# Speech Enhancement based on Linear Prediction Error Signals and Spectral Subtraction

*A. Álvarez, V. Nieto, P. Gómez, and R. Martínez*

Departamento de Arquitectura y Tecnología de Sistemas Informáticos
Facultad de Informática, Universidad Politécnica de Madrid
Campus de Montegancedo, s/n, 28660 Boadilla del Monte, Madrid, SPAIN
pedro@pino.datsi.fi.upm.es

## Abstract

Speech processing and recognition are key technologies to produce smart user interfaces in an increasing number of devices. Moreover, robust speech recognition is considered mandatory for a reliable operation of such elements in realistic working conditions. Through this paper, a method of processing speech degraded by noise and reverberation is proposed. This approach involves analyzing the prediction error signals from the *Gradient Adaptive Lattice* algorithm in order to produce a valid estimator suitable for being combined with *Spectral Subtraction* techniques. The paper includes an evaluation of the performance of the algorithm for several speech recognition experiments in a car environment.

## 1. Introduction

Speech produced and captured in real environments is perturbed by noise and reverberation. The resulting signal may exhibit a negative SNR and may contain voice activity arriving from other sources. Several multimicrophone methods have been proposed for enhancement of speech degraded by reverberation. Usually, *Array Beamforming* is combined with other techniques as *Independent Component Analysis* [1], *Spectral Subtraction* [2] or *Linear Prediction Analysis* [3]. A significant number of approaches take advantage of the LP residual signal [4], [5]. For clean voiced speech, LP residuals have strong peaks corresponding to glottal pulses, whereas for reverberated speech peaks are spread in time. Therefore, a measure of the amplitude spread of LP residual can serve as a reverberation metric. Through this paper, a speech enhancement system based on the use of the backward-prediction error signal of the *Gradient Adaptive Lattice* algorithm for reverberation and noise estimation purposes, and its application to *Spectral Subtraction* techniques [6] is presented.

## 2. Reverberation and noise detection with GAL error signals

The detection of the amount of reverberation and noise existing in portions of speech is achieved following the steps presented in Figure 1. As it may be seen, the measure of kurtosis is not taken from the speech signal itself, but using a linear-prediction analysis error signal. On a subsequent step, kurtosis function is mapped into a weight function, which will be used as an indicator of the degree of degradation existing in the original signal.

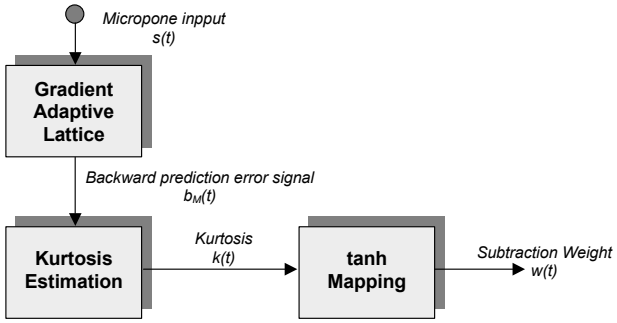


*Figure 1.* Algorithmic structure for the calculation of the subtraction weight *w(t)* from the GAL last-stage backward prediction error signal $b_M(t)$.

### 2.1. Gradient Adaptive Lattice algorithm

The *Gradient Adaptive Lattice* (GAL) algorithm [7] is a natural extension of the Least Mean Square (LMS) approach. The structure of the m[th] stage of a lattice is characterized by a single parameter, namely, the reflection coefficient or $\kappa_m$. A cost function for this stage is defined as:

$$J_m = E\left\{\left|f_m(n)\right|^2 + \left|b_m(n)\right|^2\right\} \tag{1}$$

where $E$ is the statistical expectation. On the other hand, $f_m(n)$ is the forward prediction error and $b_m(n)$ is the backward prediction error, both measured at the stage output and described by:

$$f_m(n) = f_{m-1}(n) + \kappa_m^* b_{m-1}(n-1) \tag{2}$$

$$b_m(n) = b_{m-1}(n-1) + \kappa_m f_{m-1}(n) \tag{3}$$

where $f_{m-1}(n)$ is the forward prediction error and $b_{m-1}(n-1)$ is the delayed backward prediction error, both measured at the input of the stage.

