A Comprehensive Approach for Speech Related Multimedia Applications

JALAL KARAM
Faculty of Science and General Studies
Alfaisal University
Riyadh
KINGDOM OF SAUDI ARABIA
email: jkaram@alfaisal.edu www.alfaisal.edu/faculty/cv/jalal

Abstract: - In this paper, various speech processing techniques in time, time-frequency and time-scale domains for the purposes of recognition and compression are displayed. The examination of the human cochlea is included revealing practice of Wavelet Transform representation. The interchange between theory and application is displayed in a variety of work that have been accomplished in that direction. In particular, we emphasize the role of Wavelet Transforms in recognizing and compressing speech signals.

Keywords: - Multimedia, Speech Processing, Cochlea Response, Time-Frequency, Wavelet Transform.

1 Introduction

Speech signals are processed for various applications such as recognition, compression, pitch detection and speaker identification to name a few [1] [26] [27] [28]. Each processing task is presented with different sets of challenges and limitations due to the very complex nature of speech. In the recognition problem, the complexity of a system is proportional to the size of the speech set and the speaker dependency required such as single speaker, multi speaker or speaker independent [34] [36].

For compression purposes, one would like to represent a given speech signal with the least possible number of data bits while maintaining acceptable audible reconstructed signal. In this direction, wavelet analysis plays a superior role since it concentrates speech information such as energy and perception into a few neighboring coefficients. This translates into retaining a small number of coefficients to represent a given segment of speech and ignoring the other majority of the coefficients [4] [8]. Threshold takes place at every chosen decomposition level of wavelet analysis [38]. After the segmentation and analysis of the signal, a threshold is applied to the coefficients of each of the levels. This is a lossy algorithm since one retains only those coefficients that contribute the highest energy [23].

Regardless of the task at hand, an initial segmentation procedure of the signal takes place and analysis is conducted on the resulting spectrally stable segments of speech known as subwords. This is particularly useful in the cases of speech recognition and speech compression paradigms, relying on Time-Frequency or Time-Scale analysis.

2 Speech Representations

In order to digitally process a signal $x(t)$, it has to be sampled at a certain rate. 20000 Hz is a standard sampling frequency for the Digits and the English alphabets in [29] and [30]. To make the distinction in the representation with the digitized signals, the latter is referred to as $x(m)$. Most speech processing schemes assume slow changes in the properties of speech with time, usually every 10-30 milliseconds. This assumption influenced the creation of short time processing, which suggests the processing of speech in short but periodic segments called analysis frames or just frames [36]. Each frame is then represented by one or a set of numbers, and the speech signal has then a new time-dependent representation. In many speech recognition systems like the ones introduced in [2] and [32], frames of size 200 samples and a sampling rate of 8000 Hz (i.e., $200*1000/8000 = 25$ milliseconds) are considered. This segmentation is not error free since it creates blocking effects that makes a rough transition in the representation (or measurements) of two consecutive frames. To remedy this rough transition, a window is usually applied to data of twice the size of the frame and overlapping 50% the consecutive analysis window. This multiplication of the frame data by a window favors the samples near the center of the window over those at the ends resulting into a smooth representation. If the window length is not too long, the signal properties inside it remains constant. Taking the Fourier Transform of the data samples in the window after adjusting their length to a power of 2, so one can apply the Fast Fourier Transform [5], results in time-dependent Fourier transform which reveals the frequency domain properties of the signal [31]. The spectrogram is the plot estimate of the short-term frequency content of the signals in which a three-dimensional representation of the speech intensity, in different frequency bands, over time is portrayed [34]. The vertical dimension corresponds to frequency and the horizontal dimension to time. The darkness of the pattern is proportional to the energy of the signal. The resonance frequencies of the vocal tract appear as dark bands in the spectrogram [31]. Mathematically, the spectrogram of a speech signal is the magnitude square of the Short Time Fourier Transform of that signal [3]. In the literature one can find many different windows that can be applied to the frames of speech signals for a short-term frequency analysis.Three of those are plotted in Figure 1.

