
A. Guérin,
LTSI, Université de Rennes 1
Bat 22, 7ème étage, campus de Beaulieu, 35042 Rennes Cedex, France
email: alexandre.guerin@univ-rennes1.fr

ABSTRACT
This report proposes a 2 microphone Voice Activity Detector (VAD) and a Speech Enhancer (ENH) adapted to car conditions. The two modules are derived from the well-known Magnitude Square Coherence (MSC) which expresses a normalized cross-correlation for each frequency band of the received signals by the two sensors. A global VAD is directly obtained from the MSC by adaptive threshold which ensures a quasi-constant behaviour in different environmental conditions and different relative microphones positions. The ENH filter is applied to one of the two microphones and is divided in two parts:

1. from a Modified Coherence Function including Power Cross–Subtraction of background noise estimation is based on the Wiener Filter that is used to enhance speech; the estimation of Power Spectral Densities (PSD) are optimized to prevent the emergence of musical noise as well as reverberant effect.
2. a second module extracts the pitch value of voiced sections of speech to enhance low frequency bands of main signal that are partially or even totally removed by Wiener Filtering.

1. BACKGROUND
1.1 Notations
Let’s consider the following notations, t and f parameters expressing respectively time and frequency dimensions:

\[ b_t(t), s_t(t), x_t(t) = b_t(t) + s_t(t) : \text{noise, signal (speech)}, \text{and signal picked up by } i^{th} \text{ microphone}, \]

\[ N : \text{number of frequency bands determined by the order of FFT}, \]

\[ f_i : i^{th} \text{ frequency band}, \]

\[ B_i(f_t), L_i(f_t), X_i(f_t) = B_i(f_t) + L_i(f_t) : \text{Fourier Transform of previous signals}, \]

\[ \Gamma^o_i(f_t) = E[V_i(f_t)V'_i(f_t)], \Gamma^s_i(f_t) = E[V_i(f_t)V'_s(f_t)]: V \in \{S,B,X\}, \text{auto and cross-power spectral densities of V signal (ENH)}, \]

\[ \hat{x} : \text{estimate of } x \text{ variable, } x \in [s_1, s_2], G(f_t) : \text{gain derived from ENH}, \]

1.2 Generalities
The system proposed in this paper is expressed in the spectral domain. The filter bank for spectral analysis is realized as a weighted FFT, while the synthesis filter bank is based on inverse FFT and weighted Overlap and Add techniques (OLA). Therefore, in the expression \( X(f, t - 1) \), the \((t-1)\) term means the FFT of the previous frame.

1.3 Hypothesis
We will assume the following hypothesis:

1. Independence of speech and background noise:

\[ E[B_i(f_t), S_i(f_t)] = 0, \forall (i,j) \in \{1,2\}^2 \]

2. Same PSD for each signal on the 2 microphones:

\[ \Gamma^o_i(f_t) = \Gamma^o_2(f_t), \Gamma^s_i(f_t) = \Gamma^s_2(f_t). \]

1.4 Magnitude Square Coherence
The coherence function and the MSC [1] are defined by:

\[ \rho(f) = \frac{\Gamma^s(f)}{\sqrt{\Gamma^o_1(f)\Gamma^o_2(f)}}, \quad \text{MSC} = |\rho(f)|^2 \quad (1) \text{ and (2)} \]

Using the hypothesis, we can derive another more intuitive expression of the coherence function:

\[ \rho(f) = \frac{\Gamma^s(f) + \Gamma^o_2(f)}{\Gamma^o_1(f) + \Gamma^o_2(f)} \quad (3) \]

**Interpretation:**
From equation (3), it results that in presence of completely decorrelated noise, the coherence function is simply equal to the Wiener Filter. With diffuse noise (more realistic case), the coherence distribution has a cardinal sine shape: above a threshold frequency determined by the distance between the two sensors, the coherence of these signals is quite null, whereas below this threshold, the coherence tends to 1 with decreasing frequency. This high correlation of noise tends to deteriorate the estimation of the Wiener filter by the coherence function.

1.5 PSD estimation: role of forgetting parameter \( \alpha \)
Approximation of the expectation operator is made by the recursive periodogram method (first-order IIR filtering):

\[ \Gamma^o_{\alpha}(f, t) = \alpha \Gamma^o_{\alpha}(f, t - 1) + (1 - \alpha)X_i(f, t)X'_i(f, t) \quad (4) \]

The forgetting parameter \( \alpha \) is the center point of the estimation of the PSD: a high value for \( \alpha \) is well adapted to stationary signals like previously specified noises. Nevertheless, for short-term stationary signals like speech, a too high value biases the estimation due to the smoothing effect.