The gradient of the cost function $J_m$ with respect to the reflection coefficient $\kappa_m$ is given by:

$$\nabla J_m = 2E\left\{f_m^*(n)b_{m-1}(n-1) + b_m(n)f_{m-1}^*(n)\right\} \tag{4}$$

The instantaneous estimate of the gradient may be expressed as:

$$\hat{\nabla}_m J(n) = 2\left(f_m^*(n)b_{m-1}(n-1) + b_m(n)f_{m-1}^*(n)\right) \tag{5}$$

assuming:

$$E\left\{f_m^*(n)b_{m-1}(n-1)\right\} \approx f_m^*(n)b_{m-1}(n-1) \tag{6}$$

$$E\left\{b_m(n)f_{m-1}^*(n)\right\} \approx b_m(n)f_{m-1}^*(n) \tag{7}$$

Subsequently, the estimation of the reflection coefficient at time $n$ or $\hat{\kappa}_m(n)$ may be calculated by:

$$\hat{\kappa}_m(n) = \hat{\kappa}_m(n-1) - \frac{1}{2}\mu_m(n)\hat{\nabla}_m J(n) \tag{8}$$

where $\mu_m$ denotes a time-varying adaptation factor associated with the m[th] stage. Now, combining the above equations the estimation may be expressed by:

$$\hat{\kappa}_m(n) = \hat{\kappa}_m(n-1) - \\ -\mu_m(n)\left(f_m^*(n)b_{m-1}(n-1) + b_m(n)f_{m-1}^*(n)\right) \tag{9}$$

In addition, the parameter $\mu_m(t)$ is defined as:

$$\mu_m(n) = \frac{\mu}{\xi_{m-1}(n)} \tag{10}$$

where constant $\mu$ is typically below 0.1 and $\xi_{m-1}(n)$ represents the total energy error given by:

$$\xi_{m-1}(n) = \sum_{i=1}^{n}\left(\left|f_{m-1}(i)\right|^2 + \left|b_{m-1}(i-1)\right|^2\right) = \\ = \xi_{m-1}(n-1) + \left|f_{m-1}(n)\right|^2 + \left|b_{m-1}(n-1)\right|^2 \tag{11}$$

Finally, a new parameter $\beta$ is introduced to provide the algorithm a finite memory, useful to deal with statistical variations due to the lack of stationary environments.

$$\xi_{m-1}(n) = \beta\xi_{m-1}(n-1) + \\ +(1-\beta)\left(\left|f_{m-1}(n)\right|^2 + \left|b_{m-1}(n-1)\right|^2\right) \tag{12}$$

### 2.2. Kurtosis estimation

In our approach, the linear prediction based signal is obtained applying the GAL algorithm with a number of stages *M= 32*, being *β= 0.9999 and μ= 1.0*. The signal of interest is the backward prediction error corresponding to the last stage or $b_M(t)$.

In a following step, signal $b_M(t)$ is grouped into 60 ms frames with an overlap of 40 ms. Kurstosis values $k_n$ are then estimated for every block by:

$$k_n = \frac{E\{b_M^4(t)\}}{E^2\{b_M^2(t)\}} - 3 \tag{13}$$

As, it may be noticed, that procedure allows obtaining new kurtosis estimations every 20 ms. In order to produce a new function *k(t)*, defined for each sample, a final step consisting on the repetition of kurtosis values is carried out.

