NOISE CANCELLATION IN SPEECH: COMBINATION OF TIME AND SPECTRAL DOMAIN TECHNIQUES

R.Martínez, A. Álvarez, V. Nieto, V. Rodellar, P. Gómez
DATSI, 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

This paper presents a two-microphone speech enhancement system (a primary microphone, and a reference one). 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. In the whole processing, adaptive filtering in the time domain is combined with non-linear spectral subtraction. The combination of both techniques provides a noise cancellation method with convenient qualities. The computational complexity remains under a low limit, while the amount of cancellation is quite high. Another important result is that the system shows a proper performance under non-stationary environments.

2 INTRODUCTION

Removing interference from a desired signal is not a trivial task. The situation is much more complicated when the system is continuously changing. The difficulty of the problem increases if the system works in an environment with very poor SNR (around 0 dB). In addition, the nature of the interfering signal can be quite similar to the desired one. In the particular case of speech enhancement, the possible presence of voice sounds inside the noise signals make it almost impossible to decide whether the detected voices correspond to a valid speech utterance or to an interfering one. The noise-cancelling scheme here presented was developed taking the following requirements into consideration:

- Noise continuously changing in level and spectral distribution.
- Noise levels over 85 dB SPL.
- SNR's ranging from 0 to 10 dB.
- Negative SNR at some frequencies.
- Noise spectral distribution overlapping over the speech spectrum.
- Undesired speech present in noise.
- Affordable computational costs

With such requirements the aid of a multi-microphone structure can be useful. In our case we have chosen a two-microphone scheme.

The solution here proposed presents a combination of two traditional techniques. A first linear processing stage in the time domain (an adaptive lattice ladder filter) is combined with a non-linear-processing block in the frequency domain. With this scheme, very short adaptive filters can be used, which implies a significant computational cost saving.

A logic control has been added to the filter to modify its forgetting factor accordingly with the input signals, in order to obtain shorter locking periods, and to avoid instabilities caused by very high and sudden energy differences between the input channels. An additional stability control block makes an instantaneous change in the forgetting factor if a possible instability is detected.

Finally, a further processing is performed: the output of the adaptive filter is passed through a spectral subtractor in order to acquire a higher noise cancellation.

A general framework of the proposed methodology can be seen in Figure 1.

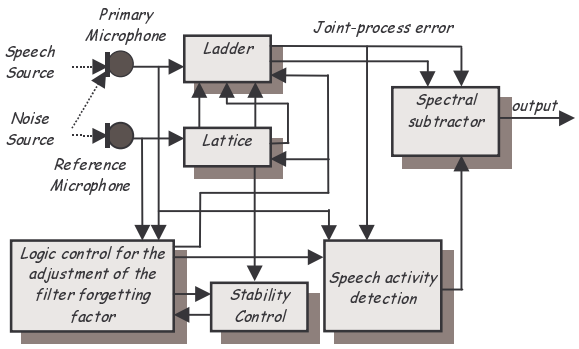

Figure 1. General framework of the proposed methodology: A Lattice-Ladder Filter is combined with a Spectral Subtractor.

3 ADAPTIVE FILTER

Adaptive processing is a well-known technique for removing non-desired signals from a given one [4]. Through the years a great number of adaptive algorithms have been developed. After several tests we determined that the algorithms which better performed for our application were the adaptive lattice ladders (joint process estimators). A comparison between three lattice ladder filters (two of then with LSL algorithms and the third with a gradient adaptation) can be fond in [2]. A Recursive LSL Filter using *a posteriori* estimation errors [1], [3] was finally selected, since this algorithm shows a good compromise between computational complexity, cancellation gain, stability and ratio of convergence in tests carried out using real records.

