Wavelets Based Modeling for Automatic Recognition of Spoken Words

Dr Jalal Karam
Department Mathematics and Computer Science
Arab Open University
Beirut Lebanon
drkaram_aou@hotmail.com

Abstract

This paper discusses the effect of relaxing the framing and windowing techniques from the analysis phase of speech processing for recognition. It is shown that the analysis of automatically generated subwords of speech using a Spectral Variation Function (SVF) and the modeling of these subwords following the Wavelet Packet Scale (WPS) makes the windowing and framing approach of Fourier based methods absolete. Radial Basis Functions Artificial Neural Network (RBF-ANN) is employed for the recognition tasks. The orthogonal db4 wavelet of the Daubechies family was used as the analyzing wavelet. The performance of the introduced system was compared with two different models and a considerable improvement in the recognition rates is attained using this approach.

1 Introduction

In this paper we employ the WPS [2] to parametrize the subwords of speech signals generated automatically by using the (SVD) of [1]. We will show that by taking the wavelet transform of the generated subwords and following the WPS of [2], a small number of simply measured parameters such as energy [12] is sufficient to secure an almost perfect recognition rate. This shows the effectiveness of wavelet analysis in overcoming the stationarity assumption used in systems that apply the traditional Fourier approach. This is accomplished by analyzing the subwords themselves instead of windowed frames. The performance of the introduced method maintains a very competitive performance and a commendable success in the recognition rates over the traditional Fourier and wavelet based algorithms that require the windowing techniques.

The next section is a brief introduction to Wavelet Transforms (WT) [5]. The Fourier Transform (FT) and the (DWT) are also derived as special cases of the CWT. Section four contains a description of the constructed subwords of the speech signals used for training and testing of the introduced method of analysis. It also compares this new approach with the traditional analysis of speech signals using the framing-windowing techniques. Section five contains the results of the experiments conducted followed by the conclusion.

2 Wavelet Transform

The Continuous Wavelet Transform (CWT) of a signal $s(t)$ with respect to a given mother wavelet $\psi(t)$ is given by:

$$CWT_{(a,b)}(s(t)) = \frac{1}{\sqrt{a}} \int s(t)\psi(t) \left(\frac{t-b}{a} \right) dt \qquad (1)$$

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 1 in two ways [8]:

(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 [5]. This leads to a poor high frequency spectral resolution.

The Discrete Wavelet Transform (DWT) and the Fourier Transform are modified versions of this general transform obtained for specified values of a and b. 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 [7].

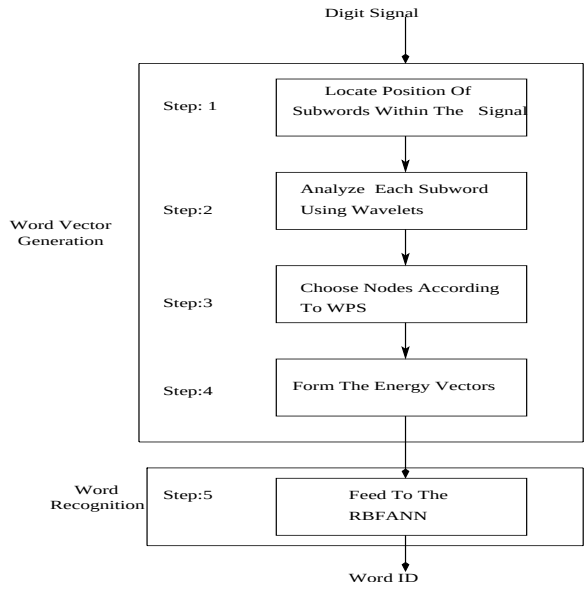

Figure 1. Flowchart of the introduced paradigm.

