Speech Enhancement Using Multi-Reference Noise Reduction in a Vehicle Environment

Abderrahman Essebbar and Tristan Poinsard

IMRA EUROPE S.A.S, 220, rue Albert Caquot, 06904 Sophia Antipolis Cedex, France
{essebbar,poinsard}@imra-europe.com

10.21437/Interspeech.2007-312

Abstract

This paper presents a multi-reference noise reduction system for use in a vehicle. It is aimed at enhancing the driver's speech in noisy driving environments for applications such as voice commands or hands free phone. First, our system objective is briefly presented as well as the problems of classical techniques, like Beamforming, may face in such a harsh environment as a vehicle cabin. Second, a brief analysis of noises aboard a road vehicle is done. Third, the noise reduction architecture comprising a linear and non linear block with two respective non acoustic noise references is presented. Finally, some results obtained in real driving conditions are presented and analysed. Both human listening tests and speech recognition tests prove our system increases global performances compared to what is obtained with classical single channel speech processing methods.

Index Terms— *Speech Enhancement, Noise Reduction, Non Acoustic Sensors, Noise Reference, Automotive*

1. Introduction

The vocal interface is one of the most natural means for the driver or the passenger to communicate with other parts of the system and recently speech-based interface in vehicles have become a popular means for improving the accessibility of in-vehicle information equipment [1][2].

We propose a noise reduction system for in-vehicle applications such as voice command or hands free phone. The system is aimed at reducing the car radio sound and the driving noise before a Vocal Human Machine Interface (VHMI). Hence, the system is aimed at improving communication through the VHMI in normal and harsh driving conditions and without turning the radio off.

Noise aboard cars can be very disturbing and often vocal interfaces fail to work properly. Noise is non-stationary and is a combination of many different sources with totally different properties. Moreover, Signal to Noise Ratio (SNR) can be relatively low in some conditions (open window, passing car, motorway).

Many researches have been done to address the problem of noise reduction in a vehicle. Classical filtering techniques [3] (Wiener filter, Kalman filter, spectral subtraction, etc.) are usually used to reduce non directional interference or noise signals. These techniques generally use a single microphone approach and the noise reference is picked up during the assumed noise-only periods of the signal, i.e. when no speech is received. These periods are detected by a Voice Activity Detector (VAD). However, because these techniques assume that the noise that the noise reference doesn't contain any contribution of the signal of interest, the VAD performance rapidly becomes the weak point of such systems. An alternative approach consists in using beam forming techniques [4].

However, several problems might arise in a vehicle environment such as reverberation and multiple reflections. Furthermore, the system complexity is increased by increasing the microphones number and also another single channel noise reduction system is usually applied on the Beamformer output signal.

In this paper we focus on multi-reference noise reduction by using non acoustical sensors. The instrumentation aboard the vehicle is the following:

- One driver microphone situated close to the driver, for example near the windscreen or dashboard.
- A simple electronic front end to pick the car radio signal before loudspeakers and summing the two left and right signals.
- A vibration sensor, typically an accelerometer fixed on the car body.

A smart architecture cascading a linear and a non linear filtering process is proposed. The evaluation of our proposal was done by listening tests and comparison with a classical single channel method. We also used a speech recognizer to show that our proposed system increases the speech recognition rate.

2. Analysis of Noise Source

Noise propagation inside a car is relatively complicated. Noise sources are of different type and have different origins.
Some noise sources can be seen as localised spatially inside the car. This means their propagation between the noise source and the microphone is relatively simple and probably linear.

Typically this is the case with a car radio sound system. Although some reflections are possible, most of the signal follows a direct path from the loudspeaker to the microphone.

Most of the noise sources aboard a vehicle are however non localised sources and it is difficult, sometimes impossible to consider a simple transfer path between the noise source and the microphone. Hence, the wind noise inside the car is generated by air flow all around the car body. The engine noise for example has also a complex behaviour as vibrations propagate in different structures of the vehicle.

