ijergs
Open Access
International Journal of Engineering Research and Generic Science (IJERGS)
Available Online at www.ijergs.in
ISSN: 2455 - 1597
Volume -3, Issue-2, March – April 2017, Page No. 19 - 26

Discrete Wavelet Transform Based Feature Extraction and Denoising of Speech Signal

Swathy M S[1], Mahesh K R[2]

[1] PG Scholar, Dept of ECE, Thejus Engineering College, Thrissur, Kerala

Email: [1]ms.swathy.ms@gmail.com

[2]Assistant Professor, Dept of ECE, Thejus Engineering College, Thrissur, Kerala

Abstract

Speech is one of the ancient ways to express ourselves. Today speech signals are used for biometric recognition technologies and machine communication. The main function and intention of speech is communication and it is still the first and foremost means of communication. Speech signal is a one dimensional stream of data. Unlike other signals which are stationary in nature, speech is non stationary in nature where the frequency changes with time. Feature extraction and denoising are two major part of the speech recognition. Feature extraction is used to seperate one speech from other and this has been important area of research for many years. Selection of the feature extraction techniques plays an important role in the recognition accuracy. Which is the main criterion for a good speech recognition system. Denoising is the process of removing noise from the speech signal. Denoising is used for speech enhancement. Speech enhancement process aims to improve the speech over all quality.

Keywords: Denoising, Feature Extraction, Discrete Wavelet Transform (DWT).

1. Introduction

The fundamental purpose of speech is communication i.e., the transmission of messages. According to Shannon's information theory, a message represented as a sequence of discrete symbols can be quantified by its information content in bits, and the rate of transmission of information is measured in bits/second (bps). In speech production, as well as in many human-engineered electronic communication systems, the information to be transmitted is encoded in the form of a continuously varying (analog) waveform that can be transmitted, recorded, manipulated, and ultimately decoded by a human listener. In the case of speech, the fundamental analog form of the message is an acoustic waveform, which we call the speech signal. Speech is produced when air is forced from the lungs through the vocal cords and along the vocal tract. The vocal tract extends from the opening in the vocal cords (called the glottis) to the mouth, and in an average man is about 17 cm long. It introduces short-term correlations (of the order of 1 ms) into the speech signal, and can be thought of as a filter with broad resonances called formants. The frequencies of these formants are controlled by varying the shape of the tract, for example by moving the position of the tounge. An important part of many speech codecs is the modelling of the vocal tract as a short term filter. As the shape of the vocal tract varies relatively slowly, the transfer function of its modelling filter needs to be updated only relatively infrequently (typically every 20 ms or so). Speech signal can be classified into voiced, unvoiced and silence regions. The near periodic vibration of vocal folds is excitation for the production of voiced speech. The random, like excitation is present for unvoiced speech. There is no excitation during silence region. Majority of speech regions are voiced in nature that include vowels, semi vowels and other voiced components. The voiced regions looks like a near periodic signal in the time domain representation. In a short term ,we

may treat the voiced speech segments to be periodic for all practical analysis and processing. The periodicity associated with such segments is defined is 'pitch period T_o' in the time domain and 'Pitch frequency or Fundamental Frequency F_0' in the frequency domain. Unless specified, the term 'pitch' refers to the fundamental frequency ' F_0'. Pitch is an important attribute of voiced speech. It contains speaker-specific information. It is also needed for speech coding task. Thus estimation of pitch is one of the important issue in speech processing.

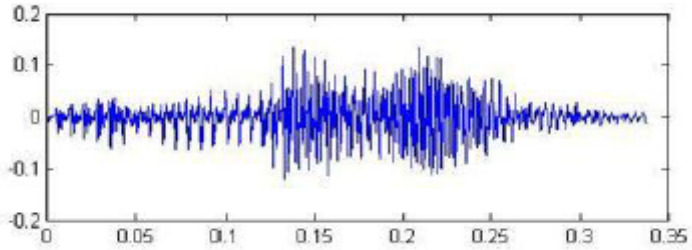

Figure 1: Block diagram of human speech production mechanism

2. Human Speech Production

Speech production in human beings is a very common phenomenon which is experienced in their day to day life and is also part of speech communication between them. Although, speech production looks very simple from outside but its inside mechanism is very complex. Human being can generate many varieties of sounds whose frequency spectrum as well as the loudness changes very rapidly. This is possible due to the very sharp and precise articulatory movement control of the organs of speech production mechanism. Figure 3.1[2] shows the block diagram of speech production mechanism The articulatory movement control is termed as Motor Control and is done by human brain through sensory nerve system connecting brain to the speech production organs such as lungs, vocal chords, tongue, jaw, lips, teeth, larynx etc.

