Applications of Wavelet Transform In Speech Processing: A Review

Sreeja.R.Nair
Dept of Electronics and Telecommunication
Fr.C.R.I.T,
Sector 9A,Vashi-400703
Navi Mumbai, India

Dr.Milind S. Shah
Dept of Electronics and Telecommunication
Fr.C.R.I.T,
Sector 9A,Vashi-400703
Navi Mumbai, India

***Abstract*— The aim of this paper is to review applications of wavelet transform in different speech processes namely denoising, compression, pitch detection, recognition, steganography,etc. The main advantage of wavelets over other transforms is its multi resolution property i.e. localization in both time and frequency which is highly suitable for speech signals. Moreover they are greatly robust to noise and provide better compression ratios.**

***Keywords*—wavelets, denoising,speech compression, pitch**

I. INTRODUCTION

The advent of technology in the field of telecommunication has lead to development of human-machine interactions. With increasing sophistication of the cellular phones and easy access to internet resources, people are more prone to use electronic and telephonic transactions by the day. A speech signal can be made an important source of information if it is noise robust, accurate and if it can be processed fast[1].Additionally, with effortless access to information the need for privacy and security arises. This concern can be addressed by controlling the access using speech itself like speaker identification and verification. This establishes the role of speech processing techniques in health care, military, automation systems, telephony, banking, etc. Speech processing techniques are based on either time analysis methods or frequency analysis methods. Time based analysis is fast and best suited for low noise conditions where the signal to noise ratio is high. However that might not be the case every time. This disadvantage is removed while using frequency domain analysis. However, while transforming the speech signal from time to frequency domain some of the information is lost that is the time at which each frequency is obtained cannot be found from frequency transform. Hence a time-frequency representation is apt for speech processing. This joint time- frequency representation is the wavelet transform.

This paper aims to explain the improvements brought about by this multi resolution analysis of wavelets in speech processing. Wavelets are successful in speech signals because of two main reasons: (1) it gives sufficient results in high noise environment with highly efficient denoising. (2) it gives a good performance even for non-stationary signals like speech unlike other transforms.

In section II, the reasons why wavelets were needed over the Fourier transforms are discussed. In section III, how waveletsare practically used for the different speech processing techniques is reviewed.

II. NEED FOR WAVELETS

Wavelet as the name suggests is a small part of a wave which increases from zero and comes back to zero i.e. it integrates to zero. It has a finite energy and hence can be put under the L2 space. Each wavelet has a characteristic location and scale. Like Fourier Transform is represented in terms of sine and cosine waves, wavelets are represented by basis function and wavelet function. Scaling function or father wavelet (ϕ) characterizes basic wavelet scale whereas wavelet function or mother wavelet (ψ) characterizes basic wavelet shape. The normalized wavelet transform of a signal *x(t)* can be given as [1]:

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

where a and b are real numbers and
a = scaling factor which determines the resolution ($a \neq 0$)
b = translation factor which determines the number of coefficients.

The traditional method of Fourier Transforms has been replaced using wavelets due to their numerous advantages over the former. Smooth sinusoids are used to represent a signal in Fourier Transform. These smooth signals though have excellent mathematical properties, are not practically achievable since they extend to infinity[2]. Wavelet functions on the other hand are compactly supported in the time as well as frequency domain and can be extended to cover the entire region of interest. Moreover Fourier transform is unsuitable for non stationary signals where the frequency components change with time. On the contrary, wavelets, because of their localization in time as well as frequency are suitable for non stationary signals. Short Time Fourier Transform (STFT) was introduced to compensate for this drawback, by dividing the signal into fixed time window and then finding its frequency transform. However in this method length of window is very

crucial. If it is chosen to be too long then some vital information might get lost and on the otherhand if it is chosen to be small then it might not suffice to extract the required parameters. Theseshortcomingsare also compensated by wavelets by providing a variable length resolution. The lower frequency components change very slowly in time hence their time resolution may be less. Higher frequency components undergo rapid fluctuation intime and hence it requires good time resolution. Wavelets provide localization in time and frequency, using thisprinciple. The objective is to minimize reconstructed error variance and maximize signal to noise ratio (SNR) [2].

