Chapter 1

Some Commonly Used Speech Feature Extraction Algorithms

Sabur Ajibola Alim and Nahrul Khair Alang Rashid

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.80419

Abstract

Speech is a complex naturally acquired human motor ability. It is characterized in adults with the production of about 14 different sounds per second via the harmonized actions of roughly 100 muscles. Speaker recognition is the capability of a software or hardware to receive speech signal, identify the speaker present in the speech signal and recognize the speaker afterwards. Feature extraction is accomplished by changing the speech waveform to a form of parametric representation at a relatively minimized data rate for subsequent processing and analysis. Therefore, acceptable classification is derived from excellent and quality features. Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC), Line Spectral Frequencies (LSF), Discrete Wavelet Transform (DWT) and Perceptual Linear Prediction (PLP) are the speech feature extraction techniques that were discussed in these chapter. These methods have been tested in a wide variety of applications, giving them high level of reliability and acceptability. Researchers have made several modifications to the above discussed techniques to make them less susceptible to noise, more robust and consume less time. In conclusion, none of the methods is superior to the other, the area of application would determine which method to select.

Keywords: human speech, speech features, mel frequency cepstral coefficients (MFCC), linear prediction coefficients (LPC), linear prediction cepstral coefficients (LPCC), line spectral frequencies (LSF), discrete wavelet transform (DWT), perceptual linear prediction (PLP)

1. Introduction

Human beings express their feelings, opinions, views and notions orally through speech. The speech production process includes articulation, voice, and fluency [1, 2]. It is a complex

naturally acquired human motor abilities, a task categorized in regular adults by the production of about 14 different sounds per second via the harmonized actions of roughly 100 muscles connected by spinal and cranial nerves. The simplicity with which human beings speak is in contrast to the complexity of the task, and that complexity could assist in explaining why speech can be very sensitive to diseases associated with the nervous system [3].

There have been several successful attempts in the development of systems that can analyze, classify and recognize speech signals. Both hardware and software that have been developed for such tasks have been applied in various fields such as health care, government sectors and agriculture. Speaker recognition is the capability of a software or hardware to receive speech signal, identify the speaker present in the speech signal and recognize the speaker afterwards [4]. Speaker recognition executes a task similar to what the human brain undertakes. This starts from speech which is an input to the speaker recognition system. Generally, speaker recognition process takes place in three main steps which are acoustic processing, feature extraction and classification/recognition [5].

The speech signal has to be processed to remove noise before the extraction of the important attributes in the speech [6] and identification. The purpose of feature extraction is to illustrate a speech signal by a predetermined number of components of the signal. This is because all the information in the acoustic signal is too cumbersome to deal with, and some of the information is irrelevant in the identification task [7, 8].

Feature extraction is accomplished by changing the speech waveform to a form of parametric representation at a relatively lesser data rate for subsequent processing and analysis. This is usually called the front end signal-processing [9, 10]. It transforms the processed speech signal to a concise but logical representation that is more discriminative and reliable than the actual signal. With front end being the initial element in the sequence, the quality of the subsequent features (pattern matching and speaker modeling) is significantly affected by the quality of the front end [10].

Therefore, acceptable classification is derived from excellent and quality features. In present automatic speaker recognition (ASR) systems, the procedure for feature extraction has normally been to discover a representation that is comparatively reliable for several conditions of the same speech signal, even with alterations in the environmental conditions or speaker, while retaining the portion that characterizes the information in the speech signal [7, 8].

Feature extraction approaches usually yield a multidimensional feature vector for every speech signal [11]. A wide range of options are available to parametrically represent the speech signal for the recognition process, such as perceptual linear prediction (PLP), linear prediction coding (LPC) and mel-frequency cepstrum coefficients (MFCC). MFCC is the best known and very popular [9, 12]. Feature extraction is the most relevant portion of speaker recognition. Features of speech have a vital part in the segregation of a speaker from others [13]. Feature extraction reduces the magnitude of the speech signal devoid of causing any damage to the power of speech signal [14].

Before the features are extracted, there are sequences of preprocessing phases that are first carried out. The preprocessing step is pre-emphasis. This is achieved by passing the signal

through a FIR filter [15] which is usually a first-order finite impulse response (FIR) filter [16]. This is succeeded by frame blocking, a method of partitioning the speech signal into frames. It removes the acoustic interface existing in the start and end of the speech signal [17].

The framed speech signal is then windowed. Bandpass filter is a suitable window [15] that is applied to minimize disjointedness at the start and finish of each frame. The two most famous categories of windows are Hamming and Rectangular windows [18]. It increases the sharpness of harmonics, eliminates the discontinuous of signal by tapering beginning and ending of the frame zero. It also reduces the spectral distortion formed by the overlap [17].

2. Mel frequency cepstral coefficients (MFCC)

Mel frequency cepstral coefficients (MFCC) was originally suggested for identifying monosyllabic words in continuously spoken sentences but not for speaker identification. MFCC computation is a replication of the human hearing system intending to artificially implement the ear's working principle with the assumption that the human ear is a reliable speaker recognizer [19]. MFCC features are rooted in the recognized discrepancy of the human ear's critical bandwidths with frequency filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to retain the phonetically vital properties of the speech signal. Speech signals commonly contain tones of varying frequencies, each tone with an actual frequency, f (Hz) and the subjective pitch is computed on the Mel scale. The mel-frequency scale has linear frequency spacing below 1000 Hz and logarithmic spacing above 1000 Hz. Pitch of 1 kHz tone and 40 dB above the perceptual audible threshold is defined as 1000 mels, and used as reference point [20].

