Proceedings of the 6th Nordic Signal Processing Symposium - NORSIG 2004
June 9 - 11, 2004
Espoo, Finland

DERIVING A WAVELET BASED SCALE FROM THE LOCALIZED RESPONSE OF THE HUMAN COCHLEA

Jalal R. Karam, PhD
Faculty of Information Technology and Computing, Arab Open University, Lebanon
P.O. Box 2058 4518 Tayyouneh, Beirut, Lebanon
email: jkaram@arabou-lb.edu.lb

ABSTRACT

In this paper we show that the response of the human ear to a pure tone is the Fourier Transform of a wavelet, thus proving that the response is localized. We also introduce a Discrete Wavelet Transform Scale (DWTS), subject to the response of the human cochlea in analyzing speech signals. The construction of this scale is accomplished based on a logarithmic shift; and its performance is tested on a subset of the English alphabet. We also show that the (DWTS) maintains a stable and competitive recognition rate over the traditional Mel scale used in Fourier based systems and over a Wavelet Packet Scale WPS.

1 INTRODUCTION

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 [9]. Unrolling and stretching the cochlea from its original spiral shape along with its basilar membrane results into a shape that is depicted in Figure 1. 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 [8]. Experiments and numerical

simulations in [2] and [3] show that the response at the level of the basilar membrane of a real tone or an excitation of the form e^{iwt} is a temporal oscillation that has the same frequency as the pure tone input. The response can be mathematically represented by:

$$R(x,t) = e^{iwt} F_w(g(x)) \qquad (1)$$

where F_w represents the dependency on ω of the response. Also in [2] and [3], F_w is described by a logarithmic shift for frequencies above 500Hz.

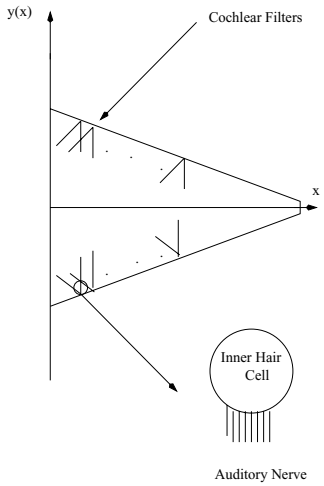

Figure 1: cochlea's model

2 LOCALIZING THE RESPONSE

We can express the response in Equation (1) as:

$$R(x,t) = F(g(x) - log\omega)e^{iwt} \quad \omega \geq 500Hz. \qquad (2)$$

In [8], it is shown that the response of the cochlea to an input speech signal $s(t)$ is:

$$R(x,t) = \frac{1}{g(x)} \int_{t=-\infty}^{\infty} s(t)\psi(\frac{t-a}{g(x)})dt. \qquad (3)$$

The change of variable $r = \frac{t-a}{g(x)}$ in Equation (3) for a pure tone input $s(t) = e^{iwt}$ allows us to reduce the response to:

$$R(x,t) = \int_{r=-\infty}^{\infty} e^{iw(rg(x)+a)}\psi(r)dr \qquad (4)$$

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 shape of the response in Figure 2 is justified to be the Fourier Transform of a wavelet, and hence should be localized in frequency.

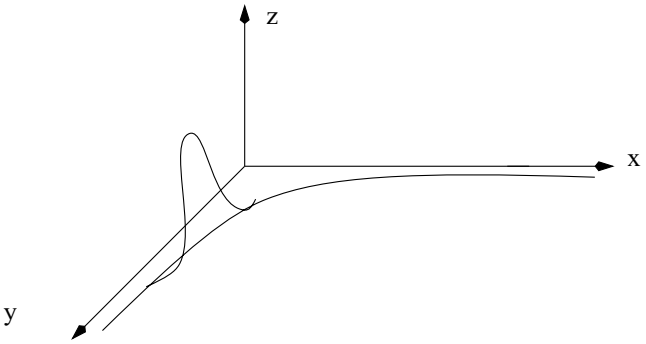

Figure 2: Localization of the response

In the next section we derive the Discrete Wavelet Transform Scale and in Section Four we construct the feature vectors that model the speech signals.

