NOISE REDUCTION USING PERCEPTUAL SPECTRAL CHANGE.

C. Beaugeant, P. Scalart
FRANCE TELECOM CNET
CNET/DIH/DIPS 2, Av. Pierre Marzin, 22307 Lannion CEDEX, FRANCE
E-mail: christophe.beaugeant@cnet.francetelecom.fr, pascal.scalart@cnet.francetelecom.fr

ABSTRACT
This article deals with the problem of noise reduction for
hands-free sound pick-up. Our context is more precisely
the sound pick-up using single microphone in cars. In such
environments, the speech is highly corrupted by engine and
aerodynamics noises. The methods proposed here try to
enhance noise reduction systems based on Wiener filtering
in the frequency domain by taking into account the short
time variations of speech. It aims in fact at adapting the
estimation of the signal spectrum and the estimation of the
noise spectrum to the spectrum variations of the noisy
signal. By preserving voice attacks, distortion on the
useful signal is reduced as shown in this article. Background residual noise and musical tone phenomenon
are also discussed about.

1. INTRODUCTION.

In moving car environment, numerous noise sources highly
degrade sound pick-up, especially at high speed. In many
applications, such as mobile voice communication or
speech recognition, the need of efficient noise reduction
system becomes mandatory. Many approaches have been
already proposed in this field of research. Concerning the
single microphone sound pick-up problem we consider in
this article, we can note power spectral subtraction
([1],[2]), Wiener filtering ([2]) or Minimum Mean Square
Error (MMSE) estimation ([3]). Two main disturbances are
usually generated by such solutions: firstly, the artefacts
called musical noise corrupt the processed signal. Many
techniques have already been proposed to limit this
undesired phenomenon ([3],[4]). Secondly, the useful
signal (the speech of the speaker) is usually far or less
degraded by the signal processing, leading to "computerized tones" in the resulting signal.

The first part of this paper recalls the solution retained and
proposed in [3]. This solution provides low musical noise
and low distortion as well. Nevertheless, higher quality can
be obtained with improved Signal to Noise Ratio (SNR)
estimation. We propose to take into account the short time
spectral changes of the noisy signal either thanks to the
SNR estimation process itself or thanks to spectral changes
measurement similar to the one proposed in [5]. Section 3
explains in details these two approaches. The resulting
enhancement, mainly during speech attacks will be shown
and discussed in section 4.

2. WIENER FILTERING.

2.1. Open-loop Wiener filtering in the
frequency domain.

Let \( y(t) \) denotes the noisy microphone signal
\[
y(t) = s(t) + n(t)
\]
(1)
where \( s(t) \) and \( n(t) \) stand for the useful speech signal and
the noise signal respectively. We assume in the following
that these signals are not correlated. The corresponding
signals in the frequency domain, computed in practice
through short-term Fourier transform are noted, for a
current frame \( m \) and the frequency \( f \), \( Y(m,f) \), \( S(m,f) \) and
\( N(m,f) \) respectively.

Considering that the estimation of \( S(m,f) \), noted \( \hat{S}(m,f) \),
is given by \( \hat{S}(m,f) = G(m,f)Y(m,f) \), the Minimization of
the Mean Square Error value \( E[(S(m,f) - \hat{S}(m,f))^2] \)
(MMSE) leads to the estimation of the filter \( G(m,f) \) as follow :
\[
G(m,f) = \frac{SNR(m,f)}{1 + SNR(m,f)}
\]
(2)
where \( SNR(m,f) \) stands for the signal to noise ratio and
corresponds to the ratio between the power spectral density
(psd) of the useful signal \( s(t) \) and the psd of the noise
signal \( n(t) \). In the case of single microphone sound pick-
up, as the only observation is the noisy signal, these psds
are unknown in advance. As far as good estimation leads
to good performance, the estimation of the signals psd
becomes a crucial point to deal with. On the other side,
biases on these estimations may create many disturbances
on the processed signal. The estimation methods explained
in the next sections, are only empirical means to smooth
the different artefacts generated by such approaches. The
processed signal does not correspond exactly to the speech
signal \( s(t) \) but a signal not too distorted and "pleasant
enough".
2.2. Estimation of Signal to Noise Ratio.