III. SPEECH PROCESSING USING WAVELETS

A. Speech Denoising

It is known that signals do not exist without noise. Sometimes it may be negligible but otherwise it can be significant. In this case it becomes necessary to remove the noise for further analysis to be done. The main aim here is to retrieve the signal by eliminating the noise without any changes in the original signal. Wavelets can be efficiently used for removal of noise from image and speech.Although the noise could be in general of any type, in most applications it can be assumed to be Additive White Gaussian Noise (AWGN) which contaminates the digital data. It can be represented as [3]:

$$y_i = f_i + \sigma\varepsilon_i \quad (2)$$

where$i$= no of samples and ranges from 1 to the length of the signal
σ = noise level which may be known or unknown
ε = iid$N(0,1)$

Wavelet denoising works for additive noise since wavelet transform is linear. Usually the wavelet coefficients of a clean speech signal have values which are effectively zero. However when noise is added to this signal, the amplitude of the coefficients increases. When the wavelet transform of a noisy signal is done, due to the orthogonal nature of the wavelets, the white noise is transformed as white noise itself. Thus the wavelet coefficients of noise will have smaller amplitude as compared to the signal's coefficients. Thus the main aim of denoising is to recover the coefficients which are relatively stronger than the white Gaussian noise present in the background. Thus the concept of thresholding comes into picture where each wavelet coefficient is compared with a threshold value to determine whether it is a part of the original signal or not. If it is found to be noise then those coefficients are set to zero. Thresholding is usually done on the detail coefficients and not the approximation coefficients [3]. This is because the approximation coefficients mostly represent the low frequency signal which consists of most of the information and is less affected by noise. The thresholded wavelet coefficients can be found using the following methods:

- *Hard Thresholding* : in this process the coefficients below the threshold is set to zero. Otherwise the coefficients are kept as it is. It can be given as [4]

$$x(k) = 0 \quad if\ |x(k)|<\lambda$$
$$= x(k)\ if\ |x(k)|>\lambda$$

where λ is the threshold value
x is the input signals and k is the sample

- *Soft Thresholding*:In this method the coefficients below the threshold is set to zero and additionally the higher amplitude coefficients are shrinked towards zero.

$$x(k) = 0 \quad if\ |x(k)|<\lambda$$
$$= x(k) - \lambda \quad if\ x(k)> \lambda$$
$$= x(k) + \lambda \quad if\ x(k)< -\lambda$$

The threshold value can be determined by universal thresholds like MinimaxandVisuSure or adaptive thresholds like RigSure and HeurSure. An implementation of denoising a signal corrupted by AWGN obtained from Matlab tool, using universal threshold VisuSure is illustrated in Fig 1.

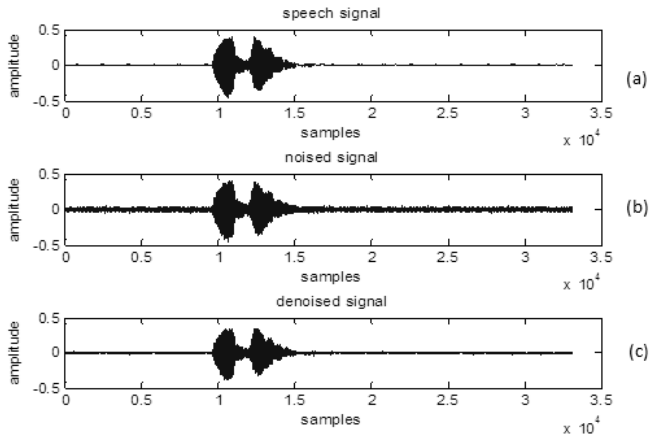


Fig1. Waveforms of (a) speech signal, (b) signal contaminated with AWGN and (c) denoised signal using global threshold.

The Mean Square Error (MSE) between original and reconstructed signal using different wavelets at level 4 are shown in table 1. It is observed here that as vanishing moment of the wavelet increases, the MSE decreases

TABLE I.
MEAN SQUARE ERROR FOR SIX DIFFERENT WAVELETS

Wavelet	Mean Square Error
Haar	0.0030
Db2	0.0024
Db4	0.0018
Db6	0.0018
Db8	0.0017
Db10	0.0018

B. Speech Compression

Speech signals are naturally redundant. For digital processing and storage keeping all the redundant samples is not feasible. This is the reason for using speech compression. Compression can be either lossy or lossless. In speech signals

a small amount of loss in the signal is mostly tolerable taking human perception into consideration. Here the main aim is to retain a small number of approximation coefficients along with even fewer detail coefficients that can accurately represent the original signal. Compression consists of five main steps: wavelet decomposition, thresholding, quantization, encoding and reconstruction by inversing the previous steps.

