INTEGRATION OF FIXED AND MULTIPLE RESOLUTION ANALYSIS IN A SPEECH RECOGNITION SYSTEM

Roberto Gemello[*] *, Dario Albesano*[*]*, Loreta Moisa*[*] *and Renato De Mori*[§]

[*] CSELT
Centro Studi e Laboratori Telecomunicazioni
via G. Reiss Romoli, 274 - 10148 Torino - Italy
roberto.gemello@cselt.it, dario.albesano@cselt.it,
loreta.moisa@cselt.it

[§] LIA CERI-IUP
University of Avignon , BP 1228
84911 Avignon Cedex 9 - France
renato.demori@lia.univ-avignon.fr

ABSTRACT

This paper compares the performance of an operational Automatic Speech Recognition system when MFCCs, J-RASTA Perceptual Linear Prediction Coefficients (J-Rasta PLP) and energies from a Multi Resolution Analysis (MRA) tree of filters are used as input features to a hybrid system consisting of a Neural Network (NN) which provides observation probabilities for a network of Hidden Markov Models (HMM). Furthermore, the paper compares the performance of the system when various combinations of these features are used showing a WER reduction of 16% w.r.t. the use of J-Rasta PLP coefficients, when J-Rasta PLP coefficients are combined with the energies computed at the output of the leaves of an MRA filter tree. Such a combination is practically feasible thanks to the NN architecture used in the system. Recognition is performed without any language model on a very large test set including many speakers uttering proper names from different locations of the Italian public telephone network.

1. INTRODUCTION

Attention has been devoted in the recent years to the use of Multi Resolution Analysis (MRA) of a signal in the time and frequency domain as opposed to popular techniques based on single resolution analysis in which a fixed time window is used for computing samples of spectral energies at different frequencies.
Experiments seem to indicate that performance improvements have been obtained by applying MRA to speech coding, while the application to Automatic Speech Recognition (ASR) appears to be more problematic.
MRA is exploited in Discrete Wavelet Transforms (DWT) which provides a compact representation with two key properties for a large class of signals and images, namely:

- smooth signal/image regions are represented by small coefficients, while edges and other singularities are represented by large coefficients; thresholding can thus be used for denoising;
- large and small coefficients cascade along the branches of a wavelet tree.

Some motivations for using Discrete Wavelet Transforms (DWT) as opposed to all Mel Frequency-scaled Cepstral Coefficients (MFCC) are [5]:

- corruption of a frequency band of speech affects all MFCCs,
- in the transition between two phonemes, some spectra may have features of both in different bands, these features are separated by different band analysis,
- a fundamental paper by Allen [1] on human speech recognition mentions that perceptual features are local in frequency and asynchronous in time; ASR systems do not respect these two principles,
- if a series expansion has to be performed, basis vectors are needed some of which have good resolution in time and some others have a good resolution in frequency,
- for coding, it is important to have signal representations with which the best reconstruction can be achieved with minimum information; moreover, for ASR, the best representation is the one which leads to the best separation of speech units,
- a simple analysis of wavelets shows that they are more concentrated in frequency and time than Discrete Cosine Transform (DCT).

Unfortunately, some problems are encountered with MRA and DWT. In fact, noise can enhance irrelevant signal coefficients and attenuate large signal coefficients. Moreover, due to the sparseness of the data, coefficient magnitudes can vary a lot, near the edges of the segment in which they are computed, for slight changes in the alignment. This can be avoided by introducing redundant DWT or complex Wavelet Trees.
Even if interesting solutions have been proposed for speech analysis and coding, attempts to improve ASR performance produced less impressive results. One of the reasons for this may be that, for ASR, the important features are not the time samples of filter outputs, but representation of energy components and their transformation. Along this line, some important attempts have been performed.
Different time and frequency resolutions can be obtained with a hierarchical time-frequency decomposition performed by a wavelet transform. The simplest way to obtain such a decomposition is to recursively use low and high pass filtering operations followed by downsampling at 0.5 the sampling frequency of the input signal. An unbalanced binary tree of low and high pass filters can be used [8]. Energy is computed at the output of each filter by summing the absolute value of N samples.
In [10] it is proposed to compute a Wavelet Packet Transform (WPT) with a complete tree of filters. 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 appear to be relatively immune to local changes within a particular band with significant advantages in speaker-dependent isolated phoneme recognition pronounced by male speakers under stress.

In [9] it is shown that, for plosive recognition, DWT coefficients lead to a much lower error rate than MFCCs.

