International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 3
ISSN: 2454-4248
279 – 284

An Approach to Extract Feature Using MFCC for Isolated Word in Speaker Identification System

Mr. Sanjaya Kumar Dash[1]
1.Assistant Professor,
Department of Computer Science And Engineering,
Orissa Engineering College ,Bhubaneswar ,Odisha,.
E_mail id-Sanjaya_145@rediff.com.

Prof.(Dr.) Sanghamitra Mohanty[2]
2. Former Professor,
P.G. Department of Computer Sc. and Application,
Utkal University,Odisha,
(E_mail id-Sangham1@rediffmail.com.

Abstract: The speech is the prominent and natural form of communication among human being. There are different aspects related to speech like speaker identification, speaker recognition, Automatic speech recognition(ASR), speech synthesis etc. The purpose of this work is to study speaker identification system using Hidden markov Model (HMM).The goal of Speaker Identification System (SIS) is to determine which speaker is speaking based on spoken information. The system uses Mel Frequency Cepstral Coefficients(MFCC) for feature extraction , HMM for pattern training and viterbi techniques. The success of MFCC combined with their robust and cost effective combination turned them into a standard choice in speaker identification system.HMM and viterbi decoding provide a highly reliable way of recognizing odia speech.

Key Words: *Mel Frequency Cepstral Coefficients(MFCC),Hidden Markov Models, speaker identification, speech recognition, Fast Fourier Transformation ,Discrete cosine transformation(DCT), Viterbi techniques,spectrum*.

*****

Introduction:

In speaker identification system, powerful tool of the information exchange using acoustic signal is used. Therefore, the speech signal for several decades is the subject of research. Speaker Identification is a technology that ables a computer to capture the words spoken by a human with help of microphone. After processing these words, the computer can identify speaker on recognizing the words of the speaker. 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[5][6].

1.1 pre processing the speech signal

The purpose of this work is to convert the speech waveform to a set of features (at a considerably lower information rate) for further analysis.

The speech signal is a slowly timed varying signal (it is called *quasi-stationary*). When examined over a sufficiently short period of time (between 5 and 100 msec), its characteristics are fairly stationary. However, over long periods of time (on the order of 1/5 seconds or more) the signal characteristic change to reflect the different speech sounds being spoken. Therefore, *short-time spectral analysis* is the most common way to characterize the speech signal.

Introducing the power spectrum of the signal $P_x(w)$,of the excitation $P_v(w)$ and the spectrum of the vocal track filter $P_h(w)$,

we have:

$$P_x(w) = P_v(w)\, P_h(w) \qquad (1.1)$$

Where w is frequency of discrete time signal. The spectrum of the filter can be obtained from power spectrum of speech $P_x(w)$ the contribution of the excitation power $P_v(w)$.

A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), and others. MFCC is perhaps the best known and most popular one.

MFCC's are based on the known variation of the human ear's critical bandwidths with frequency-filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. This is expressed in the *mel-frequency* scale, which is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. The process of computing MFCCs is described in more details afterward.

