World Academy of Science, Engineering and Technology
International Journal of Electrical and Computer Engineering
Vol:1, No:12, 2007

Various Speech Processing Techniques For Speech Compression And Recognition

Jalal Karam

***Abstract*—**Years of extensive research in the field of speech processing for compression and recognition in the last five decades, resulted in a severe competition among the various methods and paradigms introduced. In this paper we include the different representations of speech in the time-frequency and time-scale domains for the purpose of compression and recognition. The examination of these representations in a variety of related work is accomplished. In particular, we emphasize methods related to Fourier analysis paradigms and wavelet based ones along with the advantages and disadvantages of both approaches.

***Keywords*—**Time-Scale, Wavelets, Time-Frequency, Compression, Recognition.

I. Introduction

While the main focus of any speech recognition system (SRS) is to facilitate and improve the direct audible man-machine communication and provide an alternative access to machines, a speech compression system (SCS) focuses on reducing the amount of redundant data while preserving the integrity of signals.

The research in these areas dates back to the 1950's at Bell Laboratories. Since then, systems have been proposed, implemented and commercialized [12]. In the last few years, the architecture of (SRS) and (SCS) have been perturbed by the introduction of the new powerful analysis tool called wavelets. The theory of wavelets is a product of many independent developments in the fields of pure and applied mathematics, electrical engineering, quantum physics and seismic geology. The interchange between these areas in the last decade produced many new important and vital wavelet applications such as image and signal compression, turbulence, human vision, radar and earthquake prediction [7] to name a few. In the speech processing realm, this new tool is best introduced as an alternative to the classical Short Time Fourier Transform (STFT) for its effectiveness in the analysis of non-stationary signals. A survey of the wavelet literature reveals that the application of wavelets in the area of speech recognition has received a much late attention than areas such as image processing and data compression. Section 3 of this paper introduces the details of speech representations in the various domains for recognition purposes, while Section 4 and Section 5 discuss speech compression using wavelets and compare two common threshold approaches.

II. SPEECH REPRESENTATIONS

Extracting information from a speech signal to be used in a recognition engine or for compression purposes relies usually on transforming such a signal to a different domain than its original state. Although, processing a signal in the time domain can be beneficial to obtain measures such as zero crossing and others, most important properties of the signal resides in the time-frequency and time-scale domains. This section contains a review and a comparison of the different methods and techniques that allow such extractions. In this paper, $x(t)$ represents the continuous speech signal to be analyzed. 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 [14] [15]. 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 [21]. 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 [1] [17], 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 [3], results in time-dependent Fourier transform which reveals the frequency domain properties of the signal [16]. 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 [19]. 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 [16]. Mathematically, the spectrogram of a speech signal is the magnitude square of the Short Time Fourier Transform of that signal [2]. In the literature one can find many different windows that can be applied to the frames of speech signals for a short-term

Jalal Karam is the head of the department of mathematics and natural sciences at Gulf University for Science and Technology, Kuwait City, Kuwait, email: karam.j@gust.edu.kw

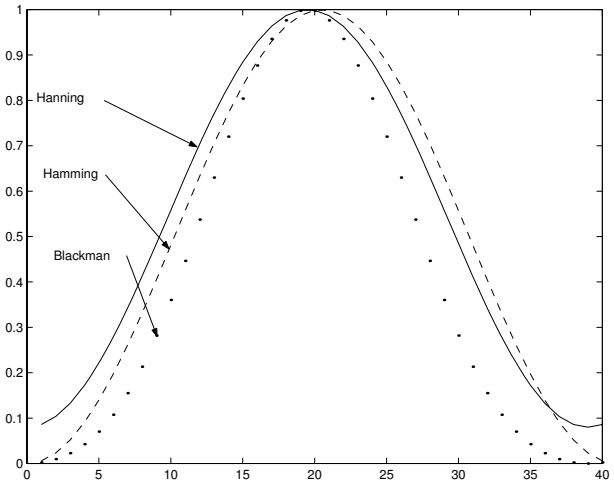

Fig. 1. Plots of window functions in the time domain.

frequency analysis. Three of them are depicted in Figure 1.

A. Time-Domain Representation

A natural and direct way to retrieve information from a speech signal is to analyze it in the time domain. The most common time domain measurements used to estimate features of speech [19] [21] are the short time: Energy, Average Magnitude Function, Average Zero-Crossing Rate, Autocorrelation. The amplitude of voiced segment is generally much higher than the amplitude of an unvoiced segments. The short-time energy of a speech signal provides a convenient representation that reflects these amplitude variations[21]. Zero crossing rate can be used as a simple measure of the frequency content of a speech signal. The work of Reddy, [22], used the zero crossing rate and the average energy to construct a large set speech recognition system. This lead Rabiner and Sambur to use the same parameters to mark the end points of a word while implementing an isolated word speech recognition system [20]. The pitch period which is a prime candidate for speaker identification is detected usually by one or more of these time domain parameters [21]. The most important advantage of time domain methods in speech analysis is their computational simplicity. Some of their disadvantages are quasi-stationary assumption on the speech signal and noise sensitivity, thus the need for complicated noise suppressing algorithms.

