A Comparative Study between MFCC and DWT Feature Extraction Technique

Pooja V. Janse
Dept. of CS & IT, Dr. B. A. M. University, Aurangabad-431004,India

Smita B. Magre
Dept. of CS & IT, Dr. B. A. M. University, Aurangabad-431004,India

Pratik K. Kurzekar
Dept. of CS & IT, Dr. B. A. M. University, Aurangabad-431004,India

R. R. Deshmukh
Dept. of CS & IT, Dr. B. A. M. University, Aurangabad-431004,India

Abstract

Past research in mathematics, acoustics, and speech technology have provided many methods for converting data that can be considered as information if interpreted correctly. In order to find some statistically relevant information from data, it is important to have mechanisms for reducing the information of each segment in the audio signal into features. These features should describe each segment in such a characteristic way that other similar segments can be grouped together by comparing their features. Pre-processing of speech signals is considered a crucial step in the development of a robust and efficient speech or speaker recognition system. This paper deals with comparative analysis of MFCC and DWT feature extraction technique.

Keywords

Speech Recognition, Mel frequency cepstral co-efficient (MFCC), DWT.

1. Introduction

Speech is one of the ancient ways to express ourselves. Today these speech signals are also used in various applications like biometric recognition technologies and communicating with machine. Speech is a unique form of audio data. It is a relatively simple and widely studied type of acoustic signal [1].

The information in speech signal is actually represented by short term amplitude spectrum of the speech wave form. This allows us to extract features based on the short term amplitude spectrum from speech. The fundamental difficulty of speech recognition is that the speech signal is highly variable due to different speakers, speaking rates, contents and acoustic conditions. So to extract feature we must use proper feature extraction technique.

1.1 Overview of Automatic Speech Recognition system

Automatic Speech Recognition (ASR) is the process of converting a speech signal to a sequence of words, by means of an algorithm. ASR system involves two phases. Training phase and Recognition phase. In training phase, known speech is recorded and parametric representation of the speech is extracted and stored in the speech database. In the recognition phase, for the given input speech signal the features are extracted and the ASR system compares it with the reference templates to recognize the utterance [2].

1.2 Modules of ASR system

Modules that are identified to develop a speech recognition system are as follows.
1) Speech Signal acquisition
2) Feature Extraction
3) Acoustic Modeling
4) Language & Lexical Modeling
5) Recognition

1.2.1 Speech signal Acquisition

Much of the success of a speech recording depends on the recording environment and microphone placement. Ideally, speech recordings should take place in soundproof studios or labs. MICROPHONES, PRAAT, AUDACITY, SPHINX, JULIUS are the various tools which are being used by researchers for recording speech database.

1.2.2 Feature Extraction

In speech recognition, feature extraction requires much attention because recognition performance depends heavily on this phase. The main goal of the feature extraction step is to compute sequence of feature vectors providing a compact representation of the given input signal. The feature analysis component of an ASR system plays a crucial role in the overall performance of the system. Many feature extraction techniques are available, these include

- Linear predictive analysis (LPC)
- Linear predictive Cepstral coefficients (LPCC),
- Perceptual linear predictive coefficients (PLP)
- Mel-frequency Cepstral coefficients (MFCC)
- Power spectral analysis (FFT)
- Mel scale Cepstral analysis (MEL)
- Relative spectra filtering of log domain coefficients (RASTA)
- First order derivative (DELTA)
- Discrete Wavelet transform (DWT)

Etc.

2. Feature Extraction Techniques

Theoretically, it should be possible to recognize speech directly from the digitized waveform. However, because of the large variability of the speech signal, it is better to perform some feature extraction that would reduce that variability. Particularly, eliminating various source of information, such as whether the sound is voiced or unvoiced and, if voiced, it eliminates the effect of the periodicity or pitch, amplitude of excitation signal and fundamental frequency etc.

Following sections are a comparative analysis of MFCC and Wavelet feature extraction techniques that are in use today, or that may be useful in the future, especially in the speech recognition area. These techniques are also useful in many areas of speech processing [3].

3. MFCC: Mel-Frequency Ceptral Coefficient

Mel Frequency Cepstral Coefficients (MFCC) is one amongst the most normally used feature extraction methodology in speech recognition. The use of Mel Frequency Cepstral Coefficients can be considered as one of the standard method for feature extraction. The use of about 20 MFCC coefficients is common in ASR, although 10-12 coefficients are often considered to be sufficient for coding speech.