MFCC is based on signal disintegration with the help of a filter bank. The MFCC gives a discrete cosine transform (DCT) of a real logarithm of the short-term energy displayed on the Mel frequency scale [21]. MFCC is used to identify airline reservation, numbers spoken into a telephone and voice recognition system for security purpose. Some modifications have been proposed to the basic MFCC algorithm for better robustness, such as by lifting the log-mel-amplitudes to an appropriate power (around 2 or 3) before applying the DCT and reducing the impact of the low-energy parts [4].

2.1. Algorithm description, strength and weaknesses

MFCC are cepstral coefficients derived on a twisted frequency scale centerd on human auditory perception. In the computation of MFCC, the first thing is windowing the speech signal to split the speech signal into frames. Since the high frequency formants process reduced amplitude compared to the low frequency formants, high frequencies are emphasized to obtain similar amplitude for all the formants. After windowing, Fast Fourier Transform (FFT) is applied to find the power spectrum of each frame. Subsequently, the filter bank processing is carried out on the power spectrum, using mel-scale. The DCT is applied to the speech signal

after translating the power spectrum to log domain in order to calculate MFCC coefficients [5]. The formula used to calculate the mels for any frequency is [19, 22]:

$$mel(f) = 2595\, x\, \log_{10}(1 + f/700) \tag{1}$$

where mel(f) is the frequency (mels) and f is the frequency (Hz).

The MFCCs are calculated using this equation [9, 19]:

$$\hat{C}_n = \sum_{n=1}^{k} \left(\log \hat{S}_k \right) cos \left[n \left(k - \frac{1}{2} \right) \frac{\pi}{k} \right] \tag{2}$$

where k is the number of mel cepstrum coefficients, $\hat{S}_k$ is the output of filterbank and $\hat{C}_n$ is the final mfcc coefficients.

The block diagram of the MFCC processor can be seen in **Figure 1**. It summarizes all the processes and steps taken to obtain the needed coefficients. MFCC can effectively denote the low frequency region better than the high frequency region, henceforth, it can compute formants that are in the low frequency range and describe the vocal tract resonances. It has been generally recognized as a front-end procedure for typical Speaker Identification applications, as it has reduced vulnerability to noise disturbance, with minute session inconsistency and easy to mine [19]. Also, it is a perfect representation for sounds when the source characteristics are stable and consistent (music and speech) [23]. Furthermore, it can capture information from sampled signals with frequencies at a maximum of 5 kHz, which encapsulates most energy of sounds that are generated by humans [9].

Cepstral coefficients are said to be accurate in certain pattern recognition problems relating to human voice. They are used extensively in speaker identification and speech recognition [21]. Other formants can also be above 1 kHz and are not efficiently taken into consideration by the large filter spacing in the high frequency range [19]. MFCC features are not exactly accurate in the existence of background noise [14, 24] and might not be well suited for generalization [23].

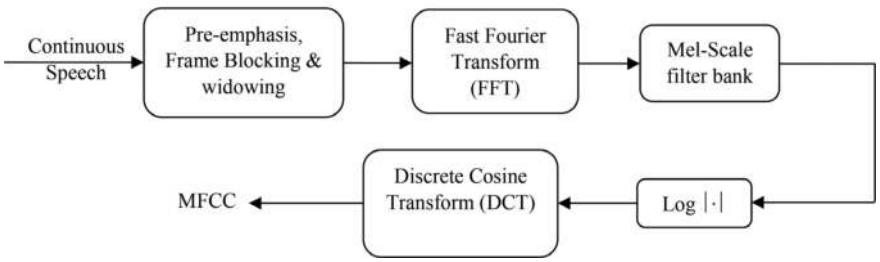

Figure 1. Block diagram of MFCC processor.

3. Linear prediction coefficients (LPC)

Linear prediction coefficients (LPC) imitates the human vocal tract [16] and gives robust speech feature. It evaluates the speech signal by approximating the formants, getting rid of its effects from the speech signal and estimate the concentration and frequency of the left behind residue. The result states each sample of the signal as a direct incorporation of previous samples. The coefficients of the difference equation characterize the formants, thus, LPC needs to approximate these coefficients [25]. LPC is a powerful speech analysis method and it has gained fame as a formant estimation method [17].

The frequencies where the resonant crests happen are called the formant frequencies. Thus, with this technique, the positions of the formants in a speech signal are predictable by calculating the linear predictive coefficients above a sliding window and finding the crests in the spectrum of the subsequent linear prediction filter [17]. LPC is helpful in the encoding of high quality speech at low bit rate [13, 26, 27].

Other features that can be deduced from LPC are linear predication cepstral coefficients (LPCC), log area ratio (LAR), reflection coefficients (RC), line spectral frequencies (LSF) and Arcus Sine Coefficients (ARCSIN) [13]. LPC is generally used for speech reconstruction. LPC method is generally applied in musical and electrical firms for creating mobile robots, in telephone firms, tonal analysis of violins and other string musical gadgets [4].

3.1. Algorithm description, strength and weaknesses

Linear prediction method is applied to obtain the filter coefficients equivalent to the vocal tract by reducing the mean square error in between the input speech and estimated speech [28]. Linear prediction analysis of speech signal forecasts any given speech sample at a specific period as a linear weighted aggregation of preceding samples. The linear predictive model of speech creation is given as [13, 25]:

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{3}$$

where $\hat{s}$ is the predicted sample, s is the speech sample, p is the predictor coefficients.

