Proceedings of the Institute of Acoustics

MODIFIED DISCRETE WAVELET FEATURES FOR PHONEME RECOGNITION

O. Farooq Department of Electronic and Electrical Engineering, Loughborough University, Loughborough, Leicestershire, LE11 3TU.
S. Datta Department of Electronic and Electrical Engineering, Loughborough University, Loughborough, Leicestershire, LE11 3TU.

1. INTRODUCTION

Last decade has seen considerable research in the area of speech recognition, however the basic feature extraction stage has not changed much. The feature extraction stage tries to simulate the characteristics of the human ear and detects features that can be used effectively for classification of speech. For this purpose Bark scale and Mel scale [1], [2] have been proposed. The Mel scale is often approximated as linear scale between 0-1000Hz and then as logarithmic scale beyond 1000Hz. The Cepstral Coefficients derived after filtering the speech signal by the Mel filter bank gives the most commonly used Mel Frequency Cepstral Coefficient (MFCC) features. The Short Time Fourier Transform (STFT) is usually used for the extraction of MFCC features. Since STFT has a fixed time-frequency resolution, it is difficult to detect sudden short burst of high frequency in a low frequency background. This problem is predominantly encountered in the case of stop (plosive) phonemes, hence the MFCC features give poor recognition performance for these phonemes.

Recently, the Discrete Wavelet Transform (DWT) has been used for feature extraction [3], [4], [5], [6], [7]. This is because DWT can be effectively used to separate out short impulses from a low frequency background easily by using its multi-resolution capability. This property of DWT has been exploited in phoneme recognition by using the high-energy wavelet coefficients as features [3], [4], [5]. However, the DWT suffers from two problems. First is the problem of shift variance i.e. if the signal is slightly shifted the wavelet coefficients will change thus, direct use of wavelet coefficients as feature is effective. The second problem inherent to DWT is that it decomposes the lower frequency sub-band obtained from the previous decomposition.

The Wavelet Packet (WP) which overcomes the second problem has also been proposed for the selection of features by using the best basis algorithm [6], [7], [8]. However, this technique also suffers form the problem of shift in the signal as it results into different set of basis for a shifted version of a signal. In this paper we propose log energy features based on DWT, which are shift resistant and give better recognition performance as compared to earlier features. Also we put forward the use of Admissible Wavelet Packets (AWP) which gives more flexibility in partitioning the frequency sub-bands for the extraction of these new features.

The paper is organised as follows. In Section 2, we give a brief introduction to the DWT, WP and AWP. Section 3 gives the details of the feature extraction process by using the above wavelet decomposition techniques. Section 4 elaborated the experimentation procedure and gives the results obtained for the phoneme recognition using the TIMIT database. The concluding remarks

on the experimental results are given in Section 5

2. WAVELET TRANSFORM

2.1 Discrete Wavelet Transform

Wavelet transform is a time-frequency analysis technique, which decomposes signal over dilated and translated wavelets. Wavelet is a function $\psi \in L^2(\Re)$ (i.e. a finite energy function) with zero mean and is normalised ($\|\psi\| = 1$) [9]. A family of wavelets can be obtained by scaling ψ by s and translating it by u.

$$\psi_{u,s}(t) = s^{-1/2}\psi\left(\frac{t-u}{s}\right) \qquad (1)$$

The Continuous Wavelet Transform (CWT) of a finite energy signal $f(t)$ is given by:

$$CWTf(u,s) = \int_{-\infty}^{+\infty} f(t).s^{-1/2}.\psi^*\left(\frac{t-u}{s}\right)dt \qquad (2)$$

where $\psi^*(.)$ is the complex conjugate of $\psi(.)$. The above equation can be viewed as convolution of the signal with dilated band-pass filters. The DWT of a signal $f[n]$ with period N is computed as:

$$DWTf[n,a^j] = \sum_{m=0}^{N-1} f[m].a^{-j/2}.\psi^*\left(\frac{m-n}{a^j}\right) \qquad (3)$$

where m and n are integers. The value of a is equal to 2 for a dyadic transform.

The signal representation is not complete if the wavelet decomposition is computed up to a scale a^j. The information corresponding to the scales larger then a^j is also required, which is computed by a scaling filter and is given by:

$$SFf[n,a^j] = \sum_{m=0}^{N-1} f[m].a^{-j/2}.\phi^*\left(\frac{m-n}{a^j}\right) \qquad (4)$$

where $\phi(n)$ is the discrete scaling filter.