1.2 Mel-frequency cepstrum coefficients (MFCC) processor

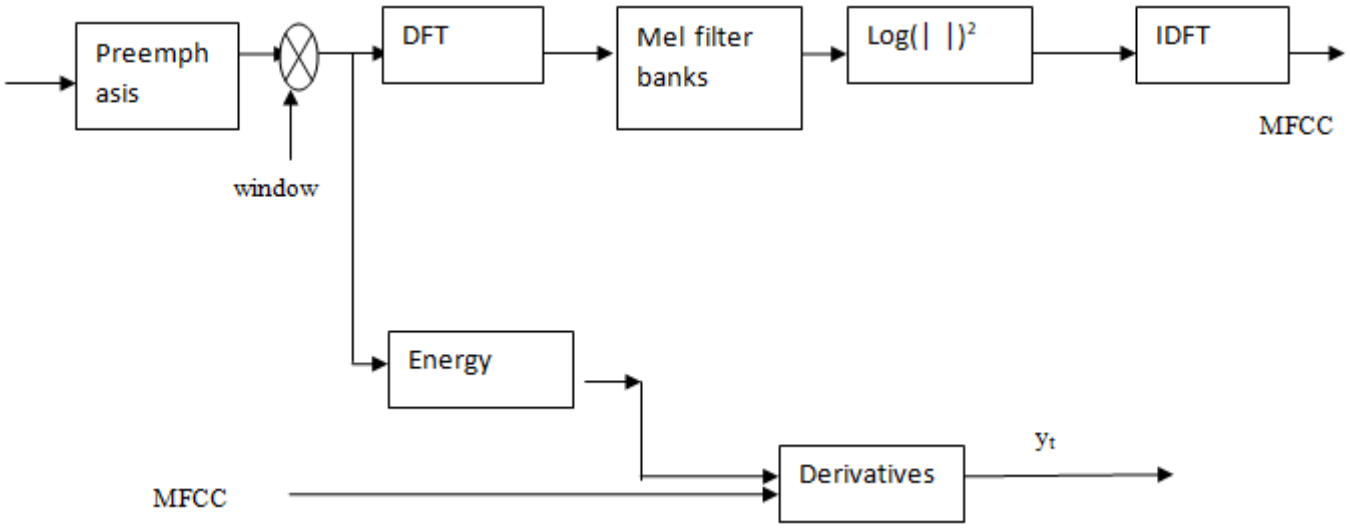
A block diagram of the structure of an MFCC processor is given in Figure-1. The speech input is typically recorded at a sampling rate above 10000 Hz. This sampling frequency was chosen to minimize the effects of *aliasing* in the analog-to-digital conversion. These sampled signals can capture all frequencies up to 5 kHz, which cover most

 ISSN: 2454-4248
279 – 284

energy of sounds that are generated by humans. As been discussed previously, the main purpose of the MFCC processor is to mimic the behavior of the human ears. In addition, rather than the speech waveforms themselves, MFFC's are shown to be less susceptible to mentioned variations.

1.3 Feature Extraction

In this section, the whole processing chain returning the signal features is described. Although many other processing schemes are possible, the one depicted here has generally obtained a better accuracy and a minor computational complexity with respect to alternative processing[2].



(Figure-1 Block Diagram Of MFCC Processor)

1.3.1 Signal preprocessing

The characteristics of the vocal tract define the current uttered phoneme. Such characteristics are evidenced in the frequency domain by locations of formants, i.e. the peaks given by resonances of the vocal tract. Although possessing relevant information ,high frequency formants have smaller amplitude with respect to low frequency formants. A preemphasis of high frequencies is therefore required to obtain similar amplitude for all formants. Such processing is usually obtained by filtering the speech signal by filtering the speech signal with a first order FIR filter whose transfer function in the z-domain [1] is:

$$H(z) = 1 - a.z^{-1} \qquad 0 \le a \le 1 \qquad (1.1.2)$$

Here a is preemphasis parameter.

In the time domain ,the preemphasized signal is related to input signal by the relation:

$$x'(n) = x(n) - ax(n-1) \qquad (1.1.3)$$

A typical value for n is 0.95, which gives rise to more than 20 dB amplification of high frequency spectrum. HMM based SIS may experience a significant reduction in performance if temporarily long silences are not removed from speech. Since these silences should not be processed by SIS, effective speech detectors are required.