### 2.3. Gradient Adaptive Lattice weight

In a following step, a GAL based weight *w(t)* suitable to be combined with a subsequent spectral subtraction module is produced by:

$$w(t) = \frac{1}{2} + \frac{\tanh\left[\lambda \cdot \left(|k(t)| - \theta\right)\right]}{2} \tag{14}$$

*w(t)* being the output weight factor at time *t*, $\lambda$ a gain factor and $\theta$ a threshold indicating the minimum kurtosis value linked with speech segments.

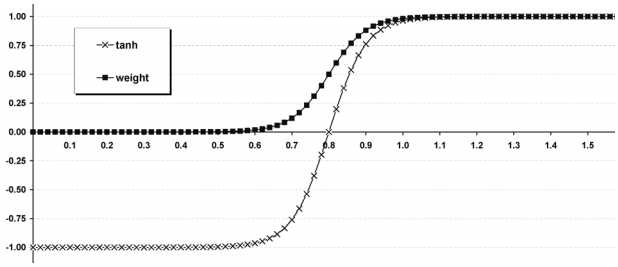


*Figure 2.* Mapping function applied to estimate the subtraction weight signal *w(t)*, being *λ= 10.0* and *θ= 0.8*.

## 3. Spectral subtraction

To implement the filtering in the spectral domain, the GAL based weight will be considered a relevant estimator. The proposed procedure may be seen in Figure 3.

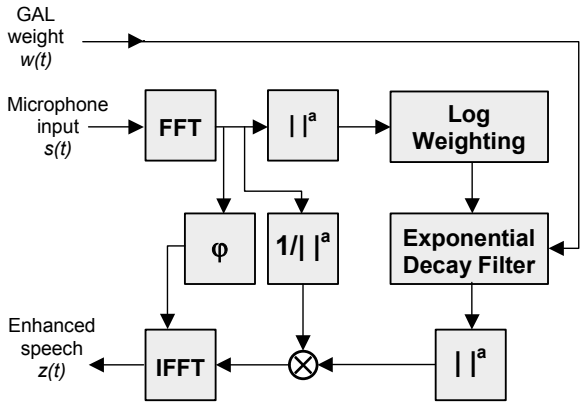


*Figure 3*. Structure of the spectral subtraction module which exploits the GAL weight estimator.

Firstly, the log power of the signal is computed for every FFT frequency channel, being and *M* the window size and *m* the frequency index: In our application *M* will be equal to 512 samples.

$$S_{\log}(m) = \log_{10}\left(\left\|S(m)\right\|^a\right); \quad 0 \le m \le M/2 - 1 \tag{15}$$

These values are passed to a filter with exponential decay given by:

$$g_S(m) = \alpha\gamma S_{\log} + (1-\alpha)g_{n-1}(m) \tag{16}$$

where $\alpha$ is a coefficient that controls the log-power rate update and $\gamma$ is a gain factor, close to 1.0, which allows increasing the amount of cancellation.

Once we have adapted the incoming signal energy, the calculation of the subtracting-signal at frame index *n* and frequency index *m* or $g_n(m)$, is accomplished by a new exponential decay filter controlled by the GAL weight *w(t)*:

$$g_n(m) = (1-w_n)g_S(m) + w_n(m)g_{n-1}(m) \tag{17}$$

where $w_n$ is the value $w(t_n)$ for a time $t_n$ which corresponds to the center of the n[th] frame.

The above expression implies that a weight equal to 1.0 prevents from updating the estimation of $g_n(m)$ at all, whereas a weight close to 0.0 will produce a fast adaptation.