The most notable downside of using MFCC is its sensitivity to noise due to its dependence on the spectral form. Methods that utilize information in the periodicity of speech signals could be used to overcome this problem, although speech also contains a periodic content [4]. Following figure shows steps involved in MFCC feature extraction.

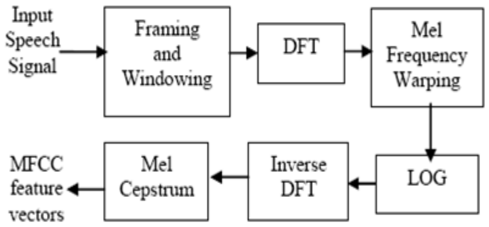

Figure 1: MFCC Feature Extraction

3.1 Methodology

The Mel frequency filter bank may be a series of triangular band-pass filters. The filter bank relies on a non-linear frequency scale referred to as the mel-scale. A thousand Hz tone is outlined as having a pitch of1000 mel. Below a thousand Hz, the Mel scale is more or less linear to the linear frequency scale.

The MFCC features correspond to the cepstrum of the log filterbank energies. To calculate them, the log energy is first computed from the filterbank outputs as

$$S_t\ [\mathrm{m}] = \ln\left(\sum_{n=0}^{N-1} |X_t\ [n]\ |^2\ H_m\ [n]\right) \qquad 0 \le \mathrm{m} < \mathrm{M},$$

Where Xt[n] is the DFT of the tth input speech frame, Hm[n] is the frequency response of mth filter in the filterbank, N is the window size of the transform and M is the total number of filters. Then, the discrete cosine transform (DCT) of the log energies is computed as

$$\overrightarrow{c_t}[\mathrm{m}] = \sum_{n=0} S_t[\mathrm{n}] \cos\left(\pi\, m\, (n - v.\text{ɔ}|M)\right) \quad 0 \le \mathrm{m} < \mathrm{M}$$

Since the human auditory system is sensitive to time 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.

$$\Delta\overrightarrow{c_t} = \vec{c}_{\mathrm{t+2}} - \vec{c}_{\mathrm{t-2}}$$

$$\Delta\Delta\overrightarrow{c_t} = \Delta\,\vec{c}_{\mathrm{t+1}} - \vec{c}_{\mathrm{t-1}}$$

These dynamic coefficients are then concatenated with the static coefficients these dynamic coefficients are then concatenated with the static coefficients $\vec{c}_k$ according to making up the final output of feature analysis representing the tth speech frame according to making up the final output of feature analysis representing the tth speech frame.

$$\vec{X}_{\mathrm{t}} = [\overrightarrow{c_t}\ \ \Delta\overrightarrow{c_t}\ \ \Delta\Delta\overrightarrow{c_t}\ ]^{\mathrm{T}}$$

Advantage

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

Disadvantage

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 [5].

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.

4. Discrete wavelet transform

The speech is a non-stationary signal. The Fourier transform (FT) is not suitable for the analysis of such non-stationary signal because it provides only the frequency information of signal but does not provide the information about at what time which frequency is present. The windowed short-time FT (STFT) provides the temporal information about the frequency content of signal. A drawback of the STFT is its fixed time resolution due to fixed window length. The WT, with its flexible time-frequency window, is an appropriate tool for the analysis of non-stationary signals like speech which have both short high frequency bursts and long quasi-stationary components also.

WT decomposes signals over translated and dilatedmother wavelets. Mother wavelet is a time function with finite energy and fast decay. The different versions of the single wavelet are orthogonal to each other. The continuous wavelet transform (CWT) is given by following equation where the function ψ(t), a, and b are called the (mother) wavelet, scaling factor, and translation parameter, respectively.

$$W_x(a, b)= \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\ \varphi(t-b|a)\ dt$$

As CWT is a function of two parameters, it contains high redundancy while analysing the signals. Instead of this, analysis of the signal using small number of scales with varying number of translations at each scale, i.e., discretizing scale and translation parameters as a = 2j and b = 2jk gives DWT. DWT theory requires two sets of related functions called scaling function and wavelet function given by

$$\Phi(t) = \sum_{n=0}^{N-1} h[n]\sqrt{2}\ \Phi(2t-n)$$

and

$$\varphi(t) = \sum_{n=0}^{N-1} g[n]\sqrt{2}\ \Phi(2t-n)$$

Where function Φ(t) is called scaling function, h[n] is an impulse response of a low-pass filter, and g[n] is an impulse response of a high-pass filter. The scaling and wavelet functions can be implemented effectively using a pair of filters, i.e., h[n] and g[n]. These filters are called a quadrature mirror filters that satisfy the property g[n] = (-1)(1-n) h[1-n]. The input signal is low pass filtered to give the approximate components and high-pass filtered to give the detail components of the input speech signal.

