COMBINING LINEAR AND NON-LINEAR PROCESSING IN THE TIME AND IN THE SPECTRAL DOMAIN FOR NON-STATIONARY NOISE FILTERING

R.Martínez, A. Álvarez, V. Nieto, V. Rodellar, P. Gómez
Depto. de Arquitectura y Tecnología de Sistemas Informáticos, Facultad de Informática
Universidad Politécnica de Madrid, Campus de Montegancedo, s/n
Boadilla del Monte, 28660, Madrid, SPAIN
e-mail: pedro@pino.datsi.fi.upm.es

1 ABSTRACT

Through this paper, the combination of linear and non-linear techniques for noise filtering is proposed. A first linear processing stage in the time domain (an adaptive lattice ladder filter)[2],[5] is followed by a non-linear processing block in the frequency domain. The association of both procedures provides a much higher level of cancellation than their individual application. An added result is that the application of the frequency domain non-linear processing makes it possible to use very short adaptive filters, with the resulting savings in computational power requirements.

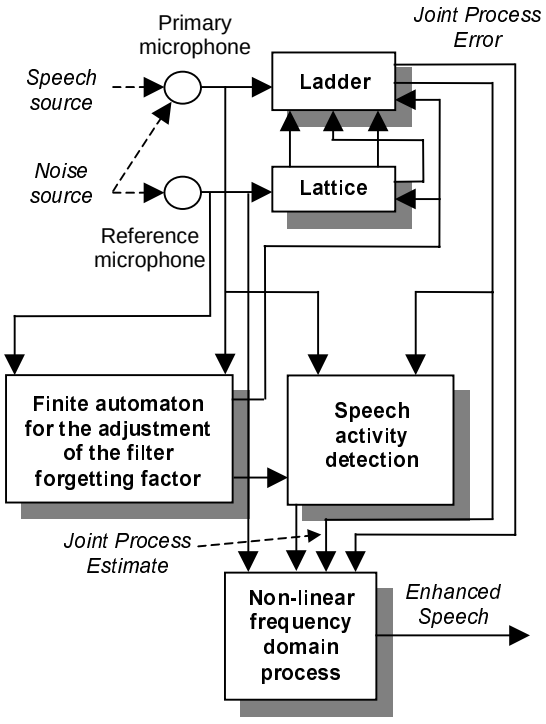

Figure 1: General framework for the proposed methodology: A lattice-ladder filter is combined with a non-linear filter in the spectral domain

The cancellation scheme proposed, which may be seen in Fig. 1, is based in a two-microphone array (Speech Source or *primary*, and Noise Source or *reference*) [3]. One of the microphones (primary) will be placed close to the speaker, and the second one, at a certain distance to acquire a good estimation of the noise received by the primary microphone and avoid the recording of speech at the same time.

2 TIME DOMAIN FILTER

Adaptive processing is a well-known technique for removing non desired signals from a given one [6]. The particular filter selected for the application is a *time-domain joint-process estimator* implemented as a *time-domain lattice-ladder filter* with a least-squares estimation algorithm (Recursive Least Squares Lattice using *a posteriori* estimation errors)[2],[5]. This filter has two inputs: the lattice part is fed with the noise source (reference), and the primary signal is inserted through the ladder part. The lattice is therefore adapted with the noise characteristics and a set of backward prediction errors is obtained. This group of backward prediction errors constitute an orthogonal base which is used to build an estimator of the noise. The so obtained estimation is compared with the signal of the ladder part in order to detect the orthogonal components of the signal that are shared in common. The addition of these components form the joint process estimate (JPEs), and the consecutive subtractions of the joint process estimate orthogonal components from the primary input give the Joint Process Error (JPEr). This JPEr contains the signal which is not in common (the speech) plus the orthogonal components of noise which are not still estimated. An important property shown by the lattice filter is that the influence of its stages decrease with the stage order (the orthogonal basis points to the minimum error in a least squares sense, so the first elements, which correspond to the first stages of the filter, are more important than the last ones). Of course, the longer the filter, the more exact the estimation carried out, but as it will be explained, some undesired effects appear. As a conclusion, it may be assured that it is possible to obtain a good joint process estimation (and consequently a good

filtering) with very short filters. In addition, it is possible to use the joint process estimate output to build an approximation of the noise still present in the joint process error output