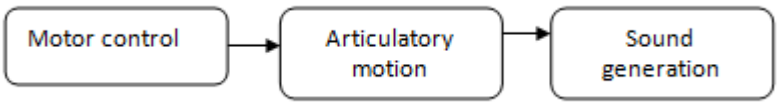

Figure 2: Block diagram of human speech production mechanism

When we speak the air expelled from the lungs moves up through trachea and enters larynx. In the larynx the air is restricted by a pair of lip like tissues called vocal cords. These are very important membranes of vocal apparatus which decide the pitch of the speech produced. Vocal cords are pearly white in colour fixed at one end attached to the arytenoids cartilages at the back and to the thyroid cartilage at the front. Generally the males have low pitch voice with large larynx as compared to the high pitch voice for females with small larynx. The length of vocal folds for males varies from 17 to 25 mm and for females it varies from 12.5 to 17.5 mm. The vocal folds vibrate to produce voiced speech and provide momentarily restriction to produce unvoiced speech. In fact there are various other types of speeches called phonemes for which vocal cords open or close in different fashion to let the air pass through it and send it to the upper part of the vocal tract. Vocal tract is the tube like passage which runs from glottis at one end and with two openings, oral and nasal cavity, at the other end. It is of non uniform cross section whose approximate length for males is about 17 cm. It branches out at

soft palate (velum), just at half way of the tract, and opens up at nostril as second branch. This part of the vocal tract is approximately 13 cm long. Air, after leaving the vocal cords enters into the pharyngeal, mouth and nasal cavities which provide necessary resonation to the sound as per word to speak by amplifying some of the frequencies and attenuating others. Other organs in the mouth such as soft palate, teeth, tongue, lips, jaw change their shape and move accordingly providing blockage or allowance to the air for the mouth and nasal exit and thus modulate the sound to give necessary shape and amplitude. Because of the difference in size and shape of different speech production organs, speech of an individual is unique. The beauty of the whole articulatory movement system is that even after having so many complexities it is able to react very fast catering to the fast changing speech parameters. Epiglottis and false vocal cords below the pharynx has an important role of preventing food to enter into the larynx and isolate the esophagus acoustically from the vocal tract.

3. Feature Extraction

Speaker recognition is the process of recognizing automatically who is speaking on the basis of individual information included in speech waves. This technique uses the speaker's voice to verify their identity and provides control access to services such as voice dialing, database access services, information services, voice mail, security control for confidential information areas, remote access to computers and several other fields where security is the main area of concern. At the highest level, all speaker recognition systems contain two main modules . feature extraction and feature matching. Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing extracted features from his/her voice input with the ones from a set of known speakers.

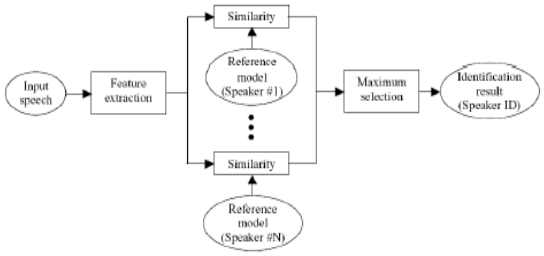

Figure 3: Speaker identification

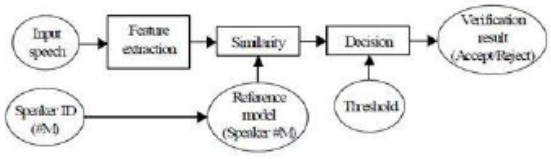

Figure 4: Speaker verification

1**. Speaker identification:** It is the process of determining which registered speaker provides a given utterance.

2. **Speaker verification:** It is the process of accepting or rejecting the identity claim of a speaker. Figure 2 & and figure 3 in the next page illustrate the basic differences between speaker identification and verification systems.

Feature Extraction: The purpose of this module is to convert the speech waveform into a set of features or rather feature vectors used for further analysis. This is often referred to as the signal-processing front end. The speech signal is a called a quasi-stationary signal i.e. a slowly timed varying signals.

