LINEAR TRANSFORMATIONS IN SUB-BAND GROUPS FOR SPEECH RECOGNITION

B. Doherty, S. Vaseghi, P. McCourt
Queen’s University of Belfast
Ashby Building, Stranmillis Rd., Belfast BT9 5AH, N. Ireland,
(b.doherty, s.vaseghi, pm.mccourt)@ee.qub.ac.uk
http://www.ee.qub.ac.uk/dsp/research/speech

ABSTRACT
Linear transforms have been demonstrated to successfully achieve on-line speaker and environmental adaptation for robust recognition. This paper explores the gains in computational speed, speaker adaptation convergence rate and recognition performance obtained through the use of multi-resolution sub-band linear transforms in speech recognition. A useful feature of multi-resolution processing is that significant savings can be attained as regards transform calculation. In this paper we appraise the relative merits of multi-band processing over that of full-band and present evaluation results on the WSJCAM0 continuous speech database.

1. INTRODUCTION
Speech processing in sub-bands has attracted renewed interests since the publication of a review paper by J. Allen of the seminal work of Fletcher on human speech recognition [1, 2, 3]. In most speech recognition systems, 20 to 30 non-linearly spaced filter banks, such as mel-frequency filter banks, form the basis for front end feature processing. The log filter bank power spectrum is conventionally transformed by a DCT to encode the spectral data into fewer relatively decorrelated cepstral coefficients. In [5] we proposed a multi-resolution speech processing method which falls between the two extremes of using cepstrum features derived over the full voice bandwidth, and using spectral features that belong to the individual critical bands. In the multi-resolution method the voice bandwidth is divided into a small number of filter bank groups each of which spans a sufficiently large number of critical bands to contain useful information for modelling and recognition.

A useful feature of multi-resolution processing is that instead of a relatively large linear matrix transformation for the full spectrum, a number of smaller transformations are used for multi-resolution processing. Linear transformations are indispensable in speech processing, and are used for; (i) encoding of the speech log spectral envelop into a relatively small number of cepstral features, (ii) linear prediction modelling, (iii) adaptation to speaker characteristics, and (iv) adaptation to noise and environmental characteristics. In its most familiar form a linear transformation has the simple form of

\[ y = Ax \]  

where \( A \) is the transformation matrix, \( x \) is the vector to be transformed and \( y \) is the transformed vector. In a multi-resolution representation with for example a set of 3 sub-band groups equation (1) assumes the form

\[ \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} A_1 x_1 \\ A_2 x_2 \\ A_3 x_3 \end{bmatrix} \]

therefore instead of using a relatively large matrix \( A \), we have a set of 3 smaller matrices \( [A_1, A_2, A_3] \) which in total have fewer parameters, require less computation, and as in the case of LPC modelling can have better signal to noise ratio properties. We present experiments in the use of sub-band linear transformations for the following cases :

(i) Linear Feature Extraction : A DCT is conventionally used on the full voice-band log power spectrum to obtain a set of about 12 cepstral coefficients to which the first and second order dynamic features are appended. The space of the cepstral vectors in each state are modelled by a mixture Gaussian with diagonal covariance matrices. Diagonal matrices are used due to the prohibitively higher
computational and training data costs required by full covariance matrices. With multi-resolution features instead of a set of large covariance matrices we have a set of smaller matrices making it practical to use full covariance matrix Gaussian distributions.

(ii) Maximum Likelihood Linear Regression (MLLR) Speaker Adaptation. An important issue with the use of MLLR in real-time applications is the speed of computation and the speed of convergence. The adaptation of speaker characteristics in sub-band groups offers advantages in both computation and speed. Furthermore speaker characteristics are more manifested in the lower sub-band groups where the voiced speech energy is mostly concentrated. We present experimental results on sub-band MLLR comparing recognition rate, computational complexity, speed and convergence in sub-bands with that of full band speaker transformation.

The layout of this paper is as follows. Section 2 describes the implementation of the linear feature extraction. Section 3 discusses the possible gains to be obtained from multi-band processing in the context of using MLLR. Section 4 presents the evaluation results and section 5 completes the paper with a discussion and conclusions.