In a first approach, a fixed value of α=0.9999 was chosen for the *forgetting factor* of the filter, but after several tests it was found that the filter, which showed a very good behaviour with low SNR, became unstable with very loud and sudden word utterances. A logic control has been added to modify the forgetting factor accordingly with the input signals, in order to obtain shorter locking periods, and to avoid instabilities after those sharp and high-energy differences between channels [4]. The main advantage of this method is that no matter the noise level, if the reference signal is good enough, a considerable amount of cancellation is achieved. Nevertheless one of its major limitations is the computational complexity. As explained before, the longer the filter, the higher the cancellation gain, but the number of operations grows accordingly. So we find that although long filters in the time domain would achieve higher cancellations than short ones, that is a limited solution due to the computational costs. Long filters show also long locking periods which is not a desirable effect (especially when the environment noise is highly non stationary). Another fact to be considered is that the longer the filter order the higher the possibility of instabilities to show up. For all these reasons, the length of the filters is limited and consequently the amount of cancellation achieved using this method. There is also an additional limitation in the amount of cancellation obtained by adaptive processing. This is due to the existing non-linearities between the noise present in the signal and the reference produced by acoustical factors (unbalanced response of both acquisition channels, room reverberation, etc).

3 SPECTRAL DOMAIN FILTER

The frequency-domain filter here proposed is very efficient, and produces a level of cancellation of the same order, or higher than the adaptive filtering. Its major limitation is that it requires a sufficient SNR (which is not always possible). Otherwise, two effects will appear: firstly, spectral lines completely buried into noise (local negative SNR) will be removed, and secondly, it will be very difficult to make a decision on the presence of speech (especially if we consider that the noise in itself can contain speech sounds). The determination of non-speech periods is imperative to make an accurate comparison of both channels as this has to be carried out when speech is not present. The previous processing block offers an enhanced version of the signal and a first estimation for speech detection. By combining both techniques, we obtain a much higher level of cancellation with a reasonable computational complexity, and an immediate response under non-stationary environments. The objective of this filter is twofold:

- It has to eliminate the noise linearly related with the reference still present in the enhanced speech. This particular component of noise has not been cancelled due to the short length of the time filter.
- This filter tries to equalise the non-linear relationship between channels.

The effects produced by non-linearities can be clearly perceived as frequency lines, which are not cancelled at all. Although those spectral components are not removed, their time variations resemble the time variations of the spectral components contained in the JPEs output (although at a different power level). The objective of the processing is to calculate the ratio between every spectral line of the JPEr and the JPEs, and modify the JPEr accordingly to produce an estimate of the noise. Fig. 2 shows the general framework of this spectral processing .

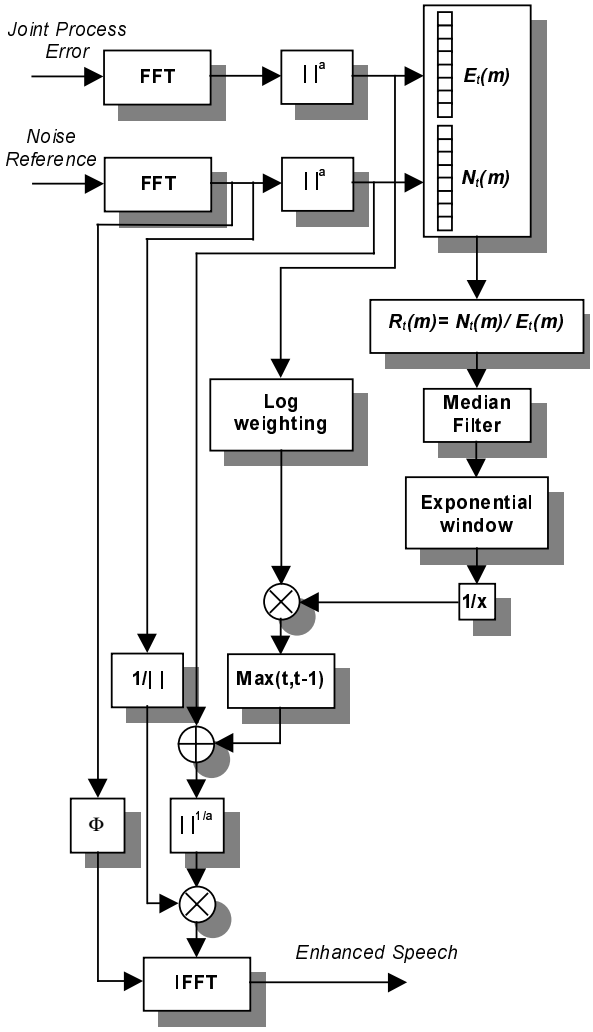

Figure 2. Detailed sketch of the non-linear frequency-domain process.