This variability of signals statistics implies that the forgetting parameter \( \alpha \) is determined depending on the signal to be analyzed (3.1) and on the use of PSD (2.2).

1.6 Noise conditions
In car conditions, noise sources are numerous: principally aerodynamic, rolling, blowing (open window) and engine. Observations show that these noises can be considered as diffuse. From a spectral distribution point of view, the first three ones may be...
seen as relatively wide-band equally distributed, on the contrary to the engine noise whose energy is concentrated in the low frequency bands.

2. VOICE ACTIVITY DETECTOR

The VAD exploits the physical characteristics of real car noises which are globally diffused. This fact implies a weak spatial coherence, compared to punctual source signals like speech. The VAD automaton structure stands in:

→ the adaptive threshold. This parameter expresses the maximum coherence of the environmental noise. It is logically computed in noise only mode.

→ the decision parameter. It expresses a global coherence characteristic of the observed signal. This parameter is computed every frame, in voice activity and noise only modes. It is then compared to the threshold to determine the presence (decision parameter greater than threshold) or the absence (resp. lower) of speech.

The aim of the VAD is to detect every period of speech. As the binary result (speech or no-speech) will be blindly exploited by the ENH module, the strategy implies no mis-detection, to the prejudice of some false-alarms.

2.1 Forgetting parameter $\alpha$

In VAD case, there is no point in considering the problem of bias in the estimation of PSD. The smoothing factor $\alpha$ should only guarantee a discriminative decision parameter.

2.2 Decision Parameter

Interpretation of equation (3) gives one main condition to achieve so as to derive an efficient VAD: with decorrelated noise, the $\text{MSC}(f)$ remains relatively low in noise only conditions and reaches $\text{SNR}(f)$ value in presence of speech.

The idea is to determine the domain which guarantees a sufficient low noise coherence:

Analyzing car noise recording and even if the two sensors are very close (that factorizes coherence), a quasi-complete decorrelation of noise signals, above a threshold frequency $f_{\text{threshold}}$ that depends on the microphones distance, can be assumed.

Moreover, spectral distribution of car noise being rather low frequency, the $\text{SNR}(f)$ in high frequencies are favorable, even in very adverse conditions like open window and with a non-energetic voice (female voice).

2.3 Adaptive threshold

As explained in the introduction, the VAD has to be efficient in all car conditions, including different relative microphones position and direction (installation of such system may be variable).

A fixed threshold is not realistic for two main reasons:

1/ relative distance and direction of microphones influence the mean coherence of noise, and also the variance,

2/ for given direction and position of microphones, the behavior of the decision parameter is different in open or closed window: mean and especially variance are higher in open window (cf. figure 2). A fixed threshold implies a too strong compromise between numerous false-alarms in open window conditions and the delay of the rising-edge decision in closed window conditions: the behaviour of the VAD would not be optimal in one of these two conditions.

The decision parameter is mobilized by its first and second order moments: mean and standard deviation. Their computation is obtained by a recursive method by first-order IIR filtering:

\[
\frac{\text{MSC}(t) - \alpha_{\text{mean}} \cdot \text{MSC}(t - 1) + (1 - \alpha_{\text{mean}}) \cdot \text{MSC}(t)}{\text{MSC}(t - 1) - \alpha_{\text{dev}} \cdot \text{MSC}(t - 1) + (1 - \alpha_{\text{dev}}) \cdot |\text{MSC}(t) - \text{MSC}_{\text{mean}}(t)|}
\]

Figure 2 exhibits the behavior of the $\text{MSC}$ for closed and open window conditions. The periods of presence of speech may be easily detected.
\( \text{MSE}_{\text{corrected}}(t) = \text{MSE}_{\text{meas}}(t) + \text{coeff} \cdot \text{MSE}_{\text{dec}}(t) \)

The term coeff is a constant used to make the threshold express a maximum coherence of noise; the probability its coherence exceeds this limit is quasi-null.

### 2.3 Results

The figure (3) depicts an example of the behavior of the VAD in adverse conditions (negative SNR).

![Figure 3. VAD](image)

The overall results are:

- **Mis-detection**: no, in every conditions.
- **False-alarms**: about 35% of detections (depending on conditions) are false alarms. Nevertheless, the majority of the detection errors shows a duration which does not exceed more than 4 frames (32ms). The remaining false-alarms correspond to punctual source noises.

### 3. SPEECH ENHANCER

ENH module is an estimator of signal amplitude, the phase of the noise-free signal being approached by the noisy signal one. The aim is to estimate the zero-phase Wiener filter:

\[
G_{\text{eye}}(f) = \frac{\hat{F}_S^2(f)}{\hat{F}_N^2(f) + \hat{F}_W^2(f)}
\]

Comparing equations (3) and (5), the only difference stands in the presence of the cross-correlation of noise.

