Speech Enhancement for a Car Environment using LP Residual Signal 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

Handsfree speaker input is mandatory to enable safe operation in cars. In those scenarios robust speech recognition emerges as one of the key technologies to produce voice control car devices. Through this paper, we propose a method of processing speech degraded by reverberation and noise in an automobile environment. This approach involves analyzing the linear prediction error signal to produce a weight function suitable for being combined with spectral subtraction techniques. The paper includes also an evaluation of the performance of the algorithm in speech recognition experiments. The results show a reduction of more than 30% in word error rate when the new speech enhancement front-end is applied.

1. Introduction

Speech produced and captured in a running car is perturbed by noise and reverberation. The resulting signal may exhibit a negative SNR, being such conditions a real challenge for currently available speech recognition algorithms. An additional difficulty is that for speakers is desirable not to wear proximity microphones. Moreover, in a car may be more than one active speaker at a given time (e.g. driver and copilot).

When dealing with reverberant signals the main objective of processing is to increase the contribution of the direct component relative to the reverberant component [1]. Several multimicrophone methods have been proposed for enhancement of speech degraded by reverberation in car environments [2]. Microphone array based methods attempt to enhance the signal in a particular direction and suppress signals arriving from the other directions. Usually, *Array Beamforming* is combined with other techniques as *Independent Component Analysis* [3], *Spectral Subtraction* [4] or *Linear Prediction Analysis* [5].

A significant number of approaches take advantage of the LP residual signal [6], [7]. 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 amplitude spread of LP residual can serve as a reverberation metric. Besides, manipulation of the residual signal is more appropriate for short segments, as the residual samples are nearly uncorrelated. On the other hand, when speech signals are handled directly, the choice of window size and shape affects significantly the performance.

Through this paper, a speech enhancement system based on the use *Linear Predictive* error signal for reverberation and noise estimation purposes, and its application to *Spectral Subtraction* techniques [8] is presented. This method is intended to be a pre-processing stage of a binaural system [9] devoted to *Robust Speech Recognition* in automobile scenarios, as presented in Figure 1.

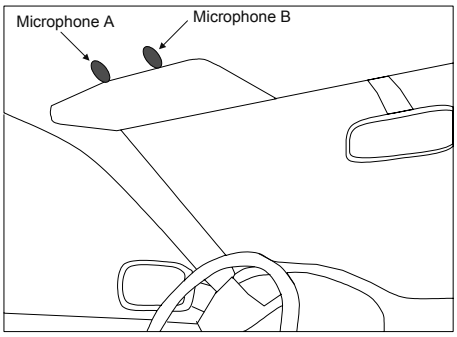

Figure 1. General framework of a two-microphone system devoted to robust speech recognition with directional source separation capabilities.

2. Reverberation and noise detection with LP residual signals

The detection of the reverberation and noise existing in portions of speech is achieved following the steps presented in Figure 2. Firstly, a measure of the kurtosis of the signal is estimated. The kurtosis function in not calculated from the speech signal itself, but using the linear-prediction analysis error signal. On a second step, kurtosis function is smoothed and mapped into a final weight function that represents the degree of degradation existing in the original signal.

2.1. Kurtosis estimation

The linear prediction residual signal $e_{LP}(t)$ is calculated applying the autocorrelation method to a speech frame of 20 ms. Previously, a Hamming window is applied. The order of the analysis depends on the sampling frequency. In our study, the sampling frequency is 8kHz so that 10^{th} order is enough. In Figure 4 may be seen the resulting signal extracted from a corrupted speech utterance presented in Figure 3.

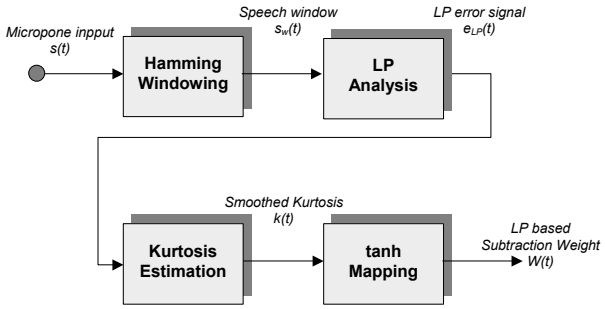

Figure 2. Algorithmic structure for the calculation of the *LPSW* from the LP residual signal.

In a following step, $e_{LP}(t)$ is blocked into 40 ms frames with an overlap of 20 ms. Kurstosis values k_n are then estimated for every block by:

$$k_n = \frac{E\left\{e_{LP}{}^4(t)\right\}}{E^2\left\{e_{LP}{}^2(t)\right\}} - 3 \tag{1}$$

As it is shown in Figure 5, the value is $|k_n|$ is then repeated as many times as the number of input samples contained in a 20 ms frame. This will produce a new function $k(t)$ defined for each sample. Finally, $k(t)$ is smoothed by mean filter with 512 samples (see Figure 6).

