IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 4 Issue 4, April 2017
ISSN (Online) 2348 – 7968 | Impact Factor (2016) –5.264
www.ijiset.com

Analysis of Feature Extraction Methods for Speech Recognition

Athira menon.G[1], Anjusha.V.K[2]

[1] Department of ECE, A.P.J Abdul Kalam Kerala technical university /Thejus engineering college, Thrissur, India
[2] Department of ECE, A.P.J Abdul Kalam Kerala technical university /Thejus engineering college, Thrissur, India

Abstract

Speech is the nicely recognized and natural form of verbal communication among human beings. On the planet, there are different languages that are spoken by human being for efficient communication. Then people groups also like to interact with machine. This should be possibly by speech recognition. In speech recognition, feature extraction is the most imperative phase. It is considered as the heart of the structure. The work of this is to extract those features from the input speech that help the system in identifying the speech. Fundamental target of this paper is analysis and summarize most broadly utilized element extraction procedures like linear predictive coding (LPC).Linear predictive cepstral coefficient (LPCC), Perceptual linear prediction (PLP) and Mel frequency cepstral coefficient (MFCC).

Keywords:Speech recognition, Feature extraction, Linear predictive coding, Linear predictive cepstral coefficients, Perceptual linear prediction, Mel frequency cepstral coefficient..

1. Introduction

Speech recognition means that the capacity to hear talked words and extracts the exact sound, and recognizes them as word of some known language. Speech signals are quasi stationary signals means that when speech signals inspected over a brief time frame, its attributes are stationary, but for a long period of time the signal characteristics changes. Feature extraction is the most critical stage in speech recognition framework, this converts the discourse waveform in to some type of parametric representation which is utilized for further analysis and processing [1].

This Paper gives the summarization and analysis of most widely used feature extraction methods. The paper is divided as follows: section 2 describe about feature extraction. Section 2.1 describe about commonly used features. Section 3 describe about 4 types of feature extraction methods, section 4 describe comparison table of different feature extraction method .conclusion is mention in section 5

2. Feature Extraction

Feature extraction is an essential and basic step of speech recognition. It is an exceptional type of dimensionality lessening technique which is used to reduce the information which is extensive to be prepared by calculation [1]. In speech recognition, Feature extraction is the way towards holding valuable data of the signal while disposing of repetitive and undesirable information. It is the parameterization of the speech signal. It is the procedure of changing the signal to digital form, measuring some imperative character of the signal, for example, energy and frequency responses. The main objective of feature extraction is obtaining the set of features with low rate of change in order to keep the computational feasible [1].

Extracted feature should meet few criteria while consulting with the speech signal such as [4]

- Large between-speaker and small within-speaker variability
- Not change over time
- Be difficult to impersonate/mimic
- Not be affected by background noise nor depend on the specific transmission medium
- Occur naturally and frequently in speech.

It is not possible that a single feature would meet all the criteria listed above. Thus, a large number of features can be extracted and combined to improve the accuracy of the system.

2.1 Features

In speaker acknowledgment, the features got from the vocal tract trademark are most ordinarily used. These features can be gotten from the spectrogram of the speech signal, in this way are classified as Short Term Spectral Features. Formants are useful for assessment of text to speech systems. Peaks indicate dominant frequency components in the speech signal. Vocal tract resonances, additionally called formants are the peaks of the spectral envelope. The resonances frequencies (formants) are contrarily corresponding to the vocal tract length Formants convey the character of the sound.

Pitch is another sort of highlight. It begins from the vocal strings. When air flow from glottal through the vocal cords, the vibration of the vocal cords convey pitch harmonics. Rate at which the vocal folds vibrate is the frequency of the pitch. So when the vocal folds sway at 300 times each second, they are said to create a pitch of 300 Hz .Some different features are voiced and unvoiced information, short term energy and zero crossing rate etc.[6]

3. Feature extraction methods

In speech recognition, the primary point of the feature extraction is to compute a sequence of feature components providing a reduced portrayal of the given input signal. Commonly LPC, MFCC, LPCC and PLP are used as feature extraction techniques for speech recognition system.