After applying wavelet transform to the speech signal, these coefficients are thresholded. Coefficients whose magnitude is lesser than the threshold are set to zero. The threshold is generally defined using Birge-Massart formula [5] where the approximation coefficients at the last level is retained and the number of detail coefficients to be kept is determined by [5]

$$n_i = \frac{M}{(J+2-i)^{\alpha}} \quad (3)$$

where α = compression factor which is greater than 1. It is typically taken to be 1.5.
i= levels from1 to J
J = total number of levels chosen
M= measure of scarceness and is equal to length L of coarse approximation coefficient if it is highly scarce i.e. close to zero.Afterthresholding, a uniform or adaptive quantization can be applied followed by an efficient coding technique. Most frequently Huffman or Run Length Encoding is used. Once this is done the original signal can be reconstructed by decoding, de quantization and inversing the wavelet transforms. A compressed signal using the above thresholding method is implemented in this paper and its output can be seen in Fig 2. The spectrogram of the original and compressed signal obtained using Praat software is shown in Fig 3. It is seen from the spectrogram that the formant are retained even after compression.

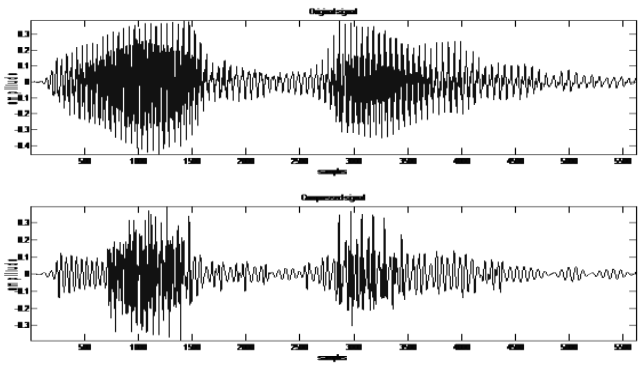
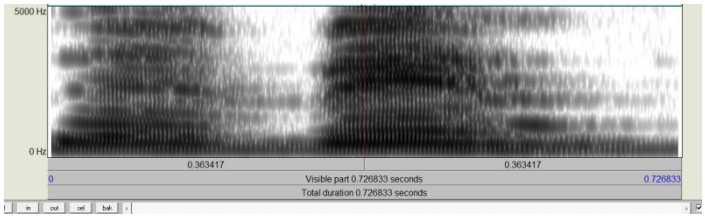
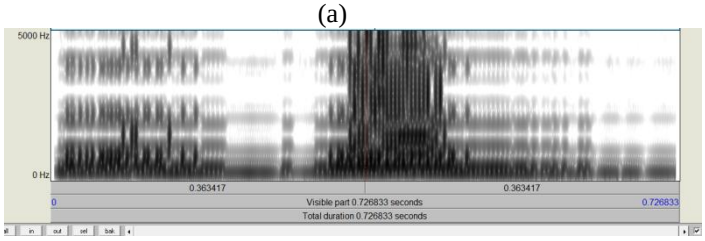


Fig2: The original and compressed signal using Birge-massart threshold



(a)



(b)

Fig 3: The spectrogram of (a)original signal (b)compressed signal

When the spectrograms were observed for formant values using the Praat software for both the original and compressed signals, it was foun that for the unvoiced sections all the formants had more or less the same value. Though it was not the case for voiced sections were the higher formants were considerably distorted.

The metrics to determine the quality of compression are Signal to Noise Ratio (SNR), Peak Signal To Noise Ratio (PSNR) and Normalized Root Mean Square Error (NRMSE) given by [6]

$$SNR = 10\log_{10}\left(\frac{\sigma_x^2}{\sigma_e^2}\right) \quad (4)$$

where σ_x^2 = mean square of speech signal
σ_e^2= mean square difference between original and reconstructed signal.

$$PSNR = 10\log_{10}\frac{NX^2}{\| x - x' \|} \quad (5)$$

where N = length of reconstructed signal
X= maximum absolute square value of the signal
$\| x - x' \|$ = energy of the difference between original and reconstructed signal.

$$NRMSE = \sqrt{\frac{\sum_n (x(n) - x'(n))^2}{\mu_n (x(n) - \mu_x(n))^2}} \quad (6)$$

where x(n) = speech signal
x'(n) = reconstructed signal
$\mu_x(n)$ = mean of the speech signal.

In [1], Najih and Syed have computed these metrics for different wavelet functions like Haar, Db2, Db4, Db6, Db8 and Db10 and it was found that the SNR and PSNR was highest and NRMSE was lowest for Db10 for both male and female speakers. They also found that wavelet based compression a signal to noise ratio of upto 17.98 db with a compression ratio of 4.31 could be obtained using Db10. In [8], James and Thomas computed these metrics for FFT, DCT and different DWT and compared the results only to find that wavelets were superior to the other conventional methods. In [9], Korning compared the performances of Bior, Haar and Meyer wavelet in speech compression and found

that Bior wavelet provides the best performance in time as well as frequency domain.