The method used to derive the Wiener filter from the coherence operator is inspired from Power Spectral Subtraction: to remove the noise coherence, a Power Cross–Spectral Subtraction is applied to the coherence numerator:

\[
G(f) = \frac{|F_{12}^x(f)|^2 - |F_{12}^\phi(f)|^2}{|F_{12}^x(f)|^2 - |F_{12}^\phi(f)|^2}
\]

The equation (7) gives two main distinct points to be considered:

1. estimation of the different PSDs that depends on the forgetting factor \( \alpha \).
2. estimation of the cross-correlation of noise \( \hat{r}_{12}^\phi(f) \).

### 3.1 Forgetting parameter \( \alpha \)

Firstly, it is important to point out that the hypothesis (quite representative of physical reality) made on speech and noise induces a correspondence between signal (resp. noise signal), short-term stationary characteristic (resp. long-term).

Simulations showed that, in high SNR conditions (> 20dB) and having perfect estimation of noise cross-correlation, the use of a medium \( \alpha \) to calculate PSD and filter defined by equation (7) was creating an annoying reverberant effect (\( \alpha \) too high).

On the contrary for low SNR, not only this value does not create reverberant effect, but also some musical noise appears (\( \alpha \) too low).

These observations lead to the conclusion that, to estimate relevant PSD, \( \alpha \) has to take into account the degree of stationarity of the signal picked up by the microphones. Using the correspondence established above, \( \alpha \) must finally depend on the SNR:

- for noise only (low SNR), a high value gives a relevant PSD estimation.
- for speech only (high SNR), a small value is necessary to follow the rapid variations of voice.
- in intermediate situations (medium SNR), a soft-decision is applied

Secondly, due to the wide-band spectrum characteristic of car noise compared to narrow-band spectral distribution of speech (presence of speech reduced to the formants, the pitch and its harmonics), a global SNR on the whole spectrum is not representative of each frequency. This implies to derive one \( \alpha(f) \) for each frequency band \( f \) depending on an estimation of \( \text{SNR}(f) \).

We carried out simulations (with separate noise and speech files to estimate the SNR) in order to determine the dependence of \( \alpha(f) \) on the \( \text{SNR}(f) \).

These simulations exhibited a simple law:

\[
\alpha(f, t) = 1 - \sqrt{\text{SNR}(f, t)}
\]

It remains to estimate the SNR in each frequency band. Based on the hypothesis that the SNR does not much vary from one frame to the other due to the stationary characteristic of the signals, \( \text{SNR}_{\text{eye}}(f, t) \) is estimated by \( \text{SNR}_{\text{eye}}(f, t - 1) \), which is simply \( G(f, t - 1) \).

In order to avoid too long or short term estimation, two thresholds, \( \alpha_{\text{min}} \) and \( \alpha_{\text{max}} \) were imposed.

The forgetting parameter is finally expressed by the formula:

\[
\alpha(f, t) = \alpha_{\text{min}} \text{ if } (1 - \sqrt{\text{SNR}(f, t - 1)}) \geq \alpha_{\text{min}},
\]

\[
\alpha(f, t) = \alpha_{\text{max}} \text{ if } (1 - \sqrt{\text{SNR}(f, t - 1)}) \geq \alpha_{\text{max}},
\]

\[
\alpha(f, t) = 1 - \sqrt{\text{SNR}(f, t - 1)}, \text{ in other cases.}
\]

### 3.2 Noise Cross-Correlation Estimation

The used method to estimate the Wiener filter (Power Subtraction like) has a major drawback: with under-estimated noise Power Spectral Densities, some disturbing musical noise is introduced. This point will influence the estimation of noise PSD.

Indeed, with the adaptive forgetting coefficients \( \alpha(f) \), variances of the estimations of the PSD are more important; especially the cross-PSD estimations (cross-correlation) which, with small \( \alpha \), approach instantaneous power estimations. To avoid cross-PSD noise under-estimation, whose main consequence is the emergence of musical noise, these will be approximated by PSD. This may be expressed as, instead of using equation (4) applied to noise:

\[
|r_{12}^n(f, t)| = \alpha |\hat{r}_{12}^n(f, t - 1)| + (1 - \alpha) b_1(f, t) B_2^2(f, t)
\]

The biased following formula is used:

\[
|\hat{r}_{12}^n(f, t)| = \alpha |\hat{r}_{12}^n(f, t - 1)| + (1 - \alpha) [b_1(f, t) B_2^2(f, t)]
\]

This strategy makes the estimation coherent with coherent noise signals, typically in low frequencies bands. In these bands, the over-estimation is small and prevents from too much speech distortion because of low SNR (the critical point in the low frequencies).