3.1 Linear predictive coding

Linear predictive coding (LPC) is a device used generally in audio signal processing and speech processing for representing the spectral envelope of speech in compressed form, using the information of a linear predictive model. It is one of the most powerful speech analysis techniques. The linear predictive model is based on mathematical approximation of the vocal tract represented by this tube of a varying diameter. At a particular time n, the speech sample S(n) is represented as a direct whole of the P previous samples.

The essential idea driving linear predictive model is that a given speech sample at time n ,S(n) can be approximated as a linear combination of the past P speech samples such that,

$S(n)=a_1 s(n-1)+a_2 s(n-2)\cdots\cdots\cdots\cdots\cdots a_p s(n-p)$ (1)

Where the coefficients $a_1, a_2, a_3 \cdots\cdots\cdots a_p$ are assumed constant values over the speech analysis frame.

Convert equation to an equality by including the excitation Gu(n), such that

$S(n)=\sum_{i=1}^{p} a_i s(n-i)+Gu(n)$ (2)

Where u(n) is a normalized excitation and G is gain of excitation ,By expressing equation 2 in the Z domain we get the relation and linear predictive model shown in the figure.

$S(z)=\sum_{1=1}^{p} a_i z^i s(z)+Gu(z)$ (3)

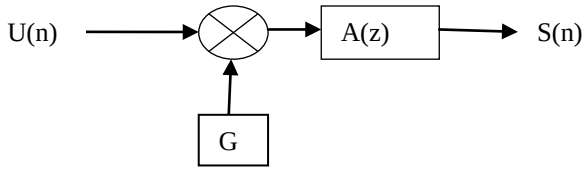

Fig.1 Linear predictive model

LPC begin with the presumption that the speech signal is produced by buzzer at the end of the tube. The space between the vocal folds called glottis produce buzz, which is featured by its intensity and frequency, which depend on the pitch of the sound. The vocal tract, which is described by its resonance frequency called formants. LPC analysis and characterize the speech signal frames by assessing the formants, removing their effects from the speech signal and estimating the intensity and frequency of the remaining sound. The process of removing the formants is called inverse filtering, and the remaining signal is called as residue. The fundamental issue of LPC is to find out the formants from the speech signal called a linear predictor, which expresses each sample of the signal as linear combination of previous samples. The coefficients of the difference equation, the prediction coefficients, characterize the formant. The coefficients are assessed by limiting the mean square error between the predicted signal and actual signal.

3.2 Linear predictive cepstral coefficients

Linear predictive cepstral coefficients are linear predictive coding coefficients presented in cepstral domain. For evaluating the fundamental parameters of a speech signal, LPCC has become one of the predominant techniques.A cepstrum is the inverse Fourier transform of the estimated spectrum of the signal.Linear Predictive Coding is used to obtain the LPC coefficients from the speech tokens. The LPC coefficients are then translated to cepstral coefficients. Once LPC vector is acquired, then possible compute linear predictive cepstral coefficients using recursion formula.The input signal is pre emphasised first using first order high pass filter. Since the energy

contained within a speech signal is distributed more in the lower frequencies than in the higher frequencies. In order to boost up the energies in high frequencies, pre-emphasis of the signal is done. Then this signal is blocked in to frames. In order to reduce the signal discontinuities at the edges of the frame, windowing of the signal is performed. Last stage of this algorithm is cepstral analysis which refers to the process of finding out the cepstrum of speech sequence. Basically there are two types of cepstral approaches: FFT cepstrum and LPC cepstrum. In the former case the real cepstrum is defined as the inverse FFT transform of the logarithm of the speech magnitude spectrum. However, one more method for estimating these cepstral coefficients is from the LPC via a set of recursive procedure and the coefficients thus obtained are known as linear prediction cepstral coefficients. [11], [12]

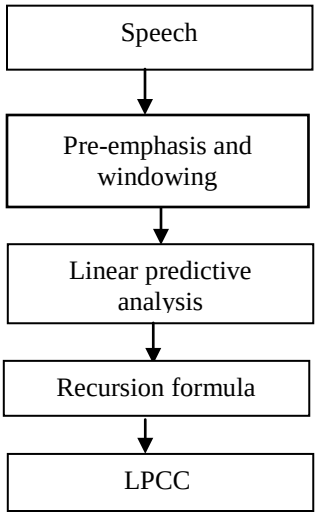

Fig 2 Linear predictive cepstral coefficients

Recursion formula:

$c_0=\ln G^2$ (4)

$c_m=a_m+\sum_{k=1}^{m-1}\frac{k}{m}c_k a_{m-k}$ $1\leq m\leq p$ (5)

$c_m=\sum_{k=1}^{m-1}\frac{k}{m}c_k a_{m-k}$ $m>p$ (6)

$G^2$is gain term in LPC model.$c_m$is cepstral coefficient,a_m is predictor coefficient. From a finite number of LPC coefficient, an infinite number of LPC coefficients, an infinite number of cepstral coefficient can be calculated.

3.3 Mel Frequency Cepstral coefficients

Mel Frequency Cepstral Coefficients (MFCC) is the robust and dynamic technique for speech feature extraction [1]. The MFCC are based on the known variation of the human ear's critical bandwidth frequencies with filters spaced linearly at low frequencies and logarithmic at high frequency. Basically MFCC models the spectral energy distribution of speech signal on the perceptual meaningful way. It is based on human hearing perception which cannot perceive frequencies over 1Khz.MFCC has two type filter, which are linearly spaced at low frequency below 1000 Hz and logarithmic spacing above 1000Hz which is called Mel scale.Fig.3 shows the complete block diagram of MFCC.The following formula is used to compute the Mels for a particular frequency.[2]

$Mel(f)=2595*\log10(1+f/700)$ (7)

Pre-emphasis

The Pre-emphasis means the passing of signal through filter which emphasizes higher frequencies. This will increase the energy of signal at higher frequency. Pre-emphasis is needed because high frequency comments of speech signal have small amplitude with respect to low frequency components. It is basically boosting energy in the high frequencies. The spectrum for voices segments has more energy at low frequency than high frequencies. This is called spectral tilt .spectral tilt is caused by the nature of glottal pulse. Boosting high energy gives more information to acoustic model.

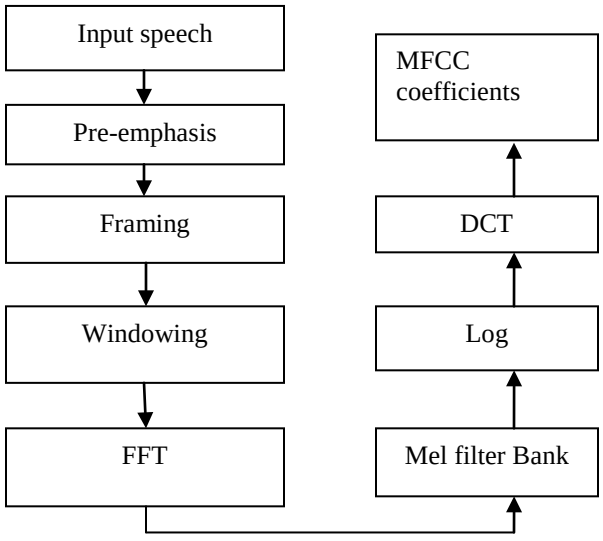

Fig 3 Block diagram of MFCC

Framing

The second step is framing. The width of the frame is generally about 30ms with an overlap of about 20ms.If the frame is much shorter we don't have enough samples to get reliable spectral estimate. If it is longer, the signal changes too much throughout the frame.

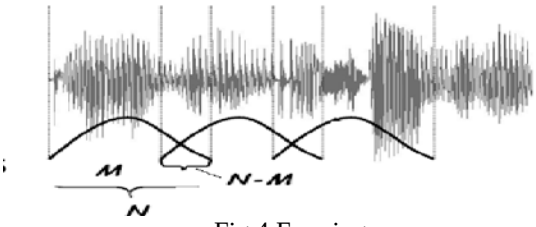

Fig.4 Framing

Windowing

The next step is windowing. The window function is used to smooth the signal for the computation of the FFT. The discontinuity in the frame is prevented. Due to windowing attenuate both ends of the signal towards zero, so this unwanted discontinuity can be avoided. Windowing functions commonly used: Hamming, Hanning, Blackman, Gauss, rectangular, and triangular .The hamming window is usually used in speech signal spectral analysis, spectrum falls off rather quickly so the resulting frequency resolution is better, which is suitable for detecting formants.