The performance of the introduced scale is compared in Section Five with the Fourier based model of the Mel scale as described in [5] and with that of the WPS as described in [6] and [7]. The last section contains the conclusion.

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 then constructed according to desired nodes in the tree decomposition as in Figure 3. It is plotted and compared with the Wavelet Packet Scale WPS and the Mel scale in Figure 4. 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 [4] and [10] 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 and the WPS that use 20 frequency selected bands each.

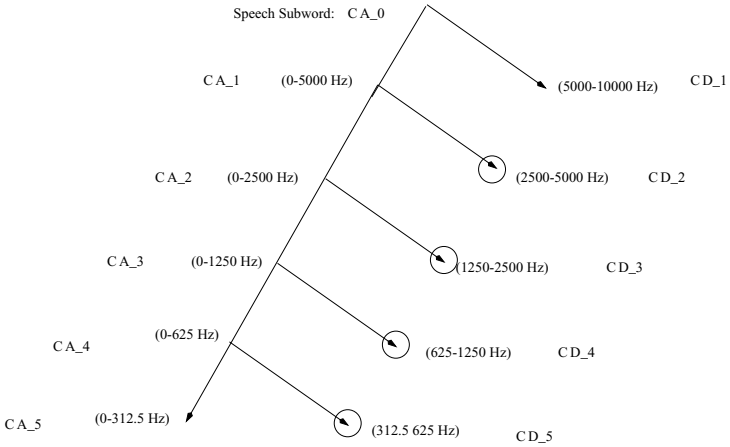

Figure 3: Selection of the DWTS frequency bands

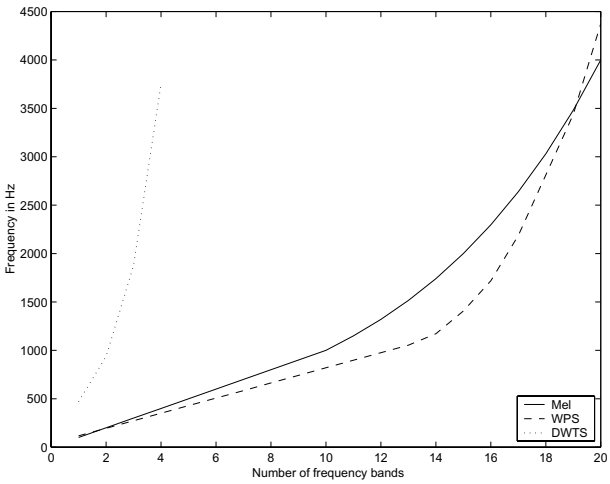

Figure 4: The DWTS, WPS and the Mel scale

3 SEGMENTATION

Traditionally, the processing of a speech signal starts with choosing the size of the frame and the kind and length of the analyzing window [5] [9]. When features are extracted and parameters are measured, an averaging technique is applied on adjacent frames that originally constitute a stable spectral part of the speech signal. These stable segments are called subwords. They are used as basic units in speech recognition systems because they represent a solution to the time-alignment problem in pattern recognition [1].

word	TM1	TM2	TM3
a	wb	we	
j	wb	ch	we
k	wb	ch	we

Table 1: Selection of subwords based on visible changes in the spectrograms

Word	Sub1	Sub2
a	***	
j	*	***
k	*	***

Table 2: * implies subword is not empty and * implies maximum energy**

The A-set of the English alphabet contains the letters a, j and k. The TI46 database [11] contains 8 male and 8 female speakers. Each speaker has 10 tokens per letter from the A-set for a total of 540 speech file. Figure 6 contains a sample of the Waveforms and thier corresponding spectrograms of the letters a, j, and k spoken by the female f5 of the database. The segmentations of the signals into subwords were manifested by inspecting the time domain waveform with their corresponding spectrograms. The list of abbreviation used in Table 1 is:
b : begin
ch : changes
e : end
w : word
(i.e. wb in column 1 represent word begin).
It is clear that the three time marks (TM1, TM2 and TM3)of Table 1 correspond to two subwords per speech signal.