2. LINEAR FEATURE EXTRACTION

A DCT was applied to 2 sub-bands of the full voice-band log power spectrum to obtain a set of 13 cepstral coefficients. 7 coefficients were extracted from the lower sub-band, which extended from 0-2.28 KHz approximately. The upper sub-band encompassed the remainder of the spectrum (2.28-8 KHz), from which the remaining 6 coefficients were extracted. These features were chosen to enable a direct comparison with conventional 13 MFCC’s (Mel Frequency Cepstral Coefficients). The sub-band features were supplemented with both delta and acceleration coefficients calculated using the HTK (Hidden Markov Model) Toolkit. For the purposes of modelling the cepstral vectors using HTK, the resulting input observation sequence was regarded as being two independent data streams. In this way stream 1 modelled the lower sub-band (including static, delta and acceleration coefficients) and stream 2 the upper sub-band. Figure 1 portrays a graphical representation of the feature extraction technique.

3. LINEAR TRANSFORMATION (MLLR)

MLLR [6, 7, 8] is a speaker adaptation technique which uses a set of regression based transforms to tune a set of HMM mean parameters to a new speaker. It produces improvements with small amounts of adaptation information and by the sharing of transformations and data. Each of the transformations are applied to a number of HMM mean parameters and estimated from the corresponding data.

Adaptation was performed for both full-band and independently in both sub-band streams as a means of establishing to some degree where the most discernible speaker characteristics are resident.

3.1 Use of Regression Classes

Regression classes were chosen according to broad phonetic class divisions. The purpose of the division was to ensure that there was enough adaptation data for each class so that a robust estimation of the regression matrix could be made. Optimal division occurs in the case where the trade off between specificity of the regression classes and robustly estimating the transform is balanced.

3.2 Computational Speed

When calculating the speaker adaptive transform with diagonal covariance matrices, it is necessary to invert an (n+1) by (n+1) matrix for each of the dimensions of the transform matrix [8], where n is the number of components in the vector to be
adapted. This inversion may be performed in $O(n^3)$ operations. Hence the total cost, is approximately
$O(n^4)$ per transform. Taking the conventional instance where $n = 39$ and the equivalent sub-band case of two separate feature vectors of length 21 and 18, inversion can now be reduced to $((21/39)^3 + (18/39)^3)$ or 25.43% of the previous calculation. Correspondingly an equivalent diminishment in computation to 12.93%, can be attained for estimation of the total transform.

For the case of full band full covariance adaptation $O(n^6)$ calculations are required per transform inversion. Again a meaningful saving can be acquired with only $((21/39)^6 + (18/39)^6)$ or 2.9% of the previous calculation needed. In terms of computational capacity a comparison of full band diagonal covariance adaptation to that of sub-band full covariance adaptation yields a ratio of computation of $1:44$. In other words for every operation using a full-band diagonal covariance model , a sub-band full covariance requires 44. This is weighted against a previous full-band ratio of $1:1521$, thus warranting a significant investigation of its merits.

### 3.3 Speed of Convergence

A comparison was made between the speed of convergence for both full-band and sub-band adaptation using diagonal covariance matrices. The latest version of HTK provides a tool for adaptation however this only works for the single stream case. Therefore while HTK was used to adapt models trained on full-band MFCC’s code was developed to perform sub-band adaptation. It was found however that as the amount of adaptation data increased the full-band adapted much more readily than the sub-band case.

### 4. RESULTS

The experiments are based on the WSJCAM0 database using 3-state left-right HMM monophones. Table 1 presents results for both sub-band and full band baseline HMM’S. It can be seen that the sub-band features compare well with those of full band.