2.2. Linear Predictive Subtraction Weight

To calculate the Linear Predictive Subtraction Weight (LPSW), the smoothed function is mapped in the range [0.0, 1.0] using the following expression:

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

$w(t)$ being the desired output weight factor at time t, λ a weight gain factor and θ a threshold indicating the minimum kurtosis value linked with speech segments. This mapping function is shown in Figure 8.

3. Spectral subtraction

To implement the filtering in the spectral domain, the LPSW will be considered a relevant estimator. The procedure we proposed may be seen in Figure 9.

First of all, the input signal $s(t)$ is segmented in overlapped windows and transformed into the frequency domain using the short-time Discrete Fourier Transform $F\{.\}$:

$$S = S(m) = F\{s(n)w(n)\} \tag{3}$$

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

In a first step, the log power of the signal is computed for every frequency channel:

$$S_{\log}(m) = \log_{10}\left(\|S(m)\|^a\right); \quad 0 \le m \le M/2 - 1 \tag{4}$$

being and M the window size and m the frequency index.

After that, these values are passed to a filter with exponential decay given by:

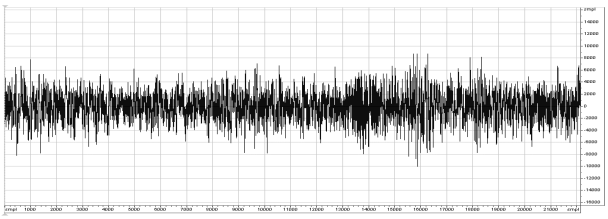

Figure 3. Segment of reverberant and noisy speech which has embedded an utterance produced of female speaker.

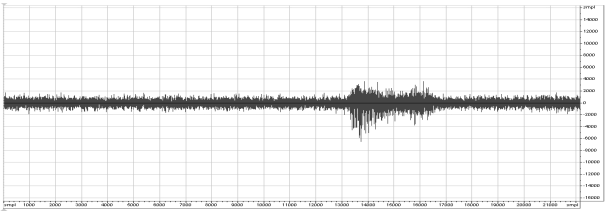

Figure 4. Linear Predictive error signal associated to the signal represented in Figure 3.

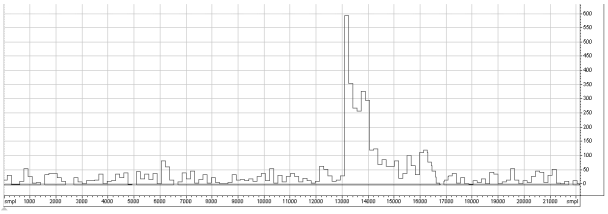

Figure 5. Estimation of kurtosis associated to the signal in Figure 4 (vertical axis values multiplied by 100).

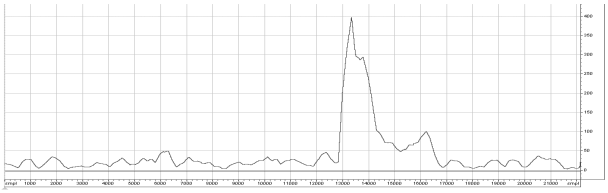

Figure 6. Smoothed kurtosis function (vertical axis values multiplied by 100).

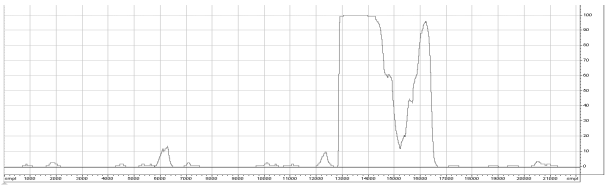

Figure 7. Weight function computed from the mapping of the smoothed kurtosis function (vertical axis values multiplied by 100).

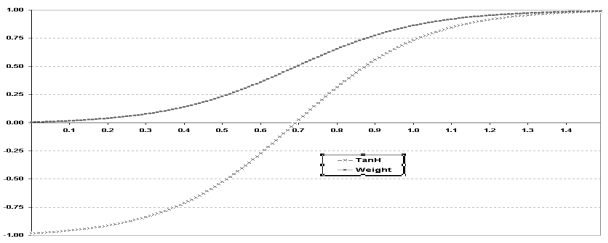

Figure 8. Mapping function used for estimation of LPSW with $\lambda = 5.0$ and $\theta = 0.7$.

$$g_S(m) = \alpha\gamma S_{\log} + (1-\alpha) g_{n-1}(m) \quad (5)$$

where α is a coefficient that controls the log-power rate update and γ 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 LPSW previously studied:

$$g_n(m) = (1-W_n) g_S(m) + W_n(m) g_{n-1}(m) \quad (6)$$

As may be noticed, the above expression implies that a weight equal to 1.0 prevents from updating the estimation of $g_n(m)$ at all. On the other hand, a weight close to 0.0 produces a fast adaptation.