The prediction error is given as [16, 25]:

$$e(n) = s(n) - \hat{s}(n) \tag{4}$$

Subsequently, each frame of the windowed signal is autocorrelated, while the highest autocorrelation value is the order of the linear prediction analysis. This is followed by the LPC analysis, where each frame of the autocorrelations is converted into LPC parameters set which consists of the LPC coefficients [26]. A summary of the procedure for obtaining the LPC is as seen in **Figure 2**. LPC can be derived by [7]:

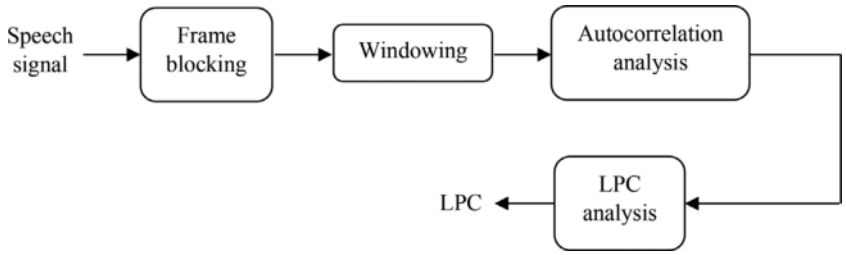

Figure 2. Block diagram of LPC processor.

$$a_m = \log\left[\frac{1-k_m}{1+k_m}\right] \tag{5}$$

where a_m is the linear prediction coefficient, k_m is the reflection coefficient.

Linear predictive analysis efficiently selects the vocal tract information from a given speech [16]. It is known for the speed of computation and accuracy [18]. LPC excellently represents the source behaviors that are steady and consistent [23]. Furthermore, it is also be used in speaker recognition system where the main purpose is to extract the vocal tract properties [25]. It gives very accurate estimates of speech parameters and is comparatively efficient for computation [14, 26]. Traditional linear prediction suffers from aliased autocorrelation coefficients [29]. LPC estimates have high sensitivity to quantization noise [30] and might not be well suited for generalization [23].

4. Linear prediction cepstral coefficients (LPCC)

Linear prediction cepstral coefficients (LPCC) are cepstral coefficients derived from LPC calculated spectral envelope [11]. LPCC are the coefficients of the Fourier transform illustration of the logarithmic magnitude spectrum [30, 31] of LPC. Cepstral analysis is commonly applied in the field of speech processing because of its ability to perfectly symbolize speech waveforms and characteristics with a limited size of features [31].

It was observed by Rosenberg and Sambur that adjacent predictor coefficients are highly correlated and therefore, representations with less correlated features would be more efficient, LPCC is a typical example of such. The relationship between LPC and LPCC was originally derived by Atal in 1974. In theory, it is relatively easy to convert LPC to LPCC, in the case of minimum phase signals [32].

4.1. Algorithm description, strength and weaknesses

In speech processing, LPCC analogous to LPC, are computed from sample points of a speech waveform, the horizontal axis is the time axis, while the vertical axis is the amplitude axis [31].

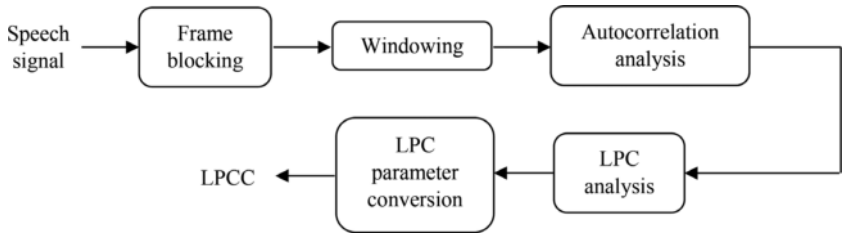

Figure 3. Block diagram of LPCC processor.

The LPCC processor is as seen in **Figure 3**. It pictorially explains the process of obtaining LPCC. LPCC can be calculated using [7, 15, 33]:

$$C_m = a_m + \sum_{k=1}^{m-1} \left[\frac{k}{m}\right] c_k a_{m-k} \tag{6}$$

where a_m is the linear prediction coefficient, C_m is the cepstral coefficient.

LPCC have low vulnerability to noise [30]. LPCC features yield lower error rate as compared to LPC features [31]. Cepstral coefficients of higher order are mathematically limited, resulting in an extremely extensive array of variances when moving from the cepstral coefficients of lower order to cepstral coefficients of higher order [34]. Similarly, LPCC estimates are notorious for having great sensitivity to quantization noise [35]. Cepstral analysis on high-pitch speech signal gives small source-filter separability in the quefrency domain [29]. Cepstral coefficients of lower order are sensitive to the spectral slope, while the cepstral coefficients of higher order are sensitive to noise [15].

5. Line spectral frequencies (LSF)

Individual lines of the Line Spectral Pairs (LSP) are known as line spectral frequencies (LSF). LSF defines the two resonance situations taking place in the inter-connected tube model of the human vocal tract. The model takes into consideration the nasal cavity and the mouth shape, which gives the basis for the fundamental physiological importance of the linear prediction illustration. The two resonance situations define the vocal tract as either being completely open or completely closed at the glottis [36]. The two situations begets two groups of resonant frequencies, with the number of resonances in each group being deduced from the quantity of linked tubes. The resonances of each situation are the odd and even line spectra correspondingly, and are interwoven into a singularly rising group of LSF [36].