4. Discrete Wavelet Transform

The wavelet transform is very well suited for speech processing because of its similarity to how the human ear processes sound and it is a multi-resolutional and multi-scale analysis..Many natural signals, however, can be better analyzed with different time-frequency resolutions in subbands.As an examble , speech signals require a high frequency resolution at lower frequencies in order to resolve pitch and forman frequencies.Although the discretized continuous wavelet transform enables the computation of the continuous wavelet transform by computers, it is not a true discrete transform. As a matter of fact, the wavelet series is simply a sampled version of the CWT, and the information it provides is highly redundant as far as the reconstruction of the signal is concerned. This redundancy, on the other hand, requires a significant amount of computation time and resources. The discrete wavelet transform (DWT), on the other hand, provides sufficient information both for analysis and synthesis of the original signal, with a significant reduction in the computation time.The DWT is considerably easier to implement when compared to the CWT. Wavelet transforms were introduced to address the problems associated with non-stationary signals like speech. A Wavelet transform decomposes a signal into a set of basic functions called wavelets.

DWT is a special case of the wavelet transform that provides a compact representation of a signal in time and frequency that can be computed efficiently. They are well suitable for processing signals like speech because of their efficient time-frequency localization and the multi-resolutional, multi-scale analysis characteristics of the wavelet representations. The DWT, which is based on subband coding, is found to yield a fast computation of Wavelet Transform. It is easy to implement and reduces the computation time and resources required. In continuous wavelet transform (CWT), the signals are analyzed using a set of basis functions which relate to each other by simple scaling and translation. In the case of DWT, a time scale representation of the digital signal is obtained using digital filtering techniques. The signal to be analyzed is passed through filters with different cutoff frequencies at different scales. In the discrete wavelet transform, a signal can be analyzed by passing it through an analysis filter bank followed by a decimation operation. When a signal passes through these filters, it is split into two bands. The low pass filter, which corresponds to an averaging operation, extracts the coarse information of the signal. The high pass filter, which corresponds to a differencing operation, extracts the detail information of the signal. The output of the filtering operations is then decimated by two. Filters are one of the most widely used signal processing functions. Wavelets can be realized by iteration of filters with rescaling.

The DWT is computed by successive low pass and high pass filtering. At each decomposition level, the half band filters produce signals spanning only half the frequency band. This doubles the frequency resolution as the uncertainty in frequency is reduced by half. In accordance with Nyquist's rule if the original signal has a highest frequency of ω, which requires a sampling frequency of 2ω radians, then it now has a highest frequency of $\omega/2$ radians. It can now be sampled at a frequency of ω radians thus discarding half the samples with no loss of information. This decimation by 2 halves the time resolution as the entire signal is now represented by only half the number of samples. Thus, while the half band low

pass filtering removes half of the frequencies and thus halves the resolution, the decimation by 2 doubles the scale. The filtering and decimation process is continued until the desired level is reached. The maximum number of levels depends on the length of the signal. The DWT of the original signal is then obtained by concatenating all the levels.

The Discrete Wavelet Transform is defined by the following equation.

$$W(j,K) = \sum_{j}\sum_{k} X(k)2^{-j/2}\psi(2^{-j}n-k) \quad \text{.....(1)}$$

where $\Psi(t)$ is the basic analyzing function called the mother wavelet. Other functions are derived from this mother wavelet by translation and dilation operations. In DWT, the original signal passes through two complementary filters, namely low-pass and high-pass filters, and emerges as two signals, called approximation coefficients and detail coefficients. In speech signals, low frequency components known as the approximations h[n] are of greater importance than high frequency signals known as the details g[n] as the low frequency components characterize a signal more than its high frequency components . The successive high pass and low pass filtering of the signal can be obtained by the following equations.

$$Y_{high}[k] = \sum_{n} x[n]g[2k-n]$$

$$Y_{low}[k] = \sum_{n} x[n]h[2k-n] \quad \text{.....(2)}$$

Where Yhigh (detail coefficients) and Ylow (approximation coefficients) are the outputs of the high pass and low pass filters obtained by sub sampling by 2. The filtering is continued until the desired level is reached according to Mallat algorithm [8]. The main advantage of the wavelet transforms is that it has a varying window size, being broad at low frequencies and narrow at high frequencies, thus leading to an optimal time–frequency resolution in all frequency ranges.

