International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.9, No.11, (2016), pp.99-108
http://dx.doi.org/10.14257/ijsip.2016.9.11.10

A Comparison and Contrast of the Various Feature Extraction Techniques in Speaker Recognition

Jaison Joshy[1] and Koj Sambyo[2]

Department of Computer Science and Engineering,
National Institute of Technology, Arunachal Pradesh, India
[1]jaisonjoshy3@gmail.com, [2]sambyo.koj@gmail.com

Abstract

Speaker Recognition and Verification is becoming one of the widely used forms of biometric authentication in today's scenario where remembering strings of textual passwords and numbers are becoming a hassle. Authentication of users using voice offers many advantages and easy to use techniques. In this paper a comparison is drawn among the most commonly used feature extraction techniques in Speaker Recognition and Verification. Extracting useful and unique features from the user's voice forms the backbone of an efficient Speaker Recognition System. Here, the most commonly used methods for Feature Extraction viz. MFCC (Mel Frequency Cepstral Coefficient), LPC (Linear Predictive Coefficient), PLP (Perceptual Linear Prediction) are discussed, compared and an attempt is made to deduce which one performs best.

Keywords: *Speaker Recognition, Speaker Verification, MFCC (Mel Frequency Cepstral Coefficient), LPC (Linear Predictive Coefficient), PLP (Perceptual Linear Prediction)*

1. Introduction

Speaker Recognition is the process of recognizing a person from a given audio input of speech. Speech is one of the natural occurring forms of human communication and due the unique shape of vocal tract, larynx and due to other unique speech characteristics like rhythm, pronunciation, accent, intonation styles *etc.*, each person is known to have a unique voice which can be differentiated from others. Thus, due to this unique sound of speech produced by each individual, Speaker Recognition can be used to uniquely identify individuals. It finds many applications in the fields of forensics and remote authentication purposes where the identity of the a person can be validated just by the sound of his/ her speech. Authentication can be easily carried out through telephonic means or through any such medium that can transmit audio even without the physical presence of the person. Telephonic banking is one such area where this system can be used to validate users through their voices. But to make this system foolproof so as to prevent it from being misused by impersonators, the system has to be designed very carefully. To achieve this, the first step of Speaker Recognition *i.e.*, "Feature Extraction" from the voice samples is a very crucial step. It is these features that will decide the entire performance of the system up to a certain extent. Feature Extraction essentially means extracting important features from voice samples which can be used to discriminate one user from another. There are many techniques that have been developed till date for Feature Extraction. This paper discusses the three main techniques that are most widely used and attempts to draw a comparison between them and choose which one is better. The next section gives a brief idea of the various steps in Speaker Recognition and the following sections will discuss the Feature Extraction techniques viz. Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coefficients (LPC) and Perceptual Linear Prediction (PLP) in detail.

2. Steps in Speaker Recognition

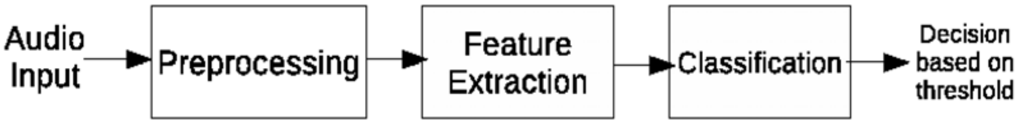

Figure 1. Steps in Speaker Recognition

The steps in Speaker Recognition consists of getting the audio input of speech, processing it to convert it to form ready for feature extraction by using any of the algorithms and classifying the samples based on the extracted features to decide the output or the result. The result could mean either to recognize the identity of the person if the system is trying to perform Speaker Recognition or to verify if the identity of a person is indeed the one that has been claimed, if the system is trying to perform Speaker Verification as in the case of authentication. A pre-decided threshold value would be used to decide if the sample audio is matched with any of the samples available in the database.