The LSF representation was proposed by Itakura [37, 38] as a substitute to the linear prediction parametric illustration. In the area of speech coding, it has been realized that this illustration has an improved quantization features than the other linear prediction parametric illustrations

(LAR and RC). The LSF illustration has the capacity to reduce the bit-rate by 25–30% for transmitting the linear prediction information without distorting the quality of synthesized speech [38–40]. Apart from quantization, LSF illustration of the predictor are also suitable for interpolation. Theoretically, this can be inspired by the point that the sensitivity matrix linking the LSF-domain squared quantization error to the perceptually relevant log spectrum is diagonal [41, 42].

5.1. Algorithm description, strength and weaknesses

LP is established on the point that a speech signal can be defined by Eq. (3). Recall

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k)$$

where k is the time index and p is the order of the linear prediction, $\hat{s}(n)$ is the predictor signal and a_k is the LPC coefficients.

The a_k coefficients are determined in order to reduce the prediction error by method of autocorrelation or covariance. Eq. (3) can be modified in the frequency domain with the z-transform. As such, a small part of the speech signal is anticipated to be given as an output to the all-pole filter $H(z)$. The new equation is

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-1}} \tag{7}$$

where $H(z)$ is the all-pole filter and $A(z)$ is the LPC analysis filter.

In order to compute the LSF coefficients, an inverse polynomial filter is split into two polynomials $P(z)$ and $Q(z)$ [36, 38, 40, 41]:

$$P(z) = A(z) + z^{-(p+1)} A\left(z^{-1}\right) \tag{8}$$

$$Q(z) = A(z) - z^{-(p+1)} A\left(z^{-1}\right) \tag{9}$$

where $P(z)$ is the vocal tract with the glottis closed, $Q(z)$ is the LPC analysis filter of order p.

In order to convert LSF back to LPC, the equation below is used [36, 41, 43, 44]:

$$A(z) = 0.5[P(z) + Q(z)] \tag{10}$$

The block diagram of the LSF processor is as seen in **Figure 4**. The most prominent application of LSF is in the area of speech compression, with extension into the speaker recognition and speech recognition. This technique has also found restricted use in other fields. LSF have been investigated for use in musical instrument recognition and coding. LSF have also been applied to animal noise identification, recognizing individual instruments and financial market analysis. The advantages of LSF include their ability to localize spectral sensitivities, the fact that

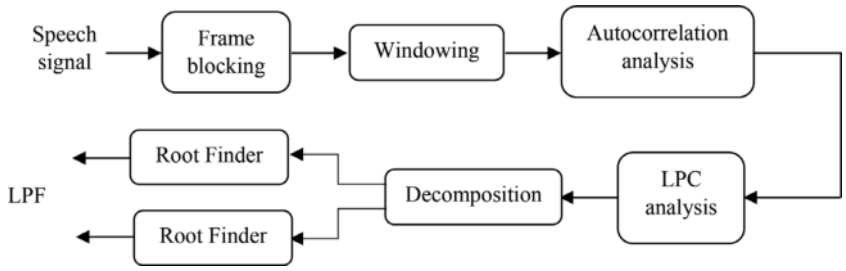

Figure 4. Block diagram of LSF processor.

they characterize bandwidths and resonance locations and lays emphasis on the important aspect of spectral peak location. In most instances, the LSF representation provides a near-minimal data set for subsequent classification [36].

Since LSF represents spectral shape information at a lower data rate than raw input samples, it is reasonable that a careful use of processing and analysis methods in the LSP domain could lead to a complexity reduction against alternative techniques operating on the raw input data itself. LSF play an important role in the transmission of vocal tract information from speech coder to decoder with their widespread use being a result of their excellent quantization properties. The generation of LSP parameters can be accomplished using several methods, ranging in complexity. The major problem revolves around finding the roots of the P and Q polynomials defined in Eqs. (8) and (9). This can be obtained through standard root solving methods, or more obscure methods and it is often performed in the cosine domain [36].

6. Discrete wavelet transform (dwt)

Wavelet Transform (WT) theory is centered around signal analysis using varying scales in the time and frequency domains [45]. With the support of theoretical physicist Alex Grossmann, Jean Morlet introduced wavelet transform which permits high-frequency events identification with an enhanced temporal resolution [45–47]. A wavelet is a waveform of effectively limited duration that has an average value of zero. Many wavelets also display orthogonality, an ideal feature of compact signal representation [46]. WT is a signal processing technique that can be used to represent real-life non-stationary signals with high efficiency [33, 46]. It has the ability to mine information from the transient signals concurrently in both time and frequency domains [33, 45, 48].

Continuous wavelet transform (CWT) is used to split a continuous-time function into wavelets. However, there is redundancy of information and huge computational efforts is required to calculate all likely scales and translations of CWT, thereby restricting its use [45]. Discrete wavelet transform (DWT) is an extension of the WT that enhances the flexibility to the decomposition process [48]. It was introduced as a highly flexible and efficient method for sub band breakdown of signals [46, 49]. In earlier applications, linear

discretization was used for discretizing CWT. Daubechies and others have developed an orthogonal DWT specially designed for analyzing a finite set of observations over the set of scales (dyadic discretization) [47].

6.1. Algorithm description, strength and weaknesses

Wavelet transform decomposes a signal into a group of basic functions called wavelets. Wavelets are obtained from a single prototype wavelet called mother wavelet by dilations and shifting. The main characteristic of the WT is that it uses a variable window to scan the frequency spectrum, increasing the temporal resolution of the analysis [45, 46, 50].