Fast Fourier transform

Then fast Fourier transform which convert each frame of N samples frame time domain into frequency domain .it is used for calculate power spectrum. Identify which frequencies are presented in each frame.

Mel filter bank

The important step is Mel filter bank processing. Human hearing is not equally sensitive to all frequency bands. Less sensitive at higher frequencies roughly greater than 1000Hz that means human perception of frequency is nonlinear. In Mel scale, Mel (melody) is a unit of the pitch. Mel scale is linear up to the frequency of 1 KHz and then become close to logarithmic for higher frequencies. Human ear act as a filter that concentrate only on certain frequency components. These filters are non-uniformly spaced on the frequency scale, with more filters in the low frequency regions and less filter in the high frequency regions. Mel scale is proposed by stevens,volkman and newman.[3]

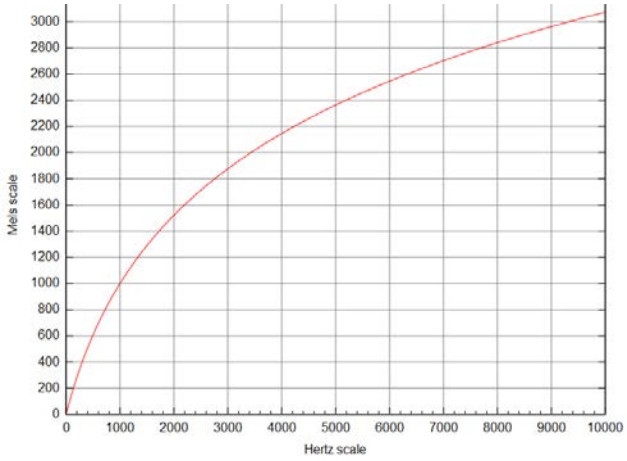

Fig.5 Mel scale[3]

Discrete cosine transform

The last step is discrete cosine transform. Thus is the process to convert the log Mel spectrum in to time domain using discrete cosine transform. The DCT de -correlate the energies. The result of the conversion is called Mel frequency cepstrum coefficients. Advantages of MFCC are the recognition accuracy is high. That means the performance rate of MFCC is high. MFCC captures main characteristics of phones in speech[6], [14].

3.4Perceptual Linear Prediction

Perceptual linear prediction model developed by Hermansky. The objective of the PLP model is to depict the psychophysics of human hearing more accurately in the feature extraction process. In contrast to pure linear predictive analysis of speech, perceptual linear prediction modifies the short term spectrum of the speech by several psychophysically based transformations.PLP inexact following main perceptual aspects namely: [9]

- Power spectrum computed from windowed signal using FFT.
- Then Bark scale is applied in it .Bark scale is another type of perceptual meaning full scale.
- An equal-loudness means that the filter-bank outputs to simulate the sensitivity of hearing**.**
- The equalized values are transformed according to the power law of Stevens by raising each to the power of 0.33

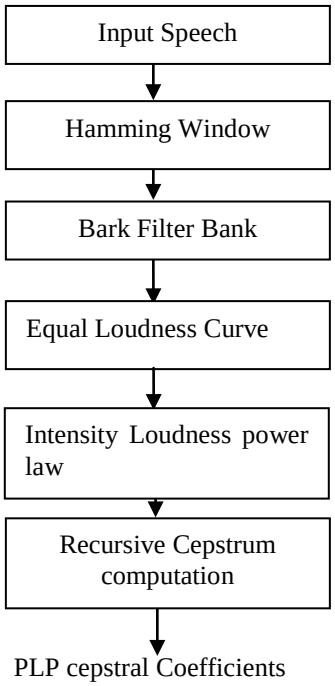

Fig.6 Block Diagram of PLP

Mel scale cepstral analysis is very similar to perceptual linear predictive analysis of speech, where the short term spectrum is modified based on psychophysically based spectral transformation. In MFCC, the spectrum is warped according to the Mel scale, whereas in PLP the spectrum is warped according to the Bark scale.