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

$$G(m) = 10^{g_n(m)} \quad (7)$$

$$\|Z(m)\|^a = \|S_{en}(m)\|^a - G(m) \quad (8)$$

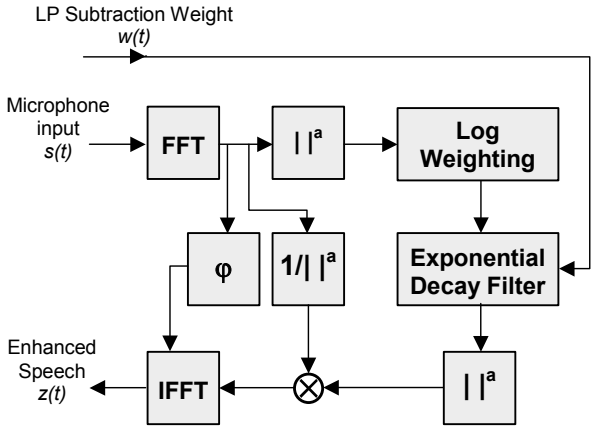

Figure 9. Structure of the spectral subtraction module that exploits the LPSW estimator.

4. Results and discussion

In a practical experiment shown in Figure 10 through Figure 14, the processing of a speech trace produced by a male speaker in a car environment is presented. Figure 10 shows a speech utterance corrupted by reverberation and noise. Figure 11 contains its associated spectrogram. After a first step, the LPSW is extracted (see Figure 12). As it may be seen, the measure provided by the weight derived from the LP residual signal is not completely accurate. However, it constitutes a reliable estimator for a further spectral subtraction process. The output of the frequency domain processing produces a new trace given in Figure 13, being its power spectrum the representation shown in Figure 14. These two figures give a clear idea of the potential both methods when combined.

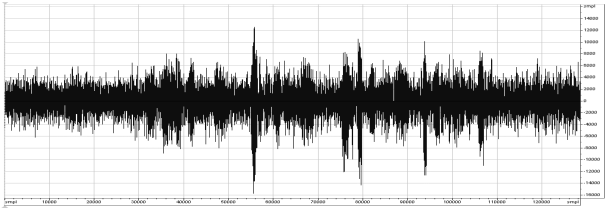

Figure 10. Degraded utterance of several connected digits produced by a male speaker.

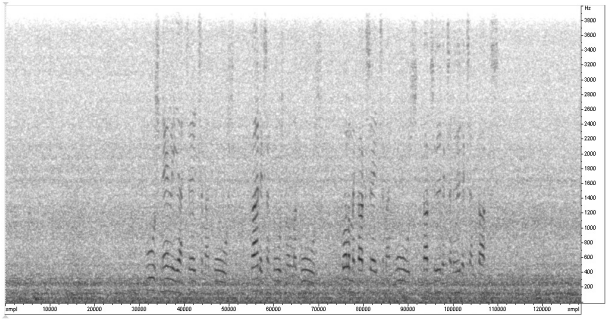

Figure 11. Power spectrum of the signal presented in Figure 10.

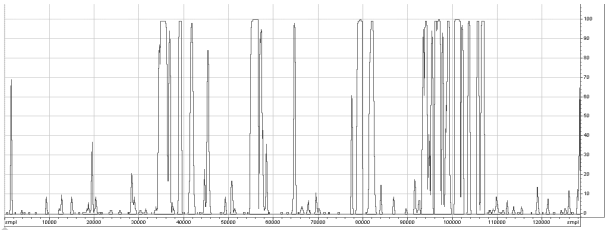

Figure 12. LPSW estimation associated to the speech signal contained in Figure 10 (vertical axis values multiplied by 100).

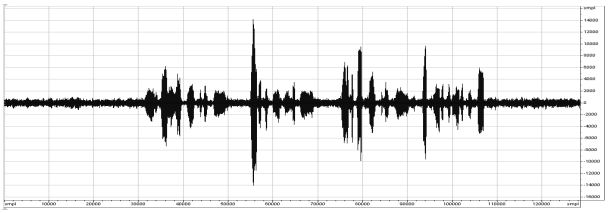

Figure 13. Enhanced output signal obtained applying the spectral subtraction method proposed after LPSW estimation.