3. Noise Reduction Architecture

3.1 System Architecture

The presented architecture depicted in Figure 1 is composed of two filtering blocks:

- A first linear filter block is used to reduce the car radio noise (localised source)
- A second non linear block is used to remove the noise generated by the car motion (non localised source)

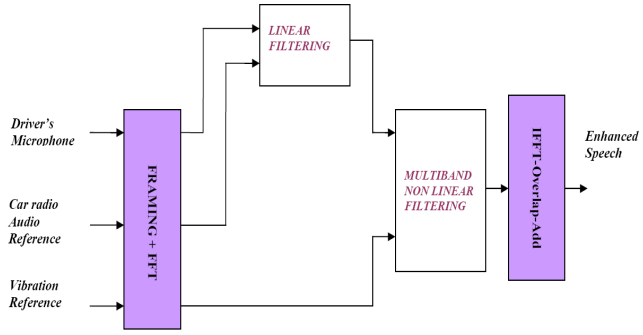

Figure 1: *Multi-reference Noise Reduction Architecture*

A simple model of the received noisy speech at sampled time *n* is the following:

$$m(n) = s(n) + g_L(a(n)) + f_{NL}(v(n)) + b(n) \qquad (1)$$

Where:

- *m(n)* is the received noisy speech by the driver microphone,
- *s(n)* is the original clear speech,
- *a(n)* is the audio reference signal picked before the loudspeakers,
- *v(n)* is the vibration reference signal received by the vibration sensor,
- $f_{NL}(n)$ is a non linear transfer function,
- $g_L(n)$ is a linear transfer function,
- *b(n)* is additive noise.

After Fast Fourier Transform (FFT) on overlapping frames of 15 ms duration we have for each frame *i* and for each frequency bin *k*:

$$M_i(k) = FFT(m_i) \quad , \quad S_i(k) = FFT(s_i) \quad ,$$
$$A_i(k) = FFT(a_i) \, , V_i(k) = FFT(v_i) \text{ and } G_i^L(k) = FFT(g_L^i)$$

3.2 Linear noise reduction block

The audio reference signal is picked at the output of the car radio system before the loudspeakers. In theory, a dual reference should be used for stereo signals. However, a reference comprising a sum of both left and right signals proves to give very good results for mono as well as for stereo signals. Therefore this sum is used as the audio reference. High coherence between the microphone and the reference signal enables the use of a linear filter based on the Wiener theory [5] [6].

Our algorithm is not adaptive which means that the filter update is independent of the output and only depends on the inputs $A_i(k)$ and $M_i(k)$. We note $\hat{S}_i^L(k)$ the enhanced noisy speech signal after this linear filtering block and $\hat{G}_i^L(k)$ an estimation of the transfer function by the Wiener filter algorithm. Below in Figure 2 is depicted the first filtering block used for the car radio system.

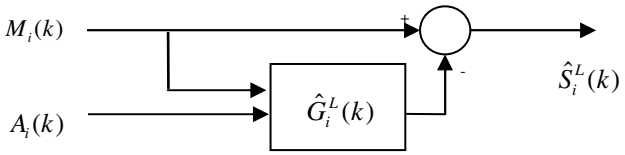

Figure 2: *Linear filtering block for reducing car radio noise*

Without going into details which can be found in [6], below is explained how the linear transfer function $G_i^L(k)$ is estimated.

In the following equations $\gamma_i^{MA}(k)$ and $\gamma_i^{AA}(k)$ are the estimation of respectively the cross-spectrum between the noisy speech signal and the audio reference signal, and, the spectrum of the audio reference signal on each frame *i* and for each frequency bin *k*.

$$\gamma_i^{AA}(k) = \lambda\gamma_{i-1}^{AA}(k) + (1-\lambda)(A_i(k)A_i^*(k)) \qquad (2)$$

$$\gamma_i^{MA}(k) = \lambda\gamma_{i-1}^{MA}(k) + (1-\lambda)(M_i(k)A_i^*(k)) \qquad (3)$$

where: λ is a forgetting factor with 0< λ <1 and * is the symbol for complex conjugate then:

$$\hat{G}_i^L(k) = \frac{\gamma_i^{MA}(k)}{\gamma_i^{AA}(k)} \qquad (4)$$

Finally and to avoid some echoes in the estimated speech signal, the magnitude and phase of the desired signal are estimated separately.