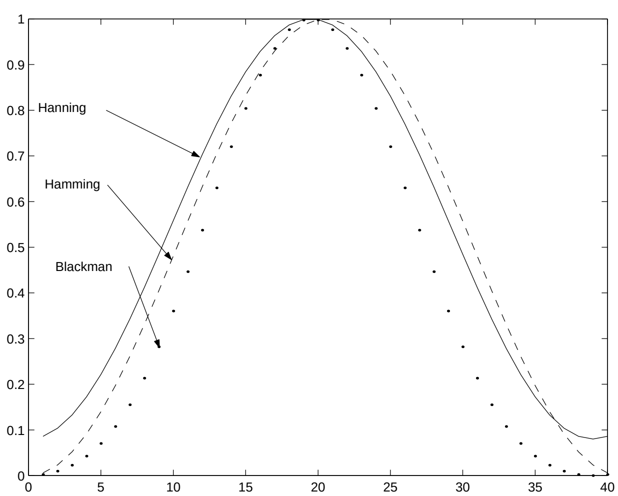

Figure 1: Plots of window functions in the time domain.

2.1 Time-Frequency Representations

Broadly speaking, there are two classes of time-frequency representations, linear and non-linear. The Wigner Distribution is an example of the non-linear class. It was first introduced by Wigner in quantum physics [37]. Gabor introduced the Short Time Fourier Transform (STFT) in 1946 to analyze finite duration signals [13]. The STFT of a signal $x(m)$ as defined in [34] is:

$$X_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-j\omega m}. \qquad (1)$$

where $w(n-m)$ is a real window sequence which determines the portion of the input signal that receives emphasis at the particular discrete time index m. The frequency ω is a normalized frequency with value $2\pi m/F_s$ with F_s representing the sampling frequency of the signal. The properties of the STFT include: homogeneity, linearity, time shift variant and has an inverse. Proofs of these properties can be found in [31] and [36] along with many applications of the STFT in estimating and extracting speech parameters such as pitch and formants. This time-frequency representation allows the determination of the frequency content of a signal over a short period of time by taking the FT of the windowed signal. It also has the ability to capture the slowly varying spectral properties of an analyzed signal. The signal is assumed to be quasi-stationary within the analysis window [36]. Thus the width of the analyzing window has to be carefully chosen. In this time-frequency analysis there are two conflicting requirements. Since the frequency resolution is directly proportional to the width of the analyzing window, good frequency resolution requires a long window and good time resolution, needs a short time length window. This is an immediate disadvantage of the STFT analysis since the window length is kept constant. Hence, there is a time-frequency resolution trade off. This is captured in the uncertainty principal [3] which states that for the pair of functions $x(t)$ and its Fourier Transform $X(w)$ one has:
$\Delta_t \Delta_w \geq 1/2$, Where Δ_t^2 and Δ_w^2 are measures of variations of spread of $x(t)$ and $X(w)$. If one start analyzing with a window of size 20 ms and needed to shorten its size to 10 ms for rapid variation detection, then there will be a loss of frequency resolution. This also increases the computational complexity of the STFT. Another interpretation, is that it can be viewed as the convolution of the modulated signal $x(m)e^{-j\omega m}$ with the analysis filter $w(m)$. Based on this interpretation, the STFT can be implemented by the filter bank approach where the signal is passed through a bank of filters of constant bandwidth since the length of the window is fixed. Thus, the temporal and spectral resolutions are fixed. Filter banks are popular analysis methods of speech signals [33] [34]. In this spectral analysis approach, a digitized speech signal $x(m)$ is passed through a bank of P bandpass filters (or channels) that covers a frequency

range of interest (e.g., $P = 20$ channels covering 78 Hz to 5000 Hz [17]). In a filter bank, each filter processes the signal independently to produce a short-time spectral representation $X_m(e^{j\omega})$ at time m through a filter i that has ω_i as its center of frequency. The center frequency and bandwidth of each filter are normally determined based on a scale model that mimics the way the human auditory system perceives sounds.

2.2 Time-Scale Representations

Another two dimensional signal processing tool that remedies problems arising from time frequency domain methods such as trade off in time frequency resolutions and limitations in analyzing non-stationary signals is the time-scale representation. The Wavelet Transform (WT) accomplishes such representation. It partitions the time-frequency plane in a non-uniform fashion and shows finer frequency resolution than time resolution at low frequencies and finer time resolution than frequency resolution at higher frequencies. This type of transform decomposes the signal into different frequency components, and then analyzes each component with a resolution that matches its scale [16]. The Continuous Wavelet Transform (CWT) [6] of a signal $x(t)$, is given by :

$$CWT_{(a,b)}(x(t)) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\psi\left(\frac{t-b}{a}\right) dt \qquad (2)$$

Where a and b are the real numbers that represent the scale and the translation parameter of the transform respectively. The function $\psi(t)$ is called the mother wavelet and has to have the following two properties:

(1) $\int_{-\infty}^{\infty} |\psi(t)|^2 dt < \infty$. This is equivalent to having $\psi(t) \in L^2(\Re)$ the space of finite energy functions.