Finally, the exact amount to be subtracted is generated and the subtraction in itself is performed, thus producing an enhanced signal in the time domain *z(t)* given by:

$$G(m) = 10^{g_n(m)}; 0 \le m \le M/2 - 1 \quad (18)$$

$$\|Z(m)\|^a = \|S_{en}(m)\|^a - G(m); 0 \le m \le M/2 - 1 \quad (19)$$

## 4. Results and discussion

The three main steps comprised in the enhancement system proposed in this paper are summarized in Figure 4 through Figure 10. In particular, Figure 4 shows a noisy and reverberant utterance of several Finnish digits produced by a male speaker. The backward prediction error signal $b_M(t)$, extracted using the GAL algorithm is presented in Figure 5. On a further step, the kurtosis function is obtained, as it may be seen in Figure 6. That function allows estimating the GAL weight, which as shown in Figure 7 is mapped in the range [0.0, 1.0]. The result after completing the spectral subtraction procedure is an enhanced speech signal (see Figure 8). That fact may be clearly seen comparing the spectrograms of the original and clean signals, which are presented in Figure 9 and Figure 10, respectively.

In order to examine the validity of the proposed method, several speech recognition systems were built and tested. A subset of the Aurora3-SpeechDat Car Finnish database [8] is used for these purposes. That corpus, which contains realizations of connected digits uttered in a realistic automobile environment, is divided in two groups: train and test, and three different categories related with the amount of distortion contained in the recordings: quiet, low, and high. In our experiments, we used the recordings associated to channels *ch2* and *ch3* (see Figure 11).

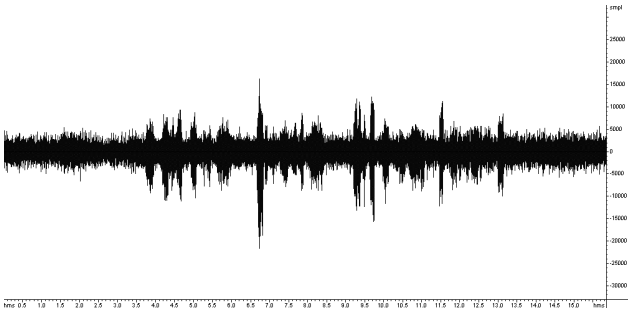

*Figure 4.* Segment of reverberant and noisy speech that contains an utterance of 16 Finnish connected digits produced by a male speaker.

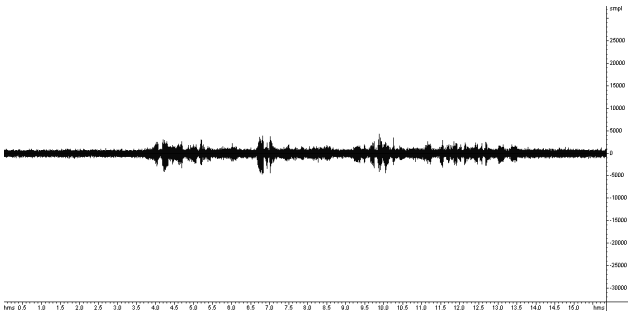

*Figure 5.* Gradient Adaptive Lattice backward-prediction error $b_M(t)$, associated to the signal appearing in Figure 4.

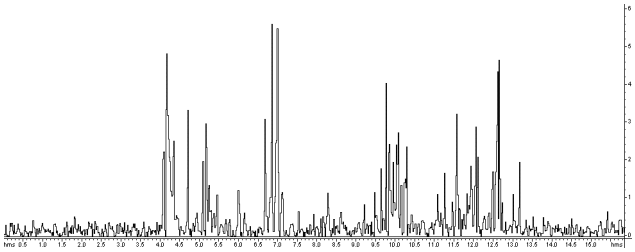

*Figure 6.* Kurtosis function computed from the GAL backward prediction error signal (see *Figure 5*).

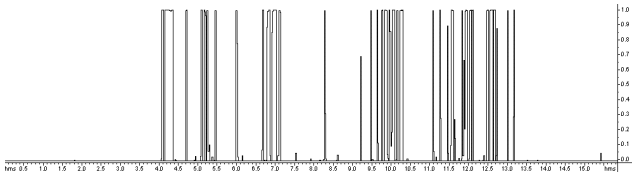

*Figure 7.* Weight function computed from the mapping of the kurtosis function.