This paper compares, in section 4, the performance of an operational ASR system when MFCCs, J-RASTA Perceptual Linear Prediction Coefficients (J-Rasta PLP) [6] and energies obtained with WPT are used as input features to a hybrid system consisting of a Neural Network (NN) which provides observation probabilities for a network of Hidden Markov Models (HMM). Furthermore, it compares the performance of the system when various combinations of these features are used showing a WER reduction of 16% w.r.t. the use of J-Rasta PLP coefficients, when J-Rasta PLP coefficients are combined with the energies computed at the output of the leaves of a tree of filters performing WPT. Such a combination is practically feasible thanks to the NN architecture used in the system.

Recognition is performed without any language model on a very large test set including many speakers uttering proper names from different locations of the Italian public telephone network.

Section 2 provides some background and section 3 describes the WPT front-end.

2. BACKGROUND

A function g(t) can be expanded as follows:

$$g(t) = \sum_k c_{j_0}(k)\varphi_{j_0,k}(t) + \sum_k \sum_{j=j_0}^{\infty} d_j(k)\psi_{j,k}(t) \qquad (1)$$

with a set of functions $\{ \varphi(j_0, k), \psi(j, k) \}$ such that:

$$\varphi_{j,k}(t) = 2^{\frac{j}{2}} \varphi(2^j t - k)$$

where k is a *translation* index and j is a *scaling* index.The function φ(t) must satisfy the following property:

$$\varphi(t) = \sqrt{2}\sum_n h(n)\varphi(2t - n) \qquad (2)$$

φ (t) is called *scaling function* while ψ(t) is called *mother wavelet*. h(n) represents the sequence of samples of the impulse response of a suitable filter.

The *wavelet functions* are defined starting from the following function, called *mother wavelet*:

$$\psi(t) = \sqrt{2}\sum_n h_1(n)\varphi(2t - n) \qquad (3)$$

where the collection $\{ h_1(n) \}$ can be expressed as:

$$h_1(n) = (-1)^n h(1 - n)$$

which, for a finite even length N becomes:

$$h_1(n) = (-1)^n h(N - 1 - n)$$

In a similar manner as for $\varphi_{j,k}(t)$, the following functions

$$\psi_{j,k}(t) = 2^{\frac{j}{2}} \psi(2^j t - k)$$

are expressed as the dilatated, normalized and translated versions of $\psi(t)$.

The coefficients $c_j(k)$ and $d_j(k)$ of the (1) are defined as follows:

$$c_j(k) = \int g(t)\varphi_{j,k}(t)dt$$

$$d_j(k) = \int g(t)\psi_{j,k}(t)dt$$

When the scaling function is well behaved, at high scales it is similar to a Dirac function and the inner product simply samples the function.

Two approximations of g(t) are usually performed in wavelet series expansions, namely truncation w.r.t. j, and with thresholding, which implies eliminating the coefficients whose absolute value is less than a threshold.

The coefficients of the expansion can be computed as follows:

$$c_j(k) = \sum_m h_l(m - 2k)c_{j+1}(m)$$

$$d_j(k) = \sum_m h_h(m - 2k)c_{j+1}(m)$$

In practice, $h_l(n)$ and $h_h(n)$ are the impulse responses of a low pass and a high pass digital filter. Multiplying by two the sampling period k is equivalent to down-sampling. For high values of j, coefficients $c_j(m)$ are obtained by multiplying the signal by a Dirac function, and are, in practice, signal samples.

The DWT is thus performed by expanding a full binary tree (*Wavelet Packet* WP) in which the operators for node expansion are filtering and down-sampling. WP implements MRA. The root node contains the samples of the signal to be analyzed. The tree expansion is performed by applying the low pass filter when descending left, and the high pass filter when descending right. Each node covers a different range of frequencies and contains the signal filtered and downsampled along the path reaching the node. The nodes at different levels have a different time/frequency resolution, with time resolution decreasing and frequency resolution increasing when going from the root to the leaves.

From this tree, it is possible to extract several *basis*, that are sets of functions covering all ranges of frequencies. A possible base is the *wavelet base*, obtained from the DWT by expanding the nodes only leftmost; another is the *(near) Mel* one, obtained by choosing the nodes of the full WP in order to approximate the Mel-scale frequency division [7]. This set of functions, indicated in the following as *Mel base*, has been used, together with the full WP, in the experiments described in this paper.

Other sets of functions can be chosen using some optimization criteria as discussed in [12].

3. WAVELETS BASED FRONT-END

A binary tree of filters is used to implement a WP in order to achieve MRA under the constraints imposed by the Heisenberg principle which states that the product of time and frequency resolution is always greater than a constant.

A segment of the original signal is propagated through a binary tree of a predefined depth. At each node, filtering and downsampling are performed.

