MFCC Feature Extracation

H Kesava Sravan, Kavya Balaji, N Hari Charan, K Sujan Surya
CB.EN.U4ELC20023, ELC20029, ELC20041, ELC20031
Department of Electrical and Computers Engineering
Amrita School of Engineering, Amrita Vishwa Vidyapeetham
Coimbatore, India 641112
cb.en.uelc20023@cb.students.amrita.edu, cb.en.uelc20029@cb.students.amrita.edu
cb.en.uelc20041@cb.students.amrita.edu, cb.en.uelc20031@cb.students.amrita.edu

***Abstract*—This paper presents an approach to extract features from speech signal of a noise signal using the Mel-Scale Frequency Cepstral Coefficients.**

***Index Terms*—Mel frequency cepstral coefficients (MFCC), spectrum, cepstrum, frequency, qrefuency.**

I. Introduction

Speech recognition is the process of automatically recognizing certain word which is spoken by a particular speaker based on some information included in voice sample. It conveys information about words, expression, style of speech, accent, emotion, speaker identity, gender, age, the state of health of the speaker etc. There has been a lot of advancement in speech recognition technology, but still it has huge scope. Speech based devices find their applications in our daily lives and have huge benefits especially for those people who are suffering from some kind of disabilities. We can say that such people are restricted to show their hidden talent and creativity. We can also use these speech based devices for security measures to reduce cases of fraud and theft.

The most popular feature extraction technique is the Mel Frequency Cepstral Coefficients called MFCC as it is less complex in implementation and more effective and robust under various conditions. MFCC is designed using the knowledge of human auditory system. It is a standard method for feature extraction in speech recognition.

II. Methodology

The most prevalent and dominant method used to extract spectral features is calculating Mel-Frequency Cepstral Coefficients (MFCC). MFCCs are one of the most popular feature extraction techniques used in speech recognition based on frequency domain using the Mel scale which is based on the human ear scale. MFCCs being considered as frequency domain features are much more accurate than time domain features

Mel-Frequency Cepstral Coefficients (MFCC) is a representation of the real cepstral of a windowed shorttime signal derived from the Fast Fourier Transform (FFT) of that signal. The difference from the real cepstral is that a nonlinear frequency scale is used, which approximates the behaviour of the auditory system. Additionally, these coefficients are robust and reliable to variations according to speakers and recording conditions. MFCC is an audio feature extraction technique which extracts parameters from the speech similar to ones that are used by humans for hearing speech, while at the same time, deemphasizes all other information. The speech signal is first dividedinto time frames consisting of an arbitrary number of samples.

Evolution of the spectral content of the signal, an effort is often made to include the extraction of this information as part of feature analysis. In order to capture the changes in the coefficients over time, first and second difference coefficients are computed as respectively.

$$\Delta c(t) = c(t+2) - c(t-2) \tag{1}$$

$$\Delta\Delta c(t) = \Delta c(t+1) - \Delta c(t-1) \tag{2}$$

These dynamic coefficients are then concatenated with the static coefficients according to making up the final output of feature analysis representing the speech frame according to making up the final output of feature analysis representing the speech frame.

A. Advantages

As the frequency bands are positioned logarithmically in MFCC, it approximates the human system response more closely than any other system.

B. Disadvantages

MFCC values are not very robust in the presence of additive noise, and so it is common to normalize their values in speech recognition systems to lessen the influence of noise.

C. Applications

MFCCs are commonly used as features in speech recognition systems, such as the systems which can automatically recognize numbers spoken into a telephone. They are also common in speaker recognition, which is the task of recognizing people from their voices. MFCCs are also increasingly finding uses in music information retrieval applications such as genre classification, audio similarity measures, etc.

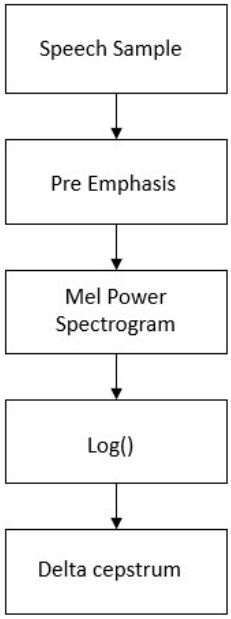

Fig. 1. MFCC Derivation.

