ABSTRACT

This paper describes the use of denoising techniques in the time domain applied to the outputs of filters corresponding to a Multi Resolution Analysis. The fact that energies of denoised samples are used for Automatic Speech Recognition (ASR) makes soft thresholding particularly attractive especially if Principal Component Analysis (PCA) is applied to the whole tree of energy features. This consideration is supported by experimental results on a very large test set including many speakers uttering proper names from different locations of the Italian public telephone network. The results show that soft thresholding outperforms J-Rasta PLP with a WER reduction, after denoising, of 26%.

1. INTRODUCTION

In a previous paper [4], the use of Multi Resolution Analysis (MRA)(see [10] for an overview) in Automatic Speech Recognition (ASR) systems has been discussed. MRA generates a signal representation in the time and frequency domains with resolutions depending on frequencies. A discussion on the advantages of this approach w.r.t. Discrete Wavelet Transforms (DWT) is also provided with reference to the pertinent literature.

Different time and frequency resolutions are obtained with a hierarchical time-frequency decomposition. The simplest way to obtain such a decomposition is to recursively use a low and a high pass filtering operations followed by downsampling at 0.5 the sampling frequency of the input signal. Design choices are discussed in [4].

A time-domain filtering is performed with a signal representation obtained from frequency components within each subband. It is shown that, if for each frame the subband signal energies are computed, a reduced variability is observed in each band. Features appears to be relatively immune [7] to local changes within a particular band.

Experiments on an Italian town test set indicated that:

- MRA shows performance almost equivalent to J-Rasta PLP at the expense of a larger number of parameters without any preprocessing like spectral subtraction.
- the synergy of fixed and multiple-resolution analysis leads to a significant WER reduction w.r.t. the situation in which the most effective features (J-Rasta PLP) are used alone.

The objective of the research described in this paper is to assess the impact of denoising techniques applied to MRA when energies of signals at tree nodes are used as features for ASR.

The investigation to be discussed deals with feature transformation using Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA).

Experiments carried on a town name test set show that PCA is slightly more effective than LDA in reducing the number of acoustic features leading to performance comparable to that of JRASTA Perceptual Linear Prediction Coefficients (JRASTAPLP) [5] and significantly better than that of Mel Frequency-scaled Cepstral Coefficients (MFCC), making MRA followed by PCA attractive for ASR front-ends.

The acoustic models in the application considered are embedded into a sophisticated Neural Network (NN) described in [3]. Models are not adapted to new corpora nor to new environment conditions, because the lexicon is continuously updated and speakers may use any telephone. As background noise has a very large variability, model training is performed with phonetically balanced sentences uttered by a large population of speakers over the telephone in normal conditions, i.e. without any particularly strong additive noise. Signal of this type are considered clean.

Recognition performance thus diminishes in noisy conditions in which real-life background noise affects the speech signal in such a way that the training conditions are considerably different w.r.t. test. Spectral subtraction or methods inspired by it are usually applied for speech enhancement when fixed resolution analysis is performed. For MRA, speech enhancement is achieved by soft thresholding of speech output samples. The literature on soft thresholding is rich of methods for increasing the signal to noise ratio (SNR).

Spectral subtraction operates in the frequency domain, while soft-thresholding operates in the time domain. The main contribution of this paper is the comparison of the two approaches to show that, as MRA enhancement is performed at the sample level, the effect of possible distortions is attenuated by energy computation. Experiments support these findings by showing a considerable advantage when soft thresholding is applied to MRA followed by PCA in comparison with the use of spectral subtraction with other methods.

Another important consideration is that, for ASR, the main objective is that of reducing the difference between train and test conditions and parameter estimation has to be formulated in this framework. Furthermore, improvements should be observed when soft thresholding is applied only when the tree of energies measured at the output of the MRA filters in
non-speech segments exhibit a substantial difference w.r.t. the tree of energies measured on non-speech segments of clean signals. Preliminary experiments not described in this paper support these last two considerations.

Section 2 describes the MRA system used for the experiments reported in this paper. After a short introduction on dimensionality reduction in Section 3, experiments with PCA and LDA are described in Section 4. Soft thresholding and denoising techniques are introduced in Section 5 followed by experimental results and comparison with spectral subtraction for JRASTA PLP.

