ABSTRACT

Phoneme recognition is a difficult task in speech recognition as it is variable in length and its acoustic properties change due to co-articulation and variation in dialects. The performance of the speech recognition system is heavily based on features extracted for the phonemes. The conventional technique of Short Time Fourier Transform (STFT) has a serious limitation in resolving the stop (plosive) sounds. This shortcoming can be overcome by using the multi-resolution capability of Wavelet Analysis. In this paper we perform a comparative study of Discrete Wavelet Transform (DWT) and Wavelet Packet (WP) for new dynamic features extraction of phonemes.

1. INTRODUCTION

The basic elements used in the speech recognition system are acoustic model, language model and search. The acoustic model consists of the feature extractor and a Hidden Markov Model (HMM). HMM is used to encode the temporal evolution of the features provided by the feature extractor. Language model gives the probability of occurrence of a word sequence. Beamed Viterbi search is used to give the likelihood of a word sequence for a given set of features. There has been lot of progress in HMM modelling, language modelling and search techniques. A good overview of the developments in these areas can be found in [1], [2] and [3]. Despite of much research in the area of speech recognition Short Time Fourier Transform (STFT) is still used to derive the features used for the recognition. Due to the poor time-frequency resolution of STFT recently Wavelet Transform has been proposed for this purpose in [4] and [5]. Further refinement in these wavelet features was suggested in [6].

Discrete Wavelet Transform (DWT) has been used in [6] for the feature extraction, but the DWT splits the lower frequency band only in dyadic manner. In this work we try to overcome this problem by using the Wavelet Packets (WP) which decomposes the lower as well as higher frequency band into two. This gives over-complete set of basis functions. Earlier best basis selection criterion was applied to select optimum number of basis [5] from an over-complete set of basis. The process however, may give different basis selection for the same signal if it is shifted slightly. Hence it creates problem in recognition; to overcome this we propose a fixed frequency axis tiling resulting into admissible wavelet packet binary tree.

The features are extracted by DWT as well as by the WP decomposition. Section 2 gives the brief introduction to WP analysis and section 3 discusses the method of feature extraction from the phonemes. Linear Discriminant Analysis is carried out for the classification purpose and the results obtained are compared in section 4.

2. WAVELET PACKET ANALYSIS

DWT performs the recursive decomposition of the lower frequency band obtained by the previous decomposition in dyadic fashion. Thus speech signal sampled at 16kHz when decomposed once will give two bands (0-4kHZ & 4-8kHz). The second decomposition will partition the lower frequency band of 0-4kHz further into a band of 0-2kHz and 2-4kHz. 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. By increasing the number of decomposition using DWT the frequency partitioning of the lower frequency band is performed which is not very useful for the classification of the phonemes. In order to have desirable partitioning of the frequency axis Wavelet Packet decomposition can be applied which is generalisation of DWT. In WP decomposition lower as well as higher frequency 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](image)
The two wavelet packet orthogonal bases generated from a parent node \((W^j_p)\) are defined as:

\[
\psi_{2^p j+1}^2(k) = \sum_{n=-\infty}^{\infty} h[n] \psi^j_p(k-2^n j) \tag{1}
\]

\[
\psi_{2^p j+1}^1(k) = \sum_{n=-\infty}^{\infty} g[n] \psi^j_p(k-2^n j) \tag{2}
\]

where \(h[n]\) is the low pass (scaling) filter and \(g[n]\) is the high pass (wavelet) filter. \(\psi\) is a wavelet function with finite energy, zero mean and is normalised \((\|\psi\| = 1)\). A family of wavelets can be obtained by scaling and translating \(\psi\).

Wavelet packet decomposition results in over-complete basis. For a full \(j\) level wavelet packet decomposition there will be over \(2^{2j-1}\) orthogonal bases. From the above library of bases (also called as packet table) best basis is to be selected. A wavelet packet basis divides the frequency axis into interval of varying sizes and a wavelet packet covers each interval with a uniform translation in time. Selection of the best basis tries to have best frequency partitioning by reducing a cost function. However, application of best basis algorithm to the pattern recognition problem is difficult, as they are not translation invariant [7]. 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.

For speech recognition if full wavelet packet analysis is applied then the energy of the each frequency band will not be good as features for recognition because the higher frequency bands have very little discriminatory information. Due to the above reasons best basis criterion as well as full wavelet packet decomposition cannot be used effectively for the extraction of features from a phoneme.

In order to overcome the above problems we use a wavelet packet decomposition, which gives a fixed admissible tree structure. Here the frequency band that is to be split is predefined taking into consideration the spectrum of the speech signal. Thus the frequency band between 300-4000Hz can be partitioned into more number of bands. This will contribute to more number of features coming from this band. As most of the discriminatory acoustic information is present in this band it is expected to provide better features for classification.

From a complete WP decomposition many admissible wavelet tree structures can be derived. Figure 2 shows the tiling of the time-frequency axis by one of the admissible wavelet 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. The splitting of the frequency band can be chosen such that it suits the requirement of selecting discriminatory feature from the phoneme under consideration.

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 use a wavelet packet decomposition, which gives a fixed admissible tree structure. Here the frequency band that is to be split is predefined taking into consideration the spectrum of the speech signal. Thus the frequency band between 300-4000Hz can be partitioned into more number of bands. This will contribute to more number of features coming from this band. As most of the discriminatory acoustic information is present in this band it is expected to provide better features for classification.

From a complete WP decomposition many admissible wavelet tree structures can be derived. Figure 2 shows the tiling of the time-frequency axis by one of the admissible wavelet 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. The splitting of the frequency band can be chosen such that it suits the requirement of selecting discriminatory feature from the phoneme under consideration.