3. Feature Extraction

Feature Extraction refers to the process of extracting useful features from speech which can be used for comparing and discriminating one sample from another. There are various techniques used for feature extraction. Three of the commonly used techniques are discussed below:

4. MFCC (Mel Frequency Cepstral Coefficient)

MFCC is the most commonly used feature extraction technique in Speaker Recognition and Verification. This method was first introduced by Bridle and Brown in 1974 and was further developed by Mermelstein in 1976. MFCC is based on human hearing system which cannot exactly perceive frequencies above 1 KHz [1]. It has two filters which are spaced linearly at a low frequency below 1000 Hz and logarithmic spacing above 1000 Hz. The steps involved in MFCC are discussed below:

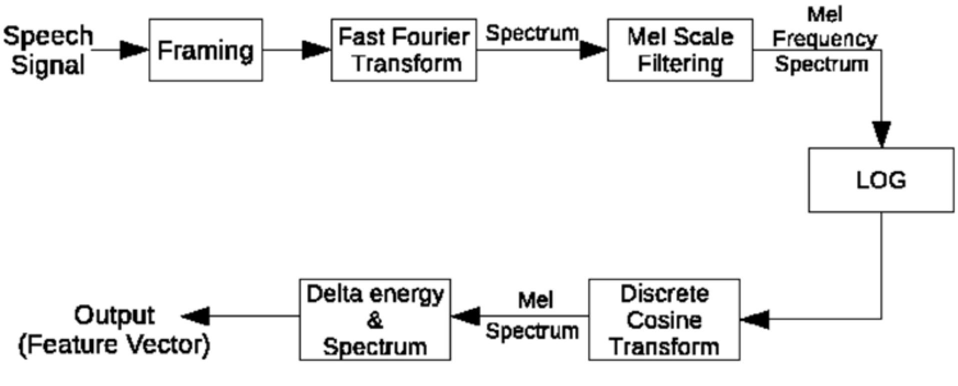

Figure 2. Steps in MFCC

4.1. Framing

Framing is the process of segmenting the digitally converted speech signal received as input into small frames usually in the length range of 20 ms to 40 ms. The voice signals are divided into N frames and adjacent frames separated with M where (M < N). Typically used values of M and N are M=100 and N= 256. [1]

4.2. Fast Fourier Transform

The next step is to convert the obtained voice samples from time domain to frequency domain. For this, Fast Fourier Transform is used. [2]

$$c_{\tau,k} = \left|\frac{1}{N}\sum_{j=0}^{N-1} f\,(i)exp[-i2\pi\frac{jk}{N}]\right| k = 0,1,....(N/2) - 1 \qquad (1)$$

where N is the number of sampling points within a speech frame and the time frame T.

4.3. Mel Scale Filtering

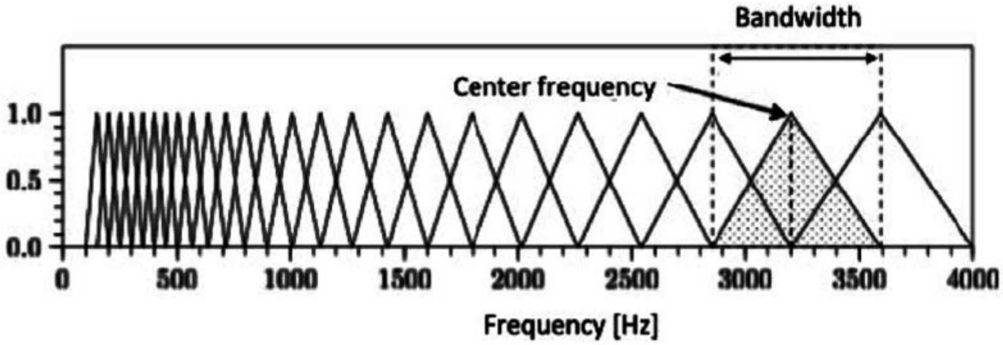

Figure 3. Filterbank with 25 Triangular Bandpass Filters to Compute the Mel Frequency Spectrum [2]

The frequencies obtained after Fast Fourier Transform are very wide and voice signals do not follow a linear scale. The Mel Scale Filtering is applied so as to obtain frequency ranges resembling that of normal human voice signals. The spectrum obtained after FFT is filtered using N_d band-pass filters and the power of each frequency band is computed. This resembles human auditory system because it uses power over frequency band for further processing. This can be explained using the equation below: [2]

$$c_{\tau,k} = \sum_{k=0}^{N/2-1} d_{j,k}\, c_{\tau,k}^{(1)} j = 0,1,2,...., N_d \qquad (2)$$

where d is the amplitude of the band-pass filter with the index j at the frequency k.