There is a specific characteristic of this filter that makes it be particularly suitable for the application required. Granted that the reference signal is a good estimate of the noise that corrupts the speech signal, and taking care for not introducing that desired speech signal in the reference channel, the cancellation gain is rather acceptable even for extremely high levels of noise. Besides, the amount of noise cancellation obtained in our experiments is almost independent of the original signal SNR. The obtained cancellation is in all cases in the order of 10 dB, and is much more dependent of factors like microphone separation or the level of cross-talk than of the SNR. This is a very important fact when the SNR is very low (other enhancement methods precise a minimum SNR to have a proper performance). However, these filters demand a great quantity of computing power and that computer complexity grows of course with the filter length. There are some added disadvantages for using long filters instead of short ones: longer locking periods, more instability problems (errors are propagated through the structure and amplified by the consecutive taps), limitations to follow the signal in non-stationary conditions.

The utilisation of these filters is as follows: the lattice part is fed with the noise source (reference), and the primary signal is inserted through the ladder part. A set of backward prediction errors (which constitute an orthogonal base) is so obtained for the noise reference. These orthogonalised components are used to build an estimator of the noise which is compared with the signal

of the ladder part 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 these components from the primary input give the Joint Process Error (JPEr). This JPEr contains the signal that is not in common (the speech) plus the orthogonal components of noise that are not estimated yet. An important property shown by the lattice filter is that the energy contribution of its stages decrease with the stage order. It is clear that the consecutive addition of more filter stages will provide a deeper noise cancellation (as more orthogonal components are deleted), but there is a point in which the increment in the SNR achieved is not significant compared with the increment in the computational requirements. 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.

For all those reasons short filters are used. 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.

3.1 Filter Forgetting Factor Adjustment

It is well known that forgetting factors close to unity minimise the estimation errors, but this choice has the handicap of a poor adaptation capability in case of non-stationary signals. On the other hand, smaller values for these forgetting factors will conduct to a faster adaptation, but with higher estimation errors, and tighter stability limits. Hence it is desirable to have a mechanism which is aware of the variations in signal conditions and changes the forgetting factor accordingly. The optimum value of the forgetting factor of the adaptive filter is dependant of the instantaneous relationship between the energies of the primary and the reference signals. That is the reason why we have added a logic control in order to change the value of the forgetting factor accordingly with the input signals. With such control shorter locking periods for the filter can be achieved at the same time that an optimal cancellation is ensured.

In normal operation, with no speech present, or with a level of speech similar to the level of noise, a forgetting factor around 0.9999 has shown to be the best choice, as it is the best compromise between estimation error and adaptation speed. But sudden changes in the SNR produce unlockings in the filters. Measurements have shown that in such situations the filter tends to introduce parts of the process estimate (noise) into the process error (enhanced signal). This can be observed as a significant noise tail in the enhanced signal that can last for several hundreds of milli-seconds. Those noise tails follow the spectral zones with high local SNR (strong formants clearly separated from noise).

In such cases the ideal operation of the filter should be as follows:

As soon as a high SNR condition is detected (energetic utterance of a word), the filter should change its forgetting factor to a value closer to unity. Otherwise noise is introduced in parts with low level of speech spectral contents. When the end of the word is detected after a high SNR condition, it is mandatory to re-adapt the filter as soon as possible. In such cases, lower values of the forgetting factor will ease this operation.

Those changes in SNR can be obtained with the difference in the power envelopes of the input signals. The estimation of the energy of the signals is computed as the sum of the squared values of the samples in a window of length 128.

$$E_P(i) = \sum_{n=i*128}^{(i+1)128} p^2(n) \qquad (1)$$

$$E_R(i) = \sum_{n=i*128}^{(i+1)128} r^2(n) \qquad (2)$$

Where $E_P(i)$ and $E_R(i)$ are the calculated energies of the primary signal *p(n)* ad of the reference one *r(n)*.

The window length is chosen such that the estimate of the envelope of the energy does not follow rapid variations in the wave form, in particular those corresponding to the pitch period in voiced speech, although being short enough to reflect real time variations in the energy of the signal. It also enables easy synchronisation with the block operations of the other algorithms in the system, in particular for FFT based frequency domain processing.