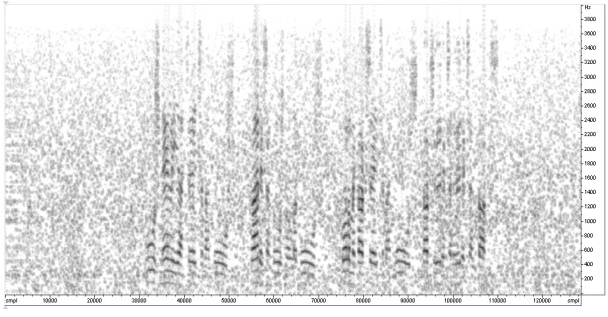

Figure 14. Power spectrum of the signal in Figure 13.

In order to examine the validity of the method proposed, several speech recognition systems are built and tested. A subset of the Aurora3-SpeechDat Car Finnish database [10] is used for these purposes. The corpus, which contains realizations of connected digits uttered in a realistic automobile environment, is divided in two different groups: train and test. Each group has three different categories related with the amount of distortion contained in the recordings: quiet, low, and high. In our experiments, we use the recordings associated to channels *ch2* (microphone placed at the ceiling of the car in front of the speaker behind the sunvisor) and *ch3* (microphone installed at the ceiling of the car over the mid-console and near the rear mirror). As it may be noticed, that configuration is exactly the same we previously introduced in Figure 1.

The recognition experiments are established by selecting different material from the training set of the database: 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 threes cases and consists on 3126 files. The front-end extracts energy plus 36 MFCCs (12 cepstrum, 12 delta cepstrum and 12 delta-delta coefficients). The HMMs are 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.

Table 1 and Table 2 present accuracy results when the method proposed in this paper is applied as a pre-processing stage to the same front-end. As it may be seen, the improvement is significant for both channels and the three training sets.

Original	Deletions	Substitutions	Insertions	WER
Train set A	104	491	916	48.34%
Train set B	62	71	234	11.74%
Train set C	62	53	196	9.95%

Enhanced	Deletions	Substitutions	Insertions	WER
Train set A	85	186	701	31.09%
Train set B	41	44	116	6.43%
Train set C	37	43	163	7.77%

Reduction	Deletions	Substitutions	Insertions	WER
Train set A	18.27%	62.12%	23.47%	35.67%
Train set B	33.87%	38.03%	50.43%	45.23%
Train set C	40.32%	18.87%	16.84%	21.86%

Table 1. Recognition results for microphone ch2.

Original	Deletions	Substitutions	Insertions	WER
Train set A	271	204	275	23.99%
Train set B	80	47	172	9.56%
Train set C	70	51	208	10.52%

Enhanced	Deletions	Substitutions	Insertions	WER
Train set A	87	130	287	16.12%
Train set B	36	35	110	5.79%
Train set C	29	41	141	6.75%

Reduction	Deletions	Substitutions	Insertions	WER
Train set A	67.90%	36.27%	-4.36%	32.80%
Train set B	55.00%	25.53%	36.05%	39.46%
Train set C	58.57%	19.61%	32.21%	35.87%

Table 2. Recognition results for microphone ch3.

5. Conclusions

The combination of *Spectral Subtraction* techniques with the *Linear Predictive Subtraction Weight* (LPSW), extracted from LP residual signal, constitutes an efficient approach to the speech enhancement problem in noisy and reverberant environments, as no a priori knowledge of the working framework is required. Speech recognition experiments 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 developed under grant TIC2002-02273 from the Programa Nacional de las Tecnologías de la Información y las Comunicaciones (Spain), and a collaboration contract between Universidad Politécnica de Madrid and the Centre Suisse d'Electronique et de Microtechnique.

7. References

[1] Yegnanarayana, B., Satyanarayana Murthy, P., Avendano, C., Hermansky, H., "Enhancement of Reverberant Speech Using LP Residual", *Proc. of ICASSP'98*, 12-15 May, 1998, pp. 405-408.

[2] Nordholm, S., Claesson, I., Bengtsson, B., "Adaptive array noise suppression of handsfree speaker input in cars", *IEEE Transactions on Vehicular Technology*, Vol. 42, No. 4 , November 1993, pp. 514- 518.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] A. Álvarez, P. Gómez, R. Martínez, V. Nieto and, V. Rodellar, "Speech Enhancement and Source Separation based on Binaural Negative Beamforming", *Proc. of Eurospeech 2001*, September 2001, pp. 2615-2618.

[10] 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.