tract) which is excited by an input (i.e., vibration of vocal folds). The separation of these two i.e., deconvolving them is necessary to process the speech.

If there is prior knowledge about the input excitation, then it is possible to separate the system component and construct it by exciting the system with the input and finally collecting its responses. If there is knowledge about the response of the system, then it is possible to recover input excitation by using the concept of inverse filter theory. The excitation is recovered using Linear Prediction analysis of speech. There is another method of deconvolution wherein the assumptions are input excitations and which has unknown system responses. The proposed work is based on this type where the responses and the input excitations of the system are not known.

Speech is composed of excitation source and vocal tract system components. In order to analyze and model the excitation and system components of the speech independently and also use that in various speech processing applications, these two components have to be separated from the speech. The objective of *cepstral analysis* is to separate the speech into its source and system components without any prior knowledge about source and / or system.

If the excitation sequence is e(x) and the vocal tract filter sequence is s(x) then the speech sequence h(n) can be written as

$$s(x)=e(x)*h(x) \quad (1)$$

In frequency domain this can be written as

$$S(w)=E(w).H(w) \quad (2)$$

DOI: 10.32377/cvrjst1002

The above equation (2) conveys that the multiplication of excitation and system components in the frequency domain gives the convolved sequence of the same in the time domain. The deconvolution involves the speech to be deconvolved into the excitation and the vocal tract in the time domain. Thus, cepstral analysis is carried out for transforming the multiplied source and system components in the frequency domain to linear combination of the two components in the cepstral domain.

From the Eqn. (2) the magnitude spectrum of given speech sequence can be represented as,

$$\left|S(w)\right|=\left|E(w)\right|\left|H(w)\right| \tag{3}$$

To linearly combine the E(ω) and H(ω) in the frequency domain, logarithmic representation is used. So the logarithmic representation of Eqn. (3) will be,

$$\log\left|S(w)\right|=log\left|E(w)\right|+\log\left|H(w)\right| \tag{4}$$

In the above equation (4), the logarithmic operation is used which transforms the magnitude of the speech spectrum. Here the excitation component and the vocal tract component are multiplied. By using this logarithmic operation, multiplication of components is converted into summation components in the frequency domain. By using the inverse discrete fourier transform (IDFT) the separation of excitation and vocal tract system components is done. IDFT of linear spectra transforms back to the time domain but the IDFT of log spectra transforms to quefrency domain or the cepstral domain which is similar to time domain. The following equation (5) gives the mathematical concept of IDFT. In cepstral domain, the vocal tract component representation is shown as slowly varying components concentrated near the low quefrency region and excitation components by fast varying components at higher quefrency region.

$$\begin{gathered} log\left(E(w)+log\left|H(w)\right|\right) \\ c(n)=IDFT\left(log\left|S(w)\right|\right)=IDFT \end{gathered} \tag{5}$$

Reason for higher order coefficients representing excitation characteristics:

Voice signal basically involves periodic impulse sequences impinging on the vocal tracts, i.e. a train of impulse is the excitation repeating at periodic intervals say τ

$$x(t) =s(t)+\alpha s(t-\tau) \tag{6}$$

The Fourier spectral density (spectrum) of such a signal is given by

$$|X (f)|2 =| S(f)|2[1+\alpha 2 +2\alpha\cos(2\pi f \tau)] \tag{7}$$

Thus, from (7) it is clear that the spectral density of a signal with an echo has the form of an envelope (the spectrum of the original signal) that modulates a periodic function of frequency (the spectrum contribution of the echo).

Hence after every few samples pertaining to τ time interval, one can observe a peak in spectrum which is due to the train of impulses. Hence the upper quefrequency samples represent the excitation characteristics.

III. PROPOSED WORK

A. Pitch and Formants

Pitch [2] is the frequency at which one can observe maximum amplitude of vibration of vocal folds. It is also the fundamental frequency of the vibration. As per the above theory it's clear that vocal fold characteristics can be observed from higher order coefficients say after 20 samples of cepstrum.

A formant can be defined as a concentration of acoustic energy around a particular frequency in the speech wave. Since they represent the filter (chords) characteristics, these frequencies can be obtained from the lower quefrency components of cepstrum.

The work is carried out by first implementing a noise filter to remove low frequency ambient noise. This filter code is then used in emotion detection to filter ambient noises while recording live speech samples giving provision for live testing apart from recorded samples. The code first calculates the pitch of the signal and then the formants. Pitch has been used for identifying angry emotion, whereas formants are used to identify happiness [3].