### 2. SYSTEM DESCRIPTION

A binary tree of filters is used having the structure depicted in Figure 1. The total bandwidth spanned is 4000 Hz.

![Figure 1: The MRA filter tree. Note the LP-HP filter inversions due to aliasing (gray code effect). The left columns shows the time-frequency resolution at each level of the tree.](image)

The filters used in the tree are half-band (low-pass LP and high pass HP) perfect-reconstruction wavelet filters. For this reason, the analysis performed by the tree of filters is called Wavelet Packet (WP) analysis. Filters are 19th order Orthogonal IIR Wavelet filters as proposed by Selesnick in [8]. Their design is discussed in [4].

A frame synchronous, variable resolution feature computation is performed at each node $v$ of the tree on a variable number of samples $N_v$ depending on the tree level:

$$M_v(n) = \sum_{k=0}^{N_v} \mu[u_v(k)]; \mu(0) = 0$$

Different types of function have been considered for $M_v(n)$, including energy per sample, entropy per sample, Teager operator, theoretical dimension. It has been found that the most suitable of them is the logarithm of the energy per sample:

$$M_v(n) = \log \left( \frac{1}{N_v} \sum_{k=0}^{N_v} [u_v(k)]^2 \right)$$  \hspace{1cm} (1)$$

As the time resolution halves at each down-sampling (while the frequency resolution doubles at each half band filtering) the product of the time and frequency resolutions is always the same at each level of the tree. The only exception takes place for the highest sub-bands where the averaging window is never smaller than one frame (10 ms). As a result, at the nodes of a tree with 6 level depth, energies are computed on different time intervals, from 48 ms at the 6th level, where the frequency resolution is 125 Hz, to 10 ms at the root where the total energy of the signal is computed with a frequency resolution of 4 kHz (the whole band, as 8 kHz sampling rate).

### 3. DIMENSIONALITY REDUCTION

In order to reduce dimensionality and increase robustness, PCA and LDA have been performed on the whole set of tree features at each time frame with a 10 ms frame rate.

PCA has been performed by transforming the 63 features in order to have zero mean and unit variance. Then the covariance matrix $C$ has been obtained. High correlations have been observed between nodes and their fathers in accordance with the theory.

It has also been observed that node 32, corresponding to the 0-125 Hz band is uncorrelated with other nodes. As it does not carry useful information on telephone data, it can be neglected.

The eigenvalues of $C$, normalized w.r.t. the total variance are plotted, ordered by their magnitude, in Figure 2. The first four eigenvectors of $C$ are shown in figure 3 where it is important to notice the repetitive pattern due to the presence of the same kind of spectral information at increasing levels of details in the different levels of the MRA tree.

![Figure 2: Eigenvalues of C, normalized w.r.t. the total variance ordered by their magnitude.](image)

![Figure 3: First four Eigenvectors of C.](image)

As the recognizer is based on a hybrid NN-HMM system described in [3], the 686 NN outputs have been used to identify the classes for LDA. A linear transformation matrix has been obtained in order to optimize separation among these classes. Experimental results are reported in the next section.
4. EXPERIMENTAL RESULTS ON REDUCTION OF DIMENSIONALITY

4.1 Experimental Setup
Separate train and test corpora were used. Both corpora are made of telephone speech, collected in Italian language from different cities of the Italian Telephone Network. The signal bandwidth is 300-3400 Hz and the sampling frequency is 8 kHz.

Speakers were evenly distributed among males and females coming from many Italian regions and with different accents. Training was performed on 2274 speakers uttering a total of 4875 phonetically balanced sentences with a vocabulary of 3653 words.

The test set corpus consists of 14473 isolated word utterances, from 1050 speakers, containing 475 city names. These words were not included in the training set.

4.2 Results
Table 1 reports the experimental results obtained with different front-ends (MFCC, J-Rasta PLP) and with different transformations obtained from the Wavelet Packet (PCA, LDA). Row headings indicate the type of parameters supplied as input to the NN.