1.3.2windowing

Traditional methods for spectral evaluation are reliable in the case of a stationary signal (i.e. a signal whose statistical characteristics are invariant with respect to time). For voice, this holds only within the short time intervals of articulatory stability, during which a short time analysis can be performed by "windowing" a signal $x'(n)$ into a succession of windowed sequences $x_t(n)$,t=1,2,....,T, called frames, which are then individually processed:

$$x'_t(n) \equiv x'(n - t.Q), \quad 0 \le n < N, \ 1 \le t \le T \qquad 1.1.4$$

$$x_t(n) \equiv w(n).x'_t(n) \qquad 1.1.5$$

where $w(n)$ is the impulse response of the window. Each frame is shifted by a temporal length Q. If Q = N, frames do not temporally overlap while if Q<N, N-Q samples at the end of a frame $x'_t(n)$ are duplicated at the be-ginning of the following frame $x'_{t+1}(n)$. We recall that Fourier analysis is performed through the Fourier transform that for a discrete time signal $x_t(n)$ is:

$$X_t(e^{jw}) = \sum_{n=0}^{N-1} x_t(n)e^{-jwn} = F\{x_t(n)\} \qquad 1.1.6$$

where w is the continuous frequency axis. Introducing the Fourier transform of $w(n)$ and $x_t'(n)$:

 ISSN: 2454-4248
279 – 284

we can have Fourier Transform of $w(n)$ and $x'_t(n)$:
$W(e^{jw})= F\{w(n)\}$ $X'_t(e^{jw})=F\{x'_t(n)\}$, a product in the time domain as (1.1.5) becomes a convolution in the frequency domain:

$$X_t(e^{jw}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X'_t(e^{j\theta}).W(e^{j(w-\theta)})d\theta = F\{x'_t(n)w(n)\} \quad 1.1.7$$

Considering (1.1.4) and (1.1.5), (1.1.6) can be written as:

$$X_t(e^{jw}) = \sum_{n=-\infty}^{n=\infty} x'(n - t.Q).w(n)\, e^{-jwn} \quad (1.1.8)$$

Formula (1.1.8) is also referred to as Short Time Fourier Transform (STFT) or Windowed Fourier Transform (WFT) of x'(n).
The simplest window has a rectangular shape. This window is implicitly used when a sequence of N samples is retrieved from a signal:

$$w(n) = \begin{cases} 1 & 0 \le n \le N-1 \\ 0 & \text{otherwise} \end{cases} \quad (1.1.9)$$

The presence of a window provokes a distortion on the estimated spectrum, since $X_t(e^{jw})$ is the convolution of the spectrum of $x'_t(n)$ and of the Fourier transform of the rectangular window w(n). $W(e^{jw})$ is composed of a higher energy main lobe centered at the zero frequency and of lower energy side lobes centered at higher frequencies. The main lobe spreads out in a wider frequency range the narrow band power of the signal $x'_t(n)$ that in our case is represented by the formants. This phenomenon reduces the local frequency resolution. Moreover, the side lobes of $W(e^{jw})$ swap energy from different and distant fre-quencies of $x'_t(n)$ This problem is called leakage.
To reduce such effects, $x_t(n)$ is multiplied by a properly shaped window w(n). The choice of w(n) is a trade-off between several factors:

- the window shape may reduce distortion, but it may increase signal shape alteration
- the length N is proportional to the frequency resolution and inversely proportional to the time resolution
- the overlap N-Q is proportional to the frame rate, but it is also proportional to the correlation of subsequent frames.

In ASR, the most-used window shape is the Hamming window, whose impulse response is a raised cosine impulse:

$$w(n) = \begin{cases} 0.54-0.46\cos\left(\frac{2\pi n}{N-1}\right) & n=0, ...,N-1 \\ 0 & \text{otherwise} \end{cases} \quad (1.1.10)$$