WT decomposes signals over translated and dilated mother 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 [33, 45, 50]:

$$W_x(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\psi^*\left(\frac{t-b}{a}\right)dt \tag{11}$$

where $\psi(t)$ is the mother wavelet, a and b are continuous parameters.

The WT coefficient is an expansion and a particular shift represents how well the original signal corresponds to the translated and dilated mother wavelet. Thus, the coefficient group of CWT (a, b) associated with a particular signal is the wavelet representation of the original signal in relation to the mother wavelet [45]. Since CWT contains high redundancy, analyzing the signal using a small number of scales with varying number of translations at each scale, i.e. discretizing scale and translation parameters as $a = 2^j \; and \; b = 2^j k$ gives DWT. DWT theory requires two sets of related functions called scaling function and wavelet function given by [33]:

$$\phi(t) = \sum_{n=0}^{N-1} h[n]\sqrt{2}\phi(2t-n) \tag{12}$$

$$\psi(t) = \sum_{n=0}^{N-1} g[n]\sqrt{2}\phi(2t-n) \tag{13}$$

where $\phi(t)$ is the scaling function, $\psi(t)$ is the wavelet function, h[n] is the an impulse response of a low-pass filter, and g[n] is an impulse response of a high-pass filter.

There are several ways to discretize a CWT. The DWT of the continuous signal can also be given by [45]:

$$(DWT)(m,p) = \int_{-\infty}^{+\infty} x(t)\cdot\psi_{m,p}dt \tag{14}$$

where $\psi_{m,p}$ is the wavelet function bases, m is the dilation parameter and p is the translation parameter.

Thus, $\psi_{m,p}$ is defined as:

$$\psi_{m,p} = \frac{1}{\sqrt{a_0^m}} \psi\left(\frac{t - pb_0a_0^m}{a_0^m}\right) \tag{15}$$

The DWT of a discrete signal is derived from CWT and defined as:

$$(DWT)(m,k) = \frac{1}{\sqrt{a_0^m}} \sum_n x[n] \cdot g\left(\frac{n - nb_0a_0^m}{a_0^m}\right) \tag{16}$$

where $g(*)$ is the mother wavelet and $x[n]$ is the discretized signal. The mother wavelet may be dilated and translated discretely by selecting the scaling parameter $a = a_0^m$ and translation parameter $b = nb_0a_0^m$ (with constants taken as $a_0 > 1$, $b_0 > 1$, while m and n are assigned a set of positive integers).

The scaling and wavelet functions can be implemented effectively using a pair of filters, h[n] and g[n], called quadrature mirror filters that confirm with the property $g[n] = (-1)^{1-n}h[n]$. The input signal is filtered by a low-pass filter and high-pass filter to obtain the approximate components and the detail components respectively. This is summarized in **Figure 5**. The approximate signal at each stage is further decomposed using the 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 [33].

The DWT parameters contain the information of different frequency scales. This enhances the speech information obtained in the corresponding frequency band [33]. The ability of the DWT to partition the variance of the elements of the input on a scale by scale basis is an added advantage. This partitioning leads to the opinion of the scale-dependent wavelet variance, which in many ways is equivalent to the more familiar frequency-dependent Fourier power spectrum [47]. Classic discrete decomposition schemes, which are dyadic do not fulfill all the requirements for direct use in parameterization. DWT does provide adequate number of frequency bands for effective speech analysis [51]. Since the input signals are of finite length, the wavelet coefficients will have unwantedly large variations at the boundaries because of the discontinuities at the boundaries [50].

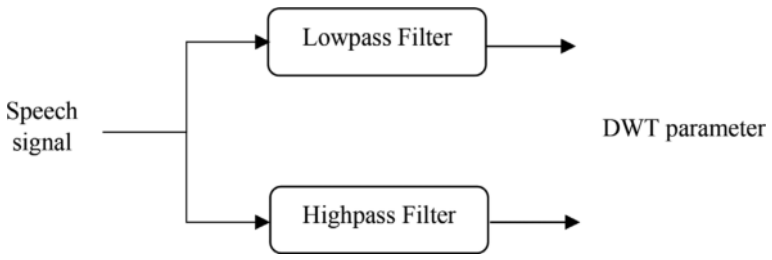

Figure 5. Block diagram of DWT.

7. Perceptual linear prediction (PLP)

Perceptual linear prediction (PLP) technique combines the critical bands, intensity-to-loudness compression and equal loudness pre-emphasis in the extraction of relevant information from speech. It is rooted in the nonlinear bark scale and was initially intended for use in speech recognition tasks by eliminating the speaker dependent features [11]. PLP gives a representation conforming to a smoothed short-term spectrum that has been equalized and compressed similar to the human hearing making it similar to the MFCC. In the PLP approach, several prominent features of hearing are replicated and the consequent auditory like spectrum of speech is approximated by an autoregressive all–pole model [52]. PLP gives minimized resolution at high frequencies that signifies auditory filter bank based approach, yet gives the orthogonal outputs that are similar to the cepstral analysis. It uses linear predictions for spectral smoothing, hence, the name is perceptual linear prediction [28]. PLP is a combination of both spectral analysis and linear prediction analysis.