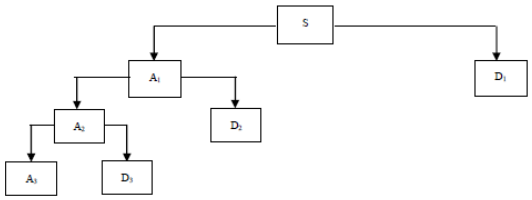

Figure 5: DWT

5. Denoising Of Speech Signal Using Wavelet Transform

During transmission and reception signals are often corrupted by noise which is unwanted signal. There are many forms of noise. One of the most common sources of noise is background noise which is always present at any location. Other types of noise include channel noise which affects both analog and digital transmission, quantization noise which results from over compression of speech signals, multi talker babble, reverberation noise or delayed version of noise are also present in some situations. To overcome these problems we are introduce enhancement terminology in speech processing.

6. Algorithm Implementation

In Automatic 1-D Denoising, Denoising is performed using one dimensional Denoising function. It performs an automatic de-noising process of a one-dimensional signal using wavelets and returns a de-noised version of input signal obtained by

thresholding the wavelet coefficients. The Denoising objective is to suppress the noise part of the signal and to recover the original one. The de-noising procedure proceeds in three steps.

a. Decomposition: Choose a wavelet, and choose a level N. Compute the wavelet decomposition of the signal s at level N.
b. Detail coefficients thresholding: For each level from 1 to N, select a threshold and apply soft/hard thresholding to the detail coefficients
c. Reconstruction: Compute wavelet reconstruction based on the original approximation coefficients of level N and the modified detail coefficients of levels from 1 to N.

a) Decomposition In this algorithm, the Daubechies and Symlets wavelet with a decomposition tree of level 3 is used; because it can provide a well orthogonality to high frequency noise with a given number of vanishing moments. The decomposition tree with wavelet coefficients at different levels, in which the boxes of approximations cA1,cA2,cA3 represents the low frequency components obtained by low pass filter, and the boxes of details cD1, cD2,cD3 represents the high frequency components obtained by high pass filter and is represented in below figure,

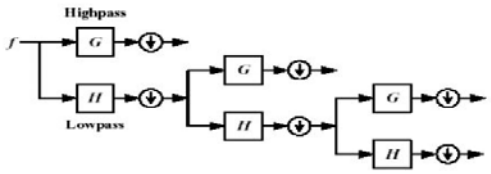

Figure 6: Decomposition tree of DWT

b) Threshold Selection: Wavelet thresholding is a signal estimation technique that exploits the capabilities of wavelet transform for signal Denoising. As one may observe, threshold selection is an important question when denoising. A small threshold may yield a result close to the input, but the result may still be noisy. A large threshold on the other hand, produces a signal with a large number of zero coefficients. This leads to a smooth signal. Paying too much attention to smoothness, however, destroys details and in image processing may cause blur and artifacts. Thresholding is a simple non-linear technique, which operates on one wavelet coefficient at a time. In its most basic form, each coefficient is threshold by comparing against threshold, if the coefficient is smaller than threshold, set to zero; otherwise it is kept or modified. Replacing the small noisy coefficients by zero and inverse wavelet transform on the result may lead to reconstruction with the essential signal characteristics and with less noise. Since the work of Donoho & Johnstone there has been much research on finding thresholds. There are two types of thresholding available.

They are as:

1. Hard Thresholding
2. Soft Thresholding

1. Hard thresholding

In this method input is kept only if it is larger than the threshold T, otherwise it is set to zero. Y =T (X, Y) = X; for $|X| > \lambda$

0; for |X| ≤ λ In the hard thresholding scheme given in equation 1, the input is kept, if it is greater than the threshold λ, otherwise it is set to zero. The hard thresholding procedure removes the noise by thresholding only the wavelet coefficients of the detailed sub bands, while keeping the low-resolution coefficients unaltered.

