NEW WAVELET PACKET MODEL FOR AUTOMATIC SPEECH RECOGNITION SYSTEM

Jalal R. Karam, William J. Phillips, William Robertson, Maen M. Artimy

Department of Engineering Mathematics, Dalhousie University

ABSTRACT

This paper introduces an automatic speaker-independent speech recognition system. We investigate the performance of the Wavelet Packet in the analysis of automatically generated subwords of single digits. The modeling of the subwords is accomplished using multi-energy levels of a derived Mel-like scale. A Radial Basis Function Artificial Neural Network (RBF-ANN) is employed for the recognition task. The proposed model is compared with two systems, one uses manual segmentation, the other segments words based on energy levels extracted from a filter bank. A comparison is made between the performance of systems using two orthogonal wavelets from the Daubechies set and two biorthogonal wavelets.

1. INTRODUCTION

The advantages of the Mel-frequency scale modeling in speech recognition systems are well known and well documented in the literature [7, 4].

In this paper we derive a Mel-like Band scale using Wavelet Packets and investigate its success in modeling spoken digits using each of the following mother wavelets "db2", "db4", "bior2.2", and "bior6.8" respectively.

The digits are automatically decomposed into subwords utilizing a spectral variation function. Vectors of energy parameters extracted from Mel like bands in each subword are then fed to the neural network for the recognition. Figure 1 shows the block diagram of the proposed system.

Section 2 provides a definition of subwords. Sections 3 and 4 provide an introduction to the Continuous and Discrete Wavelet Transforms [2, 6, 8] followed by the Wavelet Packet Decomposition and its implementation of the Mel-like Band [10]. The implementation of the introduced system is described in section 5 while its performance is evaluated in section 6 where we compare it with the Fourier based models where either an automatic segmentation of subwords is used [1] or a manual segmentation [10]. The last section contains the conclusion.

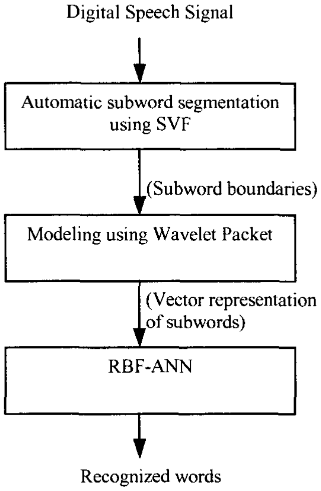
Digital Speech Signal
Automatic subword segmentation using SVF
(Subword boundaries)
Modeling using Wavelet Packet
(Vector representation of subwords)
RBF-ANN
Recognized words

Figure 1. Block diagram of the proposed system

2. SUBWORDS

Subwords are smaller sections of words that represent a spectrally or linguistically distinct part of speech [7]. For this reason, the subwords are preferable over words as basic units in speech recognition. The subwords used here are created by identifying acoustically distinct segments of speech within one word (digit). To identify these segments, we relied on an automatic algorithm that uses a spectral variation function (SVF). This function calculates the spectral changes between consecutive speech frames based on an Euclidean distance as described in [1].

3. WAVELET TRANSFORMS

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

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

Where a and b are the real numbers that represent the scale and the translation parameter of the transform respectively.

A complex wavelet function $\psi(t)$ has to have:

1. Finite energy, i.e.,

$$\int_{-\infty}^{\infty}|\psi(t)|^2 dt < \infty$$

2. Zero mean, i.e.,

$$\int_{-\infty}^{\infty}\psi(t)dt = 0$$

3. Only positive frequency components, i.e.,

$$\int_{-\infty}^{\infty}\psi(t)e^{j\omega t}dt = 0 \qquad \text{for all } \omega < 0$$

and so the real and imaginary parts are the Hilbert transform of one another for all $\omega < 0$.

The third condition is relaxed in this work since we are considering orthogonal Daubechies wavelets and spline biorthogonal wavelets, which are all real.

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 $exp(it)$, $a=1/\omega$, and $b=0$, then the CWT is reduced to the traditional Fourier Transform with the scale representing the inverse of the frequency [5].

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 $L^2(R)$ [3], and the DWT of a digitized signal $s(k)$ is then given by:

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

Where m and n are integers.

The DWT, which is obtained by this particular sampling of the CWT, possesses the following properties:

- Linear,
- Real, if both the wavelet and the signal are real, and
- Conserves energy.

There are many advantages for choosing the DWT over the CWT:

- The DWT reduces the computational complexity of the CWT.
- It reduces the redundancy of the CWT.
- Most mathematical expositions of wavelet theory develop octave decomposition.