By using the DWT the problem in recognition of stop phonemes is expected to be overcome as higher frequency burst can be easily detected by going up high in frequency and reducing the time window. Thus high frequency burst within the phonemes which were undetectable under STFT can be detected by using the DWT.

The DWT performs the recursive decomposition of the lower frequency sub-band obtained by the previous decomposition in dyadic fashion. Hence the DWT gives a left recursive binary tree structure where the left child represents the lower frequency sub-band and the right child represents higher frequency sub-band. In WP decomposition, the lower as well as higher

frequency sub-bands are decomposed into two sub-bands thereby giving a balanced binary tree structure as shown in Figure 1. Each node W_j^p, in the tree represents the depth j and the number of node p to the left of it.

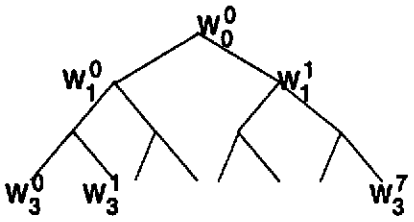


Figure 1: Balanced binary tree achieved by the full Wavelet Packet decomposition.

The two wavelet packet orthogonal bases generated from a parent node (W_j^p) are defined as:

$$\psi_{j+1}^{2p}(k) = \sum_{n=-\infty}^{\infty} h[n]\psi_j^p(k - 2^j n) \qquad (5)$$

$$\psi_{j+1}^{2p+1}(k) = \sum_{n=-\infty}^{\infty} g[n]\psi_j^p(k - 2^j n) \qquad (6)$$

where **h[n]** is the low pass (scaling) filter and **g[n]** is the high pass (wavelet) filter.

The WP decomposition results in over-complete basis. For a full j level decomposition there will be over $2^{2^{j-1}}$ orthogonal bases. From the above library of bases (also called as packet table) best basis is to be selected. Selection of the best basis tries to have best frequency partitioning by reducing a cost function [5]. However, application of the best basis algorithm to the pattern recognition problem is difficult, as they are not translation invariant. For a shift in the signal, the wavelet packet decomposition will give modified coefficients, thereby yielding different basis when the cost function is minimised. Since energy based features are used, therefore this may result into different number of features, which may further create problems at the classification stage.

2.2 Admissible Wavelet Packets

For speech recognition if full WP decomposition is applied, it will cause the features to be distributed uniformly over the entire frequency band. The speech recognition research shows that features from higher end of the frequency spectrum have very little discriminatory information. Due to this reasons full wavelet packet decomposition cannot be used effectively for the extraction of features from phonemes.

In order to overcome the above problems we use a modified wavelet packet decomposition, which is in-between DWT and WP and gives the liberty to partition the lower frequency sub-band or the higher frequency sub-band. This is known as Admissible Wavelet Packet (AWP), which gives an admissible binary tree structure. Figure 2 shows an example of tiling of the time-frequency plane by one of the admissible wavelet binary tree structure for a four level of decomposition. The corresponding admissible binary tree structure is shown in Figure 3 giving

details of splitting of the frequency bands. By using the AWP we can have more number of bands in the frequency region carrying more discriminatory information. Thus the features derived form these frequency sub-bands will have better classification ability. Also, since the partitioning of the frequency axis is fixed, therefore problem encountered in the best basis algorithm is not encountered here.

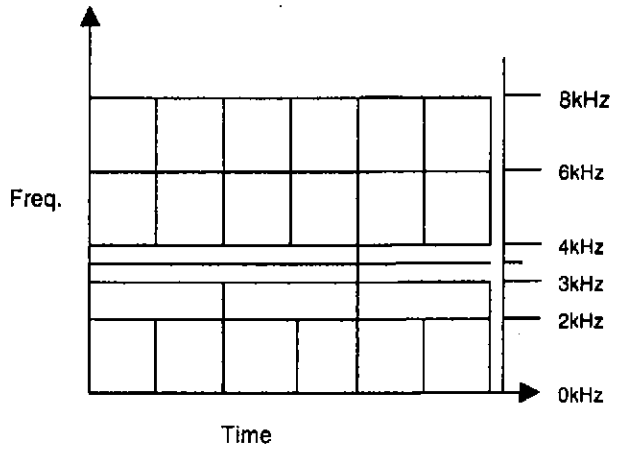


Figure 2: An example of tiling by wavelet packet of the time-frequency plane.