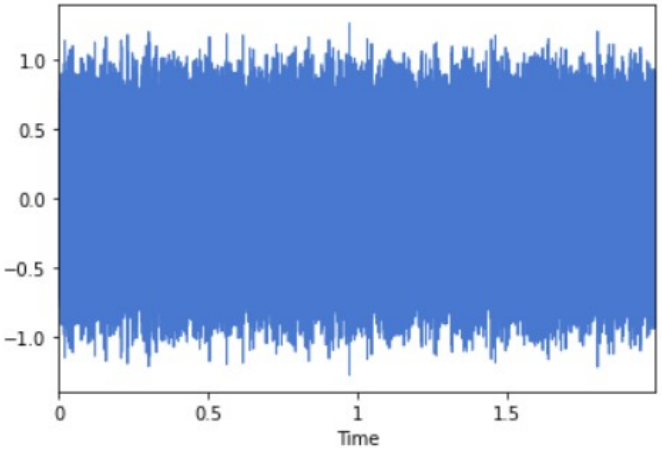

Fig. 2. Sample audio.

III. Results and Analysis

A compact representation would be provided by a set of mel-frequency cepstrum coefficients (MFCC), which are the results of a cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale. The MFCCs are proved more efficient. The calculation of the MFCC includes the following steps.

A. Mel-frequency wrapping

Human perception of frequency contents of sounds for speech signal does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale. The melfrequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000Hz .As a reference point ,the pitch of a 1 KHz tone ,40dB above the perceptual hearing threshold, is defined as 1000 mels. Therefore we can use the following approximate formula to compute the mels for a given frequency f in Hz.

$$Mel(f) = 2595 * log10(1 + f/700) \quad (3)$$

Ours approach to simulate the subjective spectrum is to use a filter bank, one filter for each desired mel-frequency component. That filter bank has a triangular band pass frequency response and the spacing as well as the bandwidth is determined by a constant mel-frequency interval. The mel scale filter bank is a series of l triangular band pass filters that have been designed to simulate the band pass filtering believed to occur in the auditory system. This corresponds to series of band pass filters with constant bandwidth and spacing on a mel frequency scale.

B. Mel Spectrogram

Sound is usually visualized as an airwave that is a two-dimensional representation of amplitude and time. Figure 3 shows an example of a sound signal in the time domain. Sound can also be represented as a frequency spectrum of an audio signal as it varies with time. This is called a spectrogram. A spectrogram of sound is created from a time signal using the fast Fourier transform (FFT). "Mel" is short for melody. It implies that this is a perceptual scale measurement based on the comparison of the pitches.

A Mel spectrogram, a combination of the Mel scale and the spectrogram, is a visual representation of a drill sound in both frequency and amplitude by the time domains. The amplitude of a particular time is represented by colors. Brighter colors up through orange correspond to progressively stronger amplitudes, as shown in Figure 5(b). The horizontal axis presents the time from left to right. The vertical axis presents the frequency from low to high.

Frequencies in a sound signal change over time. Hence, the use of Fourier transforms on the entire audio signal results in a loss of meaningful frequency information in the time domain. Supposing the frequency of the sound signal is uniform for a very short period of time, each sound is divided into short time frames of 20 ms (2000-points windows) with a 512-point overlap between successive frames. The FFT length is 2000 points. Implementing Fourier transforms on these consecutive frames can help us obtain a good approximation of frequencies across the time domain.

$$w(n) = \alpha - (1 - \alpha)cos(2\pi nN - 1), 0 <= n <= N - 1 \quad (4)$$

C. Cepstrum

The cepstrum is a representation used in homomorphic signal processing, to convert signals combined by convolution (such as a source and filter) into sums of their cepstra, for linear separation. In particular, the power cepstrum is often

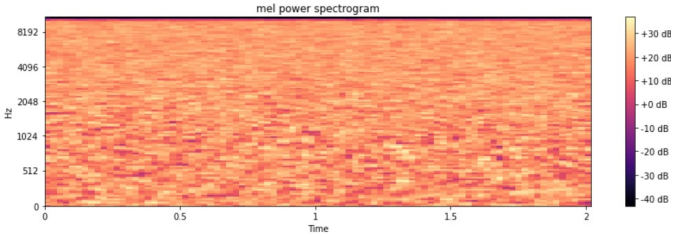

Fig. 3. Mel Power Spectrogram.

used as a feature vector for representing the human voice and musical signals. For these applications, the spectrum is usually first transformed using the mel scale. The result is called the mel-frequency cepstrum or MFC (its coefficients are called mel-frequency cepstral coefficients, or MFCCs). It is used for voice identification, pitch detection and much more. The cepstrum is useful in these applications because the low-frequency periodic excitation from the vocal cords and the formant filtering of the vocal tract, which convolve in the time domain and multiply in the frequency domain, are additive and in different regions in the quefrency domain.