Twelve MFCC and J-Rasta PLP parameters are used with the total signal energy. A frame is represented by a vector of these parameters and their first and second time derivatives computed by linear regression.

MRA WP63 has a feature vector for the n-th frame whose elements are computed with the (1) for all the nodes of WP. PCA20 and LDA20 indicate the use of the first 20 coefficients of, respectively, PCA and LDA transforms. (MRA PCA20 no32) indicates that node 32, corresponding to the 0-125 Hz band, has been neglected before performing PCA. The heading rasta indicates that a Rasta filtering has been applied at each WP node before performing PCA. First and second time derivatives of these parameters are also included in the feature vector.

<table>
<thead>
<tr>
<th>Speech Parameterization</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>3.34</td>
</tr>
<tr>
<td>J-Rasta PLP</td>
<td>4.38</td>
</tr>
<tr>
<td>MRA WP63</td>
<td>4.99</td>
</tr>
<tr>
<td>MRA LDA20</td>
<td>4.97</td>
</tr>
<tr>
<td>MRA PCA20</td>
<td>4.73</td>
</tr>
<tr>
<td>MRA PCA20 no32</td>
<td>4.45</td>
</tr>
<tr>
<td>MRA PCA20 no32 rasta</td>
<td>4.28</td>
</tr>
</tbody>
</table>

Table 1: Recognition word error rates (WER) with different types of features and dimensionality reduction

It can be observed that (MRA PCA20 no32 rasta) has performance slightly better than (J-Rasta PLP) and both outperform MFCCs even if the number of features is larger for PCA20 (20 basic features instead of 12; energy, first and second time derivatives being added in both cases).

Slight degradations were observed by reducing the number of PCA coefficients to 12 and it was concluded that this was not worth doing since the NN for handling 20 coefficients could be trained and used with the same computational effort as the one for 12.

The LDA transformation seems to be less effective than PCA in our framework. This can be due to the fact that the attempt to linearly separate the input classes is not helpful for the NN, that performs the same task in a powerful non-linear way.

5. DENOISING
State of the art in speech enhancement is described in [1]. Some of the findings relevant to ASR are:

- a distortion measure which is based on the mean squared error of the log spectra is more suitable for speech processing,
- denoising in the wavelet domain is based on the principle of selective wavelet reconstruction which finds a subset of coefficients to be multiplied by an appropriate gain,
- soft thresholding techniques are effective whenever few wavelet coefficients contribute to the noiseless signal; these techniques compute a denoised sample at each node as follows:

\[ y_v(k) = \begin{cases} u_v(k) - \alpha_v \sigma_v & \text{if } |u_v(k)| > \alpha_v \sigma_v \\ 0 & \text{otherwise} \end{cases} \tag{2} \]

where \( \alpha_v \) is a threshold and \( \sigma_v \) is an estimation of the power of additive noise affecting node \( v \) assuming that it is not correlated with the signal;
- suppressing noise by adaptively thresholding the empirical wavelet coefficients (�urenhrinkis) is noticeably effective;
- a priori SNR has to be tracked for all nodes in the MRA tree,
- shift invariance obtained by shifting the window, estimating parameters, then realigning and averaging does not necessary lead to improvements,
- undesired effects are smoothed by a decision directed approach to prior SNR estimation,
- in principle, MRA-based Wiener estimator outperforms the Discrete Fourier Transform (DFT) based one as well as when Discrete Cosine Transform (DCT) is performed on certain filter outputs.

These findings have been formulated in the framework of speech enhancement for coding and reconstruction. In the case of ASR, the purpose of denoising is not to achieve the maximum SNR, but to have the energies in noisy conditions as close as possible to those observed in the training conditions.

We can thus obtain optimal values for \( \alpha_v \) at each node \( v \) by minimizing the mean square error between the energies of the clean signal and the energies obtained after soft-thresholding.