C. Pitch Detection of Speech

Pitch period is the period with which the vocal cord or glottis vibrates when air is pushed through it inorder to produce a voiced sound. Pitch period detection is one of the most difficult problems in speech processing due to variations of the vocal tract from person to person and also for the same person at different emotional state. Pitch detection can be event based like autocovariance method, taking derivatives of glottal air flow, etc. or it can be non event based like autocorrelation, cepstrum method, etc. But these methods are unsuitable for non stationary period. When wavelets are used to determine pitch either by event based or by non event based method it becomes possible to determine non stationary pitch.

In [10], Kadambe and Srinivasan have used wavelets in an event based method to determine the pitch. The concept used by them was that if a signal or its derivatives have any discontinuities then the modulus of the Dyadic Wavelet Transform $|DWT(b,2^j)|$ will have its local maxima around the point of discontinuity where $a=2^j$ is the scale This discontinuity gives the pitch as it denotes the sudden closing of glottis. If for three successive scales, we obtain the local maxima at the same instant then that indicates glottal closure and the difference between two such maximas give the pitch period.

Gody, in [1], has implemented pitch detection using cross correlation of wavelet coefficients. In this method the speech signal is first low pass filtered and classified as voiced or unvoiced. Only the voiced section is used for this algorithm which is segmented and divided into frames. The wavelet coefficientsfor each frame are then computed. Then the crosscorrelation between two adjacent frames with frequency range between 100 Hz to 1000 Hz is done. The peaks of the correlation function represent the fundamental pitch and its harmonics. The output of this method is then compared with other conventional methods and it was found that pitch variation could be determined better using wavelets even in low signal to noise conditions. The pitch contour generated using wavelets are also smoother than that generated by other methods.

D. Speech Recognition

Speech recognition is a process by which computer recognizes human speech. Speech recognition consists of two main stages. One of which is feature extraction and the other is recognition module. Out of these two,wavelets can be used infeature extraction of complex speech signals. The wavelet parameters extracted from thespeechsignaladequately represents the original signal. The lower band coefficients or approximation coefficients represent the essential low frequency information. In order to further reduce the parameters some statistical characteristics of the wavelet coefficients are used instead of the coefficient itself. Some of these statics derived from the signal are [12]:

- mean of the absolute value of coefficients in each sub band
- the standard deviation of the coefficients in each sub band.
- energy of each sub band
- kurtosis of each sub band
- skewness of each sub band

The recognition module can be mainly of three types. Dynamic Time Warping finds the minimum distance between the parameter vectors of test and reference signal by warping one on the other. Hidden Markov Model which is the second method is a statistical model which is designed using the feature parameters. The third and the latest method is the neural networks which consists of massive interconnected networks representing the working of neurons in human brain. The inputs to all these networks are the extracted features which in this case will be obtained using wavelets. These features can also be used along with speech emotion recognition systems like Support Vector Machine (SVM) and k-Nearest Neighbour (k-NN). SVM classifies the feature vectors into different classes each defining the different emotions. Thus the emotions can be recognized from the feature vectors. While in kNNthe feature vectors are assigned different classes.K is a user defined boundary parameter which is used to select the most frequently occurring class.

E. Other Applications

Some other applications where wavelets can be successfully used in speech are speech steganography, clinical diagnosis of speech disorders and speech separation. Speech Steganography deals with hiding a secret speech signal within another cover signal such that any person not intended will not even know the presence of a secret data in the main or cover signal. In [14],Rekik and Guerchihaveused DWT-FFT based approach for hiding the signal. Here the DWT splits the signal into low and high frequency parts and the FFT of higher frequency components or the detailed coefficients are taken inorder to obtain magnitude and phase spectrum. Then the last few elements of the magnitude spectrum is replaced by the feature vector of the secret signal that has to be hidden. Another approach is to hide the secret data in the LSB bits of the detailed coefficients obtained from the wavelet transform[15].In these two approaches detailed coefficients are tampered with because a change in the high frequency components negligibly affects the original signal.

The clinical diagnosis of speech disorders using wavelets is based on the principle that the amplitude of wavelet coefficients for pathological speech, which is nothing but the speech spoken differently due to some pathological reasons, are significantly lesser and disturbed than a normal speech. Pitch detection of such a speech using wavelets also provide better assessment of roughness, hoarseness and breathiness [16].

