Robust Multiple Resolution Analysis For Automatic Speech Recognition

Roberto Gemello[*], *Franco Mana*[*], *Dario Albesano*[*] *and Renato De Mori*[§]

[*] LOQUENDO
roberto.gemello, franco.mana, dario.albesano@loquendo.com

[§] LIA CNRS
University of Avignon
renato.demori@lia.univ-avignon.fr

Abstract

This paper investigates the potential of exploiting the redundancy implicit in Multi Resolution Analysis (MRA) for Automatic Speech Recognition (ASR) systems. Experiments, carried with data collected from home telephones and in cars, confirm the proposed approach for exploiting this redundancy.

Comparisons with the use of Mel Frequency-scaled Cepstral Coefficients (MFCC)s, JRASTA Perceptual Linear Prediction Coefficients (JRASTAPLP) indicate that executing Principal Component Analysis (PCA) on MRA features result in performance superior to the use of MFCCs and competitive with the use of JRASTAPLP features.

Experiments in noisy conditions, using the Italian component of the AURORA3 corpus, show a WER reduction of 15.7% when SNR-dependent Spectral Subtraction (SS) is performed on MRA-PCA features compared to when it is performed on JRASTAPLP features. Furthermore, SS appears to be better than Soft Thresholding (ST).

1. Introduction

Many recent research efforts in speech analysis are inspired by the assumption that components of the information conveyed by a speech signal can be distinguished because energies are localized in the time-frequency (TF) plan. In [1] a method is presented to perform frameless analysis by considering amplitude and frequency modulation. In [2] it is proposed to decompose a noiseless signal by partitioning the TF plane wirh a set of radial basis functions, in such a way that each segment contains a single component. Parameters can be estimated using the maximum likelihood criterion. In [3] Gabor filters are introduced to model the characteristics of neurons in the auditory system as they do in the visual system. There is evidence that in primary auditory cortex each individual neuron is tuned to a specific combination of spectral and temporal modulation frequencies. A number of Gabor functions are used as two-dimensional filters. Other approaches use relative spectral (RASTA) features [4] and temporal patterns (TRAP) features [5] to represent dynamic properties of the evolution of spectral energies in perceptually significant frequency bands.

As there is experimental evidence that human perception has a sensitivity which is variable with frequency, TF patterns obtained with Multi Resolution Analysis (MRA) have been widely investigated and interesting applications have been found in speech coding and, more recently ([6], [7]), for Automatic Speech Recognition (ASR). In [8] it is shown that, for plosive recognition, DWT coefficients lead to a much lower error rate than Mel Frequency-scaled Cepstral Coefficients (MFCC). In these approaches, Discrete Wavelet Transform (DWT) is obtained from MRA which can be obtained by a Wavelet Packet Transform (WP) implemented by a tree of filters.

These transforms provide 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 other motivations for using Discrete Wavelet Transforms (DWT) as opposed to all MFCCs can be found in [7].

Unfortunately, when MRA or DWT are performed, 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.

Even if interesting solutions with DWT have been proposed for speech analysis and coding, attempts to improve ASR performance produced less impressive results.

This paper investigates the potential of exploiting the redundancy implicit in MRA for ASR systems and describes an experimental evaluation of the use of MRA for ASR in noisy conditions.

ASR performance is evaluated with an operational system described in [9]. This is a hybrid system consisting of an Artificial Neural Network (ANN) which provides observation probabilities for Hidden Markov Models (HMM). Experiments

with data collected in cars indicate what type of MRA features and transformations on them are competitive with other features and what type of denoising techniques make them particularly attractive for ASR.

Experiments show that, when dimensionality reduction is performed with Principal Component Analysis (PCA), performance is comparable to that of JRASTAPLPs and significantly better than that of MFCCs, making MRA followed by PCA an attractive solution for ASR front-ends. This aspect is briefly discussed in Section 2.

There are two family of methods for speech enhancement, namely time-domain and spectral-domain methods. In the case of MRA, an example of the first type of methods is soft thresholding (ST) and an example of the second method is spectral subtraction (SS). Both methods do not require any model parameter training. It is possible to apply the same type of SS to JRASTAPLP features and to MRA energies before applying PCA. In the first case, SS is applied to spectral samples, from which energies are computed. Errors due to the limits of SS can be averaged by energy computation. In the case of MRA, energies are computed directly with signal samples and SS is applied on the computed energies. Nevertheless, errors affecting these energies can be attenuated by PCA. Errors due to SS are reduced if the SS parameters are made dependent from the Signal-to-Noise Ratio (SNR). Section 3 shows that when SNR-dependent SS is performed on MRA features, it provides much better results than SNR-dependent SS on JRASTAPLP features.

