MEL-SCALED WAVELET FILTER BASED FEATURES FOR NOISY UNVOICED PHONEME RECOGNITION

O. Farooq and S. Datta

Department of Electronic and Electrical Engineering
Loughborough University, Loughborough, United Kingdom
{O.Farooq, S.Datta}@lboro.ac.uk,

ABSTRACT

In this paper we propose a filter bank structure derived by using admissible wavelet packet transform. These filters have Mel scale spacing and have an advantage of easy implementation with higher resolution in time-frequency domain because of wavelet transform. The features are obtained by first calculating the energy in each filter band and then applying the Discrete Cosine Transform (DCT) to the energy vector. We evaluate the recognition performance of the features derived from the Mel-Scaled Wavelet Filter (MSWF) bank structure and compare it with that derived from Mel Frequency Cepstral Coefficients (MFCC). Experimental results on the phoneme recognition from the TIMIT database show that, features derived by using MSWF performs better as compared to MFCC features for unvoiced stops and unvoiced fricatives. Further the noise performance of these features are also found to be better as compared to MFCC features.

1. INTRODUCTION

In order to have a robust speech recognition system, feature selection and its extraction becomes a very important task. The MFCC has been the most widely used features for speech recognition for the last two decade. Short Time Fourier Transform (STFT) and a bank of triangular filter with overlapping pass-bands are used to derive these features. The filter band spacing follows the Mel scale in order to have similarity with the human hearing system. The drawback of the MFCC is that it uses STFT, which has fixed time-frequency resolution. Due to this reason it is difficult to detect high frequency burst from a low frequency background. In addition to this STFT can be used for the transformation of the stationary signal only, while speech may not be strictly stationary during the window duration. This problem becomes severe when stop (plosive) phonemes are encountered.

Wavelet transform has capability of processing non-stationary signals and can also give multi-resolution decomposition of the signal. It performs a constant ‘Q’ (quality factor) analysis of a signal by projecting it on a set of basis functions scaled and translated to each other. According to digital signal processing approach wavelet transform decomposes the signal into low pass and high pass components and then down-samples it by 2; the inverse transform performs the reconstruction. These filters are perfect reconstruction and orthogonal with finite impulse response and are called conjugate mirror filters [1].

Earlier approach of using wavelet transform for feature extraction [2] used wavelet coefficients with high energy as features. These features suffered from the drawback of shift variance in the signal and were speaker dependent. Later [3, 4] used wavelet packets and applied the best basis to determine the optimum tiling of the time frequency plane. The limitation of the best basis selection algorithm is its shift variance and speaker dependence. Wavelet transform has also been applied instead of Discrete Cosine Transform for feature extraction [5] because of its better time-frequency resolution. Recently, we in [6] proposed features based on Mel-scaled wavelet filters that showed better recognition performance than MFCC.

In this paper we further investigate the use of above features for the recognition performance of unvoiced phonemes under noisy conditions.

2. FILTER DESIGN USING WAVELET TRANSFORM

2.1. Discrete Wavelet Transform

Wavelet transform decomposes signal over dilated and translated wavelets. A wavelet is a function \( \psi \in L^2(\mathbb{R}) \) (i.e. a finite energy function) with zero mean and is normalised (\( \| \psi \| = 1 \)). A family of wavelets can be obtained by scaling \( \psi \) by \( s \) and translating it by \( u \).

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

(1)

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

\[
CWTf(u,s) = \int_{-\infty}^{+\infty} f(t) s^{-1/2} \psi^*(t-u/s) dt
\]

(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 Discrete Wavelet Transform (DWT) of a signal \( f[n] \) with period \( N \) is computed as:

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

(3)

where \( m \) and \( n \) are integers. The value of \( a \) is equal to 2.
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 required, which is computed by a scaling filter and is given by:

$$SF\{n,a^j\} = \sum_{m=0}^{N-1} f\{m\} a^{-j/2} \frac{m-n}{a^j}$$

(4)

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

### 2.2. Admissible Wavelet Packets

DWT performs the recursive decomposition of the lower frequency band obtained by the previous decomposition in dyadic fashion. Hence DWT gives a left recursive binary tree structure where the left child represents the lower frequency band and the right child represents higher frequency band. In Wavelet Packet (WP) decomposition lower as well as higher frequency bands are decomposed into two sub-bands thereby giving a balanced binary tree structure. Each node $W^P_j$, in the tree represents the depth $j$ and the number of node $p$ to the left of it. For a full $j$ level of decomposition a frequency band will be split into $2^j$ sub-bands of equal bandwidth. The use of energy in each different sub-band is not an effective feature as the sub-bands do not follow the natural Mel scale.