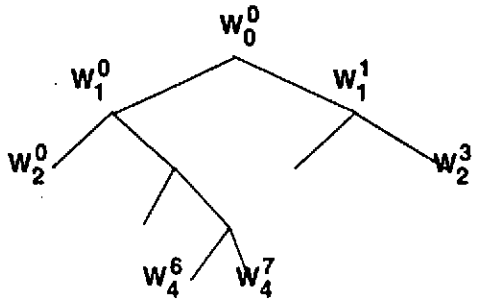


Figure 3: Admissible wavelet packet binary tree for tiling of time-frequency plane of Figure 2.

3. Feature Extraction

Here in this work a frame of 32ms (512 samples) is formed for analysis of a phoneme and different levels of DWT decomposition are applied by using 'Daubechies 6' wavelet filter. A 'k' level of discrete wavelet decomposition will split the frequency band into 'k+1' sub-bands. First of all, the total energy of the wavelet coefficients in each frequency sub-band is calculated. This is normalised by dividing the total energy by the number of wavelet coefficients in each sub-band. The lograthim of normalised energy in each sub-band is used as a feature vector. Although this technique overcomes the problem of shift in the signal, the second problem discussed in Section 2 still remains. To overcome this problem the AWP is used to tile the time-frequency plane giving more frequency sub-bands in 300Hz to 4kHz range. Once the sub-bands are obtained the process of freature extraction is similar to the one explaned above.

4. EXPERIMENTATION

Vowels, unvoiced fricatives and unvoiced stops from the dialect region DR1 (New England region) and DR2 (Northern part of USA) of the TIMIT database were extracted for training and testing the classifier. A total of 151 speakers were used out of which 114 were used for training and the rest for testing the classifier. There were 49 female speakers in all out of which 37 speakers were used for training the classifier. For classification a Linear Discriminant Analysis is used [10], [11].

In the first experiment the DWT was used to decompose the phoneme from 4 to 7 levels, thereby giving 5 to 8 frequency sub-bands. The energy features as well as the proposed log enegy features were calculated and the result obtained for the unvoiced stops is shown in Figure 4. It can be clearly seen that lograthimic compression when applied to the energy features gives a better recognition performance. Also it can be seen that by incearsing the level of decomposition, recognition performance does not improves much. This is due to the fact that more features are derived from the lower frequency end of the signal spectrum which has less discriminatory information.

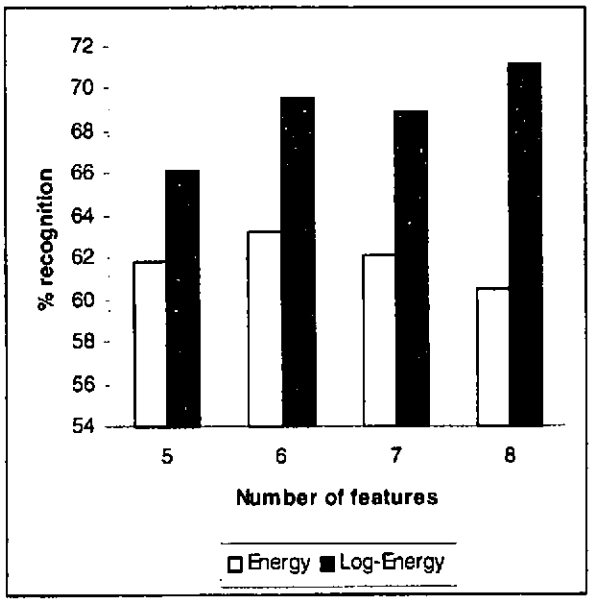
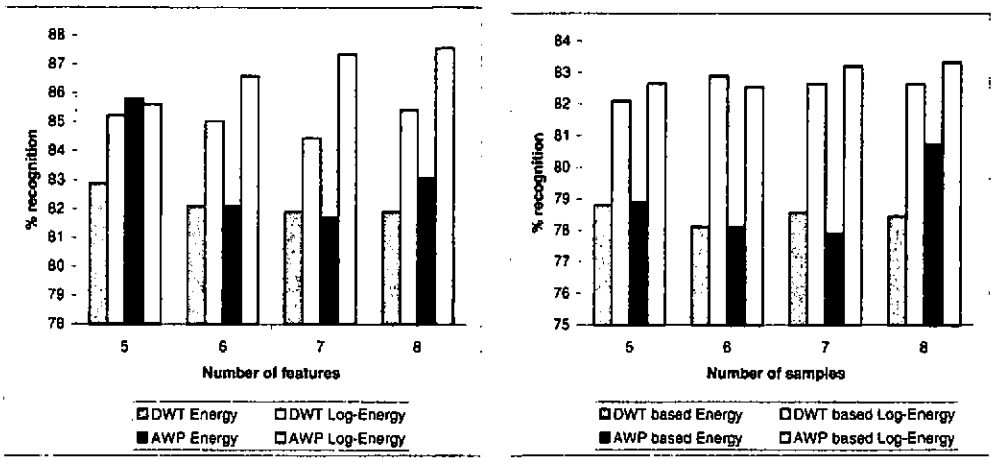