Firstly, the magnitude of the speech signal is estimated by subtracting the magnitude of the estimated contribution of the radio noise from the magnitude of the microphone signal. Then the phase is directly recovered from the microphone signal as shown in the following equation:

$$\hat{S}_i^L(k) = \left(|M_i(k)| - |\hat{G}_i^L(k).A_i(k)|\right)\exp(\, j.angle\,(M_i(k))) \qquad (5)$$

With $|.|$ representing the modulus and *angle* representing the phase of the signal.

3.3. Non Linear noise reduction block

In Figure 3 is presented the second processing block which is non linear. The input of this block is the output of the preceding block hence $\hat{S}_i^L(k)$. The vibration reference signal $V_i(k)$ is also used. We note $\alpha_i(k)$ a calibration gain (in the frequency domain), $G_i^{NL}(k)$ a spectral gain and $\hat{S}_i(k)$ the enhanced noisy speech after both linear and non linear filtering blocks.

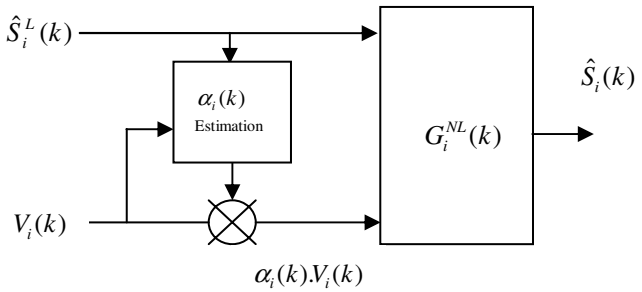

Figure 3: *Non Linear filtering algorithm Concept*

3.2.1 Spectral Gain Estimation

If we assume that much of the noise received by the microphone (engine, rolling noise) is also received by the vibration sensor then the idea is to use the vibration sensor as a noise reference to suppress the noise contained in the noisy signal at the output of the linear filtering block by using a spectral subtraction based method. The point here is that the noise components at the output of the linear block and the vibration reference signal components are correlated in the spectrum power but without being coherent.

The spectral gain $G_i^{NL}(k)$ has two entries and is a non linear function:

$$G_i^{NL}(k) = G_i^{NL}\left[\hat{S}_i^L(k), \alpha_i(k).V_i(k)\right] \qquad (6)$$

$G_i^{NL}(k)$ may be similar to the gain used in single channel spectral subtraction [7]. The estimated speech signal $\hat{S}_i(k)$ is calculated as follows:

$$\hat{S}_i(k) = G_i^{NL}\left[\hat{S}_i^L(k), \alpha_i(k).V_i(k)\right]\hat{S}_i^L(k) \qquad (7)$$

3.2.2 Calibration gain estimation

The calibration gain $\alpha_i(k)$ is an estimation of a gain between $|V_i(k)|^2$ and $\left|\hat{S}_i^L(k)\right|^2$. Estimating $\alpha_i(k)$ is equivalent to estimating the power of the noise received by the microphone that is also received by the vibration sensor. One main feature of the calibration gain is that it is slow varying in time: in normal driving conditions, the signal's power variation in the microphone is roughly proportional to the signal's power variation in the vibration sensor. This is due, as said previously, to the correlation in the power spectrum.
Experiments proved that on short time periods of around 1 s, this assumption is verified. If it isn't the case, it means that the signal's power in the microphone has changed independently of the signal's power in the vibration sensor. A dramatic sudden mismatch between the two power levels generally happens in identified cases:

- Noise Environment has changed due to sudden events not identified by the vibration sensor: passing car, horn, sudden window opening.
- Presence of speech

$\alpha_i(k)$ is not updated if the variation of signal's have a too high mismatch.