To ensure that the system is independent of the noise level and of the overall acquisition gain, a normalisation is necessary. The value E_N for this normalisation is obtained from the reference:

$$E_N = \sum_{i=1}^{20} E_R(i) \qquad (3)$$

And the energy values of *(1)* and *(2)* are normalised as follows:

$$\overline{E_P}(i) = E_P(i) / E_N \qquad (4)$$

$$\overline{E_R}(i) = E_R(i) / E_N \qquad (5)$$

A direct comparison of the values so obtained is not valid, for different gains are possible for both channels, so it is mandatory an equalisation between them. A way to do so is to calculate the factor E_E in which the reference signal is different from the primary one. This can only be done in the periods in which voice is not present or at least it is below the level of noise. This processing is so asynchronous with the window processing:

$$E_E = \frac{\sum_{n}^{n+128*50} r^2(n)}{\sum_{n}^{n+128*50} p^2(n)} \qquad (6)$$

Finally the energy of the primary source is normalised with the so obtained energy:

$$\breve{E}_P(i) = \overline{E_P}(i) E_E \qquad (7)$$

The first estimation of the presence of voice is given by the difference between the normalised and equalised energies of the primary and the reference channels E_{dif}. (A much more refined speech detector based in the enhanced speech is afterward performed).

$$E_{dif} = \breve{E}_P(i) - \overline{E_R}(i) \qquad (8)$$

If this E_{dif} is under a certain threshold, it indicates the required absence of speech (or a very low SNR), therefore the square values of the samples of the primary and the reference signals are summed in *(6)* and the relationship between them evaluated.

This value E_{dif} is also used by a finite state automaton, which changes states accordingly with the changing conditions, and chooses the adequate filter forgetting factor. A high value of the power envelope difference triggers a state with an associate value for α close to the unity (high SNR condition). If this situation is followed by a sudden decrease in the power envelope difference, it means that quick adaptation is required, and the value of α is reduced accordingly.

In some situations, when a speaker is too close to the microphones, or with plosive pronunciations, a pressure wave can be generated. Its effect may be observed as a huge oscillation with a long period. Due to the energy present in this signal, it might corrupt the estimate of instantaneous SNR and be detected

as a high energy burst. It possibly could trigger the forgetting factor control to go into an unstable mode of operation. To avoid such behaviour this pressure wave signal could be eliminated by means of a high pass filter, although care has to be taken that the filter will pass the relevant low frequency content of speech signals. Another way of detecting these low frequency oscillations is by counting the number of zero crossings in a window. An exceptionally low number of these crossings will reflect the presence of a pressure wave.

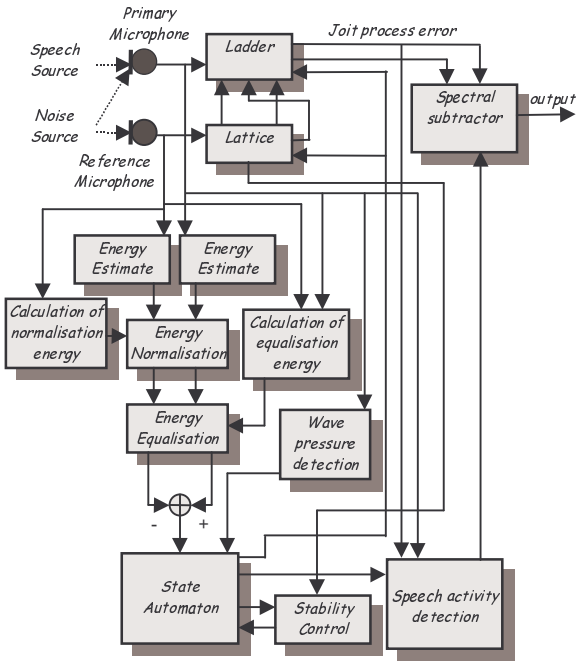

Figure 2: General framework with the forgetting factor adjustment block detailed.