The two wavelet packet orthogonal bases generated from a parent node $(W^P_j)$ are defined as:

$$\psi^{2j+1}_{j+1} (k) = \sum_{m=-\infty}^{\infty} h[n] \psi^j_{j} (k - 2^jn)$$

(5)

$$\psi^{2j+1}_{j+1} (k) = \sum_{m=-\infty}^{\infty} g[n] \psi^{j+1}_{j} (k - 2^{j+1}n)$$

(6)

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

Wavelet packet decomposition results in over-complete basis. For a full $j$ level wavelet packet decomposition there will be over $2^{2j^2}$ 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 [3, 7]. However, application of best basis algorithm to the pattern recognition problem is difficult, as they are not translation invariant [1]. For a shift in the signal, the wavelet packet decomposition will give modified coefficients thereby yielding different basis when the cost function is minimised. If the energy in each band was used as feature, this may result into different number of features, which may further create problems in recognition.

In order to overcome the above problems, we propose the use of admissible wavelet packet decomposition, which is in-between DWT and WP and gives the liberty to partition the lower frequency band or the higher frequency band. Figure 1 shows an example of tiling of the time-frequency axis by one of the admissible wavelet tree structure for a four level of decomposition.

![Figure 1: An example of tiling by wavelet packet of the time-frequency axis and the corresponding admissible binary tree structure.](image)

### 3. FEATURE EXTRACTION

The phonemes are passed through the 24 MSWF bank and energy in each frequency sub-band is calculated. If $C_{jk}$ is the $j^p$ coefficient in the $k^b$ sub-band then the total energy ($E_p$) in the sub-band $p$ is given by:

$$E_p = \sum_{j=1}^{N_p} (C_{j,p})^2 \quad p = 1,2,\ldots,L$$

(8)

$$F_p = \frac{E_p}{N_p} \quad p = 1,2,\ldots,L$$

(9)

where $N_p$ is the number of wavelet coefficients in the $p^b$ sub-band and $L$ is the number of sub-bands. The calculated energy is then divided by the number of wavelet coefficients in the corresponding sub-band thereby giving average energy per wavelet coefficients per sub-band ($F_p$). In order to have features with emphasis on the lower frequency sub-bands, Daubechies wavelet filter with 12 taps was selected instead of 32 [9]. The relative emphasis on the lower sub-bands by 12 tap and 32 tap Daubechies filter can be seen from Figure 2, where the
weights in the higher frequency sub-bands (sub-band number 21 to 24) has been taken as reference equal to unity. Logarithmic compression is then applied to the 24 coefficients obtained and then DCT is applied. First 13 coefficients are selected as feature. This process is similar to that of the MFCC feature calculation however; unlike the STFT where overlapping windows are used to reduce the side-lobe energy, due to compact support of the wavelets there is no need of using window functions. This also saves a lot of computation during the feature extraction phase.

Linear Discriminant Analysis (LDA) is used for the purpose of classification of the phonemes. LDA is a tool used for multi-group data classification and dimensionality reduction. It tries to minimise the ratio of within-class scatter to between-class scatter thereby attempting to achieve maximum separation. A within-class scatter matrix defines the scatter of samples around their respective class mean. Between-class scatter matrix defines the spread of the mean vectors around the global mean. LDA tries to separate the different group data by forming a linear decision boundary between them. LDA may not perform well if the mean of different classes is same or if the classes are not linearly separable.

### 4. EXPERIMENTAL RESULTS

Dialect regions DR1 and DR2 of the TIMIT database are used to extract the phonemes from all the possible context. A total of 151 speakers were used for training and testing the classifier. 32ms phoneme duration is selected and passed through the bank of MSWF. The features are extracted as explained earlier in Section 3. The phonemes selected from the unvoiced stops were /p/, /t/ & /k/, the unvoiced fricatives were /f/, /sh/ & /s/. The features derived by using MSWF gives better recognition performance than MFCC features for the case of fricatives and stops (see Figure 3). There is an improvement of about 2.3% in fricative recognition and 0.2% in stops recognition. However, for vowels set (/iy/, /ax/ & /aa/) the MFCC features give slightly better (1%) result because of the periodicity of the signal (since STFT is more suitable to detect periodicity because of its basis function).