In other cases such as speed variation in normal driving conditions, road condition change, acceleration, etc., the power of both signals will vary slowly in time and globally proportionally and $\alpha_i(k)$ will be updated regularly to take account of the non linearity of the sensors and long term environment change.

Experiments show that update should be performed if the mismatch is under a 20 % variation. There are several possibilities to estimate the calibration gain. They are based on energy and power estimation and can also take into account other information such as vehicle speed. It is important to note that these methods avoid using a VAD and in presence of speech, $\alpha_i(k)$ will not be updated. However even when $\alpha_i(k)$ is not updated on some frames, continuous processing is performed, hence the vibration sensor information is continuously used on all frames for the noise reduction.

Splitting signals into bands has proved increased performance for classic spectral subtraction. Several experiments on our system confirmed this tendency. On each band, the non linear filtering block described in the preceding paragraph is applied and for each band a calibration gain is calculated.

4. Results

This section analyses the effectiveness of the global architecture in a real driving situation where the car is rolling on a main road at around 80 km/h with the radio on. The driver speaks four times during this 30s acquisition at 12kHz frequency sampling. First, temporal signals and spectrograms are presented to compare the raw audio signal with classical noise reduction Ephraim-Malah[7] and the proposed multi-reference noise reduction Secondly, speech recognition tests are presented.

4.1. Temporal and Spectrogram Analysis

In Figures 4 and 5 we can see that our system performs better than a classical Ephraim-Malah noise reduction system. Non stationary noises such as radio or transient noises are not well suppressed with the Ephraim-Malah algorithm.

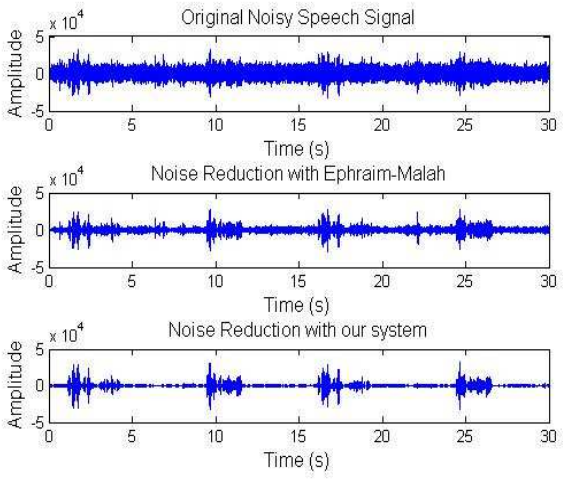

Figure 4: *Temporal signals before and after noise reduction*

Listening tests show that the proposed system comprising two different noise references (car radio noise and vibration noise) is able to reduce such noises and enhance the driver's speech compare to single microphone noise reduction system.

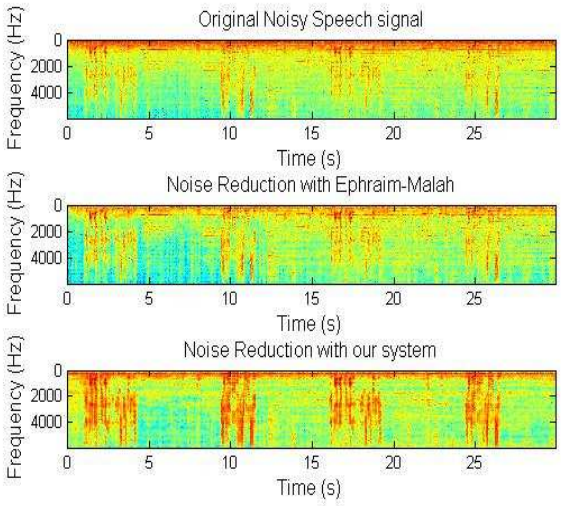

Figure 5: *Spectrogram before and after noise reduction*

4.2. Speech Recognition scores

Our objective is not to show the ASR (Automatic Speech Recognition) performances but that the multi-reference noise reduction system increases the speech recognition rate. A Speaker Dependant (SD) data base with voice commands for vehicle [8] was built by mixing voice with noises recorded in different driving conditions. The HTK ASR engine is a freeware originally developed by the University of Cambridge [9] and was used to evaluate our system (Figure 6).