The TU_BERLIN (Technical University of Berlin) database has been used to get recorded samples. The database consists of samples of voice spoken by professional actors with various sentences and emotions. These samples and the samples from movie clips had been used to test and identify the pitch and formant ranges of the emotions based on which the code had been written.

A low pass butterworth filter had been used to remove noise. The order of the filter can be increased or decreased based on type of noise to be removed. The code removes ambient noise and car horn noise. A 10th order filter had been used for car horn sound removal.

B. Pitch : Used to detect anger

Vocal folds vibrate at a very faster rate when a person is angry, thus making the interval τ very small and thereby increases the pitch of the speech. It has been observed that the pitch of neutral speech is much less when compared to the pitch of angry speech. Hence the pitch of the speech is considered to detect the emotion anger [4] [5].

Table I. below gives the pitch frequencies for six different speakers, each in various languages like one in English, three in Telugu and two in German when the

DOI: 10.32377/cvrjst1002

emotional state is anger. Table II. gives the pitch frequencies of four speakers in German, Telugu and English in neutral state. It is found that the frequency range for neutral state is less as compared to that of anger state. Also the pitch frequencies [6] lie in approximately the same range of frequencies when the emotional state is anger, irrespective of the languages used and the speakers. Pitch frequencies for neutral and anger emotion states are found to occur at approximately 300Hz and 200Hz respectively.

TABLE I.
OBSERVATIONS FOR HAPPY EMOTIONS

Number of Speakers	Language	F1	F2	F3
Speaker 1	German	363	915	1800
Speaker 2	Telugu	432	979	2267
Speaker 3	Telugu	419	819	1253
Speaker 4	Telugu	296	671	1195
Mean		**377**	**846**	**1629**

TABLE II
OBSERVATIONS FOR NEUTRAL EMOTIONS

Number of Speakers	Language	Pitch Frequency
Speaker 1	German	230
Speaker 2	German	139.13
Speaker 3	Telugu	189.04
Speaker 4	English	216.216
Mean Frequency		**193.6**

C. Formant: Used to detect happiness

Since they represent the filter (chords) characteristics, these frequencies can be obtained from the lower quefrecny components of cepstrum. The fundamental frequency (pitch) when the emotion is happy is less. Under such conditions the formants coincide with the spectral peaks of the filter coefficients. Hence, taking the FFT of the lower quefrency samples and finding spectral peaks gives the formant frequencies enabling us to detect the happy emotion [7].

Table III. below gives the observations for happy emotions. Here formant frequencies are extracted and three different frequencies f1, f2 and f3 are used to detect the emotional state. The analysis is performed on four different speakers, first speaker being German and remaining three are Telugu. The mean values are also calculated and it is found that the formant frequencies occur at three different bands. F1 occurs at around 300Hz, f2 at around 800Hz and f3 at 1600 Hz.

Pitch and formant frequencies are both calculated for every sample. It is observed that pitch of angry samples is different from pitch of neutral and formants of angry and happy are also different.

IV. RESULTS

A. Plots of Noise Filtration

Normalized plot of the spectrum brings in all the frequency components within range of [0 ,1] samples, considering only a half plot of the spectrum. Figure 1 clearly shows the noise part of the signal at higher samples.

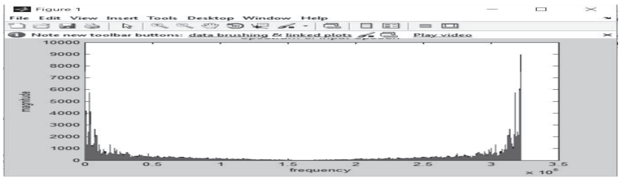


Figure 1. Spectrum of Input Speech Signal

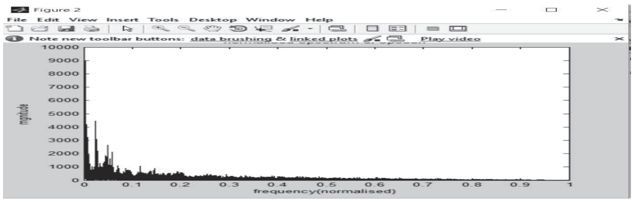


Figure 2. Normalized Spectrum of Input

Since a Low Pass Filter is used, all the higher frequency range values of the spectrum are removed. Figure 3 clearly depicts the removal of noise.