The side lobes of this window are much lower than the rectangular window (i.e. the leakage effect is decreased) although resolution is appreciably reduced. This is because the Hamming main lobe is wider.
The Hamming window is a good choice in speech recognition, because a high resolution is not required, considering that the next block in the feature extraction processing chain integrates all the closest frequency lines. In contrast, leakage has a negative effect since vocal tract characteristics are obtained considering the location and the amplitude of the peaks at distant frequencies. Regarding the length N, widely used windows have 10-25 ms length. The window length is chosen as a compromise solution between the required time and frequency resolution. Short time (3-5 ms) windows allow the detection of the amplitude decay of formants, but have a too great impact on the frequency resolution that is required to estimate formant positions and therefore phoneme characterization. The length N of the window defines the spectral resolution of the Fourier representation. Considering the sampling period $T_c= 1/ f_c$, it follows that by sampling the transformed sequence on the w-axis at $2\pi/N$ equally spaced points, the analog frequency resolution Δf is:

$$\Delta f = \frac{\Delta w}{2\pi Tc} = \frac{2\pi}{2\pi TcN} = \frac{fc}{N} \quad (1.1.11)$$

Once sampling frequency fc is fixed, the spectral resolution is inversely proportional to the sequence length N. A narrow-band spectrum is one obtained when resolution is high, while a wide-band one is obtained when the resolution is low .
Increasing resolution is equivalent to using a longer sequence and this is in contrast to the requirement to analyze stationary signal segments. A trade-off between these two requirements is necessary. For instance, in the case of, $fc =$ 20 kHz, the longest sequence compatible with stationarity should be composed of at most 512 samples (512/20 =25.6 ms),while the shortest one compatible with resolution should be composed of 64 samples (64/20 = 3.2 ms).
Moreover, larger windows (about 70 ms) have a higher frequency resolution. This allows identification of each single harmonic. However, in such a case, fast transitions in the spectrum (as for instance the pronunciation of stop consonants) are not detected. Narrow windows have been proposed to estimate the fast varying parameters of the vocal tract; while large windows are used to estimate the

 ISSN: 2454-4248
279 – 284

fundamental frequency. A 20-30 ms long window is generally a good compromise.

1.3.3 Mel frequency cepstrum computation

The final procedure for the Mel frequency cepstrum computation (MFCC) consists of performing the inverse DFT on the logarithm of the magnitude of the filter bank output:

$$y_t^{(m)}(k)= \sum log \{ \left| Y_t(m) \right| \}.cos\left(k(m - \frac{1}{2})\frac{\pi}{m}\right) \quad , \quad k = 0, \ldots, L \qquad (1.1.12)$$

The procedure has great advantage. First,we find that since the log power spectrum is real and symmetric then the inverse DFT reduces to a Discrete Cosine Transform (DCT). The DCT has the property to produce highly uncorrelated features $y_t^{(m)}(k)$ [7]. Therefore, the stochastic characterization of the feature process is simpler and in the probability density functions of the features, generally modeled by linear combinations of Gaussian functions, diagonal covariance matrices can be used instead of full covariance matrices.

1.3.4 CEPSTRUM ANALYSIS

The complex cepstrum (the name is an anagram of spectrum) $x(n)$ for a discrete signal $x(n)$ is the inverse fourier transform of the complex logarithm log $X(e^{jw})$ [1]:

$$\boldsymbol{x}(n) = F^{-1}\{log\, X(e^{jw})\} \qquad (1.1.13)$$

The logarithm of the spectrum has the effect of reducing the component amplitudes at every frequency. This logarithmic scale is also a characteristic of the human hearing system. Therefore, those signals that, are characterized by a combination of harmonics are better analyzed by the cepstrum rather than by the spectrum or the autocorrelation.

The use of the cepstrum was first introduced to discriminate voiced (vowels, sonorants) and unvoiced (plosives), affricates, etc.) speech segments. In fact, the cepstrum emphasizes the formants of the vocal tract, even with noise. In contrast, the cepstrum is flat for sounds that lack a clear harmonic structure. By exploiting these properties, the cepstrum coefficients have been used to classify voice segments, determining an evolution of the cepstrum technique. Indeed the cepstrum analysis, that is an homomorphic analysis with a logarithm as the intermediate function, allows deconvolution of the speech signals as explained below.