Figure 3, shows the bank of Mel Scale filters which are performed on the spectrum obtained after FFT. It shows a set of 25 triangular filters that are used to compute a weighted sum of filter spectral components so that the output of the process resembles to a Mel scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters [1]. After this the below equation is used to calculate the Mel for the given frequency f in Hertz.

$$F(mel) = [2595 * log_{10}[1 + f\,]/700] \qquad (3)$$

4.4. Logarithm

Experiment shows that humans perceive loudness on a logarithmic scale. So this step is used to compute the logarithm of the signal so as to mimic the human perception of loudness [2].

$$c_{\tau,k} = log(c_{\tau,j})j = 0,1,2,.....,N_d \tag{4}$$

4.5. Discrete Cosine Transform

The obtained log Mel Spectrum from the previous step has to be converted to time domain. For this Discrete Cosine Transform is used. The resultant conversion is called Mel Frequency Cepstral Coefficients. These sets of coefficients are called acoustic vectors. Thus after this step each input utterance is transformed into a set of acoustic vectors [1].

4.6. Delta Energy and Spectrum

The voice signal and the frames undergo changes, such as the slope of a formant at its transitions. So it is needed to add features related to the change in cepstral features over time . 13 delta or velocity features (12 cepstral features plus energy), and 39 features a double delta or acceleration feature are added. The energy in a frame for a signal x in a window from time sample t1 to time sample t2, is represented by the equation below:

$$Energy = \sum X^2\,[t] \tag{5}$$

Each of the 13 delta features represents the change between frames in the equation (6) corresponding to cepstral or energy feature, while each of the 39 double delta features represents the change between frames in the corresponding delta features [1].

$$d(t) = \frac{c(t+1)-c(t-1)}{2} \tag{6}$$

5. LPC (Linear Predictive Coding)

Linear Predictive Coding is another famous method of feature extraction used in Automatic Speaker Recognition. LPC consists of two main components viz. Analysis or Encoding and Synthesis or Decoding. The analysis part involves examining the speech signal and breaking it down into segments or blocks. Each segment is then examined further to find whether the segment is voiced or unvoiced, determine the pitch of the segment, find out what parameters are needed to build a filter that models the vocal tract for the current segment. The Synthesis part of LPC tries to imitate human speech production. Using the answers provided by the Analysis part, the Synthesis part tries to build a filter that when provided the correct input source will be able to reproduce the original speech signal. The main steps involved in the LPC process is given below: [3]

5.1. Preemphasis

The speech signal is passed through a low pass filter with a bandwidth of 1 kHz and then it is determined whether the signal is voiced or unvoiced. This is important because voiced sounds have a different waveform than unvoiced sounds. Voiced sounds are usually vowels and they have a periodic waveform with high energy and large amplitudes. While unvoiced sounds are usually no-vowels or consonants and have low energy, smaller amplitude and a very chaotic or random waveform. After this step the digitized speech signal, s(n), is put through a low order digital system to spectrally flatten the signal and to make it less susceptible to finite precision effects later in the signal processing. The output of the preemphasizer network is related to the input network, $\tilde{s}$(n), by difference equation as given below:

$$\tilde{s}(n) = s(n) - a\tilde{s}(n-1) \tag{1}$$

5.2. Frame Blocking

According to standard, the input signal is sampled at a rate of 8000 samples per second. This input signal is then broken up into blocks. The 8000 samples in each second of speech signal are broken into 180 sample segments. [4]. In general, the output of the preemphasis step is blocked into frames of N samples with adjacent frames being separated by M samples. If x_l is the l th frame of the speech and there are L frames within entire speech signal, then