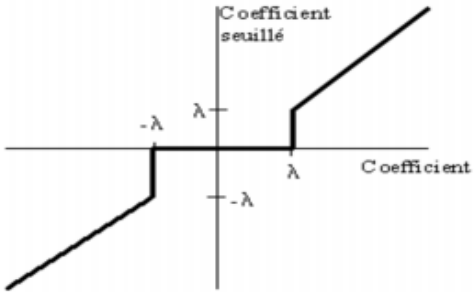

Figure 7. Characteristics of Hard Thresholding

1. Soft thresholding

It is also called shrinkage function. It takes the argument and shrinks it towards zero by the threshold T. Y =T (X, Y) = sign{X} (|X|-1); for |X| > λ = 0 ; for |X| ≤ λ.

The soft thresholding scheme shown in equation 2 is an extension of the hard thresholding. If the absolute value of the input X is less than or equal to λ then the output is forced to zero. If the absolute value of X is greater than λ then the output is |Y| = |X - 1|. When comparing both hard and soft shrinking schemes graphically from the Figures 2 and 3, it can be seen that hard thresholding exhibits some discontinuities at λ and can be unstable or more sensitive to small changes in the data, while soft thresholding avoid discontinuities and is therefore more stable than hard thresholding.

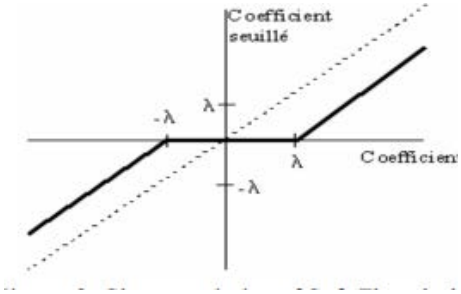

Figure 8: Characteristics of Soft Thresholding

7. Conclusion

Feature extraction produces a stream of vectors that may represent different characteristics depending on the technique used for feature extraction. Before the invention of wavelets, Fourier Transforms (FT) and Short Time Fourier Transforms (STFT) were the dominant spectral analysis tools for frequency domain analysis. The main problem with FT was that FT provided information regarding only amplitude-frequency representation and did not provide any information about the element of time .In order to overcome this limitation, the STFT, which can be considered a Windowed Fourier Transform (WFT), was introduced which was used to provide frequency-time spectrum, by providing time information. But the main limitation of STFT was that it used fixed window size by which the accuracy relied on the size and shape of the window.

This greatly affected both frequency and time resolution. So wavelet transforms (WT) were developed in order to overcome the shortcomings of FT and STFT related to its time and frequency resolution problems and it provided a concise and easier analysis of speech signals which is suitable for the non stationary nature of the speech signal. The newly proposed denoising technique which was applied during the pre-processing stage played a significant role in removing the sudden spikes due to noise.

8. References

[1] Urmila Shrawankar and Dr.Vilas Thakare," Techniques for feature in speech recognition system: A comparative study."

[2] V.S.R Kumari and Dileep Kumar Devarakonda,"A Wavelet Based Denoising of Speech Signal."

[3] Pratik K.Kurzekar,R Deshmukh,"A Comparative Study of Feaure Extraction Techniques for speech recognition system."

[4] KamyaDubey and Vikas Gupta ,"A Review On Speech Denoising Using Wavelet Techniques."

[5] Dr.Mahesh S.Chavan ., Mrs Manjusha N .Chavan,"Studies On Implementation Of Wavelet for Denoising Speech Signal."

[6] Lawrence R. Rabiner, and Ronald W. Schafer, "Introduction to Digital Speech Processing," Foundations and Trends in Signal Processing, vol. 1, nos. 1–2, pp. 1-194, Jan. 2007."

[7] Samudravijaya K., "Speech and Speaker recognition: a tutorial," in Proc. International Workshop on Technology Development in Indian Languages, Kolkata, Jan. 2003.

[8] Kenneth Thomas Schutte, "Parts-based Models and Local Features for Automatic Speech recognition," Ph.D. dissertation, Dept. of Elec. Eng. and Comp. Sci., Massachusetts Inst. of Tech., Massachusetts, 2009.

[9] O. Scharenborg, "Reaching Over the Gap: A Review of Efforts to Link Human and Automatic Speech Recognition Research," Speech Communication, vol. 49, pp. 336–347, May 2007.

[10] R. P. Lippmann, "Speech Recognition by Machines and Humans,"Speech Communication, vol. 22, pp. 1–15, Apr. 1997.

[11] Kuldeep Kumar, and R. K. Aggarwal, "Hindi Speech Recognition System Using Htk," International Journal of Computing and Business Research, vol. 2, no. 2, pp. 2229-6166, May 2011.