An efficient estimation of $\text{SNR}(m, f)$ should be the result from a good compromise between distortion and musical noise. A possible solution consists in using a ‘decision directed’ approach as proposed in ([3]):

$$\text{SNR}(m, f) = \beta \frac{\|Y(m, f)\|^2}{\gamma_N(m, f)} + (1 - \beta) \cdot P(\text{SNR}_{\text{post}}(m, f))$$  \hspace{1cm} (3)

with

$$\text{SNR}_{\text{post}}(m, f) = \frac{\|Y(m, f)\|^2}{\gamma_N(m, f)} - 1, \quad P(x) = \frac{1}{2} \left( 1 + \text{sign}(x) \right)$$  \hspace{1cm} (4)

In these relations, $\gamma_N(m, f)$ stands for an estimation of the noise PSD and $0 < \beta < 1$ corresponds to a mixing factor between present and past estimations of the SNR. Such an estimation has proven to be efficient in limiting musical tones using high values of $\beta$ ([4]). Besides, low values of $\beta$ are required to avoid distortions on the useful speech signal. $\beta$ is usually set to a value of 0.98 for optimal performance regarding musical tones and acceptable distortion ([4]).

Nevertheless, such high values of $\beta$ involve that the time domain variations of SNR remain quite slow, depending more on the previous frame PSDs than on the current ones. As a result, speech signal attacks tend to be smoothed: if we assume that a voice activity period starts at the frame $m_0$ so that $\text{SNR}(m_0, f) = 0$, the slow adaptation in Eq. (3) due to $\beta = 1$ leads to:

$$\text{SNR}(m_0, f) = (1 - \beta) \cdot P(\text{SNR}_{\text{post}}(m_0, f)) < P(\text{SNR}_{\text{post}}(m_0, f))$$  \hspace{1cm} (5)

As a result an under-estimation of SNR is computed for the current frame $m_0$ and important distortions are introduced on the useful signal attacks. A similar analysis conducted on frames following a voice activity period proves that SNR is then over-estimated. As a result the remaining noise level is more important than it should be. This phenomenon was already noticed in ([5]), named as a "trailing hiss". These two effects, at the beginning and at the end of voice activity periods, become more significant as the length of the frames increases and as the overlap between successive frames decreases.


In order to avoid the distortion on speech and artefacts on noise described in the previous section, we propose to introduce in equation (3) a time-varying parameter $\beta(t)$ (or $\beta(m)$ regarding to frames) instead of the fixed one.

3.1. SNR Improvement

The first solution tends to link $\beta(t)$ with the SNR of the frame $m-1$ as proposed in [6]:

$$\beta(m, f) = \beta_{\text{SNR}}(m, f) = \beta_{\text{max}} - \lambda \cdot \frac{\text{SNR}(m-1, f)}{1 + \text{SNR}(m-1, f)}$$  \hspace{1cm} (6)

Such expression generates a time and frequencies dependent smoother. Low values of SNR give $\beta_{\text{SNR}}(m, f) \rightarrow \beta_{\text{max}}$, which results in low musical noise if $\beta_{\text{max}} \rightarrow 1$. Conversely, high values of SNR results in $\beta_{\text{SNR}}(m, f) \rightarrow \beta_{\text{max}} - \lambda$, so lower values which mean faster adaptation of $\text{SNR}(m, f)$ in (3) and less distortion on voice.

Our experiments and informal listenings have shown that $(\beta_{\text{max}}, \lambda) = (0.98, 0.6)$ is a good couple for pertinent behavior of the estimator. An example of the variations of $\beta_{\text{SNR}}(m, f)$ is drawn in Fig 1(b). We use for this example frames of 16 ms with an overlap between two successive frames of 50%. We have drawn $\beta_{\text{SNR}}(m, f)$ for the particular frequency $f = f_s/8$ where $f_s$ is the sampling frequency ($f_s = 8kHz$ in our context).

3.2. Spectral Change Improvement.

Another possible approach to improve noise reduction is highly inspired from the solution proposed in [5]. The aim of this method is to decrease $\beta$ during transient periods of $s(t)$ (plosives or attacks of speech) to avoid smoothing on the useful signal and trailing hiss phenomenon. In the same time $\beta$ is increased during non-vocal activity periods or in steady-state speech periods (periods when speech is a voiced sound).

To integrate the spectral changes into the variations of $\beta(t)$, we used a spectral derivative distance defined for the frame $m$ by:

$$\Delta Y(m) = \left[ \frac{2}{T_f} \int_0^{T_f} |Y(m, f)|^2 \cdot df \right]^{1/2}$$  \hspace{1cm} (7)

This distance is compared with an average distance computed during non vocal activity periods, $\bar{\Delta Y}$, through the relation:

$$\Gamma_Y(m) = Q\left[ 1 - (\Delta Y(m) - \bar{\Delta Y}) \right]$$  \hspace{1cm} (8)

where $Q(x)$ is defined by $Q(x) = \begin{cases} 0, & 0 \leq x \leq 1 \\ 1, & x > 1 \end{cases}$
As the measure \( \Gamma_Y(m) \) is really erratic, the variable \( \beta(t) \) should be smoothed by a first order IIR filter such that for the frame \( m \) it becomes:

\[
\beta(m) = \beta_{SC}(m) = \alpha \beta_{SC}(m-1) + (1-\alpha) \Gamma_Y(m)
\]  (9)

We can see in Fig 1.(c) an example of temporal variations of \( \beta_{SC}(m) \) for frames of 16 ms with an overlap between successive frames of 50%.