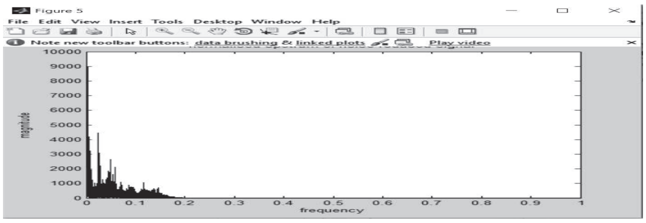


Figure 3. Filtered Output Spectrum

Cepstral analysis is applied on a 20ms speech sample. Considering the sampling frequency of 8000Hz, a 20 ms duration of speech signal has 160 samples to which cepstral analysis has been applied. Figure 4 shows the details of the cepstrum of the speech. Pitch can be identified by observing the peak after 20th sample till 80th sample.

B. Plots of Cepstral Analysis

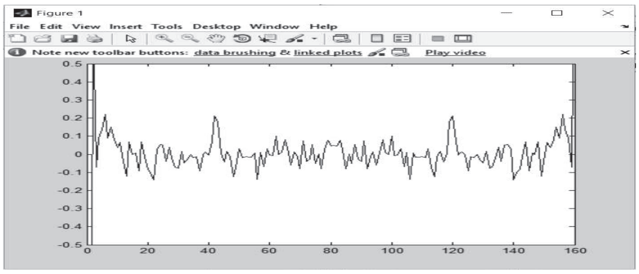
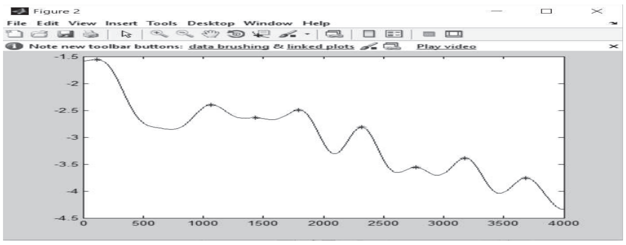


Figure 4. Cepstrum

DOI: 10.32377/cvrjst1002

Formants, as mentioned, are calculated by applying Fourier transform to lower samples of cepstrum. The peaks coincide with the formants. Figure 5 shows the formants as * mark in the spectrum.



Note new toolbar buttons: data brushing & linked plots Play video

Figure 5. Formants

V. Conclusions

It is clearly seen that cepstral method of analysis of speech is one of the best speech processing techniques developed for various applications. It was originally invented for characterizing the seismic echoes resulting from earthquakes and bomb explosions. But later it has proved to be an efficient method to evaluate human vocal tract characteristics (formants) and pitch of the speech signal. This was made possible due to application of logarithm which enabled separation of source and system components of speech (as per source-filter speech theory). Also these characteristics of pitch and formants are observed to be unique for different emotions like anger and happiness and hence, available to detect the emotion of a person through his speech with good accuracy. It is also observed that the emotional states of a person appear over a range of frequency bands. In this paper three emotions states i.e., angry, neutral and happy are recognized by finding the pitch and formant frequencies. For angry and neutral state, the pitch frequencies are centered around 300Hz and 200Hz respectively whereas for happy state the formant frequencies are centered at around 300Hz, 800Hz and 1600Hz.

The present consideration of emotion detection can be completely used to involve other states of emotion like sadness, boredom etc. with the involvement of image processing to observe the facial expressions of the individual, thus improving the accuracy.

References

[1] "Digital processing of speech signals." L.R. Rabiner, R.W.Schafer, Pearson education.

[2] "A Comparative Performance Study of Several Pitch Detection Algorithms", IEEE Transactions on acoustics, speech,and signal processing, VOL. ASSP-24, NO. 5, October 1976.

[3] "Speech Under Stress: Analysis, Modeling and Recognition".John H.L. Hansen and Sanjay Patil, M¨uller (Ed.): Speaker Classification I, LNAI 4343, pp. 108–137, 2007.

[4] "A Framework for Automatic Human Emotion Classification using Emotion Profiles" Emily Mower, IEEE Transactions on Audio, Speech and Languare Processing, Vol 19, July 2011.

[5] "Recognition of emotions in speech by a hierarchical approach", Affective Computing and Intelligent Interaction and Workshops,. ACII 2009. 3rd International Conference on 10-12 Sept. 2009 page(s): 1 - 8 Amsterdam 2009.

[6] "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," by M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, Proc of IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 805–808, May 2010.

[7] Emotion recognition from Assamese speeches using MFCC features and GMM Classifierby Aditya Kandali | Papers by Aditya "Co-authored with A. Routray and T. K. Basu" "published in Proceedings of IEEE Region 10 Conference 2008, Hyderabad, India"

DOI: 10.32377/cvrjst1002