In practice, signal windowing is not necessary as processing is done by a continuous application of the tree of filtering-downsampling operators on the input signal.

The filters used in the tree are half-band (low-pass and high pass) perfect-reconstruction wavelet filters. While for signal coding the only requested feature is perfect reconstruction, in the case of recognition, only the analysis phase is performed. Then MRA features are extracted and employed for acoustic matching. As a consequence, filters should exhibit some properties in order to obtain MRA features suitable for recognition. Some desirable properties are:

1. **No loss of information:** *perfect reconstruction filters* should be orthogonal or biorthogonal wavelet filters.
2. **Conservation of total energy:** *orthogonal filters* should be used because, by definition, they preserve the total energy, while this is not the case for biorthogonal filters. This property is fundamental as MRA features are subband energies.
3. **No spectral distorsion:** *linear phase filters* should be used in order to avoid signal distortion due to different frequency shiftings. This is in contrast with point 2) as orthogonal filters cannot, by definition, have linear phase. Nevertheless some classes of orthogonal filters exhibits a small deviation from linear phase.
4. **Accurate subband partitioning:** *sharp half-band filters* should be used with a small transition band, to avoid mixing energy contributions between different bands.

The choice of filters should satisfy, as much as possible, the above requirements as well as some other, more practical ones, like computational efficiency. The filters used for the experiments reported in this paper are *Orthogonal IIR Wavelet filters* as proposed by Selesnick in [11]. The filters order is 19.

A frame synchronous, variable resolution energy computation is performed at each node k of the tree on the same number of samples N :

$$E_k = \frac{1}{N}\sum_{i=1}^{N}[x_k(i)]^2 .$$

As the time resolution halves at each downsampling (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 subbands 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 6^{th} 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).

The energies at all the WP nodes could be used as features. This is feasible at the expenses of a bigger NN. Results using these features are indicated as *MRA WP*. Other possibilities consist in finding a set of *best basis* [12], or to choose, as proposed in [7] and [8] a set of filters that emulates the MEL scale. The results obtained with these features will be indicated as *MEL base*.

4. EXPERIMENTAL SETUP AND RESULTS

Experiments were conducted using a hybrid HMM-NN system described in [3] with a feed-forward Neural Network (NN) which computes the probability of being in a state of a Hidden Markov Model (HMM), given the observation made on a set of input frames.

The network is designed to integrate multiple features, exploiting the NN capability of mixing several input parameters without any assumption about their stochastical independence. The input window is 7 frames wide, and each frame contains the set of features extracted by the front-end along with their first and second time derivatives. The first hidden layer is divided into three feature detector blocks, one for the central frame, and two for the left and right contexts.

Each block is in its turn divided into sub-blocks to keep into account the different types of input parameters. It was empirically found that this a priori structure is generally better than a fully connected layer. The second hidden layer is fully connected with the output layer that estimates the emission probabilities associated with the HMM states. Further details on the NN architecture can be found in [4].

In the experiments described in this paper, the features used as input of the HMM-NN model are MFCC, J-Rasta PLP and WP derived energies (MEL base and whole WP).

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 1136 speakers uttering a total of 4875 phonetically balanced sentences with a vocabulary of 3653 words.

The test corpus consists of 14473 isolated word utterances of proper names, from 1050 speakers.

There was no overlap in the speakers of the train and the test corpora. All the test corpus was made of proper names which did not appear in the training set. The test set is made of isolated words belonging to a vocabulary containing the 475 most common Italian city names.

Two kinds of experiments were performed. The first one is aimed at comparing the results obtained with two classical front-ends (MFCC, J-Rasta PLP) with features obtained with MRA analysis. No language model was used for the experiments as the purpose was that of comparing acoustic features.

In the case of MFCC and J-Rasta PLP, 12 MFCCs and the total signal energy are used with their first and second time derivatives for a total of 39 parameters for each 10 ms frame. In the case of MRA, the energies of the WP are directly used. There are 18 energy samples in the MEL base, and 63 in the complete WP. First and second time derivatives of these features are also used.

The results in terms of Word Error Rates (WER) are summarized in Table 1. Feature vectors are computed every 10 msecs. and include the parameters indicated in the table plus their first and second time derivatives.

Basic features	WER
MFCC	5.31
J-Rasta PLP	4.69
MRA MEL base	5.24
MRA WP	4.58

Table 1. Comparison of individual sets of features.