(2) $\int_{-\infty}^{\infty} \psi(t)dt = 0$. This is equivalent to having the Fourier Transform of $\psi(t)$ null at zero (i.e., $\psi(t)$ has no dc components).

3 Connection to Hearing

When a sound wave hits the human eardrum, the oscillations are transmitted to the basilar membrane in the cochlea as standing waves that cause the basilar membrane to vibrate at the same frequencies as the input acoustic signal and at a place along the basilar membrane that is associated with these frequencies [34]. Unrolling and stretching the cochlea from its original spiral shape along with its basilar membrane. Each point of the basilar membrane can then be labeled by its position on a curve g(x) that represents this envelope of the cochlea [18].

Experiments and numerical simulations in [6] and [7] show that the response at the level of the basilar membrane of a real tone or an excitation of the form $e^{i\omega t}$ is a temporal oscillation that has the same frequency as the pure tone input. The response can be mathematically represented by:

$$R(x,t) = e^{iwt} F_w(g(x)) \qquad (3)$$

where F_w represents the dependency on ω of the response. Also in [6] and [7], F_w is described by a logarithmic shift for frequencies above 500Hz.
We can express the response as:

$$R(x,t) = F(g(x) - log\omega)e^{iwt} \quad \omega \geq 500Hz. \qquad (4)$$

The Fourier Transform representation [34] of a real-time signal $s(t)$ is given by the equation:

$$S(\omega) = \int_{-\infty}^{\infty} s(t)e^{-i\omega t}dt. \qquad (5)$$

The signal $s(t)$ can be synthesized from its Fourier components by the formula:

$$s(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(\omega)e^{i\omega t}dw. \qquad (6)$$

If one assumes linearity of the above cochlear model, then the response of the basilar membrane to $s(t)$ can be written as:

$$R(x,t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(\omega)e^{i\omega t}F(g(x) - log\omega)dw. \qquad (7)$$

Define the function P such that $F(z) = \hat{P}(e^{-z})$, where $\hat{P}$ is the Fourier Transform of P. Also, let $R^{'}(x,t) = R(-log(|g(x)|),t)$. The response can now be written as:

$$R^{'}(x,t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(w)\hat{P}(\omega g(x))e^{i\omega t}dw. \qquad (8)$$

This is the inverse Fourier Transform of a product of two functions which is the convolution of the inverse transform of the functions. The response can now be represented as:

$$R^{'}(x,t) = \frac{1}{g(x)} \int_{t_1=-\infty}^{\infty} s(t_1) P(\frac{t-t_1}{g(x)}) dt_1. \quad (9)$$

Now let $\psi(t) = P(-t)$ then,

$$R^{'}(x,t) = \frac{1}{g(x)} \int_{t_1=-\infty}^{\infty} s(t_1) \psi(\frac{t_1-t}{g(x)}) dt_1. \quad (10)$$

Up to a normalizing factor, the last equation can be interpreted now as the CWT of the input signal. It is identical to the Wavelet Transform equation. In this sense, the cochlea can be interpreted as a natural wavelet transformer of sound. This occurrence of the CWT in our biological processing of acoustics insinuated the construction of the Discrete Wavelet Transform Scale (DWTS) for modeling the speech waveforms. This discussion leads to the following statement:

Assume that the proposed cochlear model is linear for the envelope g(x), then its response to an acoustic signal, is up to a normalization factor, the CWT at scale g(x) of that signal.

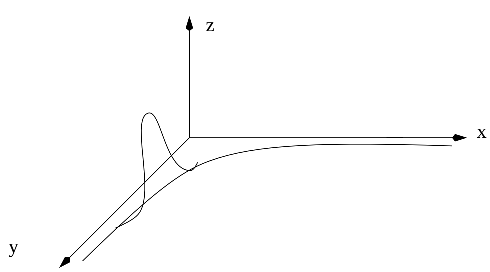

Figure 2: Localization of the response

The response of the cochlea to a pure tone input is the Fourier Transform of a wavelet, and hence should be localized in frequency. This is justified by the following:

The change of variable $r = \frac{t-a}{g(x)}$ for a pure tone input $s(t) = e^{iwt}$ allows us to reduce the response to:

$$R(x,t) = \int_{r=-\infty}^{\infty} e^{iw(rg(x)+a)} \psi(r) dr \quad (11)$$

This implies that the amplitude of the response is $\hat{\psi}(-wg(x))$ which is the Fourier Transform of the wavelet $\psi(\omega g(x))$. Thus the response is localized in frequency. Note that, the shape of the response in Figure 2 is justified to be the Fourier Transform of a wavelet, and hence should be localized in frequency. In the next section we derive the Discrete Wavelet Transform Scale (DWTS) used to construct feature vectors that model the speech signals.