Experiments in noisy conditions have been performed using the Italian component of the AURORA3 corpus. A WER reduction of 15.7% is observed when SNR-dependent SS is performed on MRA features compared to when it is performed on JRASTAPLP features. Furthermore, SS appears to be better than ST.

2. Dimensionality reduction

In order to reduce dimensionality and increase robustness, PCA and Linear Discriminant Analysis (LDA) have been performed on the whole set of WP features at each time frame with a 10 ms frame rate. As that node 32 of WP, corresponding to the 0-125 Hz band does not carry any useful information on telephone data, it was neglected.

PCA has been performed after transforming the 63 WP features in order to have zero mean and unit variance. The covariance matrix C has been obtained with the transformed data. High correlations have been observed between nodes and their fathers in accordance with the theory. LDA has been performed in similar conditions using the 686 NN to identify the classes.

The elements of the eigenvectors of C show a repetitive pattern due to the presence of the same kind of spectral information at increasing levels of details at different distances from the root of the MRA tree.

The use of the first 20 coefficients obtained with PCA after normalization and ignoring node 32 of the WP, show a performance slightly better than J-RASTA PLP while both outperform MFCCs . With PCA, 20 basic features instead of 12 are used with energy, first and second time derivatives. 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. This set of features appears to be competitive with JRASTAPLP.

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

3. Denoising

When deploying Automatic Speech Recognition (ASR) applications, a set of well-trained Acoustic Models (AM) is available. A development set is usually collected for a new application that is used for performing acoustic feature transformations or model parameter adaptation. Transformation and adaptation parameters depend on the development set used for obtaining them. Another possibility, which can be used in conjunction with feature transformation and model parameter adaptation, is the application of general purpose denoising algorithms which do not require any parameter estimation with a set of data.

Denoising algorithms can be conceived in the time domain and act to signal samples, or in the frequency domain or after other types of acoustic data transformation. Two denoising methods have been investigated, Spectral Subtraction (SS) which operates in the frequency domain and attempts to compute a denoised version of the WP node energies and Soft Thresholding (ST) which acts on the samples at the output of the WP filters.

3.1. Spectral Subtraction

Various Spectral Subtraction (SS) schemes proposed in the literature were implemented and evaluated on a corpus different from the one used for the tests described later. They all attempt to estimate $\hat{S}(n,f)$, the de-noised speech spectrum amplitude at time-frame n and at frequency f, from the correspondent spectrum Y(n,f) of the noisy signal, as follows:

$$\hat{S}(n,f)=\begin{cases} Y(n,f)-\alpha(n)\hat{N}(n,f) & \text{if } Y(n,f)-\alpha(n)\hat{N}(n,f)>\beta(n)Y(n,f) \\ \beta(n)Y(n,f) & \text{otherwise} \end{cases} \quad (1)$$

$\hat{N}(n,f)$ is the sample at time-frame and frequency f of an estimate of the noise spectrum, $\alpha(n)$ is called *overestimation* and $\beta(n)$ is called *flooring*.

A popular variant of spectral subtraction (1) is Wiener spectral subtraction, obtained by looking for the linear filter that minimizes the mean square error in the time domain:

$$\hat{S}(n,f)=\begin{cases} \dfrac{\left[Y(n,f)-\alpha(n)\hat{N}(n,f)\right]^2}{Y(n,f)} & \text{if } Y(n,f)-\alpha(n)\hat{N}(n,f)>\beta(n)Y(n,f) \\ \beta(n)Y(n,f) & \text{otherwise} \end{cases} \quad (2)$$

A special case of the (1) and (2) assumes that flooring and overestimation are constant in time. Poor results were observed with this assumption, which was abandoned.

