BEST SPEAKER-BASED STRUCTURE TREE FOR SPEAKER VERIFICATION

C. Tadj, C. Gargour and N. Badri

École de Technologie Supérieure - Electrical Engineering Department
1100, Notre-Dame Ouest, Montreal, Qc, H3C 1K3 Canada

ABSTRACT

In this paper we study the use of the Wavelet transform for text-dependent speaker verification purposes. A new algorithm to construct the best admissible tree is proposed which has been used to obtain a speaker dependent tree library. Every tree in this library corresponds to the best structure for a given speaker, therefore the extracted parameters from a given tree are well suited and discriminative to the considered speaker and hence appropriate for speaker verification purposes. Experiments have been conducted using data extracted from the Yoho database. The results that have been obtained show a good level of efficiency and robustness compared to those which can be obtained by using MFCC or other wavelets and wavelets-based methods.

1. INTRODUCTION

Several features for speech and speaker recognition have been used during the last decades. The MEL-Frequency Cepstral coefficients (MFCC) are perhaps those which have been mostly used for this purpose [5]. It should however be noted that because the Discrete Cosine Transform (DCT) covers the whole spectrum of frequencies, its use on a MEL-scaled log filter bank energies in the MFCC based feature extraction in the presence of noise on any single frequency band, could result in a corruption of all the MFCC [7]. Consequently, subband speech recognition has been utilized [1] to overcome this problem in MFCC-based recognizers. Wavelet Transform based methods have recently attracted the attention and interest of researchers [4]. Due to its good time and frequency resolution, the Wavelet Transform (WT) has been used in feature extraction for speech recognition to overcome many of the problems associated with subband based recognizers [1].

In this paper we will focus our interest on the application of the WT to Text-Dependant Speaker Verification (TDSV). The purpose of our work is to investigate the capability of the WT based multi-resolution methods to extract from the signal enough information on the speech and the speaker to permit a good discrimination between speakers.

The paper is organized as follows: Section 2 presents some background on the Wavelet Transform; Section 3 is our main contribution: it describes the selection of the best basis from the best Wavelet Packet (WP) transform; Section 4 contains the experimental work and results as well as the discussion and analysis of these results.

2. RELATED WORK

2.1. Wavelet Transform

The wavelet transform permits the use of short time-windows at high frequencies. This property permits to obtain signal representations with good resolutions in both the frequency domain and the time domain [6]. A wavelet can be seen as band pass filter and the set of its dilated version as a bank of constant filters. To avoid the need for an infinite number of such filters in a given signal representation, a scaling function can be used. The analysis of a signal with wavelets and a scaling function permits its expression in terms of the wavelets up to a given scale.

The choice of the best tree for a given application can be performed using the WT approach.

2.2. MFDWC

Mel Frequency Discrete Wavelet Coefficients (MFDWC) give a new feature vector for speaker verification. This method is used by applying WP transform. An admissible WP tree is proposed by [4] giving a new filter banks structure in which filters have frequency bands spacing approximating the Mel scale used in MFCC. The purpose of using the WPT is to benefit from its localization in the time and frequency domains [7].

The property of partitioning the frequency axis using WPT in MFDWC is obtained by using recursively a pair of conjugate mirror filters [9], which divides the frequency band into two equal halves at each given scale.

Each scale in the WP tree is indexed by its depth j and number of subspaces p below it. The two orthogonal bases at each parent node (j,p) are defined by a low pass filter (LPF) and a high pass filter (HPF) as in [4].

The features are extracted from the different nodes (j,p) of the WP tree. The bands spacing of this nodes corresponds exactly to the frequency bands, which are achieved by the approximate Mel scale as shown in Figure 1. The MFDWC coefficients are computed by taking the DCT of log-energy of each node (j,p) in the WP tree. Each couple of filters (LPF, HPF) in the WP tree is associated with the biorthogonal Daubechies compactly supported wavelets with N vanishing moments [3].

3. NEW SCHEME FOR SPEAKER VERIFICATION

3.1. Selecting the Best Basis

The best basis selection method is usually used for signal compression applications. However, we will use it here for
features selection in order to extract the most significant information from our signal with a minimum cost. The cost being a selected entropy function. This can be done by searching the best basis in which that signal is best represented. The main idea proposed by [2] is to build a library of orthonormal basis relative to a given signal or collection of signals which has the lowest information cost. The best basis relative to the best WPT binary tree are obtained by eliminating branches according to the selected entropy cost function. Here we use the (non-normalized) Shannon entropy function defined by

$$E(x) = -\sum x_i^2 \log(x_i^2)$$

Where $x$ is the signal of a certain node $(j,p)$ of the WP tree. Starting with the root node, the best tree is built as follows.