3. FEATURE EXTRACTION AND CLASSIFICATION

A new set of features based on the energy in each frequency band are used which are shift invariant for the DWT [6]. To obtain the features from the WP analysis, the best basis criteria is not applied due to its shift variance problem; instead a fixed WP decomposition is performed to give an admissible tree structure. The advantage of using admissible tree decomposition is twofold. Firstly it gives a unique decomposition, which is not dependent on the signal variations (variation occurs due to difference in speakers for the same phoneme), however this decomposition may not be an optimum one if signal compression is considered. The second advantage of having the admissible tree structure is that the tiling of frequency axis can be selected in a desirable manner. A single frame of 32ms was chosen where DWT and WP decomposition be applied to extract the features.

To extract the dynamic features from the phoneme the frame was further divided into sub-frames of duration 16ms and 8ms. Features extracted from these sub-frames were combined to
from a feature vector. This feature vector actually gives an idea of change of energy in different frequency bands with time. The number of bands in the frequency was also varied to give different number of features.

The minimum value of the frame duration was chosen to be 8ms as no variation can occur in the speech signal in shorter duration. This is due to the physical limitation of the articulators used for speech production. Thus shorter sub-frame duration will only increase the dimension of the feature without giving any discriminatory information to the classifier.

Linear Discriminant Analysis (LDA) was used for the purpose of classification of these 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 separability. 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

TIMIT [8] database was used and dialect region DR1 and DR2 were chosen for the extraction of phonemes. The details of the phonemes extracted form the database is shown in Table 1. A total of 151 speakers were used for training and testing the classifier. The training and the test set prepared were mutually exclusive of each other. Daubechies compactly supported wavelets with 6 vanishing moments was used [7]. Wavelet decomposition of the sub-frame samples is then carried out and energy of the wavelet coefficients in each frequency band is calculated. This energy was normalised by the number of samples in the band thereby giving an average energy per sample in each band.

<table>
<thead>
<tr>
<th>Phonemes</th>
<th>/aa/, /ax/ &amp; /iy/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unvoiced fricatives</td>
<td>/f/, /sh/ &amp; /s/</td>
</tr>
<tr>
<td>Voice fricatives</td>
<td>/v/, /dh/ &amp; /z/</td>
</tr>
<tr>
<td>Unvoiced stops</td>
<td>/p/, /t/ &amp; /k/</td>
</tr>
<tr>
<td>Voiced stops</td>
<td>/b/, /d/ &amp; /g/</td>
</tr>
</tbody>
</table>

Table 1: Details of the phoneme extracted from the TIMIT database for feature extraction

The first recognition test was carried out on the unvoiced stops (/p/, /t/ & /k/). 8 features were extracted by using DWT by having 7 level of decomposition using single frame duration (32ms). By using WP with 8 features different partitioning of the frequency bad is possible. The 8 set of frequency band giving the highest recognition was 0-500Hz, 500-1000Hz, 1-1.5kHz, 1.5-2kHz, 2-3kHz, 3-3.5kHz, 3.5-4kHz & 4-8kHz. The frame was then divided into sub-frames of duration 16ms and 8ms and the features were extracted for the sub-frame duration.

The results obtained are shown in Figure 4 and it can be seen clearly that the WP decomposition outperforms DWT in all the three cases. The maximum difference of about 6% is obtained when 4 sub-frames are used.

![Figure 4: Comparison of the recognition performance of features derived by DWT and WP for different number of sub-frames for unvoiced stops (/p/, /t/ & /k/) phonemes](image)

Figure 4: Comparison of the recognition performance of features derived by DWT and WP for different number of sub-frames for unvoiced stops (/p/, /t/ & /k/) phonemes

The recognition performance achieved by using the LDA on vowels, unvoiced fricatives, voiced fricatives, unvoiced stops and voiced stops is shown in Figure 5. The total number of features used per sub-frame is 8 and 4 sub-frames were used. This gave a total of 32 features in 32ms duration. The recognition performance achieved by the WP features is always better as compared to DWT except for the voiced stop phonemes.

The recognition rate reduces if the number of features is increased for DWT while there is an increase with WP features. This is because of the reason that DWT always tries to decompose the lower frequency band only while in WP we choose the best possible frequency band partitioning. The choice of dividing the frequency band is very important and if not chosen properly may result in the reduction of the recognition performance. Different admissible binary tree structures were tried and the results obtained by the best decomposition are plotted in Figure 6.

![Figure 5: Comparison of the recognition performance of features derived by DWT and WP for different set of phonemes.](image)

Figure 5: Comparison of the recognition performance of features derived by DWT and WP for different set of phonemes.

Figure 6 shows the variation of percentage recognition with the number of features per sub-frame for the case of vowels and unvoiced stops. It can be seen that the recognition rate reduces if the number of features is increased for DWT while there is an increase with WP features. This is because of the reason that DWT always tries to decompose the lower frequency band only while in WP we choose the best possible frequency band partitioning. The choice of dividing the frequency band is very important and if not chosen properly may result in the reduction of the recognition performance. Different admissible binary tree structures were tried and the results obtained by the best decomposition are plotted in Figure 6.
5. CONCLUSION

The acoustic phonetic features derived by using WP is found to be much better that the DWT in all the above cases. Since a fixed admissible binary tree structure is used the problem of selecting different basis for different speaker is avoided. The number of features is not large for classification. The use of four sub-frames increases the recognition performances of the LDA classifier. The performance achieved can be further enhanced by the use of language model. Non-linear classifier can also increase the recognition performance.

6. REFERENCES


6. Farooq O. and Datta S., “A Neural Network Phoneme Classification Based On Wavelet Features”, Accepted for Int. Conf. on Recent Advances in Soft Computing 2000, DE Montfort Univ. Leicester, UK.