$$x_l(n) = \tilde{s}(Ml + n) \tag{2}$$

where n= 0,1,.....,N and l=0,1,....,L-1

5.3. Windowing

The next step after frame blocking is to window each individual frame. This is done to minimize the signal discontinuities at the beginning and end of the frame. If the window is defined as w(n), $0 \leq n \leq N - 1$, then the result of windowing is the signal:

$$\tilde{x}_l(n) = x_l(n)w(n) \tag{3}$$

where $0 \leq n \leq N - 1$

5.4. Autocorrelation Analysis

Apart from determining whether the speech signal is voiced or unvoiced, another attribute that is required for producing an input for the LPC filter is the Pitch Period of the current speech segment. Pitch period can be defined as the time taken by a wave cycle to completely pass a fixed position.It is computationally intensive to determine the pitch period for a given segment of speech [4]. Therefore a specific algorithm is used which takes advantage of the fact that the autocorrelation of a period function r(m) will have maximum when m is equivalent to the pitch period. So each frame of the windowed signal obtained form the previous step is auto-correlated to give

$$r_l(m) = \sum_{n=0}^{N-1-m} \tilde{x}_l\,(n)\tilde{x}_l(n + m)m = 0,1,\ldots.,p \tag{4}$$

where the highest autocorrelation value, p, is the order of the LPC analysis

5.5. LPC Analysis

In this step each of the p+1 frames of autocorrelation are converted into LPC parameter set using Durbin's method. The Levinson-Durbin Algorithm is a recursive algorithm that is considered very efficient computationally as it takes advantage of the properties of r when determining the filter coefficients [4]. This can be depicted using the following steps:

$$E^{(0)} = r(0) \tag{5}$$

$$k_i = \frac{r(i) - \sum_{j=1}^{i-1} \alpha_j^{i-1} r(|i-j|)}{E^{i-1}} 1 \leq i \leq p \tag{6}$$

$$\alpha_i^{(i)} = k_i \tag{7}$$

$$\alpha_j^{(i)} = \alpha_j^{i-1} - k_i \alpha_{i-j}^{i-1} i \leq j \leq i - 1 \tag{8}$$

$$E^{(i)} = (1 - k_i^2)E^{i-1} \tag{9}$$

By solving equations (5) to (9) recursively for i=1,2,...,p, the LPC coefficient a_m, is given as

$$a_m = \alpha_m^{(p)} \tag{10}$$

5.6. Conversion of LPC Parameters to Cepstral Coefficients

The LPC Cepstral coefficients can be directly derived from the LPC coefficient set. The LPC cepstral coefficients are the features that are extracted from the voice signals and further used as the input for the classification part of Automatic Speaker Recognition. The recursion used for extracting the LPC cepstral coefficients is:

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) .c_k.a_{m-k} 1 \le m \le p \tag{11}$$

$$c_m = \sum_{k=m-p}^{m-1} \left(\frac{k}{m}\right) .c_k.a_{m-k} m > p \tag{12}$$

6. PLP (Perceptual Linear Prediction)

The Perceptual Linear Prediction developed by Hynek Hermansky is based on the psychophysics of human hearing. [5]. PLP discards irrelevant information that does not resemble to human perception of voice. It is similar to LPC except from the fact that PLP works in close resemblance to that of the human auditory system. The way in which PLP works in order to mimic the human auditory system and thus extract useful features are shown in the figure below and the steps are further explained [6,7]:

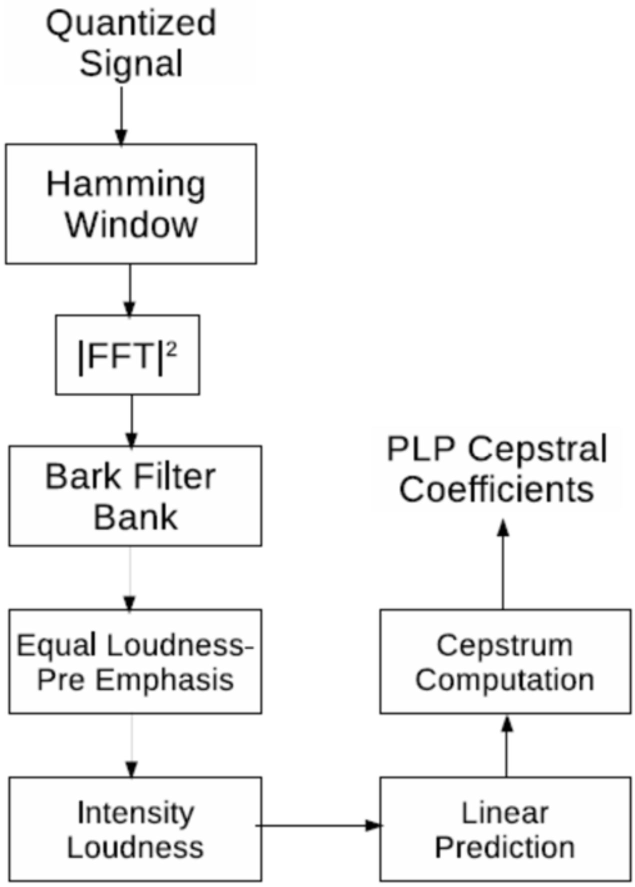

Figure 4. Working of PLP

6.1. Windowing

First of all the quantized signal is windowed in order to minimize the signal discontinuities. Typically the Hamming Window is used for this purpose as shown in the figure. This step is same as that of LPC.

6.2. Calculation of Power Spectrum

in the next step the power spectrum of the windowed signal is calculated using FFT as:

$$P(\omega) = Re(S(\omega))^2 + Im(S(\omega))^2 \qquad (1)$$

6.3. Application of Frequency Warping into Bark Scale

Frequency is converted to bark. This is done because Bark Scale is a better representation of the human hearing resolution in frequency. The bark frequency corresponding to an audio frequency is calculated as:

$$\Omega(\omega) = 6ln\left[\frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi}\right)^2 + 1\right]^{0.5}\right] \qquad (2)$$

The auditory warped spectrum is convoluted with the power spectrum of the simulated critical-band masking curve to simulate the critical-band integration of human hearing. The smoothed spectrum is down sampled at intervals of 1 Bark. The three steps frequency warping, smoothing and sampling are integrated into a single filter bank called Bark Filter Bank [6].

6.4. Equal Loudness Pre Emphasis

An equal-loudness pre-emphasis weights the filter-bank outputs to simulate the sensitivity of human hearing.

6.5. Intensity Loudness

The equalized values from the above step are transformed according to the power law of Stevens by raising each to the power of 0.33.

6.6. Linear Prediction

The auditory warped line spectrum obtained as a result of the above step is then processed by Linear Prediction. Applying Linear Prediction to the auditorily warped line spectrum means we compute the predictor coefficients of a signal that has this warped spectrum as a power spectrum. [7]

6.7. Cepstrum Computation

The last step is the computation of the Cepstral Coefficients. The Cepstral Coefficients are obtained from the predictor coefficients by a recursion that is equivalent to the model spectrum followed by an inverse Fourier transform.

7. Tabular Comparison of MFCC, LPC and PLP

Table 1. Comparison of MFCC, LPC and PLP Feature Extraction Techniques

Technique	Principle	Filtering Scale Used	Recognition Rate (%)					Merits and/or Demerits
			Clean Speech	SNR (dB)				
				30	20	10	5	

MFCC	**Mel Filter bank coefficients**	**Mel-Scale**	**98.95**	**98.45**	**97.92**	**94.87**	**90.37**	**More information about lower frequencies than higher frequencies due to Mel spaced filter banks and hence behaves more like human ear**
LPC	**Modelled by All Pole Model**	**Source-Filter Model (Linear Acoustic Filter)**	**99.95**	**98.59**	**97.63**	**93.27**	**82.73**	**Based on the principle of sound production, but Performance degrades in the presence of noise**
PLP	**Modelled by All Pair Model & works in close resemblance to human auditory system**	**Bark Scale**	**99.95**	**98.50**	**98.35**	**93.42**	**93.52**	**Works in close resemblance to human auditory system and gives best performance as compared to other techniques in presence of noise.**