### 3.2. Construction Algorithm: Speaker-Based Admissible Wavelet Packet Tree

Starting from a collection of $M$ signals $S_1, S_2, \ldots, S_M$ (Signals which correspond to the sentences pronounced by a given speaker), we perform for each of these signals a complete WP decomposition with 5 levels, and $8^{th}$ order of Daubechies wavelet. We apply then for each of these $M$ trees, the algorithm of selecting the best basis using the non-normalized Shannon criterion. Consequently we will have $M$ admissible trees. We create then an initial tree with a complete structure and 5 levels. Thus we calculate for each node $C^j_p$ of this tree, the number of its occurrences in the $M$ admissible trees as a leaf and we denote it by $O^j_p$. Note that $j$ represents the decomposition level and $p$ the position of this node in the corresponding level.

At the end of the construction process, each node $C^j_p$ of the tree is associated with a number $O^j_p$ of its occurrences in the $M$ admissible trees.

The following algorithm shows an efficient way to extract an admissible tree starting from the initial tree, based on the selection of the nodes that have the most significant number of occurrences.

### 3.3. Best Speaker-Based Tree Algorithm (BSBT)

**Step 0.** Initial step. Associate each node of the initial tree to its number of occurrences $O^j_p$ (pre-computed).

**Step 1.** Select new leaves of the initial tree. The selection is done by seeking the nodes for which their number of occurrences, are higher than their descendants.

**Step 2.** Test: $L < \text{threshold}$ ($L = \text{number of selected leaves}$)

* If yes, go to step 6.
* Else, go to step 3.

**Step 3.** Search for “sub-leaves” from the descendants of all the nodes. For each leaf, we consider its descendants as a sub-tree, in which, each node is a sub-tree if conditions of step 1 are satisfied.

**Step 4.** Replace partially or totally the leaves by their sub-leaves, to reach the pre-defined threshold.

**Step 5.** All sub-leaves replaced by their leaves in step 4 become leaves.

**Step 6.** End.

### 3.4. Numerical Examples

The following examples show how the proposed algorithm performs while constructing the admissible tree. The level of decomposition is set to three (3) and the threshold $L$ to five ($L=5$). Each node of the initial tree is materialized by the value of its corresponding number of occurrences.

#### 3.4.1. Example 1

We consider the initial tree as represented in Figure 1.

![Figure 1. Initial tree: Example 1.](image1)

As the threshold is reached (number of leaves is greater than $L=5$) the algorithm stops.

#### 3.4.2. Example 2

Figure 3 shows an initial tree with different number of occurrences.

![Figure 3. Initial tree: Example 2.](image2)
After step 1 of the algorithm, we obtain the following tree

```
1  C_0
   |   |
  40  |  41  C_1
```

Figure 4. Structure of the tree obtained from Figure 3 using BSBT algorithm.

As the threshold (L=5) is not reached, the sub-trees are considered in step 3. We start the process with

```
C_1
   |
  40  |
   |
  5   |
   |
  38  |
   |
  39  |
```

Sub-tree of $C_1$

```
C_1
   |
  39  |
   |
  39  |
```

Sub-tree of $C_1$

Figure 5. Structure of the sub-trees considered from Figure 4 at step 3 using BSBT algorithm.

At the end of the process, the following tree is obtained

```
C_1
   |
  40  |
   |
  5   |
   |
  38  |
   |
  39  |
```

Sub-leaf of $C_1$

```
C_1
   |
  39  |
```

Sub-leaf of $C_1$

Figure 6. Structure of the sub-trees considered from Figure 5 at step 3 for the leaf $C_1^0$.

The same process is repeated for the leaf $C_1^1$. The same reasoning provides the following structure:

```
41  C_1
   |
  36  |
   |
  38  |
```

Sub-leaf of $C_1^1$

```
C_1^1
   |
  36  |
   |
  38  |
```

Sub-leaf of $C_1$

Figure 7: Structure of the sub-trees considered from Figure 4 at step 3 for the leaf $C_1^1$ using BSBT algorithm.

Finally, the nodes $C_3^0$, $C_3^1$, $C_2^1$, $C_2^2$ and $C_2^3$ are the nodes of the final tree as shown in Figure 9.

```
C_3
   |
  40  |
   |
  41  C_3
```

```
C_2
   |
  5   |
   |
  38  |
```

```
C_3
   |
  36  |
```

```
C_2
   |
  38  |
   |
  39  |
```

```
C_3
   |
  39  |
```

```
C_2
   |
  39  |
```

Figure 9: Structure of the final tree obtained using BSBT algorithm.

## 4. EXPERIMENTAL SETUP AND TASK

### 4.1. Corpus Description

A subset of sixty (60) speakers (47 males and 13 females) extracted from Yoho corpus [8] has been used for all the experiments. It consists of 96 sentences uttered by each speaker for the training process and 40 different sentences for the verification task. Each speech signal contains approximately 6 seconds of speech.

### 4.2. MFCC Analysis

The features are a 24-dimensional vector consisting of 12 cepstral coefficients and 12 $\Delta$ coefficients. Analysis conditions are listed in Table 1.