7.1. Algorithm description, strength and weaknesses

In order to compute the PLP features the speech is windowed (Hamming window), the Fast Fourier Transform (FFT) and the square of the magnitude are computed. This gives the power spectral estimates. A trapezoidal filter is then applied at 1-bark interval to integrate the overlapping critical band filter responses in the power spectrum. This effectively compresses the higher frequencies into a narrow band. The symmetric frequency domain convolution on the bark warped frequency scale then permits low frequencies to mask the high frequencies, concurrently smoothing the spectrum. The spectrum is subsequently pre-emphasized to approximate the uneven sensitivity of human hearing at a variety of frequencies. The spectral amplitude is compressed, this reduces the amplitude variation of the spectral resonances. An Inverse Discrete Fourier Transform (IDCT) is performed to get the autocorrelation coefficients. Spectral smoothing is performed, solving the autoregressive equations. The autoregressive coefficients are converted to cepstral variables [28]. The equation for computing the bark scale frequency is:

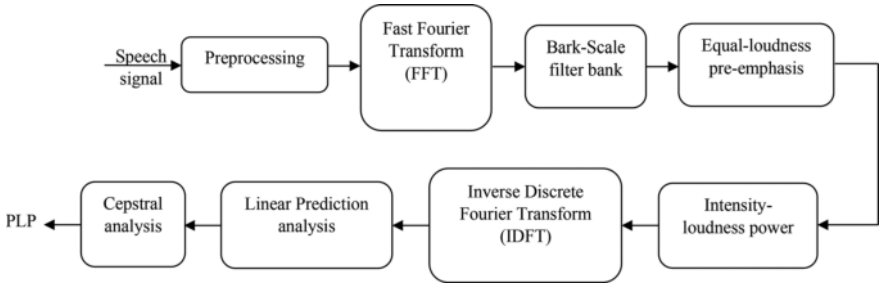

Figure 6. Block diagram of PLP processor.

	Type of Filter	Shape of filter	What is modeled	Speed of computation	Type of coefficient	Noise resistance	Sensitivity to quantization/additional noise	Reliability	Frequency captured
Mel frequency cepstral coefficient (MFCC)	Mel	Triangular	Human Auditory System	High	Cepstral	Medium	Medium	High	Low
Linear prediction coefficient (LPC)	Linear Prediction	Linear	Human Vocal Tract	High	Autocorrelation Coefficient	High	High	High	Low
Linear prediction cepstral coefficient (LPCC)	Linear Prediction	Linear	Human Vocal Tract	Medium	Cepstral	High	High	Medium	Low & Medium
Line spectral frequencies (LSF)	Linear Prediction	Linear	Human Vocal Tract	Medium	Spectral	High	High	Medium	Low & Medium
Discrete wavelet transform (DWT)	Lowpass & highpass	—	—	High	Wavelets	Medium	Medium	Medium	Low & High
Perceptual linear prediction (PLP)	Bark	Trapezoidal	Human Auditory System	Medium	Cepstral & Autocorrelation	Medium	Medium	Medium	Low & Medium

Table 1. Comparison between the feature extraction techniques.

http://dx.doi.org/10.5772/intechopen.80419

$$bark(f) = \frac{26.81\,f}{1960 + f} - 0.53 \tag{17}$$

where bark(f) is the frequency (bark) and f is the frequency (Hz).

The identification achieved by PLP is better than that of LPC [28], because it is an improvement over the conventional LPC because it effectively suppresses the speaker-dependent information [52]. Also, it has enhanced speaker independent recognition performance and is robust to noise, variations in the channel and microphones [53]. PLP reconstructs the autoregressive noise component accurately [54]. PLP based front end is sensitive to any change in the formant frequency.

Figure 6 shows the PLP processor, showing all the steps to be taken to obtain the PLP coefficients. PLP has low sensitivity to spectral tilt, consistent with the findings that it is relatively insensitive to phonetic judgments of the spectral tilt. Also, PLP analysis is dependent on the result of the overall spectral balance (formant amplitudes). The formant amplitudes are easily affected by factors such as the recording equipment, communication channel and additive noise [52]. Furthermore, the time-frequency resolution and efficient sampling of the short-term representation are addressed in an ad-hoc way [54].

Table 1 shows a comparison between the six feature extraction techniques that have been explicitly described above. Even though the selection of a feature extraction algorithm for use in research is individual dependent, however, this table has been able to characterize these techniques based on the main considerations in the selection of any feature extraction algorithm. The considerations include speed of computation, noise resistance and sensitivity to additional noise. The table also serves as a guide when considering the selection between any two or more of the discussed algorithms.

8. Conclusion

MFCC, LPC, LPCC, LSF, PLP and DWT are some of the feature extraction techniques used for extracting relevant information form speech signals for the purpose speech recognition and identification. These techniques have stood the test of time and have been widely used in speech recognition systems for several purposes. Speech signal is a slow time varying signal, quasi-stationary, when observed over an adequately short period of time between 5 and 100 msec, its behavior is relatively stationary. As a result of this, short time spectral analysis which includes MFCC, LPCC and PLP are commonly used for the extraction of important information from speech signals. Noise is a serious challenge encountered in the process of feature extraction, as well as speaker recognition as a whole. Subsequently, researchers have made several modifications to the above discussed techniques to make them less susceptible to noise, more robust and consume less time. These methods have also been used in the recognition of sounds. The extracted information will be the input to the classifier for identification purposes. The above discussed feature extraction approaches can be implemented using MATLAB.