3.2 Control of the Filter Stability

Quick adaptation moves the filter to its stability limits, therefore an additional *stability control block* would be required. As it was mentioned before a low value of α can carry the filter to an unstable condition. This situation may be detected tracing a conversion factor, which actuates on the *reflection coefficients* of the lattice filter. The usual value of this coefficient is close to the unity, but it begins to oscillate just before the filter becomes unstable. This oscillation may be easily noticed, and a correction of α towards the unity puts the filter back under stable conditions.

4 SPECTRAL SUBTRACTION

As mentioned previously the order of the adaptive filter, and consequently its performance, is limited by several reasons. Although low-order filters still ensure high cancellation gains, some other phenomena, such as reverberations, could only be treated using extremely long filters. Consequently, if low-order filters are used, the residuals of the filter, namely the *joint-process estimation error*, and the *joint-process estimate* may still be correlated. Besides, the *joint-process estimation* grants the removal of noise only to a certain level. If more noise is to be removed, *spectral subtraction* may be used.

This further processing could not be done in any SNR condition: it needs the speech signal to be sufficiently higher than the noise. Otherwise although a perfect estimation of the modulus of the noise signal contained in speech was done, the indetermination of the value of its phase would be enough to contribute as a noise source. Besides, a signal almost buried in noise would make it impossible to determine the periods when speech is present, and that is essential for the proper operation of spectral subtraction. After adaptive filtering a sufficient SNR is achieved to ensure the proper performance.

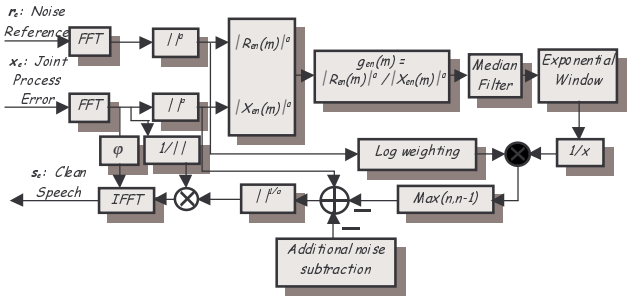

Figure 3 Specific Spectral Subtraction method being proposed

The *joint-process estimation error* output of the lattice-ladder filter $\boldsymbol{x_e}$ is used as the primary signal, and an estimate of the noise formed with the *joint process estimate* $\boldsymbol{r_e}$ is used as the reference. The original noise reference signal could also be considered to produce an estimate of the power spectrum of the noise remaining in the enhanced signal, but the *joint process estimate* $\boldsymbol{r_e}$ has an important advantage: it is normalised in power with respect to the *joint process estimation error* $\boldsymbol{x_e}$. Both signals are segmented in overlapped windows and transformed into the frequency domain using the short-time Discrete Fourier Transform $F\{.\}$:

$$\boldsymbol{X_{en}} = X_{en}(m) = F\{x_e(n)w(n)\} \quad (9)$$

$$\boldsymbol{R_{en}} = R_{en}(m) = F\{r_e(n)w(n)\} \quad (10)$$

where $w(n)$ is the window function, and n and m are the time and frequency indices.

The relationship between the power spectra of the Joint Process Estimation Error and the Joint Process Estimate is calculated for every frequency during the segments when speech is not present:

$$g_n(m) = \frac{\|R_{en}(m)\|^a}{\|X_{en}(m)\|^a}; \quad 0 \le m \le M/2\text{-}1 \quad (11)$$

Where M is the size of the window used.