To implement this filtering, the JPEr output of the lattice-ladder filter *e(n)* is used as the primary signal, and a noise estimate *n(n)* obtained from the JPEs output is used as the reference. This two signals are segmented in overlapped windows and transformed into the frequency domain using the short-time Discrete Fourier transform.

$$E(m) = TF\{e(n)w(n)\} \tag{1}$$

$$N(m) = TF\{n(n)w(n)\} \tag{2}$$

Where *w(n)* is a windowing function, and with $0 \le m \le N/2-1$, where N is the length of the window.

The length and function of the window and the degree of overlapping are important factors to minimise the presence of artificial tones after subtraction. We have observed that these tones are distinguishable with *23 msec.* windows (*256* samples for a sampling frequency f_s = *11 KHz*) and an overlap of *50%*. With longer windows and higher overlapping levels this tones almost disappear. Nevertheless the length of the window is limited by the quasi-stationarity hypothesis. With *46 msec.* windows (*512* samples for a sampling frequency f_s = *11 KHz*) and *75%* of overlapping the effect is not perceived anymore.

These longer windows imply a higher spectral resolution. Therefore it is possible to use window functions with wide spectral main lobes, but lower secondary lobes (the range of frequencies affected by the main lobes decrease with the increase in the length of the window). These windows have smoother temporal transitions, which reduce the edge effects when adding consecutive windows to recover the signal in the time domain after subtracting.

As it has been commented before, the proportion of overlapping has also influence in the reduction of artificial tones. High overlapping reduce the errors in the estimation of noise, as they are averaged by the addition of overlapped consecutive windows when going back into the time domain, and these errors are supposed to be un-biased.

From this point of view, every frequency present in the input signals of the frequency domain filter (*E(m)* and *N(m)*) is considered a separate channel, so we have *N/2* channels. The input rate of the elements of these channels is:

$$Ir = \frac{f_S}{N}\frac{100}{100-v} \tag{3}$$

Where v is the percentage of overlapping.

The next step is to calculate the relationship between the power spectrum of the primary and the reference signals for every frequency channel:

$$R_n(m) = \frac{\|N_n(m)\|^a}{\|E_n(m)\|^a} \tag{4}$$

where the sub-index n shows the time variation of these magnitudes.

The relation $R_n(m)$ measures the instantaneous power ratios between both channels. They have to be calculated in periods when speech is not present, therefore a speech activity detector is required.

There is a slight variation in the values computed for consecutive windows. Nevertheless, window effects, and the random character of noise may produce occasional peaks in those ratios, specially when the noise remaining in the principal input *e(n)* is low, as it implies larger relative errors. These peaks have to be removed as they do not correspond to real relations between $N_n(m)$ and $E_n(m)$. For doing this, $R_n(m)$ is processed by a bank of median filters.

$$\overline{R}_n(m) = Med(R_n(m), R_{n-1}(m), R_{n-2}(m)); \quad 0 \le m \le N/2-1 \tag{5}$$

In the ideal case of a stationary noise and a stationary environment, these inter-channel relations would be constant (this point is not exactly true as it also depends on speech). However, in real situations, neither the noise, nor the environment, nor the speech are stationary, so these relations have to be continuously re-evaluated. Fortunately their variation is not fast.

To smoothen the variation of the median filter outputs, a filter with exponential decay is used:

$$\breve{R}_n(m) = \alpha\overline{R}_{n-1}(m) + (1-\alpha)\overline{R}_n(m); \quad 0 \le m \le N/2-1 \tag{6}$$

The cascade combination of the median and the exponential-decay filters shows important properties:

- A peak overestimation in the calculation of $R_n(m)$ is eliminated by the median filter.
- An underestimation of the ratios is smoothed by the exponential-decay filter.