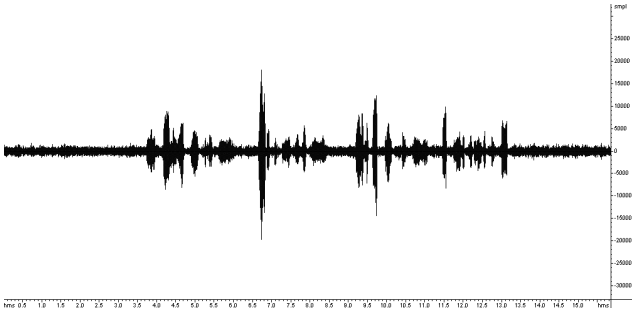

*Figure 8.* Enhanced output signal obtained after the application of the Spectral Subtraction method proposed.

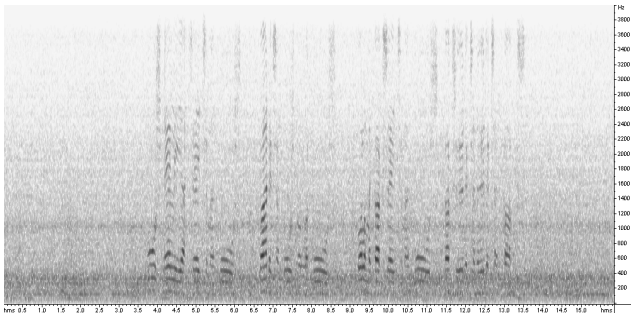

*Figure 9.* Power spectrum of signal introduced in *Figure 4*.

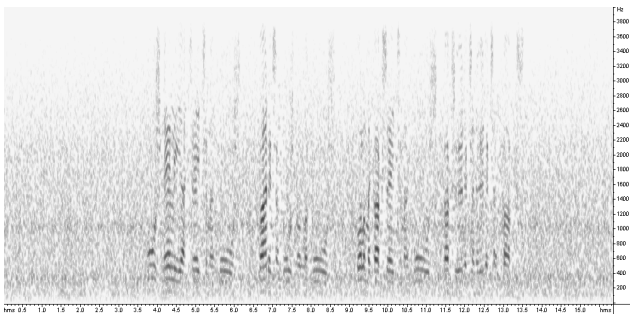

*Figure 10.* Power spectrum of signal contained in *Figure 8*.

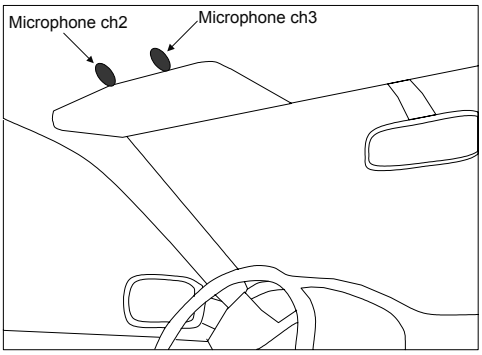


*Figure 11.* Recording framework for the Aurora3 speech corpus. The position of microphones linked to channels *ch2* and *ch3*, is also indicated.

Recognition experiments are established by selecting different materials from the training set of the database. More exactly, s*et A* includes files labeled as quiet, s*et B* incorporates also files with low distortion and, finally, *set C* comprises all the training material available. The test material is the same for the three cases and consists on 536 files per channel. The front-end extracts energy plus 36 MFCCs (12 cepstrum, 12 delta cepstrum and 12 delta-delta coefficients). The HMMs are then built with 16-state whole word models for each digit in addition to a begin-end model and a word-separation one. Finally, models have 3 diagonal Gaussian mixture components in each state.

