Audio/Speech Signal Analysis for Depression
Arpit Parikh1 Sameena Zafar2 Mukesh Saini3

1PG Research Scholar 2Assistant Professor and Head of Department 3Assistant Professor
1,2,3Department of Electronics and Communication Engineering
1,2,3Patel Institute of Engineering & Science, RGPV University, Bhopal, India

Abstract— The word “depressed” is a common everyday word. People might say "I am depressed" when in fact they mean "I am fed up because I have had a row, or failed an exam, or lost my job", etc. These ups and downs of life are common and normal. Most people recover quite quickly. Depression is identified by different methods. Here we are identified depression by MFCC (Mel Frequency Cepstral Coefficient) method. There are different parameters used for the identification of depressed speech and normal speech, but MFCCs based parameter is the most applicable information then other parameter because depressive speech or audio signal can contain more information in the higher energy bands when compared with normal speech.

Keywords: Mel Frequency Cepstral Coefficients (MFCC), Speech Emotion Recognition (SER), discrete Cosine Transform (DCT), Depression

I. INTRODUCTION
Speech Emotion recognition through speech is an area which is increasingly attracting attention within the engineers in the field of pattern recognition and speech/Audio signal processing in recent years. Speech recognition is a technique that enables a device to recognize and understand spoken words. Automatic emotion recognition paid close attention in identifying emotional state of speaker from voice signal. Emotions play an extremely important role in human life. To recognize human’s emotions and feelings, various physiological signals have been widely used to classify emotion because signal acquisition by non-invasive sensors is relatively simple and physiological responses induced by emotion are less sensitive in social and cultural difference. Depression is classified blood volume pulse, electrocardiogram (ECG), and Galvanic skin response, Skin temperature and Gaussian Mixture Model. Speech signal is also recognize with Wiener filter and Kalman filter. [1]

MFCC (Mel Frequency Ceptral Coefficients) is used to analysis of depression. The main point to understand about speech is that the sounds generated by a human are filtered by the shape of the vocal tract including tongue, teeth etc. This shape determines what sound comes out. If we can determine the shape accurately, this should give us an accurate representation of the phoneme being produced. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and the job of MFCCs is to accurately represent this envelope. Mel Frequency Cepstral Coefficients (MFCCs) are a feature widely used in automatic speech and speaker recognition. They were introduced by Davis and Mermelstein in the 1980’s, and have been state-of-the-art ever since. Prior to the introduction of MFCCs, Linear Prediction Coefficients (LPCs) [8] and Linear Prediction Cepstral Coefficients (LPCCs) and were the main feature type for automatic speech recognition (ASR) [5].

II. MFCC METHODOLOGY
Now there are some steps to calculate MFCCs [7].

(1) Frame the signal into short frames.
(2) For each frame calculate the period gram estimate of the power spectrum.
(3) Apply the mel filter bank to the power spectra, sum the energy in each filter.
(4) Take the logarithm of all filter bank energies.
(5) Take the DCT of the log filter bank energies.
(6) Keep DCT coefficients, discard the rest.

The first step is framing. The speech signal is split up into frames typically with the length of 10 to 30 milliseconds. The frame length is important due to the trade-off between time and frequency resolution. If it is too long it will not be able to capture local spectral properties and if too short the frequency resolution would degrade. The frames overlap each other typically by 25% to 70% of their own length. Overlapping is used to make sure that each speech sound is approximately centered at some frame. After the signal is split up into frames each frame is multiplied by a window function. A good window function has a narrow main lobe and a low side lobe. A smooth tapering at the edges is desired to minimize discontinuities. The most common window used in speech processing is the Hamming window.

The second step is to apply the Discrete Fourier Transform on each frame. The fastest way to calculate the DFT is to use FFT which is an algorithm that can speed up DFT calculations by hundred-folds [3].

\[
X_k = \sum_{n=0}^{N-1} x_n e^{-2\pi i k n / N} \quad k = 0, 1, ..., N-1 \quad (1)
\]

Spectral analysis shows that different timbres in speech signals corresponds to different energy distribution over frequencies. Therefore we usually perform FFT to obtain the magnitude frequency response of each frame.

When we perform FFT on a frame, we assume that the signal within a frame is periodic, and continuous when wrapping around. If this is not the case, we can still perform FFT but the in continuity at the frame’s first and last points is likely to introduce undesirable effects in the frequency response. To deal with this problem, we have two strategies. Multiply each frame by a Hamming window to increase its continuity at the first and last points. Take a frame of a variable size such that it always contains an integer multiple number of the fundamental periods of the speech signal [6].