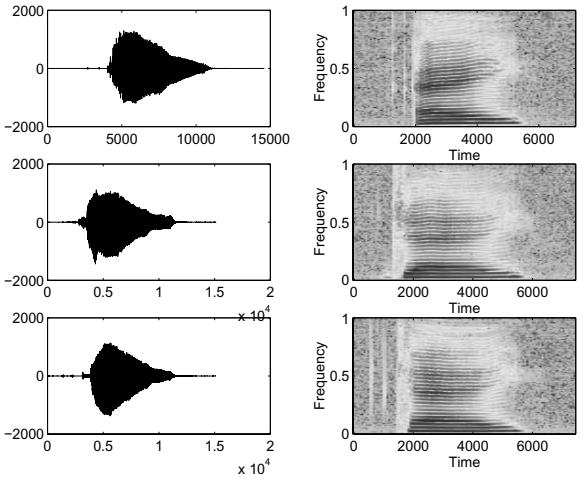

Figure 6: From top to bottom: Waveforms and corresponding spectrograms of the letters a, j, and k spoken by female f5.

4 CONSTRUCTING THE FEATURE VECTORS

To extract energy parameters of the DWTS, we apply the wavelet analysis to each subword. In this phase of the analysis, the Daubechies orthogonal mother wavelet "db6" were used. To know more about this mother wavelet, refer to [10] The frequency bands are chosen according to Figure 3. The next step is to compute the average absolute values of the wavelets coefficients over the corresponding bands of the scale to obtain the energy values. These values are then scaled to a decibel scale of 0-60 dB.

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

$$ES(p) = 20 * log10(E(p)/E_{max}) \quad 0 \leq p \leq P-1 \tag{6}$$

$$ES'(p) = ES(p) - E_{max} \quad 0 \leq p \leq P-1 \tag{7}$$

$$ES''(p) = max(ES'(p), -60dB)+60dB \quad 0 \leq p \leq P-1 \tag{8}$$

There are 4 bands in the DWTS corresponding to $P = 4$.
Once the feature vectors were constructed, a Radial Basis Functions Neural Network was employed for recognition. Table 3 displays the recognition rates of the experiments according to the scale and an overall statistical comparison with the WPS and the Mel scale.

In Table 2, the "*" acknowledges the existence of the subword and the "***" represents the subwords with high energy.

5 CONCLUSION

In this paper we introduced the DWTS via the response of the human cochlea to speech sound. It was shown that this response is localized. The performance of the DWTS was tested and proved its competitiveness when compared with that of the WPS and the Fourier based Mel scale.

Scale	Max.	Min.	Ave.	σ
DWTS(db6)	88	78	83.0667	3.3905
WPS(db6)	88	74	81.8667	4.2572
Mel	88	71	79.2667	4.9924

Table 3: Maximum, Minimum, Average and Standard Deviation of the Experiments conducted on the A-set

References

[1] Algazi, V.R., Brown, K.L., Ready, M.J., Irvine, D.H., Cadwell, C.L. Chung,S., Transform Representation of the Spectra of Acoustic Speech Segments with Application-I: General Approach and Application to Speech Recognition, IEEE Transactions on Speech and Audio Processing 1 (2) pp: 180-195, April 1993.

[2] Daubechies I., "Ten Lectures on Wavelets", Philadelphia:SIAM,1992.

[3] 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.

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

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

[6] Karam,J.R., "Wavelet Based Modeling for Automatic Recognition of Spoken Words", 11th IEEE Mediterranean Conference on Control and Automation "MED 03" Rhodes, Greece, pp: 89-93, June 17-20, 2003.

[7] Karam,J.R.,"Automatic Recognition of Spoken Words via Wavelets Coding", IEEE International Conference on Imaging Science, Systems and Technology, Las Vegas, pp: 49-52, June 23-26, 2003.

[8] Karam, J. R., "Simulation and Analysis of Wavelets Based Scales For Speech Recognition Of Isolated Spoken Words" PhD thesis, Technical University of Nova Scotia, Halifax, 2000.

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

[10] Misiti M., Misiti Y., Oppenheim G., PoggiJ. Matlab wavelet tool box, 1997.

[11] TI46, Speech Discs, "Studio Quality Speaker-Independent Connected-Digital Corpus", NIST PB91-506592 Texas Instruments, Feb. 1991.