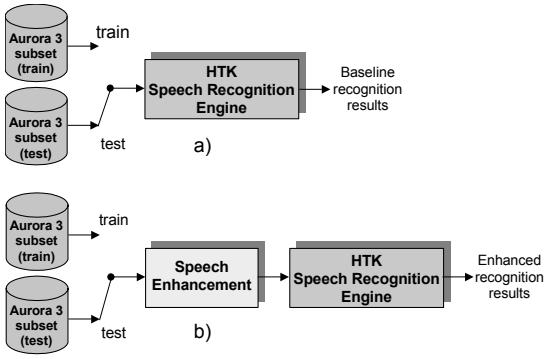


*Figure 12.* a) Baseline HTK based speech recognizer. b) Enhanced system incorporating the method proposed in this paper to the same speech recognition engine.

| Baseline system | WER ch2 | WER ch3 |
|---|---|---|
| Train set A | 48.34% | 23.99% |
| Train set B | 11.74% | 9.56% |
| Train set C | 9.95% | 10.52% |

| Enhanced system | WER ch2 | WER ch3 |
|---|---|---|
| Train set A | 29.79% | 16.76% |
| Train set B | 7.92% | 5.98% |
| Train set C | 7.25% | 7.74% |

| WER reduction | ch2 | ch3 |
|---|---|---|
| Train set A | 38.36% | 30.13% |
| Train set B | 32.52% | 37.46% |
| Train set C | 27.14% | 26.44% |

*Table 1.* Recognition results for microphones *ch2* and *ch3*.

Table 1 presents accuracy results when the method proposed in this paper is applied as a pre-processing stage to the same front-end (see Figure 12). As it may be seen, WER reduction over the baseline system is significant for both channels and the three training sets.

## 5. Conclusions

The combination of *Spectral Subtraction* techniques with a subtraction weight, extracted from the backward error signal of the *Gradient Adaptive Lattice* algorithm, constitutes an efficient approach to the speech enhancement problem in noisy and reverberant environments. Moreover, this method allows the system to operate with no a priori knowledge of the working framework. Speech recognition experiments in a car environment carried out with real data taken form the Aurora3 database show a reduction in *Word Error Rates* higher than 30% on average.

## 6. Acknowledgements

This research is being carried out under grants TIC99-0960, and TIC2002-02273 from the *Programa Nacional de las Tecnologías de la Información y las Comunicaciones* (Spain), and grant 07T-0001-2000 from the *Plan Regional de Investigación de la Comunidad de Madrid*.

## 7. References

[1] Barros, A. K., Itakura, F., Rutkowski, T., Mansour, A.; Ohnishi, N., "Estimation of speech embedded in a reverberant environment with multiple sources of noise" *Proc. of ICASSP´01*, May 7-11 2001, Vol. 1, pp: 629- 632.

[2] Mokbel, C. E, and Chollet F. A., "Automatic Word Recognition in Cars", *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 5, September 1995, pp. 346-356.

[3] Grbic, N., Nordholm, S., Johansson, A., "Speech enhancement for hands-free terminals", *Proc. of 2nd International Symposium on Image and Signal Processing and Analysis (ISPA 2001)*, pp. 435- 440.

[4] Gillespie, B. W., Malvar, H. S., Florencio, D. A. F., "Speech dereverberation via maximum-kurtosis subband adaptive filtering", *Proc. of ICASSP´01*, May 7-11 2001, Vol. 6, pp. 3701- 3704.

[5] Yegnanarayana, B., Satyanarayana Murthy, P., "Enhancement of Reverberant Speech Using LP Residual Signal", *IEEE Trans, on Speech and Audio Processing*, Vol. 8, No. 3, May 2000, pp. 267- 281.

[6] Boll, S. F., "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Trans. on ASSP*, Vol. 27, 1979, pp. 113-117.

[7] S. Haykin, *Adaptive Filter Theory*, Third edition, Prentice Hall, 1996.

[8] A. Moreno, et al., "SPEECHDAT-CAR: A Large Speech Database for Automotive Environments", Proc. of *2nd International Conference on Language Resources and Evaluation*, Athens, Greece, 2000, paper 373.