Author details

Sabur Ajibola Alim[1]* and Nahrul Khair Alang Rashid[2]

*Address all correspondence to: moaj1st@yahoo.com

1 Ahmadu Bello University, Zaria, Nigeria

2 Universiti Teknologi Malaysia, Skudai, Johor, Malaysia

References

[1] Hariharan M, Vijean V, Fook CY, Yaacob S. Speech stuttering assessment using sample entropy and Least Square Support vector machine. In: 8th International Colloquium on Signal Processing and its Applications (CSPA). 2012. pp. 240-245

[2] Manjula GN, Kumar MS. Stuttered speech recognition for robotic control. International Journal of Engineering and Innovative Technology (IJEIT). 2014;**3**(12):174-177

[3] Duffy JR. Motor speech disorders: Clues to neurologic diagnosis. In: Parkinson's Disease and Movement Disorders. Totowa, NJ: Humana Press; 2000. pp. 35-53

[4] Kurzekar PK, Deshmukh RR, Waghmare VB, Shrishrimal PP. A comparative study of feature extraction techniques for speech recognition system. International Journal of Innovative Research in Science, Engineering and Technology. 2014;**3**(12):18006-18016

[5] Ahmad AM, Ismail S, Samaon DF. Recurrent neural network with backpropagation through time for speech recognition. In: IEEE International Symposium on Communications and Information Technology (ISCIT 2004). Vol. 1. Sapporo, Japan: IEEE; 2004. pp. 98-102

[6] Shaneh M, Taheri A. Voice command recognition system based on MFCC and VQ algorithms. World academy of science. Engineering and Technology. 2009;**57**:534-538

[7] Mosa GS, Ali AA. Arabic phoneme recognition using hierarchical neural fuzzy petri net and LPC feature extraction. Signal Processing: An International Journal (SPIJ). 2009;**3**(5): 161

[8] Yousefian N, Analoui M. Using radial basis probabilistic neural network for speech recognition. In: Proceeding of 3rd International Conference on Information and Knowledge (IKT07), Mashhad, Iran. 2007

[9] Cornaz C, Hunkeler U, Velisavljevic V. An Automatic Speaker Recognition System. Switzerland: Lausanne; 2003. Retrieved from: http://read.pudn.com/downloads60/sourcecode/multimedia/audio/209082/asr_project.pdf

[10] Shah SAA, ul Asar A, Shaukat SF. Neural network solution for secure interactive voice response. World Applied Sciences Journal. 2009;**6**(9):1264-1269

[11] Ravikumar KM, Rajagopal R, Nagaraj HC. An approach for objective assessment of stuttered speech using MFCC features. ICGST International Journal on Digital Signal Processing, DSP. 2009;**9**(1):19-24

[12] Kumar PP, Vardhan KSN, Krishna KSR. Performance evaluation of MLP for speech recognition in noisy environments using MFCC & wavelets. International Journal of Computer Science & Communication (IJCSC). 2010;**1**(2):41-45

[13] Kumar R, Ranjan R, Singh SK, Kala R, Shukla A, Tiwari R. Multilingual speaker recognition using neural network. In: Proceedings of the Frontiers of Research on Speech and Music, FRSM. 2009. pp. 1-8

[14] Narang S, Gupta MD. Speech feature extraction techniques: A review. International Journal of Computer Science and Mobile Computing. 2015;**4**(3):107-114

[15] Al-Alaoui MA, Al-Kanj L, Azar J, Yaacoub E. Speech recognition using artificial neural networks and hidden Markov models. IEEE Multidisciplinary Engineering Education Magazine. 2008;**3**(3):77-86

[16] Al-Sarayreh KT, Al-Qutaish RE, Al-Kasasbeh BM. Using the sound recognition techniques to reduce the electricity consumption in highways. Journal of American Science. 2009;**5**(2):1-12

[17] Gill AS. A review on feature extraction techniques for speech processing. International Journal Of Engineering and Computer Science. 2016;**5**(10):18551-18556

[18] Othman AM, Riadh MH. Speech recognition using scaly neural networks. World academy of science. Engineering and Technology. 2008;**38**:253-258

[19] Chakroborty S, Roy A, Saha G. Fusion of a complementary feature set with MFCC for improved closed set text-independent speaker identification. In: IEEE International Conference on Industrial Technology, 2006. ICIT 2006. pp. 387-390

[20] de Lara JRC. A method of automatic speaker recognition using cepstral features and vectorial quantization. In: Iberoamerican Congress on Pattern Recognition. Berlin, Heidelberg: Springer; 2005. pp. 146-153

[21] Ravikumar KM, Reddy BA, Rajagopal R, Nagaraj HC. Automatic detection of syllable repetition in read speech for objective assessment of stuttered Disfluencies. In: Proceedings of World Academy Science, Engineering and Technology. 2008. pp. 270-273

[22] Hasan MR, Jamil M, Rabbani G, Rahman MGRMS. Speaker Identification Using Mel Frequency cepstral coefficients. In: 3rd International Conference on Electrical & Computer Engineering, 2004. ICECE 2004. pp. 28-30

[23] Chu S, Narayanan S, Kuo CC. Environmental sound recognition using MP-based features. In: IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE; 2008. pp. 1-4

[24] Rao TB, Reddy PPVGD, Prasad A. Recognition and a panoramic view of Raaga emotions of singers-application Gaussian mixture model. International Journal of Research and Reviews in Computer Science (IJRRCS). 2011;**2**(1):201-204

[25] Agrawal S, Shruti AK, Krishna CR. Prosodic feature based text dependent speaker recognition using machine learning algorithms. International Journal of Engineering Science and Technology. 2010;**2**(10):5150-5157

[26] Paulraj MP, Sazali Y, Nazri A, Kumar S. A speech recognition system for Malaysian English pronunciation using neural network. In: Proceedings of the International Conference on Man-Machine Systems (ICoMMS). 2009