4 The Discrete Wavelet Transform Scale

To follow the assumption of the logarithmic shift for frequencies above 500Hz, the level five decomposition of the Discrete Wavelet Transform was chosen as it is illustrated in Figure 3. The DWTS is constructed according to desired nodes in the decomposition tree. The frequency bands of the DWTS are the frequency contents of the four details coefficients $[CD_2, CD_3, CD_4, CD_5]$. These notations are very clear in [**?**] and [23] where a detailed treatment of the Discrete Wavelet Transform can be found. This scale leads to a more compact representation of a speech signal. It results in a 5:1 reduction of the size of the feature vectors when compared with those of the Mel scale that uses 20 frequency selected bands.

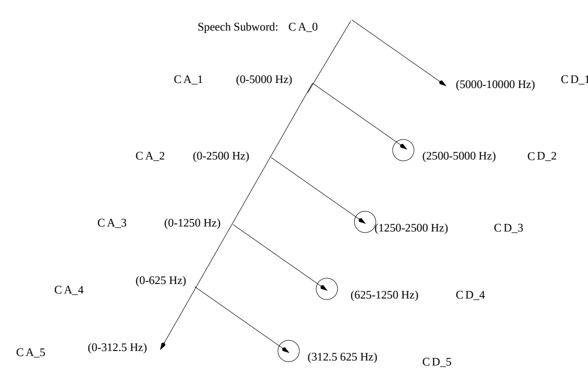

Figure 3: Selection of the DWTS frequency bands

One can interpret the Wavelet Transform integral operation in two ways [3]:

(1) It evaluates the inner product or the cross correlation of $x(t)$ with the $\psi(t/a)/\sqrt{a}$ at shift b/a. Thus it evaluates the components of $x(t)$ that are common to those of $\psi(t/a)/\sqrt{a}$. Thus it measures the similarities between $x(t)$ and $\psi(t/a)/\sqrt{a}$.

(2) It is the output of a bandpass filter of impulse response $\psi(-t/a)/\sqrt{a}$ at b/a of the input signal $x(t)$. This is a convolution of the signal $x(t)$, with an analysis window $\frac{1}{\sqrt{a}}\psi(t/a)$ that is shifted in time by b and dilated by a scale parameter a.

The second interpretation can be realized with a set of filters whose bandwidth is changing with frequency. The bandwidth of the filters is inversely proportional to the scale a which is inversely proportional to frequency. Thus, for low frequency we obtain high spectral resolution and low (poor) temporal resolution. Conversely, (This is where this type of representation is most useful) for high frequencies we obtain high temporal resolution that permits the wavelet transform to zoom in on singularities and detect abrupt changes in the signal [16]. This leads to a poor high frequency spectral resolution. The Discrete Wavelet Transform and the Fourier Transform are modified versions of the Continuous Wavelet Transform. They can be derived from the CWT for specified values of a and b. For example, if the mother wavelet $\psi(t)$ is the exponential function e^{-it} and $a = \frac{1}{w}$ and b=0 then, the CWT is reduced to the traditional Fourier Transform with the scale representing the inverse of the frequency [40]. The advantages that this new representation has over the STFT can be noticed in its efficiency in representing physical signals since it isolates transient information in a fewer number of coefficients and also in overcoming the time frequency trade off induced by STFT [16]. The properties of the CWT for real signals include: linearity, scale invariant, translation invariant, real and has an inverse. For a detailed discussion about the properties of the CWT and their proofs, refer to [6]. Some of the applications of the CWT in speech processing include: Analysis, synthesis and processing of speech and music sound in [22], Analysis of sound patterns in [24], Formant tracking in [15], Speech compression in [38] and Speech recognition in [9] [10][11][17] and [25] almost all of which base their work on one of the following databases [29] and [30].