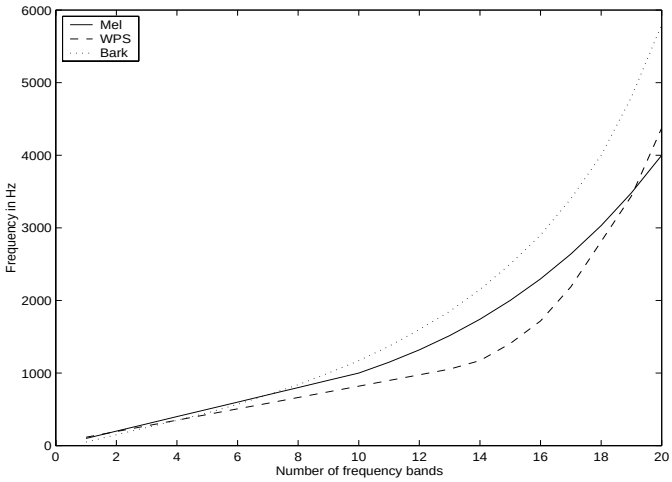

Figure 2. The twenty center frequencies of the Mel band and of the (WPS) bands of the proposed model.

In the case of the (DWT), the scale parameter "a" is sampled as : $a = a_0^m$ and the translation parameter "b" is sampled as : $b = na_0^m b_0$.
If $a_0 = 2$ and $b_0 = 1$ we obtain the Daubechies orthonormal basis of $L^2(R)$ [5], and the DWT of a digitized signal s(k) is then given by :

$$DWT_{(2^m,n2^m)}(s) = \frac{1}{\sqrt{2^m}} \sum_k s(k)\psi(2^{-m}k - n) \quad (2)$$

Where m and n are integers.

3 Parametrizing the Subwords

A subword of a speech signal is a section of that signal that represents a spectrally or a linguistically stable part [10]. The subwords generated in this work are obtained by identifying accoustically distinct segments of speech within one digit. An automatic algorithm that uses a spectral variation function (SVF) is employed to accomplish such subwording. It calculates the spectral changes between consecutive speech frames based on an Euclidean distance As described in [1].

Typically, 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

[10]. 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 [3], frames of size 200 samples and a sampling rate of 8000 Hz. This segmentation creates blocking effects that makes a rough transition in the representation 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. In this spectral analysis approach, a digitized speech signal each windowed frame is analyzed following 20 channels covering 78 Hz to 5000 Hz according to the WPS model [3]). The result is averaged according to the number of frames in each subword and energy vectors of size 20 are thus constructed to model each of the five subwords detected by the automatic algorithm of [1].

The method proposed in this paper chooses to analyze the subword itself without further segmenting them into frames.

To extract energy parameters of the WPS, we apply the wavelet analysis to each subword. The frequency bands are chosen according to the scale in use as they are set up next section. The last step is to compute the average absolute values of the wavelets coefficients over the corresponding bands of the WPS to obtain the energy values. These values are then scaled to a decibel scale of 0-60 dB for each band P of the WPS.

$$E_{max} = max(E(p)) \quad 0 \leq p \leq P-1 \tag{3}$$

$$ES(p) = 20 * log10(E(p)/E_{max}) \quad 0 \leq p \leq P-1 \tag{4}$$

$$ES'(p) = ES(p) - E_{max} \quad 0 \leq p \leq P-1 \tag{5}$$

$$ES''(p) = max(ES'(p), -60dB) + 60dB \quad 0 \leq p \leq P-1 \tag{6}$$

4 Experiments

27 experiments were conducted on a subset of the NIST database of the digits [13]. Signals are sampled at 20kHz. 17 speakers where chosen from two dialects, 5 subwords per signal were allowed to be consistent with the work in [1] [2] [3] [4].

For each experiment, we choose some of the speakers to form the training set while some or all the rest of the speakers form the test set. Once this selection is made, the WPS with windowing and then without windoing were applied to model the digits files experiments respectively. An overall comparison of the performance of these two techniques with that of the traditional Fourier one found in [4] are summarized in Table II.