In this final step, we convert the log mel spectrum backto time. The result is the Mel Frequency Cepstrum Coefficients (MFCC).The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the discrete cosine transform (DCT). In this final step log mel spectrum is converted back to time. The result is called the Mel Frequency Cepstrum Coefficients (MFCC).The discrete cosine transform is done for transforming the mel coefficients back to time domain.

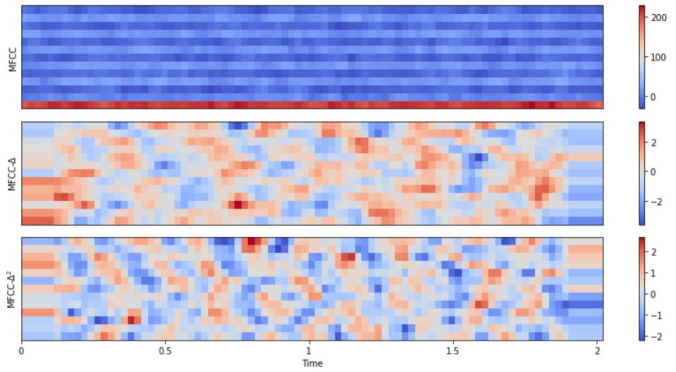

Fig. 4. Extracted Features.

IV. Inference

In this research we have successfully denoise the input sample and while extracting the MFCC coefficients we also taken into the consideration of Delta energy function and draw a conclusion that we can increase the MFCC coefficient according to our requirement. We can add velocity and acceleration to extract 39 MFCC coefficients. The MFCC feature extraction technique is more effective and robust, and with the help of this technique we can normalizes the features as well, and it is quite popular technique for isolated word recognition in English language. Features are extracted based on information that was included in the speech signal. Extracted features were stored in a cs file. In our future work we will do another breakthrough in the field of research, and will use these extracted MFCC coefficients for designing a speaker independent system type.

References

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first . . .”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes. Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

References

[1] Ahmed Salman, Ejaz Muhammad and Khurshid Khawar, ”Speaker verification using boosted cepstral features with gaussian distributions”, IEEE International Multitopic Conference 2007. INMIC 2007, pp. 1-5, 2007.

[2] M. A. amin and H. Yan, ”Sign Language Finger Alphabet Recognition from Gabor -PCA Representation of hand gestures,” presented at the Proceeding of the sixth Internaional Conference on Machine Learning and Cybernetics, Hong Kong, 2007.

[3] Anjali, A. Kumar and N. Birla, Voice Command Recognition System based on MFCC and DTW, International Journal of Engineering Science and Technology, 2(12),2010.

[4] N.N. Lokhande, N.S. Nehe and P.S. Vikhe , MFCC based Robust features for English word Recognition, IEEE, 2012.

[5] S. Dhingra, G. Nijhawan and P. Pandit, Isolated Speech Recognition using MFCC and DTW, International journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering,8(2), 2013.

[6] Goutam Saha and Malyaban Das,On Use of Singular Value Ratio Spectrum as Feature Extraction Tool in Speaker Recognition Application, CIT-2003, pp. 345-350, Bhubaneswar, Orissa, India, (2003).

[7] S. Furui, “An overview of speaker recognition technology, in Automatic Speech and Speaker Recognition (C.H. Lee, F.K. Soong, and K.K. Paliwal,eds), ch.2 pp.31-56Boston : Kluwer Academic, (1996).

[8] M.A.Anusuya, S.K.Katti, “Comparison of Different Speech Feature Extraction Techniques with and without Wavelet Transform to Kannada Speech Recognition”, International Journal of Computer Applications (0975 – 8887) Volume 26– No.4, July 2011.