The results of Table 1 show that J-Rasta PLP parameters outperform MFCCs, while the MRA MEL base show performance almost equivalent to MFCCs, and MRA WP show performance almost equivalent to J-Rasta PLP at the expense of a larger number of parameters. It is important to notice that no preprocessing like spectral subtraction, cepstral mean normalization or vocal tract length normalization were performed in order to ensure a fair comparison.
Nevertheless, J-Rasta PLP coefficients are computed using a number of perceptual and practical findings like Rasta filtering and perceptual compression, while suitable processing enhancements have not yet been attempted for MRA.
The results confirm that there is no practical advantage in replacing popular front ends with MRA front-ends. Nevertheless, the different types of analysis allow one to perform different types of pre or post processing and performance differences may appear with suitable additional feature processing.
A possibility, worth to be investigated, is whether or not the integration of different types of features provide any WER reduction or, in other words, if different types of features have different information contents. For this purpose, experiments were conducted on the integration of two kinds of features. Integration, described in [4], exploits the capability of NNs to mix different input sources without limiting constraints.

Basic features	Additional feat.	WER
J-Rasta PLP	-	4.69
J-Rasta PLP	MFCC	4.1
J-Rasta PLP	MRA MEL base	3.96

Table 2. Recognition results for the integration of single and multiple resolution analysis.

The results reported in Table 2 show that the synergy of fixed and multiple-resolution analysis leads to a significant improvement (16% WER reduction) w.r.t. the situation in which the most effective features (J-Rasta PLP) are used alone. Given the very large test corpus, the differences appearing in Table 2 are statistically significant.
The results in Table 2 have been obtained with an effort which is still in progress and indicate two main lines for future work. One consists in finding an optimal subset or transformation of the combined features which produces almost the same or better results. The other consists in exploring new features or functions which can be computed from MRA.

5 CONCLUSIONS

Now that reference systems have been established, some important conclusions about the performance of different types of features and their combinations can be formulated.
MRA does not appear to be superior to J-Rasta PLP, but whole WP is superior to MFCCs. Notice that no significant improvements were observed by increasing the numbers of MFCCs.
MRA Mel base provides additional information to the one carried by J-Rasta PLP features. This encourages continuing investigations on parameters which can be obtained from WP, such as ratios of energies for the children of the same node, mutual information of filter outputs, patterns described by trajectories of filter outputs or parameters derived from them, suitable preprocessing for MRA energies.

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 programme of the Human Language Technology Division of the European Community. The authors would like to thank Franco Mana, Donata Bonino and Paolo Pegoraro for their important contribution to the research.

REFERENCES

[1] J.B. Allen “How do humans process and recognize speech”. IEEE Transactions on Audio and Speech Processing, SAP-2(4):567-577. 1994
[2] E. Erzin, A.E. Cetin and Y. Yardimci, “Subband Analysis for robust Speech Recognition in the Presence of Car Noise”, Proc. of ICASSP 1995, pp.417-420.
[3] R. Gemello, D. Albesano, F. Mana, “Multi-source Neural Networks for Speech Recognition”, in *Proc. of International Joint Conference on Neural Networks (IJCNN’99)*, Washington, July 1999.
[4] R. Gemello, D. Albesano, F. Mana, L. Moisa, “Multi-source Neural Networks for Speech Recognition: a Review of Recent Results”, in *Proc. of International Joint Conference on Neural Networks (IJCNN-2000)*, Como, Italy, July 2000.
[5] J.N. Gowedy and Z. Tufekcil, “Mel-scaled discrete wavelet coefficients for speech recognition”, *Proc. International Conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey.
[6] Hermansky H. and Morgan N. (1994), “RASTA Processing of Speech”, *IEEE Transactions on Speech and Audio Processing*, Vol. 2, n° 4, pp. 578-589, october 1994.
[7] F. Jabloun and A. Enis Cetin, “The Teager Energy based feature parameters for robust speech recognition in car noise”, *IEEE Int. Conf. On Acoustics, Speech and Signal Processing*, Phoenix, AZ, 1999
[8] D. Kryze, L. Rigazio, T. Appelbaum and J.C. Junqua, “A new noise robust subband front-end and its comparison to PLP”, *Proc. IEEE ASRU Workshop*, Keystone, Colorado, (1999)
[9] E. Lukasik, “Wavelet packet based features selection for voiceless classification”, *Proc. International Conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey. (2000)
[10] R. Sarikaya and J.H.J. Hansen, “High resolution speech feature parameterization for monophone based stressed speech recognition”, *IEEE Signal Processing Letters*, 7(7):182-185. 2000
[11] I. W. Selesnick, Fomulas for IIR Wavelet Filters, *IEEE Trans. on Signal Processing*, 46(4):1138-1141, April 1998.
[12] M. V. Wickerhauser, “Adapted Wavelet Analysis from Theory to Software”. Boston: A.K. Peters, 1994.