Node Chosen From tree	Bandwidth in Hz	center in Hz
[7,1]	78	(78-156)
[7,3]	78	(156-234)
[7,2]	78	(234 - 312)
[7,7]	78	(312 -390)
[7,6]	78	(390 -468)
[7,4]	78	(468-564)
[7,5]	78	(564 - 625)
[7,15]	78	(625 - 703)
[7,14]	78	(703 -781)
[7,12]	78	(781 -859)
[7,13]	78	(859 -937)
[7,8]	78	(937- 1015)
[7,9]	78	(1015 -1093)
[6,5]	156	(1093 -1250)
[5,7]	312	(1250 -1562)
[5,6]	312	(1562-1875)
[4,2]	625	(1875 -2500)
[4,7]	625	(2500 -3125)
[4,6]	625	(3125 - 3750)
[3,2]	1250	(3750 -5000)

Table1:Selection of nodes according to Bandwidths and centers

The seventeen speakers are chosen such that 6 males and 6 females are from Rochester, and 2 males and 3 females are from Pittsburgh. This subset contains 2 tokens for each of the 11 digits making a total of 374 speech samples. This setup was chosen to compare the performances of the introduced model with the WPS model introduced in [2].

5 Conclusion

The experiments conducted show that by ignoring the windowing techniques a noticeable higher recognition rates were accomplished.
The goal of this paper is primarily to investigate the application of wavelet analysis in extracting features for speaker independent speech recognition systems. The research focus is on the simulation and analysis of new wavelet-based scale that can compete with or

Model	%RR
FT	91.25
WPS/db4 (With Windowing)	94.5
WPS/db4 (No Windoing)	97.8

Table II Overall preformance comparison of the three models.

replace the traditional Mel scale filter bank representation of speech. The Wavelet Packet decomposition is employed to formulate the Wavelet Packet Scale to model subwords of speech signals. This scale has a higher low frequency resolution and a poorer high frequency resolution than the Mel scale. A seven level decomposition is chosen to ensure the same size feature vectors as those of the Mel. Since by applying the Wavelet Transform one can relax the assumption of quasi-periodicity of speech, wavelet analysis was applied to different length subwords of speech signals. Thus, the popular framing and windowing operations usually required in Fourier based systems were omitted. This type of representation shows an advantage over the Fourier based methods since it reduces the amount of averaging in creating the feature vectors.

References

[1] Artimy M., Phillips W.J. and Robertsos W., "Automatic Detection of Acoustic Subword Boundaries for single digit Recognition", Proceedidngs IEEE CCECE'99, pp 751-754, May 1999.

[2] Karam J.R., Artimy M., Phillips W.J. and Robertsos W., "A New Wavelet Packet Model For Automatic Speech Recognition System", Proceedidngs IEEE CCECE'01, pp 483-486, May 2001.

[3] Karam J.R., Phillips W.J. and Robertsos W., "New Low Rates Wavelet Models For the Recognition of Single Spoken digits", Proceedidngs IEEE CCECE'00, pp 331-334, May 2000.

[4] Phillips,J.William, Caner Tosuner, William Robertson "Speech Recognition Techniques Using RBF Networks",IEEE,WESCANEX,Proceedings,1995.

[5] Daubechies,I., *Ten Lectures on Wavelets.* Philadelphia:SIAM,1992.

[6] Joseph W. Picone "Signal Modeling Techniques in Speech Recognition" IEEE,Vol.81.No.9,September 1993.

[7] Randy K. Young, *Wavelet theory and its applications*, Kluwer academic Publishers 1995.

[8] Chan, Y.T., *Wavelet Basics*, Kluwer Academic Publisher, Boston, 1995.

[9] Gilbert Strang, Truong Nguyen. Wavelets and Filter Banks, Wellesley Cambridge Press 1996.

[10] L.Rabiner, B. Juang , Fundamental of Speech Recognition, Prentice Hall 1993.

[11] M. Misiti, Y. Misiti, G. Oppenheim, J. Poggi. Matlab wavelet tool box, 1997.

[12] Lawrence R. Rabiner, Ronald W. Schafer , Digital Processing of Speech Signals, Prentice Hall 1978.

[13] NIST, Speech Discs, "Studio Quality Speaker-Independent Connected-Digital Corpus", NIST PB91-506592 Texas Instruments, Feb. 1991.