5 Analysis and Synthesis of Hypermedia Elements

A two-channel perfect reconstruction filter bank [37] [39] contains four filters and two phases. The filters are:

1. The lowpass filter $H_0(n)$ and the highpass filter $F_0(n)$ in the decomposition phase.

2. The lowpass filter $H_1(n)$ and the highpass filter $F_1(n)$ in the reconstruction phase.

The Analysis (decomposition) and Synthesis (reconstruction) phases are depicted in Figure 4 and in Figure 5 respectively. The four filters H_0, F_0, H_1, F_1 that guaranty perfect reconstruction must satisfy the following 2 conditions expressed in the Z-domain [37] [39]:

$$H_1(z)H_0(z) + F_1(z)F_0(z) = 2z^{-l} \tag{12}$$

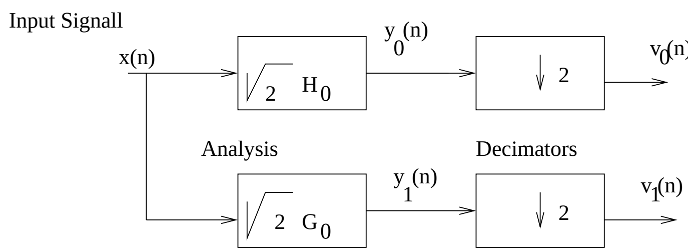

Figure 4: Analysis phase of the filter bank.

This condition removes distortion.

$$H_1(z)H_0(-z) + F_1(z)F_0(-z) = 0 \tag{13}$$

This condition cancels aliasing.

In Equation 12, l represents the over all delay. For alias cancellation we assume the following extra conditions also imposed in [37]:

$$H_1(-z) = F_0(z) \tag{14}$$

$$-H_0(-z) = F_1(z) \tag{15}$$

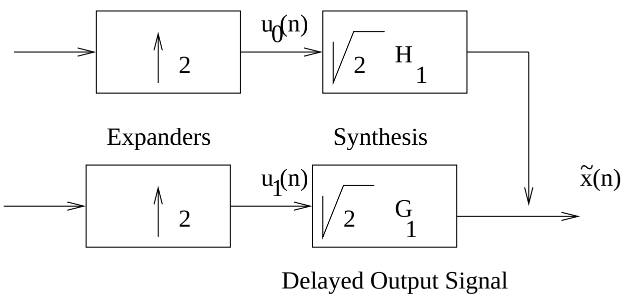

Figure 5: Synthesis phase of the filter bank.

6 Wavelets Compression

The goal of using wavelets to compress speech signal is to represent a signal using the smallest number of data bits commensurate with acceptable reconstruction and smaller delay. Wavelets concentrate speech information (energy and perception) into a few neighboring coefficients, this means a small number of coefficients (at a suitably chosen level) will remain and the other coefficients will be truncated. These coefficients will be used to reconstruct the original signal by putting zeros instead of the truncated ones.

6.1 Thresholding Techniques

Thresholding is a procedure which takes place after decomposing a signal at a certain decomposition level. After decomposing this signal a threshold is applied to coefficients for each level from 1 to N (last decomposition level). This algorithm is a lossy algorithm since the original signal cannot be reconstructed exactly [23]. By applying a hard threshold the coefficients below this threshold level are zeroed, and the output after a hard threshold is applied and defined by the following equation :-

$$y_{hard}(t) = \begin{cases} x(t), & |x(t)| > \delta \\ 0, & |x(t)| \leq \delta \end{cases} \tag{16}$$

where $x(t)$ is the input speech signal and δ is the threshold. An alternative is soft thresholding at level δ which is chosen for compression performance and defined by this equation :-

$$y_{soft}(t) = \begin{cases} sign(x(t))(|x(t)| - \delta), & |x(t)| > \delta \\ 0, & |x(t)| \leq \delta \end{cases} \tag{17}$$

where equation 16 represents the hard thresholding and equation 17 represents the soft thresholding.

7 Thresholding Methods Used in Wavelets Compression

In this section two thresholding algorithms will be introduced and later used in compressing speech signals. These two methods are, Global thresholding and Level dependent thresholding.

7.1 Global Thresholding

Global thresholding works by retaining the wavelet transform coefficients which have the largest absolute value. This algorithm starts by dividing the speech signal into frames of equal size F. The wavelet transform of a frame has a length T (larger than F). These coefficients are sorted in a ascending order and the largest L coefficients are retained. In any application these coefficients along with their positions in the wavelet transform vector must be stored or transmitted. That is, $2.5L$ coefficients are used instead of the original F samples, 8 bits for the amplitude and 12 bits for the position which gives 2.5 bytes. The compression ratio C is therefore:

$$C = \frac{F}{2.5L} \quad or \quad L = \frac{F}{2.5C} \tag{18}$$

Each frame is reconstructed by replacing the missing coefficients by zeros.