One can note on these figures that the variations of \( \beta_{SNR}(t) \) are much more erratic than the ones of \( \beta_{SC}(t) \). Nevertheless, less smoothed behavior of \( \beta_{SC}(t) \) is obtained using more selective low-pass linear filtering of \( \beta_{SC}(t) \) instead of the exponential window proposed in Eq (9). One should also note that \( \forall t, \beta_{SNR}(t) \leq 0.98 \) whereas we can have \( t_0 \) such that \( \beta_{SC}(t_0) = 1 \). This last property is important considering residual musical noise and will be discussed in section 4.2.

The influence of \( \beta \) is also illustrated by the time domain variations of the processed signals as depicted in Fig. 2. On this figure, zooms on a short period of signals processed from the example \( y_s(t) = s_s(t) + n_s(t) \) of Fig.1(a) are drawn. We have represented for the period \( t \in [14.5;14.6] \) the useful noise free speech \( s_s(t) \) (Fig. 2 (a)) corresponding to a plosive sound (the sound /po/). We have also drawn the same signal \( s_s(t) \) filtered by the three filters already described: one for each of the different estimations of \( SNR(m,f) \) using \( \beta = \text{constant}, \beta_{SNR}(t) \) and \( \beta_{SC}(t) \) respectively. These figures illustrate how much the attack can be smoothed for \( \beta = \text{constant} \) and even using \( \beta_{SNR}(t) \), whereas the use of \( \beta_{SC}(t) \) in the estimation of \( SNR(m,f) \) preserves efficiently such a plosive sound.

4. RESULTS AND DISCUSSION.

Two points will be discussed about in this section, as far as two main drawbacks of spectral subtraction processing have been pointed out: firstly the enhancement considering distortions on speech provided by the different solutions proposed in section 3 and secondly the characteristics of the residual noise still remaining after the filtering process.

4.1. Important decrease of distortions.

In order to provide information about the distortion introduced on the useful signal, we use cepstral distance measurements. These distances are computed between the noise free original speech signal \( s(t) \) and the same signal filtered by the three filters computed (with the different cases \( \beta = \text{constant}, \beta_{SNR}(t) \) and \( \beta_{SC}(t) \)). The resulting curves are given in Fig 3 for the same noisy speech as in Fig. 1 (a).

These curves clearly show that the distortion is highly decreased when using \( \beta_{SC}(t) \). The preserving of the attacks, as shown on Fig. 2, avoids efficiently the perceptual smoothing that can be heard if \( \beta = \text{constant} \). The distortion is not decreased as much when using \( \beta_{SNR}(t) \). The improvement in this last case is nevertheless relevant during long periods when \( SNR \) is important (\( t \in [14;16] \) for instance).
remaining musical tones just after vocal activity periods or during transitions of speech. During such periods, we have indeed $\beta < 1$ and no speech signal (this occurs on Fig. 1 for $t \in [8,9]$). As a result, musical noise can be heard during vocal activity periods and disappears a few frames after. The long term non stationarity of the musical tone results in a sensation of unnaturally cut signal: an alternation of silent periods, speech periods corrupted by musical tone and periods with only musical noise. Using a threshold on the gain factor hides partially this problem by giving a surrounded noise ambiance. Another tested solution consists in applying a shorter low-pass FIR filter instead of the exponential IIR solution retained in Eq (9). This solution was tested but, in the case of too selective filters, the end of speech or transitions of speech into a sentence tend to be cut as far as $\beta_{SC}(t)$ varies too rapidly and tends often to 1.

5. CONCLUSION.

This article presents two methods to enhance the performance of Wiener filtering with the Signal to Noise Ratio estimation derived from the approach proposed in [3]. These two approaches try to take into account the time variations of the smoothing constant used to estimate the SNR. They solve partially the problems remaining in noise reduction systems: $\beta_{SNR}(t)$ decreases the distortion of speech compared to the use of constant $\beta$, but musical tones persist; $\beta_{SC}(t)$ decreases efficiently distortions, eliminates musical tones but gives a unnatural sensation of cut signal due to the alternation of no noise/musical noise periods.

6. REFERENCES.