Figure 4: Comparative recognition performance of unvoiced stops using DWT based features

In the second experiment the AWP was used to decompose the phonemes and features were consequently extracted. Figure 5(a) shows the recogntion performance for the unvoiced fricatives based on the DWT and AWP for energy and log energy feautres. It is clear from the Figure 5(a) that AWP base log energy features out perform for all except the first case.

The results obtained for the recognition of the vowels is shown in Figure 5(b). Since the vowels have lower frequency componemts reason the DWT based features even at higher level of decomposition give good results (which is not seen in the case for the unvoiced phonemes). Hence, by using the AWP to have a different time-frequency tiling does not result in huge improvement in the recognition performance as found for the unvoiced stops and fricatives.



(a) (b)

Figure 5: Comparative recognition performance of (a) unvoiced fricatives and (b) vowels for features based on DWT and AWP

5. CONCLUSION

The features obtained by using the AWP are found to be superior compared to the DWT based features. The AWP further overcomes the problem arising due to the features coming from very low frequency at higher level of decomposition by DWT. Also the logarithmic compression does help in significantly improving the recognition performance of the unvoiced stops and fricatives.

REFERENCES

[1] J. Picone. Signal modelling techniques in speech recognition. In *Proc. of IEEE 81 (9), pp 1215-1247,* 1993.

[2] D. O'Shaughnessy. Speech Communication: Human and Machine. *Addison Wesley, New York, USA,* 1987.

[3] T. Tan Beng, Fu Minyue, Spray Andrew and Dermody Phillip. The use of wavelet transform for phoneme recognition. In *Proc. of 4th Int. Conf. of Spoken Language Processing Philadelphia, USA, October, Vol. 4, pp 2431-2434,* 1996.

[4] C. J. Long and S. Datta. Wavelet based feature extraction for phoneme recognition. In *Proc. of 4th Int. Conf. of Spoken Language Processing Philadelphia, USA, October, Vol. 1, pp 264-267,* 1996.

[5] C. J. Long. Phoneme discrimination using non-linear wavelets method. *Ph.D. thesis, Loughborough University, Dept. of Electronic and Electrical Engineering, February,* 1999.

[6] C. J. Long and S. Datta. Discriminant wavelet basis construction for speech recognition. In *Proc. of 5th Int. Conf. of Spoken Language Processing Sydney, Australia Nov-Dec., Vol. 3, pp 1047-1049*, 1998.

[7] Sungwook Chang, Y. Kwon and Sung-il Yang. Speech feature extracted from adaptive wavelet for speech recognition. In *Electronic Letters, Vol. 34, No. 23, 12th November, pp 2211-2213*, 1998.

[8] E. Lukasik. Wavelet packets based features selection for voiceless plosives classification. In *Proc. of ICASSP 2000, Vol. 2, pp 689-692*, 2000.

[9] Stéphane Mallat. A wavelet tour of signal processing. *Academic Press, San Diego,* 1998.

[10] S. Balakrishnama, A. Ganapathiraju, J. Picone, Linear discriminant analysis for signal processing problems, Proc. of the IEEE Southeastcon Lexington, Kentucky, USA, March 1999, pp. 36-39.

[11] J. B. Buckheit, D. L. Donoho, Improved linear discrimination using time-frequency dictionaries, Proc. of SPIE, San Diego, USA, July 1995, Vol. 2569, Pt. 2, pp. 540-551.