One should note here that the dyadic and binary sampling of the parameters a and b lead to a complete representation of a signal. The continuous scaling and shift parameters result in a redundant representation of a signal.

Sampling the scale parameter of the CWT with a factor that is finer than an octave resolution has been studied and applied in speech recognition [16].

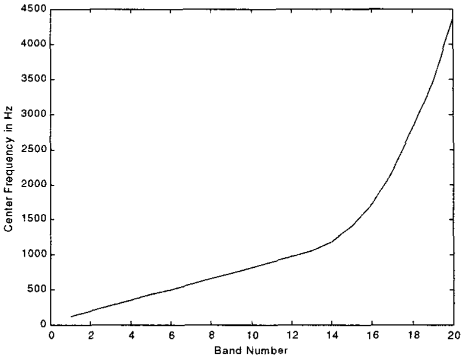

Figure 2. Mel-like Scale

4. WAVELET PACKET DECOMPOSITION

The DWT follows a one-sided dyadic tree decomposition of signals; the generic step splits the approximation coefficients into two parts. This produces two new vectors of approximation and detail coefficients at a coarser scale. The vector containing the details is kept intact while the other vector is decomposed again [8]. The wavelet packet decomposition allows any dyadic tree structure analysis [14, 8] where not only the approximation coefficients are decomposed iteratively but also the detail coefficients.

This means that the high frequency half is also recursively decomposed and a better representation of the signal is obtained. Up to the level chosen, the frequency bands are measured at every node of the tree. Level-seven decomposition is chosen here. The frequency bandwidth at this level is 78 Hz, which allows a better resolution (13 coefficients) for frequencies below 1100Hz. Seven other coefficients are chosen for the representation of the rest of the band. The flow of the frequency decomposition is not in order, i.e., consecutive nodes do not contain parameters that represent ordered frequencies.

The nodes are selected from the levels 7, 6, 5, 4, and 3 of the wavelet packet tree [10]. They are chosen to cover linearly the low frequencies and logarithmically high frequency components of the signals to secure a Mel like scale without overlapping. Figure 2 shows the center frequencies of this Mel-like perceptual scale.

5. IMPLEMENTATION

The system described in Figure 1 is implemented. In this system, a digit is segmented automatically using the SVF algorithm described in [1]. Some digits are also segmented manually into a maximum of five subwords as described in [11, 15] to provide a basis for comparison. Each subword passes through a seven-level wavelet decomposition to simulate a filter bank of 20 bands where energy of each band is extracted. This process results in a five-vector representation for the digit. These 20-element vectors are then normalized into a scale of 0-60dB. A Radial Basis Function is used for the recognition process as described in [11, 15]. The type of neural networks was considered to build a system comparable with [1, 11, 15].

The wavelet decomposition is performed using two orthogonal wavelets of the Daubechies set: "db2" and "db4". Also, two biorthogonal wavelets, "bior2.2", and "bior6.8" are used.

Train/ Test	*Fourier*	*db2*	*db4*	*bior2.2*	*bior6.8*
5/15	83	75	80	75	80
5/21	84	76	80	76	80
5/15	82	74	80	77	80
5/21	82	65	79	78	80
5/15	83	64	79	78	80
5/21	82	80	79	78	80
7/15	85	80	79	82	80
7/21	85	75	85	81	80
7/15	84	78	84	77	82
7/21	86	78	84	79	76
7/15	83	78	80	78	80
7/21	84	74	82	80	83
7/15	83	77	82	78	84
7/21	84	82	82	80	84
8/15	83	82	82	80	80
8/21	85	78	83	81	81
8/15	86	82	81	76	84
8/21	87	82	82	79	84
9/15	85	78	81	82	79
9/21	86	80	83	81	82
9/15	89	82	84	80	86
9/21	87	80	83	81	84
10/15	90	78	84	82	83
10/21	89	79	85	84	84
11/21	87	80	84	85	86
11/15	88	82	85	82	88

Table 1. Recognition rates of Fourier and Wavelet based systems with automatic segmentation.

6. EXPERIMENTS

The digit recognition rate is used to evaluate the system. We conducted sets of experiments, each aimed at evaluating the performance of a certain wavelet. The performance is also compared to previous systems that use Fourier analysis to extract energy parameters form Mel-scale filter bank [1, 11].

The first set of experiments includes 26 experiments on a subset of the NIST database [12]. This subset contains male and female speakers of four American English dialects: Philadelphia, Boston, Rochester, and Pittsburgh.