The second strategy encounters difficulty in practice since the identification of the fundamental period is not a trivial problem. Moreover, unvoiced sounds do not have a fundamental period at all. Consequently, we usually adopt the first strategy to multiply the frame by a Hamming window before performing FFT.

We multiple the magnitude frequency response by a set of 20 triangular band pass filters to get the log energy of each triangular band pass filter. The positions of these
filters are equally spaced along the Mel frequency, which is related to the common linear frequency \( f \) by the following equation:

\[
M = 2595 \log_{10} \left( \frac{f}{700} + 1 \right) \tag{2}
\]

The mel scale is based on how the human hearing perceives frequencies. It was defined by setting 1000 mels equal to 1000 Hz as a reference point. Mel-frequency is proportional to the logarithm of the linear frequency, reflecting similar effects in the human's subjective aural perception. Then listeners were asked to adjust the physical pitch until they perceived it as two-fold ten-fold and half, and those frequencies were then labeled as 2000 mel, 10000 mel and 500 mel respectively. The resulting scale was called the mel scale and is approximately linear below frequencies of 1000 hz and logarithmic above [4].

Spectrum is calculated by using a filter bank, spaced uniformly on the mel scale. That filter bank has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval. The mel frequency warping when calculating MFCCs is accomplished by the use of a triangular mel spaced filter bank. It consists of several triangular shaped and mel spaced filters, and their outputs are described by:

\[
Y(i) = \sum_{j=1}^{n} S_j H_i f \tag{3}
\]

Where \( S_j \) is the N-point magnitude spectrum and \( H_i \) the sampled magnitude response of an M-channel filter bank.

In this step, we apply DCT on the 20 log energy is obtained from the triangular bandpass filters to have L mel-scale cepstral coefficients. Apply compression by using a logarithm on the filter outputs \( Y \) and then to apply the discrete cosine transform which yields the MFCCs \( c[n] \) according to the following formula:

\[
C[n] = \sum_{i=0}^{M} \log(Y(i)) \cos \left( \frac{\pi n}{M} (i - \frac{1}{2}) \right) \tag{4}
\]

Where \( N \) is the number of triangular bandpass filters, \( L \) is the number of Mel-scale cepstral coefficients. Usually we set \( N = 20 \) and \( L = 12 \). Since we have performed FFT, DCT transforms the frequency domain into a time-like domain called quefrency domain. The obtained features are similar to cepstrum, thus it is referred to as the mel-scale cepstral coefficients, or MFCC. MFCC alone can be used as the feature for speech recognition. The energy within a frame is also an important feature that can be easily obtained. Hence we usually add the log energy as the 13rd feature to MFCC. If necessary, we can add some other features at this step, including pitch, zero cross rate, high-order spectrum momentum, and so on.

When we run the program, three messages are displayed in GUI. ‘Start record’, ‘Browse file’ & ‘Recognize’ is displayed. When we push the button ‘Start Record’, program asks to record of audio/speech signal or voice for 1 second and when we push the button ‘Browse Wave file’, program asks to browse wave file from the system and to record of audio/speech signal or voice for 1 second. After Recording When we push the button ‘Recognize’, of the Audio signal, if feature of voice matched with depressed data, then it display message ‘Depressed’.

After recording, if feature of voice is not matched with depressed data, then it displays message ‘Normal’ [2].

### III. RESULT

Result obtained during recognition of speech/audio signal. It seems to be very effective and robust to implement in real time working environment.

![Fig. 1: Output wave form of depression](image)

### IV. CONCLUSION

Finally analysis of speech/audio sample, conclude that vocal properties represented by MFCC. It is provide most effective result between depressed and normal people. The extracted features were stored in a .mat file using MFCC algorithm. A distortion measure based on minimizing the Euclidean distance was used when matching the unknown speech signal with the speech signal database. The experimental results were analyzed with the help of MATLAB and it is proved that the results are efficient. We hope this paper brings a basic understanding of the researchers motivate them to explore the community of speech/audio recognition.

### REFERENCES


[3] Yen-Ting Chen, I-Chung Hung, Chun-Ju Hou, Signal Analysis for Patients with Depression”, Department of Electrical Engineering, Southern Taiwan University, Tainan, Taiwan.