These values are passed through a bank of median filters:

$$\bar{g}_n(m) = Med(g_n(m), g_{n-1}(m), g_{n-2}(m)); 0 \le m \le M/2-1 \quad (12)$$

followed by an integrator filter:

$$\breve{g}_n(m) = a\ \breve{g}_{n-1}(m) + (1-a)\ \bar{g}_n(m)\ ; 0 \le m \le M/2\text{-}1 \quad (13)$$

with $0<\alpha<1$.

Finally, the power spectrum of the joint process estimate $R_{en}(m)$ is weighted using a logarithmic law:

$$\tilde{R}_{en}(m) = \left(1 + \beta \log_{10}(\hat{R}_{en}(m))\right)\|R_{en}(m)\|^a; \ 0 \le m \le M/2\text{-}1 \quad (14)$$

with:

$$\hat{R}_{en}(m) = \frac{\|R_{en}(m)\|^a}{\sum_{i=0}^{N/2-1}\|R_{en}(m)\|^a}; \ 0 \le m \le M/2\text{-}1 \quad (15)$$

and compensated with the relation between power spectra evaluated in *(13)* to obtain a reference of the noise still present in the signal:

$$\hat{R}_{en}(m) = \frac{\breve{R}_{en}(m)}{\breve{g}_n(m)}; \quad 0 \le m \le M/2\text{-}1 \quad (16)$$

An additional noise reduction is performed when the SRN is high. In such cases it has been observed that some noise in high frequencies may remain. A parabolic noise profile $P_{en}(m)$ is then added to the value calculated in *(16)*:

$$P_{en}(m) = \left(\sum_{i=0}^{M/2-1} \|R_{en}(m)\|^a \right) * \left[0.25 + (m/(M/2))^2 \right];$$

$$0 \le m \le M/2\text{-}1 \qquad (17)$$

$$\hat{R}_{en}(m) = \bar{R}_{en}(m) - P_{en}(m); \qquad 0 \le m \le M/2\text{-}1 \qquad (18)$$

The logarithmic operation *(14)* increases the cancellation gain in the zones where the energy of the residual noise is higher. The maximum of two consecutive values of the so calculated reference is selected and subtracted from the joint-process estimation error output (to compensate for temporal variations):

$$\tilde{R}_{en}(m) = max\left\{ \hat{R}_{en}, \hat{R}_{en-1} \right\}; \qquad 0 \le m \le M/2\text{-}1 \qquad (19)$$

$$\|S_{en}(m)\|^a = \|X_{en}(m)\|^a - \tilde{R}_{en}(m); \quad 0 \le m \le M/2\text{-}1 \qquad (20)$$

The reason for doing so is twofold: if speech is not present (or there are not speech contents at that frequency), the result is a stronger cancellation gain, and if there are spectral components of speech, as their levels are now higher than noise, they will remain almost unaffected. The window overlapping reinforces spectral components that remain in consecutive windows (as is the case of the pitch harmonics), so although a noise overestimation was done, and speech components removed, the consecutive addition of frames would recover part of the eliminated speech. This subtraction is combined with spectral flooring to limit the presence of artificial tones.

The phase of the enhanced signal is recovered from the Joint Process Estimation Error trace:

$$\varphi_{sen}(m) = \varphi_{xen}(m); \qquad 0 \le m \le M/2\text{-}1 \qquad (21)$$

5 RESULTS

In figures 4 through 7 is shown an example of the performance of the presented method. The English words */down/* and */eight/* were recorded under very noisy conditions with the two-microphone scheme explained.

In Figure 4 is represented the power spectrum of an original noisy trace, as it is recorded by the primary microphone. In this spectrum it is impossible to distinguish the voice signal, which is completely buried in noise. The power spectrum of the signal filtered by the adaptive filter (joint process error of the filter), is shown in Figure 5. In this case the voice signal can be discerned, although the level of noise is still quite high. The amount of cancellation of this first processing block is about 10 dB, as can be corroborated in Figure 7, where the upper trace represents the energy of the signal entering the primary microphone, and the middle trace represents the energy of the filtered signal.