Table 1 shows the results of our system with the recognition rate of the corresponding wavelets that are used. The first column of the table shows number of speakers in the training set and the test set respectively. Table 1 also shows a comparison with the Fourier model of [1] where subwords are segmented automatically. All the 26 experiments have identical sets of speakers in all the systems.

Train/ Test	*db2 manual*	*db2 auto.*	*Fourier manual*	*Fourier auto.*
8/4	100	77	98	87
8/5	99	78	98	87
8/4	99	80	98	87
8/5	99	85	98	87
8/4	99	80	98	87
8/5	100	84	98	87
5/7	99	75	97	82
5/5	99	75	98	82
5/7	98	71	98	82
5/5	98	78	98	82
5/7	99	70	99	82
5/5	98	79	98	82
4/13	99	72	97	80
4/13	99	69	99	79
4/13	98	70	99	80
3/14	97	64	96	74
3/14	98	66	97	75
3/14	98	65	96	58
2/15	98	52	93	58
2/15	98	52	83	59
2/15	98	51	74	55
8/9	99	80	98	83
8/9	99	78	96	85
8/9	98	74	98	91
5/12	98	71	98	80
5/12	99	70	99	79
5/12	98	75	99	82

Table 2. Recognition rates of Fourier and Wavelet systems under manual and automatic subword segmentation.

A second set of 27 experiments is also conducted on a different subset of the NIST database. This subset contains male and female speakers of two dialects: Rochester and Pittsburgh. In these experiments, subwords are segmented automatically and manually in order to compare the results with other models described in [1, 10, 11].

Table 2 shows the results of these experiments when using the mother wavelet "db2". The first column of the table shows the number of speakers in the training set and the test set respectively which indicates that the 27 experiments have identical sets of speakers in all systems.

7. CONCLUSION

In this paper we introduced a new Wavelet Packet based automatic speech recognition system. It employs four different mother wavelets in analyzing the speech signals. It models each of the five subwords of a digit using a Mel-like band scale. By comparing our results with those in [1], we see that the maximum recognition rate of the Wavelet Packet model fell 2% short. The experiments conducted also show that although some types of wavelets performed better than others, the manual segmentation produced noticeable higher recognition rates than automatic segmentation where the former achieved an average recognition rate of 98.6%.

We recommend that future work should include an automatic segmentation algorithm that is based entirely on wavelets (uniform and non-uniform length subwords), a larger database, and a wide mixture of different dialects.

8. REFERENCES

[1] Artimy M., Phillips W.J., and Robertson W., "Automatic Detection of Acoustic Sub-word Boundaries for Single Digit Recognition", *Proceedings IEEE CCECE'99*, pp. 751-754, May 1999.

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

[3] Daubechies, I., "Wavelets: A Tool for Time-Frequency Analysis", *Sixth MDSP Workshop session TP1*, 1998.

[4] Joseph W. Picone, "Signal Modeling Techniques in Speech Recognition", *Proceedings of the IEEE*, 81(9):1215-1247, September 1993.

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

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

[7] Rabiner L., Juang B., *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[8] Misiti M., Misiti Y., Oppenheim G., Poggi J. *Matlab wavelet toolbox*, The MathWorks Inc., 1997.

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

[10] Karam J.R., Phillips W.J. and Robertson W., "New Low Rate Wavelet Models for the Recognition of Single Spoken Digits", CCECE, 331-334, 2000.

[11] Phillips W.J., Tosuner C., and Robertson W., "Speech Recognition Techniques Using RBF Networks", *Proceedings of IEEE WESCANEX*, pp. 185-190, 1995.

[12] NIST Speech Discs, "Studio Quality Speaker-Independent Connected-Digital Corpus"', *NIST PB91-506592 Texas Instruments*, February 1991.

[13] Oliver Rioul and Vetterli Martin, "Wavelets and Signal Processing", *IEEE Signal Processing Magazine*, October 1991.

[14] Coifman R.R. and Wickerhauser M.V., "Entropy-based Algorithm for best Basis Selection", *IEEE Transactions on Information Theory*, Vol.38, No.2, pp. 713-718, March 1992.

[15] C. Tosuner, *Single Spoken Digit Recognition Using RBF Networks*, Master Thesis, Department of Electrical Engineering, Dalhousie University, Halifax Nova Scotia (TUNS), 1995.

[16] Richard F. Favero and Robin W. King, "Wavelet Parameterization of Speech Recognition: Variations in Translations and Scale parameters", Int. Symp. Speech and Image Processing and Neural Networks, Hong Kong, Vol. 2, 694-697, April 1994.