[27] Tan CL, Jantan A. Digit recognition using neural networks. Malaysian Journal of Computer Science. 2004;**17**(2):40-54

[28] Kumar P, Chandra M. Speaker identification using Gaussian mixture models. MIT International Journal of Electronics and Communication Engineering. 2011;**1**(1):27-30

[29] Wang TT, Quatieri TF. High-pitch formant estimation by exploiting temporal change of pitch. IEEE Transactions on Audio, Speech, and Language Processing. 2010;**18**(1):171-186

[30] El Choubassi MM, El Khoury HE, Alagha CEJ, Skaf JA, Al-Alaoui MA. Arabic speech recognition using recurrent neural networks. In: Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology (IEEE Cat. No.03EX795). Ieee; 2003. pp. 543-547. DOI: 10.1109/ISSPIT.2003.1341178

[31] Wu QZ, Jou IC, Lee SY. On-line signature verification using LPC cepstrum and neural networks. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics. 1997; **27**(1):148-153

[32] Holambe R, Deshpande M. Advances in Non-Linear Modeling for Speech Processing. Berlin, Heidelberg: Springer Science & Business Media; 2012

[33] Nehe NS, Holambe RS. DWT and LPC based feature extraction methods for isolated word recognition. EURASIP Journal on Audio, Speech, and Music Processing. 2012;**2012**(1):7

[34] Young S, Evermann G, Gales M, Hain T, Kershaw D, Liu X, et al. The HTK Book, Version 3.4. Cambridge, United Kingdom: Cambridge University; 2006

[35] Ismail S, Ahmad A. Recurrent neural network with backpropagation through time algorithm for arabic recognition. In: Proceedings of the 18th European Simulation Multiconference (ESM). Magdeburg, Germany; 2004. pp. 13-16

[36] McLoughlin IV. Line spectral pairs. Signal Processing. 2008;**88**(3):448-467

[37] Itakura F. Line spectrum representation of linear predictor coefficients of speech signals. The Journal of the Acoustical Society of America. 1975;**57**(S1):S35-S35

[38] Silva DF, de Souza VM, Batista GE, Giusti R. Spoken digit recognition in portuguese using line spectral frequencies. Ibero-American Conference on Artificial Intelligence. Vol. 7637. Berlin, Heidelberg: Springer; 2012. pp. 241-250

[39] Kabal P, Ramachandran RP. The computation of line spectral frequencies using Chebyshev polynomials. IEEE Transactions on Acoustics, Speech and Signal Processing. 1986;**34**(6):1419-1426

[40] Paliwal KK. On the use of line spectral frequency parameters for speech recognition. Digital Signal Processing. 1992;**2**(2):80-87

[41] Alang Rashid NK, Alim SA, Hashim NNWNH, Sediono W. Receiver operating characteristics measure for the recognition of stuttering Dysfluencies using line spectral frequencies. IIUM Engineering Journal. 2017;**18**(1):193-200

[42] Kleijn WB, Bäckström T, Alku P. On line spectral frequencies. IEEE Signal Processing Letters. 2003;**10**(3):75-77

[43] Bäckström T, Pedersen CF, Fischer J, Pietrzyk G. Finding line spectral frequencies using the fast Fourier transform. In: 2015 IEEE International Conference on in Acoustics, Speech and Signal Processing (ICASSP). 2015. pp. 5122-5126

[44] Nematollahi MA, Vorakulpipat C, Gamboa Rosales H. Semifragile speech watermarking based on least significant bit replacement of line spectral frequencies. Mathematical Problems in Engineering. 2017. 9 p

[45] Oliveira MO, Bretas AS. Application of discrete wavelet transform for differential protection of power transformers. In: IEEE PowerTech. Bucharest: IEEE; 2009. pp. 1-8

[46] Gupta D, Choubey S. Discrete wavelet transform for image processing. International Journal of Emerging Technology and Advanced Engineering. 2015;**4**(3):598-602

[47] Lindsay RW, Percival DB, Rothrock DA. The discrete wavelet transform and the scale analysis of the surface properties of sea ice. IEEE Transactions on Geoscience and Remote Sensing. 1996;**34**(3):771-787

[48] Turner C, Joseph A. A wavelet packet and mel-frequency cepstral coefficients-based feature extraction method for speaker identification. In: Procedia Computer Science. 2015. pp. 416-421

[49] Reig-Bolaño R, Marti-Puig P, Solé-Casals J, Zaiats V, Parisi V. Coding of biosignals using the discrete wavelet decomposition. In: International Conference on Nonlinear Speech Processing. Berlin Heidelberg: Springer; 2009. pp. 144-151

[50] Tufekci Z, Gowdy JN. Feature extraction using discrete wavelet transform for speech recognition. In: IEEE Southeastcon 2000. 2000. pp. 116-123

[51] Gałka J, Ziółko M. Wavelet speech feature extraction using mean best basis algorithm. In: International Conference on Nonlinear Speech Processing Berlin. Heidelberg: Springer; 2009. pp. 128-135

[52] Hermansky H. Perceptual linear predictive (PLP) analysis of speech. The Journal of the Acoustical Society of America. 1990;**87**(4):1738-1752

[53] Picone J. Fundamentals of Speech Recognition: Spectral Transformations. 2011. Retrieved from: http://www.isip.piconepress.com/publications/courses/msstate/ece_8463/lectures/current/lecture_17/lecture_17.pdf

[54] Thomas S, Ganapathy S, Hermansky H. Spectro-temporal features for automatic speech recognition using linear prediction in spectral domain. In: Proceedings of the 16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland. 2008