The energies are computed as follows:

\[ Y_v(n) = \sum_{k=n}^{N} |y_v(k)|^2 = \sum_{k=n}^{N} |u_v(k) - \alpha_v \sigma_v|^2 \tag{3} \]

If two versions of the same data are available, namely before adding noise (S) and after adding noise and thresholding (Y), then, expressing the energies as function of the window

\[ |u_v(k)| = \begin{cases} u_v(k) - \alpha_v \sigma_v & \text{if } |u_v(k)| > \alpha_v \sigma_v \\ 0 & \text{otherwise} \end{cases} \]
interval \( n \) in which they are computed, the following mean-square error can be computed:

\[
E_\nu(\alpha) = \sum_n Y_n(\alpha) - S_n(\alpha))^2
\]  

(4)

Mismatch between train and test conditions is reduced by finding \( \alpha \) for which \( E_\nu(\alpha) \) is minimum. 

For the experiments described in the next section, a classical speech enhancement approach has been followed, a common value of \( \alpha \) has been used for all the nodes, noise has been dynamically estimated using a method proposed in [6], and \( \alpha \) has been estimated as proposed in [2].

6. SOFT THRESHOLDING EXPERIMENTS

6.1 Experimental Setup

In this set of experiments, in order to obtain more accurate and stable models, the training set described in subsection 4.1 has been increased to include 12212 speakers for a total of 37333 training utterances.

The test set is the same one described in 4.1, but two versions of it were considered. The first one is exactly the one mentioned in sub-section 4.1. This corpus, indicated as TC, is considered as clean because the noise level in the test set is comparable to that in the training set.

An additional test set was derived from the first one by adding telephone-box background noise with a resulting average Signal-to-Noise Ratio (SNR) of 15 dB. This corpus, indicated as TN, is referred to as the noisy one.

The vocabulary used for the recognition experiment contains 475 names of Italian towns.

6.2 Results

Table 2 reports the experimental results obtained with J-Rasta PLP and the best configuration of MRA (MRA PCA20 no32 rasta) on the clean and noisy test sets.

For J-Rasta PLP the results are reported without any additional processing and with spectral subtraction.

For (MRA PCA20) the results are reported without any additional processing and with denoising based on soft-thresholding.

<table>
<thead>
<tr>
<th>Speech Parameterization</th>
<th>Denoising</th>
<th>TC</th>
<th>TN</th>
</tr>
</thead>
<tbody>
<tr>
<td>J-Rasta PLP</td>
<td>NO</td>
<td>2.59</td>
<td>26.66</td>
</tr>
<tr>
<td>J-Rasta PLP</td>
<td>Spectral Subtraction</td>
<td>2.66</td>
<td>15.95</td>
</tr>
<tr>
<td>MRA PCA20 no32 rasta</td>
<td>NO</td>
<td>2.63</td>
<td>23.30</td>
</tr>
<tr>
<td>MRA PCA20 no32 rasta</td>
<td>Soft-Thresholding</td>
<td>2.85</td>
<td>11.79</td>
</tr>
</tbody>
</table>

Table 2: Recognition Word Error Rates on clean and noisy test set

MRA appears to be superior in the presence of mismatch between training and test conditions. In the case of J-Rasta PLP, spectral subtraction reduces the WER by 40.17%, while soft thresholding reduces the WER by 49.4% and outperforms J-Rasta PLP with a WER reduction, after denoising, of 26%.

7. CONCLUSIONS

After having shown in a previous paper [4] that a complete wavelet package is more effective than a wavelet basis for ASR, it is shown in this paper that recognition performance improves if PCA is applied to all the energies of a MRA tree, leading to results comparable to those obtained by using JRASTA PLP coefficients and outperforming MFCCs.

Denoising techniques in the time domain are applied to MRA features while techniques in the spectral domain are applied to JRASTA PLP features. A substantial improvement is observed when denoising is performed in the time domain. This can be explained by the fact that non-linearities introduced by denoising in the time domain are smoothed because the features considered for recognition are energies computed at each MRA tree node.

Furthermore, energy computation is dominated by the largest samples which are moderately affected by uncorrelated additive noise and by soft thresholding.

Future efforts will be dedicated to the estimation of soft thresholding parameters in such a way that the mismatch between train and test conditions is minimized.

ACKNOWLEDGMENTS

The work described in this paper is part of a research effort carried out in the SMADA project on Telephone Directory Assistance. This project is partially funded by a program of the Human Language Technology Division of the European Community.

REFERENCES