The best results were obtained with Wiener SS in which flooring and overestimation parameters are dependent on the estimated local Signal-to-Noise Ratio (SNR), as suggested in [10], and are defined as piecewise linear functions (fig.1):

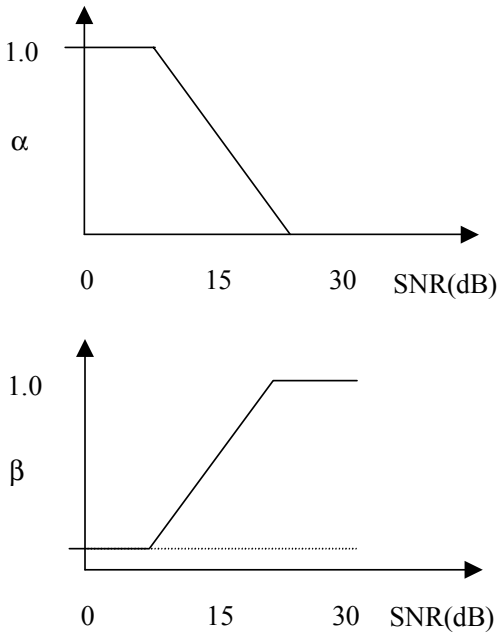

Figure 1: Example of linear-type functions defining α *(noise overestimation factor),* and β *(spectral floor)* depending on the estimated local SNR.

The local SNR is estimated using the noise estimate $\hat{N}(n,f)$ and the noisy signal spectrum amplitude Y(n,f) as follows:

$$SNR(n) = 10\log_{10}\left(\frac{\sum_{f} Y^2(n,f)}{\sum_{f} \hat{N}^2(n,f)}\right) \qquad (3)$$

This method implements an adaptive switch in which SS is practically not performed when SNR is high.

An estimate of the noise spectrum amplitude is obtained using a modified first order recursion [11] in conjunction with an energy based Voice Activity Detector (VAD) as follows:

$$\hat{N}(n,f) = \begin{cases} \gamma\hat{N}(n-1,f) + (1-\gamma)Y(n,f) & if \left\{\left\{\left|\dfrac{Y(n,f)}{\hat{N}(n,f)}\right| \le k\sigma(n)\right\} \wedge \{VAD = false\}\right\} \\ \hat{N}(n-1,f) & otherwise \end{cases} \qquad (4)$$

where Y(n,f) is the power spectrum of the signal, γ and *k* control the update speed and the allowed dynamics of noise; σ(n) is the noise standard deviation, estimated as:

$$\sigma^2(n) = \gamma\sigma^2(n-1) + (1-\gamma)\left(Y(n,f) - \hat{N}(n,f)\right)^2$$

The values employed for γ and *k* are γ=0.9 and *k*=4.0

The noise spectrum amplitude is initialized with an average of the spectra of first frames and is updated with a first order recursion only when then speech is most likely to be absent; this condition takes place when the absolute value of the difference between the noisy signal and the estimated noise $\left|Y(n,f) - \alpha(n)\hat{N}(n,f)\right|$ is below k times the estimated noise standard deviation σ(n) and a standard energy based VAD does not detects speech.

SS introduces errors in the estimation of a sample of the noise spectrum. Let e(n,f) such an error and N(n,f) be the exact noise sample, then :

$$\hat{S}(n,f) = Y(n,f) - \{N(n,f) + e(n,f)\}$$

If features are energies computed from a Fourier Transform, then spectral estimation can be performed on each spectral sample amplitude before using the cleaned samples to compute the energy in a frequency band. In such a case, one gets:

$$E_j(n) = \sum_{f_1}^{f_2} \hat{S}^2(n,f) = \sum_{f_1}^{f_2} \{Y(n,f) - N(n,f) - e(n,f)\}^2$$

In the case of WP, energies are computed by summing the squares of signal samples. The estimate of the signal energy at the ν-th node and in the time frame n is given by:

$$E_\nu(n) = \left\{\sum_{t_1}^{t_2} y_\nu^2(nT+t)\right\} - \{N_\nu(n) + e_\nu(n)\}^2$$

The way of estimating signal energies are different and filters are different, thus a comparison between the two approaches can be done only experimentally. Nevertheless, as it is well known that PCA is beneficial for denoising, it is reasonable to expect a great reduction of the residual noise due to errors in a reduction from 63 dimensions to 20 especially because there is a large redundancy in the 63 original dimensions.

3.2 Soft Thresholding

Soft thresholding was proposed in [12] and is effective for speech enhancement whenever few samples of a filter output contribute to the noiseless signal. With this technique, a denoised sample at the output of a WP node is computed as follows:

$$|s_\nu(k)| = \begin{cases} |y_\nu(k)| - \alpha\sigma_\nu^2(n) & \text{if } |y_\nu(k)| > \alpha\sigma_\nu^2(n) \\ 0 & \text{otherwise} \end{cases}$$

where $y_\nu(k)$ is the t-th time sample at the output of node ν, α is a constant set equal to 1 in our experiments and σ_ν^2 is an estimation of the power of additive noise affecting node ν assuming that it is not correlated with the signal. For the experiments described below, noise has been dynamically estimated at each WP node using the (4) as outlined in subsection 3.1.