As already pointed out, a speech wave form $x(n)$ can be considered as a convolution between the excitation produced by the vocal cords $v(n)$ and the impulse response of a filter that represents the vocal tract $h(n)$:

$$x(n) = v(n)*h(n) \qquad \ldots(1.1.14)$$

Since the phonetic information is mainly related to the shape of the vocal tract, deconvolution algorithms for speech signal are of considerable interest to isolate the response of the vocal tract. These algorithms belong to the system theory branch known as *homomorphic filtering* [1]. Resorting to the complex cepstrum we have:

$$\boldsymbol{x}(n) = F^{-1}\{log(F\{v(n)*h(n)\})\}= \qquad (1.1.15)$$
$$= F^{-1}\{log(V(e^{j\omega})) + log(H(e^{j\omega}))\} = \mathbf{v}(n) + \mathbf{h}(n)$$

Where $\boldsymbol{v}(n)$, $\boldsymbol{h}(n)$ are the complex cepstrum of v(n) and h(n) respectively. The complex cepstrum transforms the convolution (1.1.14) into a sum of the two components $\boldsymbol{v}(n)$, $\boldsymbol{h}(n)$ that can be separated by band-pass linear filters, if there is no frequency overlapping. The most popular feature extraction technique is Mel Frequency Cepstral coefficients called MFCC as it is less complex in implementation and more effective and robust under various conditions[4].

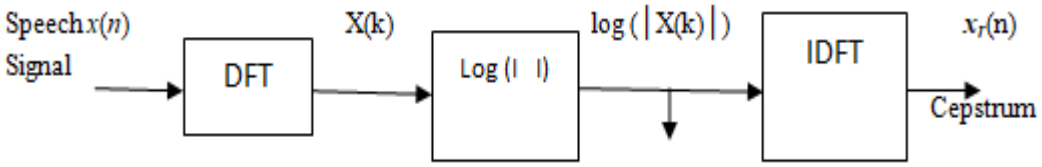
For speech signals, this is feasible since the short time spectrum shows that the envelope of the vocal tract filter h(n) changes slowly with respect to the fine structure of the harmonics that are produced by the periodic excitation of speech v(n).

For minimum phase signals or when phase information is not of interest the real cepstrum $\boldsymbol{x_r}$ (n) can be used instead of the complex cepstrum. The real cepstrum of a signal is defined by the inverse Fourier transform of the logarithm of the magnitude of $X(e^{jw})$:

$$\boldsymbol{x_r}(n) = F^{-1}\{log \left| X(e^{jw}) \right| \}$$

As shown in figure-2 , the real cepstrum can be computed using the inverse. DFT. The scheme recalls that used to compute the MFCC. The relevant difference between the two schemes is that for MFCC, the cepstrum is computed on a perceptually-weighted spectrum obtained from the filter banks.

ISSN: 2454-4248
279 – 284



(Figure- 2 Block diagram of real cepstrum signal processing analysis.)

Homomorphic deconvolution expressed in (1.1.15) may highlight relevant properties of MFCC. First,we note that if a multiplicative constant is applied to a speech signal, the logarithm of such a constant is added to all the coefficients of log $|Y_t(m)|^2$ Such a constant influences only the zero coefficient $y_t^{(m)}(0)$ of the MFCC. Therefore, MFCC are not sensitive to a gain factor, apart from $y_t^{(m)}(0)$. We find also that vocal tract response and signal excitation are combined additively in cepstrum as shown in formula (1.1.15). The vocal tract log spectrum has a smooth behavior while excitation has a highly variant quasiperiodic spectrum for voiced sounds. Thus, vocal tract response can be obtained by simply retaining the first cepstral coefficients $Y_t^{(m)}(k)$. That is why only *k*-th coefficients, $k \leq L \leq 15$, are retained. Note also that environment influence can be modeled as a linear filter. This degradation becomes a bias on the log spectrum estimation that can be evaluated and removed. The i^{th} order time-difference of a generic vector indexed in time t can be computed to capture dynamic evolution of speech signal[3].