7.2 Level Dependent Thresholding

This compression technique is derived from the Birge-Massart strategy [19]. This strategy is working by the following wavelet coefficients selection rule :
Let J_0 be the decomposition level, m the length of the coarsest approximation coefficients over 2, and α be a real greater than 1 so :

1. At level J_0+1 (and coarser levels), everything is kept.
2. For level J from 1 to J_0, the K_J larger coefficients in absolute value are kept using this formula:

$$K_J = \frac{m}{(J_0 + 1 - J)^{\alpha}} \tag{19}$$

The suggested value for α is 1 and was used in [19] [20] [21].

7.3 Algorithms Interpretation

These algorithms are used to compress speech signals and compare the quality of the reconstructed signal with the original. In this section, outlines the steps followed in implementing these algorithms.

7.4 Compression using the Global Thresholding

The following procedure is usually followed to implement the global thresholding to compress speech signals.

1. Divide the speech signal into frames of equal size. Different frame sizes are tested to see how the frame size will affect the performance of the reconstructed signal. Three different frame sizes are examined since wavelet analysis is not affected by the stationarity problem. Expanding the frame length will speed up the processing time which reduces the processing delay.
2. Apply the discrete wavelet transform to each one of these frames separately at the five decomposition levels. This level is chosen since the best performance of the reconstructed signal is obtained at this level.
3. Sort the wavelet coefficients in a ascending order.
4. Apply the global thresholding to these coefficients by choosing the compression ratio and using equation 18 to obtain the non zero coefficients.
5. Keep the retained coefficients and their positions to reconstruct the signal from them.
6. Reconstruct the compressed frames by using the non zero coefficients and their positions and replacing the missing ones by zeros.
7. Repeat steps 2 to 6 to compress all the frames.
8. Insert these reconstructed frames into their original positions to get the reconstructed signal.

7.5 Compression Using Level-dependent Thresholding

After the speech signal is divided into equal frame sizes, the following steps are to be followed to implement the level dependent thresholding.

1. Apply the wavelet decomposition to each frame separately.
2. Keep all the coefficients of the last approximation, and use equation 19 to retain coefficients from each detail level.
3. Decompose all the frames and apply step 2 to each one of the frames, then keep the non zero coefficients and their positions using 2.5 bytes as in the global thresholding.
4. Reconstruct each decomposed frame using the non zero coefficients and replace the missing ones by zeros.
5. Insert these reconstructed frames into their original positions to get the reconstructed signal.

7.6 The Compression Parameters

In this paper, four compression parameters are used. They are defined next along with their mathematical expressions.

1. Signal to Noise Ratio: $SNR = 10 * log \frac{\sigma x^2}{\sigma e^2}$

   Where σx^2 is the mean square of the speech signal and σe^2 is the mean square difference between the original and reconstructed signals.

2. Peak Signal to Noise Ratio: $PSNR = 10 * log \frac{NX^2}{||x-r||^2}$

   Where N is the length of the reconstructed signal, X is the maximum absolute square value of the signal x and $||x-r||^2$ is the energy of the difference between original and reconstructed signals.

3. Normalized Root Mean Square Error: $NRMSE = \sqrt{\frac{(x(n)-r(n))^2}{(x(n)-?x(n))^2}}$

   Where $X(n)$ is the speech signal, $r(n)$ is the reconstructed signal, and $?x(n)$ in the mean of the speech signal.

4. Retained Signal Energy: $RSE = 100 \frac{||x(n)||^2}{||r(n)||^2}$

   Where $||x(n)||$ is the norm of the original signal and $||r(n)||$ is the norm of the reconstructed one. For *db* orthogonal wavelets the retained energy is equal to the $L^2 - norm$ recovery performance.

The speech signals compressed are the Arabic digits Zero and eight. Different Compression Ratios (CR) were obtained. Different parameters were examined when simulating the code. The 8Khz sampled signals are divided into frames (0.2ms, 0.25s, and 0.5s) and decomposed up to level 5. Each frame is decomposed separately. At this stage the threshold is applied on the coefficients to truncate whatever unnecessary. The obtained coefficients are then used to reconstruct the output compressed signal. Different results were obtained allowing efficient evaluations and comparisons of the used methods and parameters.

8 Comperative Results

In [18], it was shown that in the case of speech signals of the digits $0, 1, 2, ..., 9$ and the utterance "oh", Time-Scale wavelet representations had an order of magnitude advantage in time processing and recognition rate over the Time-Frequency Fourier based representation. Also in [17] it was shown that in processing speech for recognition purposes, subword speech units were sufficient for accurate recognition rates by using Time-Scale wavelet analysis instead of frame and windowed speech units in the Time-Frequency Fourier based representation. This was also the case for subset of the alphabets containing the letters a, j, k. On the other hand, in [9][10][11], Favero have proved by accurate statistical analysis of his experiments results that Time-scale wavelet based representation of the letters $b, c, d, e, g, p, t, v, z$ produced better recognition rates than using its rivalry representation. It is worthwhile mentioning here that in [18], Radial Basis Artificial Neural Networks were used as the recognition engines whereas in the work of Favero, Hiddien Markov Model engines were employed.