Test condition	*WER CH0 (%)*	*WER CH1 (%)*	*overall WER (%)*
JRASTAPLP	1.3	41.0	20.1
JRASTAPLP + SS	0.8	24.0	12.1
MRA + PCA	0.9	38.7	19.3
MRA + PCA + SS	**0.7**	**20.2**	**10.2**
MRA + PCA + ST	0.8	23.7	11.9

Table 4: Experimental results with various denoising methods and acoustic features

3.3 Experiments

A suitable corpus for denoising experiments is SpeechDat.Car of the *Aurora 3* database. The Italian component of the *Aurora3* was thus used. It contains a set of close-talking utterances indicated as CH0 and a set of hand-free utterances, indicated as CH1. Utterances of CH0 are nearly clean, as the close-talking microphone collects little environmental noise, while utterances of CH1 are quite noisy as the hand-free microphone gathers a lot of car noise. The train corpus is made of 1466 utterances in CH0 and 1485 utterances in CH1. The test component comprises 664 utterances in CH0 and 645 utterances in CH1. The training set was not used.

Table 4 summarizes the results obtained with the test set, when SS and ST are applied to WP and when SS is applied to JRASTAPLP.

Experiments clearly show that the application of SS to the nodes of WP after dimensionality reduction is more effective than when it is applied to the energies of the filters used for JRASTAPLP.

4. Conclusions

Experiments in noisy conditions, show a WER reduction of 15.7% when SNR-dependent SS is performed on MRA-PCA features compared to when it is performed on JRASTAPLP features. Furthermore, SS appears to be better than Soft Thresholding which still slightly outperforms denoising with JRASTAPLP features.

Further denoising can then be obtained by transforming the speech spectra estimations in order to adapt training and test conditions. This possibility was not investigated because the ASR system used here was trained on a very large corpus of phonetically balanced sentences and was conceived to be used for telephone applications in a domain and environment independent way.

5. Acknowledgements

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.

6. References

[1] A.C. Bovik, P. Maragos and T. Quatieri, “AM-FM energy detection and separation in noise using multiband energy operators”, *IEEE Transactions on Signal Processing*, 41:3245-3265, 1993.

[2] L.Coates and W. Fitzgerald, “Time-frequency signal decomposition using energy mixture models”, Proc. *IEEE Intl. Conference on Acoustics, Speech and Signal Processing*, Istanbul,Turkey, 2000.

[3] M Kleinschmidt and D. Gelbart, “Improving word accuracy with Gabor feature extraction”, Proc. *International Conference on Spoken Language Processing*, Denver, CO, pp. 25-28, 2002

[4] H. Hermansky H. and N. Morgan “RASTA Processing of Speech”, *IEEE Transactions on Speech and Audio Processing*, Vol. 2, n° 4, pp. 578-589, 1994.

[5] H. Yang, S. van Vuuren and H. Hermansky, “Relevancy of time-frequency features for phonetic classification measured by mutual information.” *IEEE Intntl Conf. On Acoustics, Speech and Signal Processing*, Phoenix, AZ, 1999.

[6] 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.

[7] J.N. Gowedy and Z. Tufekci, “Mel-scaled discrete wavelet coefficients for speech recognition.” Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, 2000.

[8] E. Lukasik, “Wavelet packet based features selection for voiceless classification.” Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, 2000.

[9] R. Gemello, D. Albesano, F. Mana, ”Multi-source neural networks for speech recognition” Proc. *International Joint Conference on Neural Networks* (IJCNN 1999), Washington,D.C, 2000

[10] V. Schless, F. Class, “SNR-Dependent flooring and noise overestimation for joint application of spectral subtraction and model combination”, Proc. *Intl. Conference on Spoken Language Processing* , Sydney, Australia, 1998.

[11] H. G. Hirsch, C. Ehrlicher, “Noise estimation techniques for robust speech recognition” Proc. *IEEE Intl. Conference on Acoustics, Speech and Signal Processing*, Detroit, MI, pp.153-156, 1995.

[12] D. L. Donoho, ”Denoising by soft tresholding” *IEEE Transactions on Information Theory*, IT-41(5):613-627, 1995.