B. Time-Frequency Representations

To overcome the problems associated with the time domain methods, two dimensional signal processing tools such as time-frequency representations were introduced. This type of representation transforms a one dimensional signal into two dimensional space. 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 [24]. Gabor introduced the Short Time Fourier Transform (STFT) in 1946 to analyze finite duration signals [6]. The STFT of a signal $x(m)$ as defined in [19] 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 [16] [21] 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 [21]. 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 [2] 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 of Equation 4, 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 [23] [19]. 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 [8]). 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.

C. Time-Scale Representations

Another two dimensional signal processing tool that remedies problems arising from time frequency domain methods

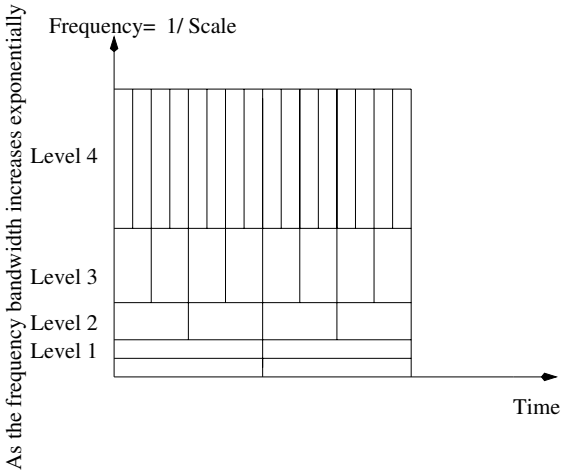

Fig. 2. The DWT coverage of the Time-Frequency plane.

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 [7]. The Continuous Wavelet Transform (CWT) [4] 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).

One can interpret the integral operation of Equation 5 in two ways [2]:

(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 [7]. This leads to a poor high frequency spectral resolution. The DWT coverage of the Time-Frequency plane is depicted in Figure 2.

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 [26]. 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 [7]. The properties of the CWT for real signals include: linearity, scale invariant, translation invariant, real and has an inverse.

III. 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 [5]. These coefficients will be used to reconstruct the original signal by putting zeros instead of the truncated ones.

A. 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 [13]. 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 this equation :-

$$y_{hard}(t) = \begin{cases} x(t), & |x(t)| > \delta \\ 0, & |x(t)| \leq \delta \end{cases} \qquad (3)$$

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 :-

$$y_{soft}(t) = \begin{cases} sign(x(t))(|x(t)| - \delta), & |x(t)| > \delta \\ 0, & |x(t)| \leq \delta \end{cases} \qquad (4)$$

where equation 3 represents the hard thresholding and equation 4 represents the soft thresholding.

IV. 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.

A. Global Thresholding

Global thresholding [10] 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 [5] . The compression ratio C is therefore:

$$C = \frac{F}{2.5L} \quad or \quad L = \frac{F}{2.5C} \tag{5}$$

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

B. Level Dependent thresholding

This compression technique is derived from the Birge-Massart strategy [11]. 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 :-

$$K_J = \frac{m}{(J_0 + 1 - J)^\alpha} \tag{6}$$

The suggested value for α is 1 and was used in [10] [11].

C. Interpretation of the two algorithms

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.

D. 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. In this thesis 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 5 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.

E. 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 6 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.

V. Conclusion

Speech processing for compression and recognition was addressed in this paper. A comprehensive examination of the different techniques used for these two purposes were examined. Various methods and paradigms based on the time-frequency and time-scale domains representation for the purpose of compression and recognition were discussed along with their advantages and draw-backs. Level dependent and global threshold compression schemes were also examined in details.

Acknowledgment

The author would like to thank Gulf University for Science and Technology for their financial support of this publication.

Dedication

For planting the seed of research curiosity and overall support and encouragement, this paper is dedicated to my friend Prof. Abdel Aziz Farrag.