<table>
<thead>
<tr>
<th>Mixtures</th>
<th>0+7+6 (SB)</th>
<th>13+0+0 (FB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>54.48/36.95</td>
<td>54.47/36.33</td>
</tr>
<tr>
<td>3</td>
<td>55.89/39.69</td>
<td>56.23/39.89</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mixes</th>
<th>Both sub-bands</th>
<th>Lower sub-band</th>
<th>Upper sub-band</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>60.71/43.08</td>
<td>56.56/37.93</td>
<td>56.44/38.21</td>
</tr>
<tr>
<td>3</td>
<td>61.98/45.60</td>
<td>57.85/40.32</td>
<td>60.76/43.81</td>
</tr>
<tr>
<td>5</td>
<td>62.86/46.59</td>
<td>59.62/42.06</td>
<td>60.76/43.81</td>
</tr>
<tr>
<td>8</td>
<td>63.51/47.57</td>
<td>60.52/43.67</td>
<td>61.31/45.00</td>
</tr>
<tr>
<td>12</td>
<td>62.46/46.66</td>
<td>58.24/43.05</td>
<td>61.78/45.92</td>
</tr>
<tr>
<td>15</td>
<td>63.23/47.74</td>
<td>61.50/45.45</td>
<td>61.48/45.66</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reg. Class</th>
<th>1 mix -A</th>
<th>3 mix -A</th>
<th>5 mix -A</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>60.71/43.08</td>
<td>61.98/45.60</td>
<td>62.86/46.59</td>
</tr>
<tr>
<td>4</td>
<td>61.12/44.50</td>
<td>60.76/44.96</td>
<td>62.96/45.04</td>
</tr>
<tr>
<td>7</td>
<td>61.54/44.03</td>
<td>62.35/45.40</td>
<td>63.63/45.45</td>
</tr>
<tr>
<td>13</td>
<td>60.85/43.48</td>
<td>60.55/43.78</td>
<td>62.06/44.49</td>
</tr>
<tr>
<td>44</td>
<td>61.15/43.72</td>
<td>62.41/44.55</td>
<td>63.62/44.66</td>
</tr>
</tbody>
</table>

### Table 1: Comparison of sub-band and full-band cepstral feature recognition.

Adaptation was performed using 18 adaptation utterances including silences. Table 2 gives the results obtained when adaptation was performed for both sub-bands together and independently for each sub-band. In this way it was possible to evaluate the individual performance of each frequency band in terms of its contribution to the recognition rate achieved. Any improvement in gain achieved from sub-band adaptation is seen to diminish as the number of mixtures used increases. It was noted however that the upper sub-band consistently out performs the lower sub-band in its contribution to recognition rate.

### Table 2: Independent sub-band adaptation

Of particular interest to note is that as the number of mixtures increase, the accuracy of the upper band approaches that of dual sub-band adaptation. Table 3 illustrates the results obtained when regression class trees were used in the adaptation process.

### Table 3: Adaptation with the use of Regression classes (A - Adapted).

In all cases regression classes show an improvement over the global transform for the single mix case. However this rate of improvement soon falls behind that of the global transformation
as the number of mixtures increase. It can be said that regression classes appeared to produce little benefit over the global transformation.

5. DISCUSSION AND CONCLUSIONS
This paper has investigated the gains in computational speed, speaker adaptation convergence rate and recognition performance obtained through the use of multi-resolution sub-band linear transforms in speech recognition. As can be appreciated adaptation in sub-bands can offer significant reductions in computation. In terms of potential advantages the gains are most apparent when using full covariance. Although this technique has used features which give comparable performance to full-band features, the increase in computational speed when adapting models is mitigated by a loss in recognition rate. Evidently, investigations have be initiated to ascertain the optimal features to be extracted for use in the adaptation process.

This technique can also be seen as being instrumental as a means of substantiating to some degree, where the most salient speaker characteristics are located, as adaptation of various frequency bandwidths can be evaluated independently. Of advantage to note is that the upper sub-band adapted on its own provides a greater increase in recognition than the lower band. This was unexpected as most of the speaker information was hypothesised to lie in the lower frequencies where most of the voiced energy is concentrated. This could be due to the cuttoff frequency or simply to the fact that the upper bandwidth is modelled better. It could also signify that the full voice bandwidth is not a prerequisite for speaker adaptation and that speaker characteristics are only manifest in certain frequencies. Thus choosing pertinent frequency bandwidths with high speaker information densities which are more amenable to adaptation could be a possible future avenue for research. In addition this work will subsequently be extended to the use of full covariance matrices in order to exploit the reduced computational complexity made possible with multi-resolution processing.

Acknowledgments
This work is sponsored by the UK research council EPSRC under grant number GR/L60463.

6. REFERENCES