9 Conclusion

In this paper, the different representations of speech signals are presented thoroughly. Time domain, time frequency domain and time scale domain methods are described. Their advantages and drawbacks are discussed along with their different applications in speech processing in general and speech recognition and compression in particular. A discussion about segmentation and subwording of speech signals was included along with their role in wavelet-based speech recognition and compression. Finally, comparative results were extracted from the literature to show the advantages of using Wavelet-Based representations of speech signals over the traditional Time-Frequency based representation. For compression, the performance of the Discrete Wavelet Transform in compressing speech signals was tested and the following points were observed. High compression ratios were achieved with acceptable SNR. No further enhancements were achieved beyond level 5 decomposition. The effect of frame size and the Level Dependent Threshold on the NRMSE is evident while this measurement remains almost constant for all experiments with negligible changes. Increasing the frame size, positively affects the overall performance in both threshold techniques used. Overall, global threshold lead to better results than the level dependent threshold technique in the case of SNR and CR. This was the case with and without framing and for both tested digits. It is worthwhile noting that we could not pinpoint the best compression wavelet.

References

[1] K. Aczel, I. Vajk, Simple and Powerful Instrument Model for the Source Separation of Polyphonic Music, WSEAS TRANSACTIONS on SIGNAL PROCESSING, Volume 5, 2009, pp: 1 - 10.

[2] M. Artimy, W. J. Phillips, W. Robertson, Automatic Detection Of Acoustic Sub-word Boundaries For Single Digit Recognition Proceeding IEEE Canadian Conference on Electrical and Computer Engineering, 1999.

[3] Y. T. Chan, "Wavelet Basics", Kluwer Academic Publisher, Boston, 1995.

[4] R.R.Coifman and M. V. Wickerhauser, "Entropy-based Algorithm for best Basis Selection", IEEE Trans.on Info. Theory, vol. 38, no. 2, pp: 713-718, March 1992.

[5] J. W. Cooley and J. W. Tukey, An Algorithm For The Machine Computation Of Complex Fourier series, Mathematics of Computation, Vol. 19. pp: 297-301, 1965.

[6] I. Daubechies, "Ten Lectures on Wavelets", Society of Industrial and Applied Mathematics, Philadelphia, 1992.

[7] Daubechies, I. and Maes, S., A Nonlinear Squeezing of the Continuous Wavelet Transform Based on Auditory Nerve Models". Wavelets in Medicine and Biology, Aldroubi, A. and Usner, M., CRC press, pp: 527-546, 1996.

[8] A.Bruce and D.Donoho, "Wavelet Analysis, IEEE Spectrum", pp: 26-35 October 1996.

[9] R. F. Favero, Comparison of mother wavelets for speech recognition, Proc. Intl. Conf. Speech Science and Technology, pp:468-472, 1994.

[10] R. F. Favero, Comparison of perceptual scaling of wavelets for speech recognition, Proc. Intl. Conf. Speech Science and Technology, pp:336-341 1994.

[11] R. F. Favero, and R. W. King, Wavelet Parameterization for Speech Recognition:Variations in Translations and Scale parameters, Int. Symp. Speech Image Processing and Neural Networks Hong Kong, Vol 2 pp:694-697, April 1994.

[12] J. L. Flanagan, Speech Analysis Synthesis And Perception Signals, Springer-Verlag, New York, 1972.

[13] D. Gabor, Theory of Communication, Journal of the IEEE No. 93, pp: 429-456, 1946.

[14] O. Ghitza, Auditory Nerve Representation as a basis for Speech Processing, Advances in Speech Signal Processing, S. Furui and M. Sondhi, Eds., Marcel Dekker, NY, 453-483, 1991.

[15] H. Goldstein, Formant tracking using the wavelet-based DST, IEEE, COMSIG, pp:183-189, 1994.

[16] A. Graps, An Introduction To Wavelets, IEEE Computational Sciences and Engineering, Volume 2, Number 2, pp: 50-61, Summer 1995.