Speech separation is a process of isolating or extracting a particular speech signal from a mixture of many signals coming from different sources. This is done by a combination of DWT and Independent Component Analysis (ICA) assuming that the sources are independent. Here DWT is used

to decompose the signal into different subbandsand give a linear representation of the signals and ICA statistically separates the signal [17]. Using wavelets with ICA has proved to give better results than using only ICA.

IV. CONCLUSION

Wavelet serves as a powerful tool in various speech processing techniques.Indenoising, wavelets help to remove additive noises, even which are present at every level, by using the level dependent thresholds. If wavelets with higher vanishing moments are chosen then a signal with order less than the vanishing moments on wavelet decomposition will yield more coefficients which are zero. Thus wavelet forms an excellent compressor of signals. Another undisputed advantage of wavelet transform is that it can work well with non stationary signals and hence is an attractive option for pith detection. The variation of pitch in each frame is represented accurately using wavelets.It has also been observed that as the level and vanishing moments of the wavelet are increased, better results are found. However a level beyond five increases computational complexity and provides no significant improvements in all these applications.

REFERENCES

[1] A. M. R. M. Gody, "Speech Processing Using Wavelet Based Algorithms," Ph.D. dissertation, Dept. of Electron. andCommun. Eng., Cairo. Univ., Giza, Egypt,1999.
[2] V. M. Gadre, "Advance digital signal processing- Multirate and Wavelets" [Online] Available: http://nptel.ac.in/courses/117101001
[3] R. Cohen, "Signal Denoising Using Wavelets" [Online] Available: http://tx.technion.ac.il/˜rc
[4] S. G. Mihov, R. M. Ivanov and A. N. Popov, "Denoising Speech Signals by Wavelet Transform," *Annual Journal of Electronics*, 2009
[5] A. M. M. A. Najih, A. R.binRamli, V. Prakash, and A. R. Syed, "Speech Compression Using Discrete Wavelet Transform", *Proc. Fourth National Conference on Telecommunication Technology*, pp. 1-4, 2003
[6] S. M. Joseph and B. Anto, "Speech Compression Using Wavelet Transform," in *Proc..of ICRTIT*, Chennai, 2011, pp 754-758.
[7] D. M. Rasetshwane, "Identification Of Transient Speech Using Wavelet Transforms," M.S. Thesis, Dept. of Elect. andComput. Eng., University of Pittsburgh, 2002.
[8] J. James and V. J. Thomas, "A Comparative Study of Speech Compression using Different Transform Techniques," *International Journal of Computer Applications,* vol. 97, no. 2, pp. 16-20, 2014.
[9] S. Kornsingand J. Srinonchat, "Enhancement Speech Compression Using Modern Wavelet Transforms," in*Symp. 2012 International Symposium on Computer, Consumer and Contro*l, 2012, pp. 393-396.
[10] S. Kadambe, and P. Srinivasan, "Application of the Wavelet Transform for Pitch Detection of Speech Signals", *IEEE Transaction on Information Theory*, vol. 38, no. 2, pp. 917-924, 1992.
[11] S. H. Chen and J. F. Wang, "Noise-robust Pitch Detection Method Using Wavelet Transform With Aliasing Compensation", *IEE Proceedings of Vision, Image and Signal Processing*, vol. 149, no. 6, pp. 327 – 334, December 2002.
[12] V. Kumar, N. Trivedi, S. Singh et all, **"**Speech Recognition by Wavelet Analysis," *International Journal of Computer Applications,* vol. 15, no. 8, pp. 27-32, 2011.
[13] M. S. Chavan and M. S. Gaikwad,"Studies on Implementation of Wavelet for Denoising Speech Signal," *International Journal of Computer Applications*, vol. 3, no. 2, pp. 1-7, 2010.
[14] S. Rekik, D. Guerchi, H. Hamam, *et al*. (2012, Aug. 8) *Speech Steganography using Wavelet and Fourier Transforms* [Online]. Available: http://asmp.eurasipjournals.com/content/2012/1/20.
[15] S. Shirali-Shahreza and M. T. Manzuri-Shalmani, "High Capacity Error free Wavelet Domain Speech Steganogaphy," in *ICASSP*, Las Vegas, NV, 2008, pp.1729 – 1732.
[16] M. H. Farouk, "Clinical Diagnosis and Assessment of Speech Disorders," *Application of Wavelets in Speech Processing* Springer Science and Business Media, New York, 2013 pp. 49-50.
[17] A. M. Mary, A. P. Kumar and A. A. Chacko, "Blind Source Separation Using Wavelets," in *IEEE International Conf. on Computational Intelligence and Computing Research*,India,2010