The approximate signal at each stage is further decomposed using same low-pass and high-pass filters to get the approximate and detail components for the next stage. This type of decomposition is called dyadic decomposition, whereas decomposition of detail signal along with the approximate signal at each stage is called uniform decomposition. Dyadic decomposition divides the input signal bandwidth into the logarithmic set of bandwidths, whereas the uniform decomposition divides it into the uniform set of bandwidths.

In speech signal, high frequencies are present very briefly at the onset of a sound while lower frequencies are presented latter for long period. DWT resolves all these frequencies well. The DWT parameters contain the information of different frequency scales. This helps in getting the speech information of corresponding frequency band. In order to parameterize the speech signal, the signal is decomposed into four frequency bands uniformly or in dyadic fashion.

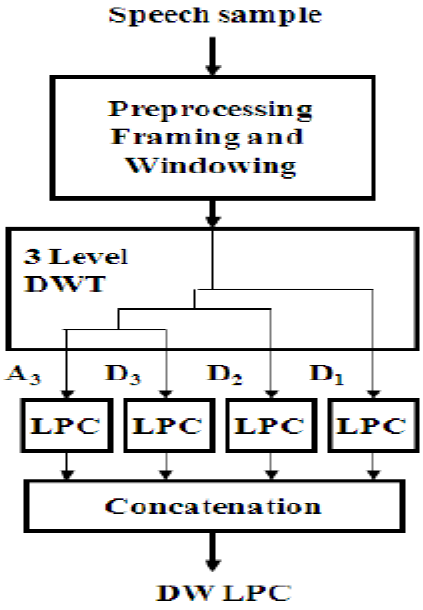

Figure 2: Block Diagram of Wavelet [6]

4.1 Advantages

- The main advantage of wavelets is that they offer a simultaneous localization in time and frequency domain.
- Wavelet transform have been used for speech feature extraction in which the energies of wavelet decomposed sub-bands have been used in place of Mel filtered sub-band energies. Because of its better energy compaction property.
- Wavelet transform-based features give better recognition accuracy than LPC and MFCC.
- Wavelets have the great advantage of being able to separate the fine details in a signal. Very small wavelets can be used to isolate very fine details in a signal, while very large wavelets can identify coarse details.
- The Wavelet Transform has a better capability to model the details of unvoiced sound portions.
- It has better time resolution than Fourier Transform.
- Wavelet is computationally very fast.
- A wavelet transform can be used to decompose a signal into component wavelets.

4.2 Disadvantage

- The cost of computing DWT as compare to DCT may be higher.
- It requires longer compression time.

5. Conclusion

Here in this paper we discussed MFCC and DWT features extraction techniques and their advantages and disadvantages. As speech recognition is used any many application some new methods can be developed using combination of more techniques to improve performance. There is a need to develop some new hybrid methods that will give better performance in robust speech recognition area.

6. References

[1] 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.

[2] Shanthi Therese S.,Chelpa Lingam, “Review of Feature Extraction Techniques in Automatic Speech Recognition”, International Journal of Scientific Engineering and Technology (ISSN : 2277-1581) Volume No.2, Issue No.6, pp : 479-484,June 2013.

[3] Urmila Shrawankar, “Techniques For Feature Extraction In Speech Recognition System : A Comparative Study”. International Journal Of Computer Applications In Engineering, Technology and Sciences (IJCAETS),ISSN, 6 May 2013.

[4] Leena R Mehta , S.P.Mahajan , Amol S Dabhade Comparative Study Of MFCC And LPC For Marathi Isolated Word Recognition System” International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering Vol. 2, Issue 6, June 2013.

[5] A. Shafik, S. M. Elhalafawy, S. M. Diab, B. M. Sallam and F. E. Abd El-samie, “A Wavelet Based Approach for Speaker Identification from Degraded Speech”, International Journal of Communication Networks and Information Security (IJCNIS) Vol. 1, No. 3, December 2009.

[6] Navnath S Nehe and Raghunath S Holambe,” DWT and LPC based feature extraction methods for isolated word recognition”, EURASIP Journal on Audio, Speech, and Music Processing 2012