For incoherent noises whose spectrum is relatively high, the over-estimation is important in absence of speech. Nevertheless, in these bands, SNR are high and the estimation of cross-PSD approaches PSD when speech appears and the noise estimation becomes relevant. Therefore, no musical noise in absence of speech (in case of false-alarm) occurs and speech is only barely distorted.

The noise estimation algorithm is structured as following:

\[
\text{VAD}=0 \text{ : it is ensured that there is no voice. The estimation of noise is simply derived from the equation (8), replacing } B_t \text{ by } X_t \text{ (signals}
\]
picked up are noise only): 
\[ |P(f,t)| - \alpha_{\text{max}}|P(f,t-1)| + (1 - \alpha_{\text{max}})|X(f,t)X^2(f,t)| \]

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{alpha_function.png}
\caption{"alpha" function}
\end{figure}

VAD=1 : voice is almost surely present (even if some few frames may contain noise only). The estimation is made by a recursive algorithm whose equation is:
\[ |P(f,t)| = \text{alpha} \left( \frac{|X(f,t)X^2(f,t)|}{|P(f,t-1)|} \right) |P(f,t-1)| \]

Based upon the a-priori SNR, the behaviour of this algorithm is coherent with equation (8) for low SNR (with same \( \alpha \)), and quasi-freezes noise estimation for high ones; for medium SNR, a soft decision is applied.

**interpretation:**
- **high SNR**: \( \alpha(SNR_{\text{rho}}) = 1 \), and \( |P(f,t)| = |P(f,t-1)| \). The estimation is frozen.
- **low SNR**: \( \alpha(SNR_{\text{rho}}) = SNR_{\text{rho}} \) and we get the estimation:
\[ |P(f,t)| = \alpha_{\text{max}}|P(f,t-1)| + (1 - \alpha_{\text{max}})|X(f,t)X^2(f,t)| \]
This equation corresponds to estimation in noise only mode (eq. (8)).
- **medium SNR**: a soft-decision is applied.

### 3.3 Low Frequencies Enhancement using Pitch

Results of simulations pointed out a great improvement of speech quality compared to mono-microphone techniques. Although, analysis showed that **low frequencies of speech are systematically under-estimated**, especially for **voiced segments** in adverse conditions (high speed). The reason of this under-estimation is relatively simple: a low SNR estimation implies high \( \alpha(f) \), therefore the estimation of the PSD is not well adapted to speech short-term stationary characteristic. The estimation of the PSD then contains much more noise information than speech one.

These under-estimations result in a “nasal” voice. The estimation of SNR in low bands, based on the only PSD and cross-PSD, is proved to be inaccurate. To mitigate the effects of these wrong estimations, a new concept is employed:

1. To detect by another mean that power information the presence of low frequencies of speech which correspond, for voiced sections, to pitch and its first harmonics.
2. To enhance these detected frequency bands with a given rule.

**1. masking effect.** A minimum gain \( G_{\text{min}} \) is computed depending on the global SNR: it stands for the a priori minimum SNR in the worst frequency bands (lowest bands) where speech signal is present.

**2. the overall spectral characteristics of signals.** Due to globally increasing SNR with growing frequency, a \( G_{\text{inf}}(f) \) is determined, depending on statistical studies: it stands for the a priori minimum SNR in the frequency band \( f \), that is to say the minimum gain to apply to this frequency band.

**Results:**
The figure (5) exhibits the efficiency of low frequencies enhancement based on pitch information.

In window closed conditions, the voice quality improvement due to the pitch enhancement is undeniable. It is characterized by two main improvements:

- **Energy increase**: the pitch enhancement reduces by 10 to 30 % the global energy error of the estimated signal.
- **Cepstral distance**: in the same time, the cepstral distance between the estimated signal and the free-noise signal is reduced by 5 to 10%.

### 4. CONCLUSION

The contribution presented a two-microphone combined Voice Activity Detector and a Speech Enhancer based on Magnitude Square Coherence. The proposed system takes advantage of the physical characteristics of the signals in a car environment to derive an efficient VAD in very adverse conditions and in different installation situations (different microphones distances were tested). Concerning the Speech Enhancer, the use of two sensors, combined with PSD estimations relevant to signal statistics brings great objective improvements compared to conventional mono-sensor techniques, confirmed by listening tests. With the Pitch Enhancer module, the problem of estimation of speech low frequencies is quite overcome: the speech is no longer nasal, but on the contrary recovers its original timber and its sultry character.

### 5. REFERENCES


**Figure 4. \( \alpha \) function**

**Figure 5. Enhancement of pitch and harmonics.**