In order to test the performance of the features in the presence of noise, additive white Gaussian noise with zero mean is simulated and injected into the phonemes. The performance of the features for 20dB, 15dB and 10dB signal to noise ratio (SNR) is tested by using the LDA classifier. The results obtained are shown in Figure 4 and Figure 5. It can be seen clearly that the performance of the features derived by using MSWF has

---

**Table 1**: Filter and their corresponding frequency bands achieved by admissible wavelet packet decomposition.

<table>
<thead>
<tr>
<th>Filter number</th>
<th>Lower cut off frequency (Hz)</th>
<th>Higher cut off frequency (Hz)</th>
<th>Band-Width (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>125</td>
<td>125</td>
</tr>
<tr>
<td>2</td>
<td>125</td>
<td>250</td>
<td>125</td>
</tr>
<tr>
<td>3</td>
<td>250</td>
<td>375</td>
<td>125</td>
</tr>
<tr>
<td>4</td>
<td>375</td>
<td>500</td>
<td>125</td>
</tr>
<tr>
<td>5</td>
<td>500</td>
<td>625</td>
<td>125</td>
</tr>
<tr>
<td>6</td>
<td>625</td>
<td>750</td>
<td>125</td>
</tr>
<tr>
<td>7</td>
<td>750</td>
<td>875</td>
<td>125</td>
</tr>
<tr>
<td>8</td>
<td>875</td>
<td>1000</td>
<td>125</td>
</tr>
<tr>
<td>9</td>
<td>1000</td>
<td>1125</td>
<td>125</td>
</tr>
<tr>
<td>10</td>
<td>1125</td>
<td>1250</td>
<td>125</td>
</tr>
<tr>
<td>11</td>
<td>1250</td>
<td>1375</td>
<td>125</td>
</tr>
<tr>
<td>12</td>
<td>1375</td>
<td>1500</td>
<td>125</td>
</tr>
<tr>
<td>13</td>
<td>1500</td>
<td>1750</td>
<td>250</td>
</tr>
<tr>
<td>14</td>
<td>1750</td>
<td>2000</td>
<td>250</td>
</tr>
<tr>
<td>15</td>
<td>2000</td>
<td>2250</td>
<td>250</td>
</tr>
<tr>
<td>16</td>
<td>2250</td>
<td>2500</td>
<td>250</td>
</tr>
<tr>
<td>17</td>
<td>2500</td>
<td>2750</td>
<td>250</td>
</tr>
<tr>
<td>18</td>
<td>2750</td>
<td>3000</td>
<td>250</td>
</tr>
<tr>
<td>19</td>
<td>3000</td>
<td>3500</td>
<td>500</td>
</tr>
<tr>
<td>20</td>
<td>3500</td>
<td>4000</td>
<td>500</td>
</tr>
<tr>
<td>21</td>
<td>4000</td>
<td>5000</td>
<td>1000</td>
</tr>
<tr>
<td>22</td>
<td>5000</td>
<td>6000</td>
<td>1000</td>
</tr>
<tr>
<td>23</td>
<td>6000</td>
<td>7000</td>
<td>1000</td>
</tr>
<tr>
<td>24</td>
<td>7000</td>
<td>8000</td>
<td>1000</td>
</tr>
</tbody>
</table>
more noise tolerance as compared to MFCC for the unvoiced stop and fricative phonemes. Also the improvement in the recognition performance by using MSWF is more in the case of unvoiced stops than for unvoiced fricatives for lower SNR. This is due to the fact that the unvoiced stops have more energy in the lower frequency end of the spectrum while in the case of unvoiced fricatives the spectrum is flat (like white noise). At lower SNR the lower frequency band of the unvoiced stops are less corrupted by noise because they have more energy in them. Thus giving the lower bands more emphasis helps in improving the recognition in the unvoiced stops, while it has no effect in the unvoiced fricatives.

5. CONCLUSION

The acoustic phonetic features derived by using MSWF is found to be better that the MFCC for stop and fricative recognition. The admissible wavelet packet transform has been successfully used to give features that are shift invariant as well as speaker independent, thereby, overcoming the problems caused by best basis algorithm. This also verifies that the higher resolution of time-frequency by wavelets is useful for detecting the sudden burst in the signal that is useful for classification. The recognition performance in the presence of noise is found to be better by features derived by MSWF for the unvoiced stops and fricatives. The recognition performance can be further improved by using the delta and delta delta features similar to that used with MFCC.

6. REFERENCES