Finally, the spectrum of the noise reference is weighted using a logarithmic law:

$$\breve{N}_n(m) = \left(1+\beta\log_{10}(\hat{N}_n(m))\right)\|N_n(m)\|^a; \quad 0 \le m \le N/2-1 \tag{7}$$

with:

$$\hat{N}_n(m) = \frac{\|N_n(m)\|^a}{\sum_{i=0}^{N/2-1}\|N_n(i)\|^a}; 0 \le m \le N/2-1 \tag{8}$$

The logarithmic operation in *(7)* increases the cancellation gain in the zones where the energy of the residual noise is higher. The normalisation value in *(8)* introduces an equalisation effect in the average energy per frame.

The reference of the noise still present in the signal $P_n(m)$ is then evaluated as the relation between the values calculated in *(6)* and *(7)*:

$$P_n(m) = \frac{\breve{R}_n(m)}{\breve{N}_n(m)};\ 0 \le m \le N/2-1 \qquad (9)$$

The maximum of two consecutive values of the so-calculated reference is selected and subtracted from the primary input of the frequency domain filter to produce the enhanced speech trace $C_n(m)$:

$$\tilde{P}_n(m) = max\{P_n(m), P_{n-1}(m)\};\ 0 \le m \le N/2-1 \quad (10)$$

$$\|C_n(m)\|^a = \|E_n(m)\|^a - \tilde{P}_n(m);\ 0 \le m \le N/2-1 \quad (11)$$

The reason for doing so is twofold: if speech is not present (or there are not speech contents at that frequency), a larger cancellation gain is obtained, and if there are spectral components of speech, as their levels are now higher than noise, they will remain almost unaffected.

After the operation in *(11)*, a half wave rectification is carried out to avoid the possibility of obtaining negative values for the norm of the vector [1].

The phase of the enhanced signal is recovered from the JPEr trace.

4 CONCLUSIONS

The importance of the combination of both methods can be clearly seen in Figure 3. In Fig. 3.a the spectrum of a noisy speech trace is represented. As we can see, the level of noise is so high that the speech signal is completely buried in it, and there is no clue of the presence of speech. Figure 3.b corresponds to the Joint Process Error output of the time domain filtering. Now we can distinguish the periods of speech activity, but the level of noise is still too high. The Joint Process Estimate output is represented in figure 3.c. Figure 3.d corresponds to the estimation of the noise contained in figure 3.b. The output of the frequency domain filter can be seen in figure 3.e. In this case, although there is still some residual noise its level is now quite low. Finally, the comparison between the energy of the signals in 3.a, 3.b and 3.e may be seen in figure 3.f.

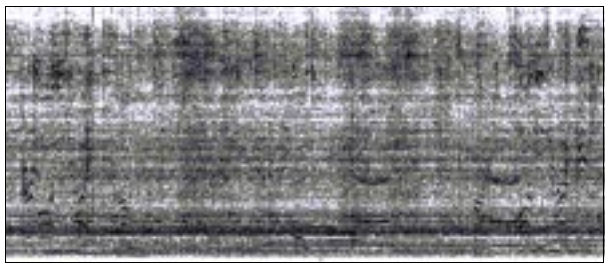

Figure 3.a. Original noisy speech trace (horizontal axis corresponding to 9 sec., vertical axis spanning 0-5,500 Hz.

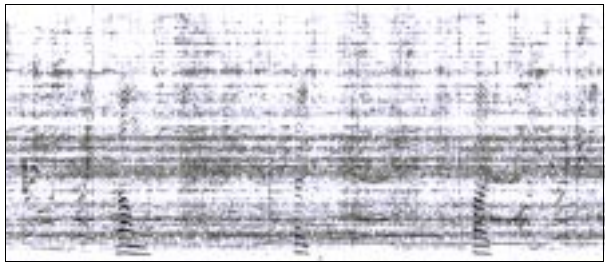

Figure 3.b: Filtered output of the lattice-ladder filter (Joint-process error).