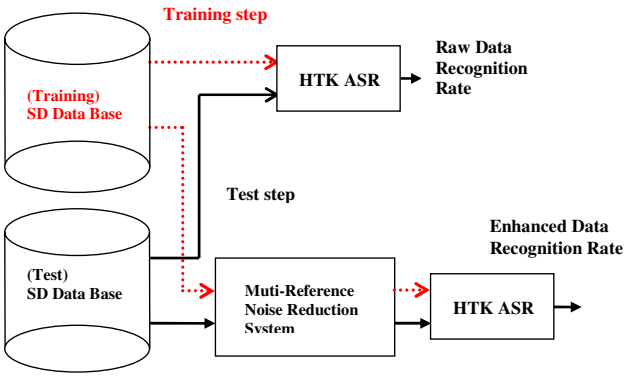

Figure 6: *Speech Recognition Evaluation Procedure*

	Raw data without Noise Reduction			With Noise Reduction ASR trained on the filtered data		
SNR (dB)	15	10	0	15	10	0
All noises (scores in %)	81	76	60	85	82	71

Table 1. *Speech recognition score results*

The VAD fails to detect voice in the car radio noises. And the classical Ephraim-Malah algorithm is not able to remove such noises. In Table 1, we present only results using the proposed system with the HTK ASR engine. The training data base contains car radio noises (talks, songs), road noises (engine, wind, tyres,..) and some transient noises happening while driving (squeaks, passing car, wiper, etc). We observe that at low SNR (0 dB), recognition score is increased by about 20% to 30%.

We could also improve the linear filtering performances by using a multi-band coherence threshold. In this case, a filter would be applied in the band only when the coherence value is higher than 0.3 for example.

A real time processing using a PC has been implemented and validated aboard a car in different driving situations.

5. Conclusions

The proposed multi-references noise reduction system can drastically reduce polluting noise from the car radio system. This is particularly useful when using voice commands, especially when disturbing noises from the car radio contains speech that might interfere with the driver's speech.

Concerning other noises generated by the car motion such as rolling or engine noise for example, our system is more adapted than single channel methods to non stationary situations and avoids using a conventional VAD.

6. Acknowledgements

The authors would like to thank Michel Gaeta, Takashi Mitsumuto and Yuchi Murakami (from AISIN Seiki) for supporting this project.

7. References

[1] T. Wakita, "Speech based interfaces in Vehicle", *R&D Review of Toyota* CRDL Vol.39 No.1, March 2004.
[2] D. Van Campernolle, "Future Directions in Microphone Array Processing" in. M.S. Brandstein and D.B. Ward, editors, *Microphone Arrays: Signal Processing Techniques and Applications*, Chapter 18, pages 389–394. Springer, Berlin, 2001.

[3] S. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, John Wiley & Sons Ltd, 2000.
[4] J. Bitzer and K.U. Simmer. Superdirective microphone arrays. In M.S. Brandstein and D.B. Ward, editors, *Microphone Arrays: Signal Processing Techniques and Applications*, Chapter 2, pages 19–38. Springer, Berlin, 2001.
[5] B. Widrow and S.D. Stearns. *Adaptive Signal Processing*. Prentice-Hall, New Jersey, 1985.
[6] A. Essebbar, T. Poinsard, Y. Murakami and Y. Suwa, "Radio Noise Reduction System for Vocal Human Interface Improvement in a Vehicle", *ITS'06 Conference London*, October 8-12, 2006.
[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443-445, April 1985.
[8] A.W. Gellatly, "The Use of Speech Recognition Technology in Automotive Applications". Dissertation, *Virginia Polytechnic Institute and State University*, 1997.
[9] http://htk.eng.cam.ac.uk/