Preemphasis from frequency 50 Hz was applied to different speech signals. Number of coefficients =12. Window length = 0.015 s. Time step =0.005 s. As the value of C0 has little significance, hence I have discarded all C0 values.
$Freq_{min}$ = 0 Hz $freq_{max}$ =4000 Hz.
Maximum number of coefficients = 38.Out of 12 frames one frame values are displayed in below table. Here Praat software is used.
Filter bank parameter:
Position of first filter(mel)= 100.0
Distance between filters(mel)=100.0

COEFFICIENTS	/cheer/- Frame1- female	/chira/- Frame1- Female	/cheer/- frame1- male	/chira/- frame1- male
C1	C1= -429.5676099163043	C1= -405.76683377390225	C1= -524.2993540582794	C1= -485.54890094553394
C2	C2= -170.59969804729425	C2= -162.6551948039764	C2= -129.60890563199888	C2= -126.78672214289683
C3	C3= 60.35912662855221	C3= 84.81666752362574	C3= 167.75311850059555	C3= 177.1145819591884
C4	C4= 20.22979613777331	C4= 49.14557237325349	C4= -54.998882058578545	C4= -27.021714327467492
C5	C5= -97.04502846665999	C5= -62.81275538943707	C5= 38.651485444702274	C5= 1.0926856057878702
C6	C6= -5.003715556006501	C6= -21.865505130094927	C6= -17.84894074513785	C6= -30.494811105288843
C7	C7= 71.82465214467291	C7= 55.229555576613436	C7= 65.58623716440904	C7= 22.407632235527974
C8	C8= -30.981916952546122	C8= -37.796357752353266	C8= -35.646690618844836	C8= -17.750892925898235
C9	C9= -3.092143145283887	C9= -9.48698218193883	C9= -17.288182618289706	C9= 28.932769362019428
C10	C10= 27.429959629789778	C10= 0.7352037912223158	C10= -1.3181106898714465	C10= 26.6909698282476
C11	C11= 19.96368308255336	C11= -19.866975957512675	C11= 45.40986864311749	C11= 2.0262075168723035
C12	C12= -14.602207026840409	C12= -30.13812077665352	C12= -27.32485256058891	C12= -41.30907396308576

(Table-1 for MFCC coefficients of male and female speaker)

ISSN: 2454-4248
279 – 284

1.4Conclusion:

These MFCC values as the feature extraction can be used for training and testing phase of SIS as these values are noise robust. The extracted MFCC's for speech samples are given to pattern trainer for training and are trained by HMM to create HMM model for each word. Then viterbi decoding can be used to select the one with maximum likelihood for the identification of speaker.

Referrence:

[1]. Oppenheim A.V.,Shafer R.W, Digital Signal Processing,Prentice Hall,1989

[2]. Davis S.B.,Mermelstein P., "Comparison of parametric representation of monosyllabic word recognition in continuously spoken sentences ",IEEE trans. Acoustics,Speech and Signal Processing,28,pp.357-366(1980)

[3]. furui s., "Cepstral analysis techniques for automatic speaker verification", IEEE Tran. On ASSP, 29, No.2, pp.254-272 (1981).

[4]. C. Poonkuzhali, R. Karthiprakash, S. Valarmathy and M. Kalamani, An Approach to feature selection algorithm based on Ant Colony Optimization for Automatic Speech Recognition, *International journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 11(2),* and 2013.

[5]. V. Sharma and P. Sharma, Discrete and continuous Mouse Motion using Vocal and Non-Vocal Characteristics of Human Voice, *International journal of Computer Science and Engineering Technology,4,*2013.

[6]. C. Ittichaichareon, S. Suksri and T. Yingthawornsuk, speech Recognition using MFCC, *International Conference on Computer Graphics Simulation and Modeling,* 2012.

[7]. Jayant N.O.s., Noll p., Digital Coding of waveforms, prentice Hall (1984).