The power spectrum of the speech after subtraction can be seen in Figure 6. The energy of the so filtered signal is represented in the lower trace of Figure 7**,** and can be compared with the energy of the original noisy signal (upper trace), and with the energy of the output of the adaptive filter (middle one).

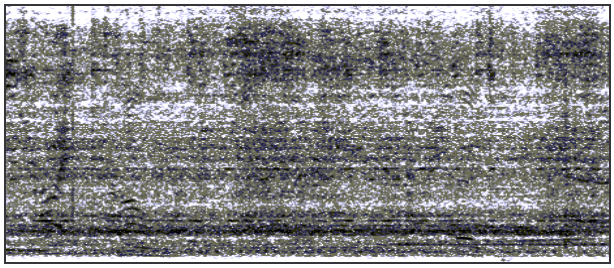

Figure 4. Power Spectrum of the Noisy Speech Trace (primary microphone)

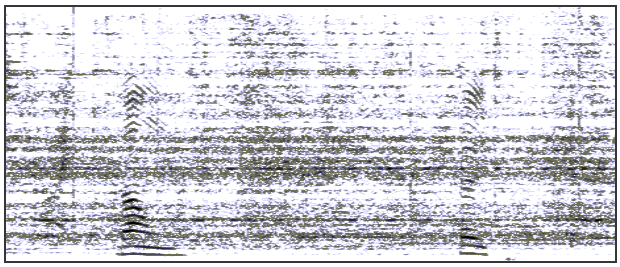

Figure 5. Power Spectrum of the Joint Process Estimation Error (Enhanced Speech).

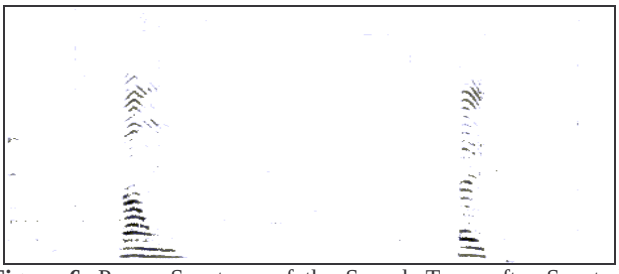

Figure 6. Power Spectrum of the Speech Trace after Spectral Subtraction

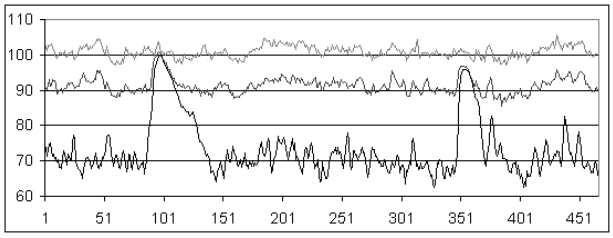

Figure 7. Instantaneous energy of the Noisy Speech Trace from Figure 4 (upper trace), of the Joint Process Estimation Error from Figure 5 (middle trace), and of the Speech Trace after Spectral Subtraction from Figure 6 (lower trace).

6 CONCLUSIONS

The combination of adaptive filtering and non-linear spectral subtraction achieve a much higher noise cancellation than the separate use of those techniques. It allows the use of very short adaptive filters, with the consequent saving in computational requirements. This technique may be used in non-stationary conditions and with very low SNR's. In experiments like the one presented in figures 4 through 7, improvements of 30 dB in SNR have been achieved.

7 ACKNOWLEDGEMENTS

This work is being funded by grants TIC96-1889-C, TIC97-1011, 07T-0001-2000.

REFERENCES

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

[2] Martínez, R., A. Álvarez, V. Nieto, V. Rodellar and P. Gómez, "ASR in Highly Non-Stationary Environments using Adaptive Noise Canceling Techniques", *Proceedings of the ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-à-Mousson, France, 17-18 April, 1997, pp. 181-184.

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

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