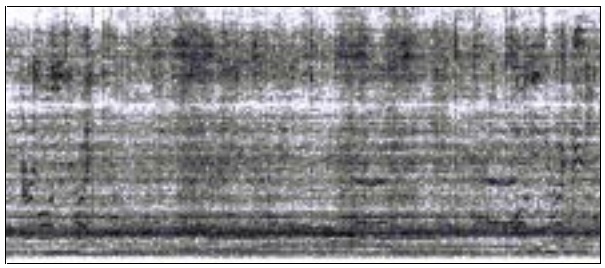

Figure 3.c: Joint process estimate.

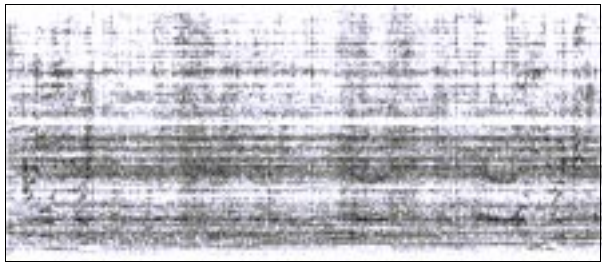

Figure 3.d: Residual noise estimation.

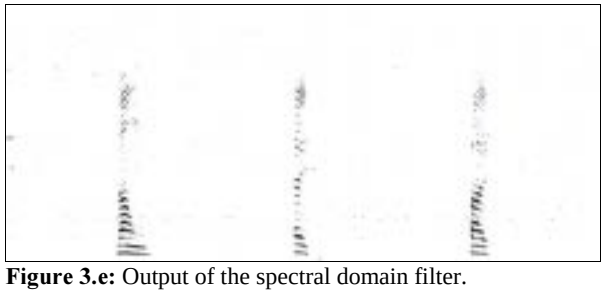

Figure 3.e: Output of the spectral domain filter.

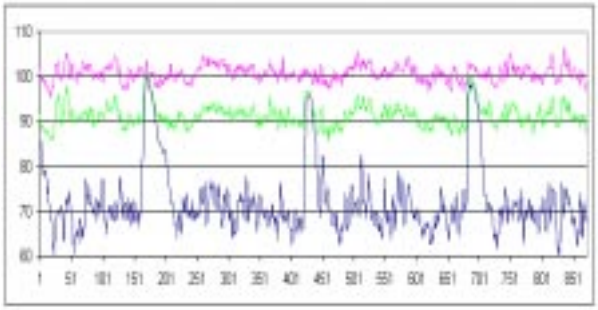

Figure 3.f: Energy corresponding to the noisy speech trace of figure 3.a (upper trace), to the cleaned speech trace of figure 3.b produced with the time-domain filter (middle trace), and to the output of the frequency domain filter of figure 3.e (lower trace). An improvement of more than 20 dB in the SNR can be observed between the noisy speech trace and the resulting enhanced speech.

5 ACKNOWLEDGEMENTS

This work is being funded by grants TIC96-1889-C, TIC97-1011, from the Comisión Interministerial de Ciencia y Tecnología and by an Agreement between UPM and the Centre Suisse d'Electronique et de Microtechnique.

6 REFERENCES

[1] Boll, S. F., "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Tran. on ASSP*, vol. ASSP-27, NO. 2, April 1979.

[2] Haykin, S., *Adaptive Filter Theory*, 3rd Ed., Prentice-Hall, Englewood Cliffs, N.J., 1996.

[3] R. Martínez, A. Alvarez, V. Nieto, V. Rodellar and P. Gómez, "Implementation of an Adaptive Noise Canceller on the TMS320C31-50 for Non-Stationary Environments", *Proc. of the 13th International Conference on Digital Signal Processing*, Santorini, Greece, 2-4 July, 1997 pp. 49-52.

[4] R. Martínez, P. Gómez, A. Alvarez, V. Nieto, V. Rodellar, M. Rubio and M. Pérez, "Dynamic Adjustment of the Forgetting Factor in Adaptive Filters for Non-Stationary Noise Cancellation in Speech", *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP98*, Seattle, Washington, USA, May 12-15, 1998. Vol 2, pp. 1009-1012.

[5] Proakis, J. G., *Digital Communications*, 2nd. Ed, McGraw Hill, 1989.

[6] Widrow, B., et al., "Adaptive Noise Cancelling: Principles and Applications", *Proc. IEEE*, vol. 63, no. 12, pp. 1692-1716, Dec. 1975.