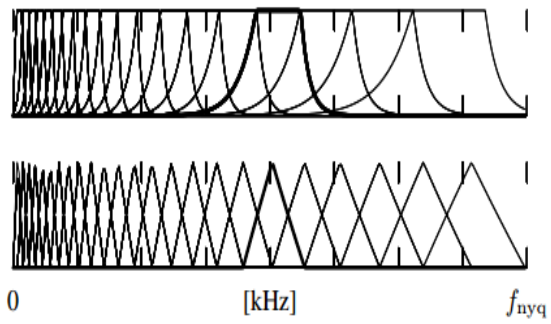

Fig.7 comparison of Bark scale and M el scale [9]

The first figure represent Bark scale based filter and second graph based on Mel scale Both allocate more filters to the lower frequencies, where hearing is more sensitive [13].

4. Analysis of feature extraction techniques

Techniques	Principle Based	Merits and demerits
Linear Predictive coding	All pole model	(a) Based on basic principle of sound production (b) Performance Degradation in presence of Noise (c)Computation speed of LPC is good. (d) useful for encoding low bit rate.
Linear Predictive cepstral coefficients	All pole model	(a)Give smoother spectral envelope and stable representation as compared to LPC
Mel –frequency cepstralcoefficient	Mel scale based	(a)Use Mel spaced filter banks hence behave more like a human ear. (b)simple process
Perceptual Linear production	Bark scale based	(a) Model is to describe the psychophysics of human hearing

5. Conclusions

Feature extraction is a urgent stride of the speech recognition process. Mel scale cepstral coefficients, Perceptual linear prediction, linear predictive cepstral coefficients and linear predictive coding are the most proposed acoustic features. By concentrate each of these techniques, conclude that they have their own advantages and disadvantages and all of them are being used for different purposes. Mel frequency cepstrum is a feature extraction technique that is used widely for many speech recognition systems as it is able to mimic the human auditory system and it gives a better performance rate.

Acknowledgments

The authors would like to thank the college Authorities for providing the infrastructure to carry out the research.

References

[1] Pratik.K.Kurzekar, "A Comparative Study of Feature extraction techniques for speech recognition system", IJIRSET, 2014.
[2] Dr,Mukesh.Rana,SaloniMiglani, "Performance analysis of MFCC and LPCC Techniques in Automatic speech recognition',IJECS,2014
[3] Sayf.A.Majeed, Hafiz Huzain"Mel frequency cepstral coefficients feature extraction enhancement in the application of speech recognition": A Comparison study Malasia,2015
[4]NidhiDesai ,ProfKinnalDhaneliya"Feature extraction and classification technique for speech recognition": Areview,IJETAE 2013
[5] UrmilaShrawanker,"Techniques for Feature Extraction In speech Recognition system:'A Comparative Study.
[6] Varshasingh,Vinay Kumar Jain, "A comparative study on Feature extraction Techniques for language Identification",IJERGS,2014
[7] ParwinderPal singh,Pushpa Rani, "An Approach To Extract Feature Using MFCC",IOSR,vol.4,2014
[8] NamrataDeve ,"Feature Extraction Methods LPC,PLP, And MFCC In Speech Recognition",ijaret,2013
[9] Florian Hong,Georgstemer ,"Revising Perceptual Linear Prediction",Interspeech,2005
[10]Easwari.N,Ponmuthuramalingam. "A Comparative study on feature extraction technique for isolated word speech recognition",ijetj,2015
[11] TaabishGulzar ,Anandsingh,sandeepsharma, "Comparative Analyis of LPCC,MFCC and BFCC for the recognition of Hindi words using artificial neural networks", International journal of computer application,2014
[12] GiulianoAntoniol,Vincenzo Fabio Rollo , "Linear predictive coding and cepstral coefficients for mining time variant information from software respositors".
[13] HynekHermansky "Perceptual linear predictive (PLP) analysis of speech."speech technology laboratory,1989
[14] ShreyaNarang,Ms.Divyagupta "Speech Feature Extraction Technique:AReview",International journal of computer science and mobile computing ,vol.4,issue 3,2015