REFERENCES

[1] Artimy, M., Phillips, W.J. and Robertson, W., Automatic Detection Of Acoustic Sub-word Boundaries For Single Digit Recognition Proceeding IEEE Canadian Conference on Electrical and Computer Engineering, 1999.
[2] Chan, Y.T., Wavelet Basics, Kluwer Academic Publisher, Boston, 1995.
[3] Cooley, J.W. and Tukey, J.W., An Algorithm For The Machine Computation Of Complex Fourier series, Mathematics of Computation, Vol. 19. pp: 297-301, 1965.
[4] Daubechies, I., The Wavelet Transform, Time Frequency Localization and Signal Analysis, IEEE Transaction on Information Theory, Vol. 36, No.5 pp: 961-1005, 1990.
[5] Feng, Yanhui,Thanagasundram, Schlindwein, S., Soares, F., Discrete wavelet-based thresholding study on acoustic emission signals to detect bearing defect on a rotating machine,Thirteenth International Congress on Sound and Vibration, Vienna, Austria July 2-6, 2006.
[6] Gabor, D., Theory of Communication, Journal of the IEEE No. 93, pp: 429-456, 1946.
[7] Graps, A., An Introduction To Wavelets, IEEE Computational Sciences and Engineering, Volume 2, Number 2, pp: 50-61, Summer 1995.
[8] Karam, J.R., Phillips, W.J. and Robertson, W., New Low Rate Wavelet Models For The Recognition Of Single Spoken Digits, IEEE, proceedings of ccece, Halifax, pp:331-334, May, 2000.
[9] Karam, J.R., Phillips, W.J. and Robertson, W., Optimal Feature Vector For Speech Recognition Of Unequally Segmented Spoken Digits, IEEE, proceedings of ccece, Halifax, pp:327-330 May, 2000.
[10] Karam, J., 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.
[11] Karam, J., Saad, R., The Effect of Different Compression Schemes on Speech Signals, International Journal of Biomedical Sciences, Vol. 1 No. 4, pp: 230 234, 2006.
[12] News and Analysis of Speech Recognition Markets, Products and Technology, Num. 73 pp: 1-32, July 1999.
[13] Misiti, M., Misiti, Y., Oppenheim, G., Poggi, J., Matlab Wavelet Toolbox, Math Works, Natick, MA, 1997.
[14] NIST, TIDIGITS, Speech Discs, Studio Quality Speaker-Independent Connected-Digital Corpus, NTIS PB91-506592, Texas Instruments, Feb. 1991.
[15] NIST, Speech Discs 7-1.1, TI 46 Word Speech Database Speaker-Dependent Isolated-Digital Corpus, LDC93S9, Texas Instruments, Sep. 1991.
[16] Oppenheim, A.V. and Schafer, R.W., Discrete-Time Signal Processing, Prentice Hall, Englewood Cliffs, New Jersey, 1989.
[17] Phillips, W.J., Tosuner, C. and Robertson, W., Speech Recognition Techniques Using RBF Networks, IEEE, WESCANEX, Proceedings, 1995.
[18] Rabiner, L., Digital Formant Synthesizer For Speech Synthesis Studies, J. Acoust. Soc. Am., Vol, 43, No. 2, pp: 822-828, April 1968.
[19] Rabiner, L. Juang, B., Fundamental of Speech Recognition, Prentice Hall, New Jersey, 1993.
[20] Rabiner, L. and Sambur, M.R., An algorithm for determining the end points of isolated utterances, Bell Systems Technical Journal, Vol.54, pp: 297-315, Feb. 1975.
[21] Rabiner, L.R. and Schafer, R.W., Digital Processing of Speech Signals, Prentice Hall, New Jersey, 1978.
[22] Reddy, D.R., Computer recognition of connected speech, Journal of the Acoustical Society of America, Vol. 42, pp:329-347, 1967.
[23] Picone, J.W., Signal Modeling Techniques in Speech Recognition, IEEE, Vol.81, No.9, September 1993.
[24] Strang, G. and Nguyen, T., Wavelets and Filter Banks, Wellesley MA, Wellesley–Cambridge Press, Wellesley, MA, 1996.
[25] Taswell, C., 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.
[26] Young, R.K., Wavelet Theory and its Applications, Kluwer Academic Publishers, Lancaster, USA 1995.

Jalal Karam has a Bsc, an Advanced Major Certificate, and an Msc in Mathematics from Dalhousie University in Halifax, Canada. In the year 2000, he finished a PhD degree in Applied Mathematics from the Faculty of Engineering at the Technical University of Nova Scotia. Currently, his is the Head of the Department of Mathematics and Natural Sciences at the Gulf University for Sciences and Technology in Kuwait. Dr. Karam is the Editor - in - Chief of the International Journal of Computational Sciences and Mathematics and a Reviewer for "Mathematical Reviews" of the American Mathematical Society. He also Chairs the task force of the World Academy of Science in the Middle East and the Near Far East Region.