8. Conclusion

The workings of the of the most widely used Feature Extraction techniques in ASR viz. MFCC, LPC and PLP have been discussed and compared. Other available Feature Extraction techniques available today are mostly a modification of the discussed techniques or a combination of two or more techniques to obtain better results. Each of these techniques has their advantages and disadvantages. MFCC and PLP are based on the human perception of voice and thus yields much more realistic and better results. LPC on the other hand does not work on the basis of human auditory reception system but is suitable for systems where the audio has to be transmitted over a large range because LPC reduces the size of the audio to a great extent thus making it a lossy but fast technique. Thus we find that each techniques has its own pros and cons and depending on the nature and need of the ASR systems to be developed, the suitable Feature Extraction technique can be applied. Also a combination of two or more techniques can be used taking the positive sides of each techniques in order to obtain results of higher accuracy.

.

References

[1] L. Muda, M. Begam and I. Elamvazuthi,"Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques,Journal of Computing", ISSN 2151-9617, vol. 2, no. 3, (**2010**) March.

[2] M. Lutter, "http://recognize-speech.com/feature-extraction/mfcc", Feature Extraction, (**2014)** November 25.

[3] Thiang and S. Wijoyo, "Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot", 2011 International Conference on Information and Electronics Engineering IPCSIT, vol. 6, **(2011).**

[4] J. Bradbury, "Linear Predictive Coding", **(2000)** December 5**.**

[5] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", Acoustical Society of America Journal, vol. 87, **(1990)** April**,** pp.1738–1752**.**

[6] N. Dave, "Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition", International Journal for Advance Research in Engineering and Technology, vol. 1, no. 6, **(2013)** July**.**

[7] F. Hoing and G. Stemmer, "Christian Hacker and Fabio Brugnara, Revising Perceptual Linear Prediction (PLP)"), Interspeech, **(2005).**

[8] T. Kinnunen and H. Li, "An Overview of Text-Independent Speaker Recognition: from Features to Supervectors", Speech Communication, **(2009).**

[9] K. Dhameliya and N. Bahtt, "Feature Extraction and Classification Techniques for Speaker Recognition: A Review", IEEE, **(2015).**

[10] J. P. Campbell Jr., "Speaker Recognition: A Tutorial", Proceedings of the IEEE, vol. 85, no. 9, **(1997)** September**.**

[11] G. Kaur, Dr. D. Singh and G. Kaur, "A Survey on Speech Recognition Algorithms", International Journal of Emerging Research in Management and Technology ISSN: 2278- 9359, vol. 4, no. 5, **(2011)** May**.**

[12] K. Kaur and N. Jain, "Feature Extraction and Classification for Automatic Speaker Recognition System –A Review", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, no. 1, **(2015)** January.

[13] V. Z. Kepuska and H. A. Elharati, "Robust Speech Recognition System Using Conventional and Hybrid Features of MFCC, LPCC, PLP, RASTA-PLP and Hidden Markov Model Classifier in Noisy Conditions", Journal of Computer and Communications, no. 3, **(2015)**, pp. 1-9.

[14] P. K. Kurzekar, R. R. Deshmukh, V. B. Waghmare and P. P. Shrishrimal, "A Comparative Study of Feature Extraction Techniques for Speech Recognition System", International Journal of Innovative Research in Science, Engineering and Technology, vol. 3, no. 12, **(2014)** December**.**

Authors

Jaison Joshy is currently pursuing his M.Tech in Computer Science and Engineering from National Institute of Technology, Arunachal Pradesh, India. He completed his B.Tech from Mizoram University and is now working on Speaker Recognition and Verification as his final year M.Tech Project in NIT Arunachal Pradesh.

Koj Sambyo received his M.Tech in Computer Science & Engineering in 2011 from Rajiv Gandhi University, Arunachal Pradesh, India. He is working as Assistant Professor in the department of Computer Science & Engineering in National Institute of Technology, Arunachal Pradesh. His research interests are Cloud Computing, Software Defined Networking.