[17] J. Karam, w. J. Phillips, W. Robertson, New Low Rate Wavelet Models For The Recognition Of Single Spoken Digits, IEEE, proceedings of ccece, Halifax, pp:331-334, May, 2000.

[18] J. Karam, Simulation and Analysis of Wavelet-Based Scales for Speaker Independent Single Word Speech Recognition Systems, PhD Thesis, Technical University of Nova Scotia, Halifax,Canada, October 2000.

[19] J. Karam, A Global Threshold Wavelet-Based Scheme for Speech Recognition, Third International conference on Computer Science, Software Engineering Information Technology, E-Business and Applications, Cairo, Egypt, Dec. 27-29 2004.

[20] J. Karam, R. Saad, The Effect of Different Compression Schemes on Speech Signals, International Journal of Biomedical Sciences, Vol. 1 No. 4, pp: 230 234, 2006.

[21] J. Karam, A New Approach in Wavelet Based Speech Compression, Proceedings of the 10th WSEAS Int. Conf. on Mathematical Methods, Computational Techniques and Intellegent Systems. pp: 228 - 233, October, Corfu, 2009.

[22] R. Kronland-Martinet, The Wavelet Transform for analysis, synthesis and processing of speech and music sound, Computer Music Journal, Vol. 12, pp: 11-20, 1988.

[23] M. Misiti, Y. Misiti, G. Oppenheim , J. Poggi, Matlab wavelet toolbox, 1997.

[24] R. Kronland-Martinet, J. Morlet, A., Grossmann, Analysis of sound patterns through wavelet transform, Recognition and Artificial Intelligence, Vol.1, pp: 273-302, 1987.

[25] J. Leonard, J. Marti, C. Nadeu, E. L. Solano, Wavelet Transforms for Non-Uniform Speech Recognition Systems, ICSLP Proc. Forth Intl. Conf. on Spoken languages, Vol.4 pp:2348-2351, 1996.

[26] A. E. Mahdi, E. Jafer, New Adaptive Speech Enhancement System Using A Novel Wavelet Thresholding Technique, WSEAS TRANSACTIONS on COMPUTERS, Issue 4, Volume 3, October 2004, pp: 1092-1096.

[27] A. E. Mahdi, E. Jafer,Second Generation and Perceptual Wavelet Based Noise Estimation, WSEAS TRANSACTIONS on COMPUTERS, Issue 4, Volume 3, October 2004, pp: 1096 - 1101.

[28] H. Marvi, Speech Recognition Through Discriminative Feature Extraction WSEAS TRANSACTIONS on SIGNAL PROCESSING, Issue 10, Volume 2, October 2006, pp: 1364-1370.

[29] NIST, TIDIGITS, Speech Discs, Studio Quality Speaker-Independent Connected-Digital Corpus, PB91-506592, T.I., Feb. 1991.

[30] NIST, Speech Discs 7-1.1, TI 46 Word Speech Database Speaker-Dependent Isolated-Digital Corpus, LDC93S9, T.I., Sep. 1991.

[31] A. V. Oppenheim and R. W. Schafer, "Discrete-Time Signal Processing", Prentice Hall, Englewood Cliffs, New Jersey, 1989.

[32] W. J. Phillips, C. Tosuner and W. Robertson, Speech Recognition Techniques Using RBF Networks, IEEE, WESCANEX, Proceedings, 1995.

[33] J. W. Picone, Signal Modeling Techniques in Speech Recognition, IEEE, Vol.81, No.9, September 1993.

[34] L. Rabiner and Juang, B., "Fundamental of Speech Recognition", Prentice Hall, New Jersey, 1993.

[35] L. Rabiner and M. R. Sambur, An algorithm for determining the end points of isolated utterances, Bell Systems Technical Journal, Vol.54, pp: 297-315, Feb. 1975.

[36] L. Rabiner and R. W. Schafer, "Digital Processing of Speech Signals", Prentice Hall, New Jersey, 1978.

[37] G. Strang and T. Nguyen, ”Wavelets and Filter Banks”, Wellesley MA, Wellesley–Cambridge Press, Wellesley, MA, 1996.

[38] C. Taswell, Speech Compression with Cosine and Wavelet packet near-best bases, IEEE International Conference on Acoustic, Speech, and Signal Processing, p.p 566-568 Vol. 1, May 1996.

[39] M. Vetterli and C. Herley, ”Wavelets and Filter Banks: Relationships and New Results”, IEEE Transactions, Signal Processing pp: 1723-1726, 1990.

[40] R.K.Young, ”Wavelet Theory and its Applications”, Kluwer Academic Publishers, Lancaster, USA 1995.