### 4.3. Wavelet-Based Analysis

WP decomposition is applied to each temporal analysis window of 25 ms of duration. a) The log energy in each of the frequency bands is computed giving a total of 20 coefficients in the case of MFDWC. b) The log energy is applied to BSBT as described in section 3. DCT is then computed and the first 12 DCT coefficients are selected as static features. Finally, the 12 corresponding $\Delta$ coefficients are computed.

We have then set the threshold (paragraph 3.3) of the minimum number of leaves to construct the admissible tree to $L=13$.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Fourier</th>
<th>Wavelet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-emphasis</td>
<td>$1 - 0.97 z^{-1}$</td>
<td>$1 - 0.97 z^{-1}$</td>
</tr>
<tr>
<td>Window length</td>
<td>25.0 ms</td>
<td>25.0 ms</td>
</tr>
<tr>
<td>Window shift</td>
<td>10.0 ms</td>
<td>10.0 ms</td>
</tr>
<tr>
<td>Number of features</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>Cepstral coefficient liftering</td>
<td>22</td>
<td>-</td>
</tr>
<tr>
<td>Cepstral mean normalization</td>
<td>yes</td>
<td>-</td>
</tr>
<tr>
<td>Hamming window</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Order of Daubechies</td>
<td>-</td>
<td>8</td>
</tr>
</tbody>
</table>
4.4. Speaker Model Estimation

As discussed, we are interested in TDSV applications. Three-state left to right with no skip phone-based HMM models were constructed for each speaker.

Each model contains 16 mixtures (8 as static and 8 as dynamic). Each of the mixture components has a diagonal covariance matrix. The background model is estimated using all the data available for training.

4.5. Experimental Results

We have set the following parameters:
- Order of Daubechies wavelets is set to 8.
- Initial tree of 5 decomposition levels is used.
- Performance of the system is defined by:
  \[ \text{Perf} = 100 - (FA + FR) \]

FA and FR are the percentages of false acceptations and false rejections.

We have used the method of selecting the Best Speaker-Based Structure Tree (BSBT), which computes the optimal sub tree from an initial tree. The results are as fellows.

5.1.1. Experiments with Yoho

Table 2 shows TDSV system performance (SP) using BSBT, MFCC and MFDWC.

Table 2: Comparison of TDSV SP between BSBT, MFCC and MFDWC using Yoho corpus.

<table>
<thead>
<tr>
<th>Type</th>
<th>FR (%)</th>
<th>FA (%)</th>
<th>Perf (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSBT</td>
<td>0.96</td>
<td>0.00</td>
<td>99.04</td>
</tr>
<tr>
<td>MFCC</td>
<td>1.79</td>
<td>0.54</td>
<td>97.67</td>
</tr>
<tr>
<td>MFDWC</td>
<td>12.53</td>
<td>0.84</td>
<td>86.83</td>
</tr>
</tbody>
</table>

The results show the efficiency of MFCC parameters compared to MFDWC. It also shows the superiority of BSBT compared to both methods.

5. DISCUSSION

From a numerical point of view, we found that there is a significant difference in the output probabilities. Let \( P_{\text{BS1}} \) and \( P_{\text{BS2}} \) be the first and second best output probabilities for a given speech signal. We define \( \mu_{P_{\text{BS}}} = (1/N) \sum P_{\text{BSi}} \) the average best output probability over the N test speech signals. We also define \( \Delta P_{\text{BS}} = (1/N) \sum |P_{\text{BSi}} - P_{\text{BSj}}| \), the average difference between the best two scores.

Table 3 shows that MFCC and MFDWC range probabilities are close as well as the \( \Delta P_{\text{BS}} \), which suggest that the system is vulnerable to errors. Simulations have shown that BSBT provides a more interesting \( \mu_{P_{\text{BS}}} \) and \( \Delta P_{\text{BS}} \) values and therefore provides more robustness to the Speaker Verification system.

6. CONCLUSIONS

In this paper we have introduced successfully a PCA-based wavelet transform to perform frequencies segmentation with levels decomposition. A speaker dependent library tree has been built. The constructed tree is specific to each speaker. Therefore the extracted parameters are more discriminative and appropriate for speaker verification applications. This technique has shown robustness and efficiency. We are currently running more experiments with larger databases.

Table 3: Average output probabilities between the first and the second best scores for BSAT and MFCC, MFDWC and BBS using Yoho corpus.

<table>
<thead>
<tr>
<th>Yoho Corpus</th>
<th>( \mu_{P_{\text{BS}}} ) (log)</th>
<th>( \Delta P_{\text{BS}} ) (log)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSBT</td>
<td>-20.93</td>
<td>1.12</td>
</tr>
<tr>
<td>MFCC</td>
<td>-55.63</td>
<td>0.28</td>
</tr>
<tr>
<td>MFDWC</td>
<td>-28.37</td>
<td>0.53</td>
</tr>
</tbody>
</table>

7. ACKNOWLEDGEMENTS

The authors would like to acknowledge the financial support of the Natural Sciences and Engineering Research Council (NSERC) of Canada.

8. REFERENCES