14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

NOISE REDUCTION USING RELIABLE *A POSTERIORI* SIGNAL-TO-NOISE RATIO FEATURES

Cyril Plapous [1], *Claude Marro* [1], *Pascal Scalart* [2]

[1] France Télécom - TECH/SSTP, 2 Avenue Pierre Marzin, 22307 Lannion Cedex, France
[2] University of Rennes - IRISA / ENSSAT, 6 Rue de Kerampont, B.P. 80518, 22305 Lannion, France
E-mail: claude.marro@francetelecom.com; pascal.scalart@enssat.fr

ABSTRACT

This paper addresses the problem of single microphone speech enhancement in noisy environments. State of the art short-time noise reduction techniques are most often expressed as a spectral gain depending on Signal-to-Noise Ratio (SNR). The well-known decision-directed approach drastically limits the level of musical noise but the estimated *a priori* SNR is biased since it depends on the speech spectrum estimated in the previous frame. The consequence of this bias is an annoying reverberation effect. We propose a method, called Reliable Features Selection Noise Reduction (RFSNR) technique, capable of classifying the *a posteriori* SNR estimates into two categories: the reliable features leading to speech components and the unreliable ones corresponding to musical noise only. Then it is possible to directly enhance speech using these reliable components thus obtaining an unbiased estimator.

1. INTRODUCTION

The problem of enhancing speech degraded by additive noise, when only the noisy speech is available, has been widely studied in the past and is still an active field of research. Noise reduction is useful in many applications such as voice communication and automatic speech recognition.

Scalart and Vieira Filho presented in [1] an unified view of the main single microphone noise reduction techniques where the process relies on the estimation of a short-time spectral gain which is a function of the *a priori* Signal-to-Noise Ratio (SNR) and/or the *a posteriori* SNR. They also emphasize the interest of estimating the *a priori* SNR with the decision-directed (DD) approach proposed by Ephraïm and Malah in [2]. Cappé analyzed the behavior of this estimator in [3] and demonstrated that the *a priori* SNR follows the shape of the *a posteriori* SNR with a one frame delay. Consequently, since the gain depends on the *a priori* SNR, it does not match anymore the current frame and thus it degrades the performance of the noise reduction system.

We propose a method, called Reliable Features Selection Noise Reduction (RFSNR) technique, that uses the *a priori* SNR estimated with the DD approach and the *a posteriori* SNR in order to classify this latter into reliable or unreliable features. This approach provides an efficient separation of speech components from musical noise ones and ensures that the enhanced speech is obtained using unbiased SNR estimator. The present paper consists in an extension of the work presented in [4], including deeper analysis and results over a large corpus of signals.

2. NOISE REDUCTION PARAMETERS

In the classical additive noise model, the noisy speech is given by $x(t) = s(t) + n(t)$ where $s(t)$ and $n(t)$ denote the speech and the noise signal, respectively. Let $S(p,k)$, $N(p,k)$ and $X(p,k)$ designate the kth spectral component of short-time frame p of the speech $s(t)$, the noise $n(t)$ and the noisy speech $x(t)$, respectively. The objective is to find an estimator $\hat{S}(p,k)$ which minimizes a given distortion measure conditionally to a set of spectral noisy features. Since there does not exist any direct solution for the spectral estimation, we first derive an SNR estimate from the noisy features. An estimate of $S(p,k)$ is subsequently obtained by applying a spectral gain $G(p,k)$ to each short-time spectral component $X(p,k)$. This gain corresponds to different functions proposed in the literature (*e.g.* amplitude and power spectral subtraction, Wiener filter, MMSE STSA, *etc.*) [5, 1, 2]. The choice of the distortion measure determines the gain behavior, *i.e.* the well-known trade-off between noise reduction and speech distortion. However, the key parameter is the estimated SNR since it determines the efficiency of the speech enhancement for a given noise power spectrum density (PSD).

Most of the classical speech enhancement techniques require the evaluation of two parameters, the *a posteriori* SNR and the *a priori* SNR, respectively defined by

$$SNR_{post}(p,k) = \frac{|X(p,k)|^2}{\mathrm{E}[|N(p,k)|^2]}, \tag{1}$$

$$SNR_{prio}(p,k) = \frac{\mathrm{E}[|S(p,k)|^2]}{\mathrm{E}[|N(p,k)|^2]} \tag{2}$$

where $\mathrm{E}[.]$ is the expectation operator. Let us define an additional parameter, the *instantaneous* SNR :

$$SNR_{inst}(p,k) = SNR_{post}(p,k) - 1. \tag{3}$$

This parameter can be interpreted as an estimation of the local *a priori* SNR in a way equivalent to power spectral subtraction and will be useful for the sake of analysis. In practical implementations, the PSDs of speech $\mathrm{E}[|S(p,k)|^2]$ and noise $\mathrm{E}[|N(p,k)|^2]$ are unknown as only the noisy speech is available, then these SNRs have to be estimated. The estimation of the noise PSD, noted $\hat{\gamma}_n(p,k)$, is beyond our scope and can be computed during speech pauses using recursive averaging [1] or continuously using the Minimum Statistics [6] to get a more accurate estimate in case of noise level fluctuations.

3. SNR ANALYSIS TOOL

In order to evaluate the behavior of speech enhancement techniques, we propose to use an approach derived from the one described by Renevey and Drygajlo in [7]. The basic principle is to consider the *a priori* SNR versus the *a posteriori* SNR in order to analyze the behavior of the features defined by the 2-tuple (SNR_{post}, SNR_{prio}).

In the additive model, the amplitude of the noisy signal can be expressed as $|X(p,k)| =$

$$\sqrt{|S(p,k)|^2 + |N(p,k)|^2 + 2|S(p,k)||N(p,k)|\cos\alpha(p,k)} \quad (4)$$

where $\alpha(p,k)$ is the phase difference between $S(p,k)$ and $N(p,k)$. The local *a posteriori* and *a priori* SNRs, assuming the knowledge of the clean speech and the noise, can be defined by

$$SNR^{local}_{post}(p,k) = \frac{|X(p,k)|^2}{|N(p,k)|^2}, \quad (5)$$

$$SNR^{local}_{prio}(p,k) = \frac{|S(p,k)|^2}{|N(p,k)|^2}. \quad (6)$$

By replacing $|X(p,k)|$ in (5) by its expression (4) and using (6), it comes $SNR^{local}_{post}(p,k) =$

$$SNR^{local}_{prio}(p,k) + 1 + 2\sqrt{SNR^{local}_{prio}(p,k)}\cos\alpha(p,k). \quad (7)$$

This relation depends on $\alpha(p,k)$ which is an uncontrolled parameter in speech enhancement techniques.

In the following, the discussion will be illustrated using a French sentence corrupted by car noise at 12dB global SNR but it can be generalized to other noise and SNR conditions. The relation expressed by (7) is illustrated in Fig. 1. When

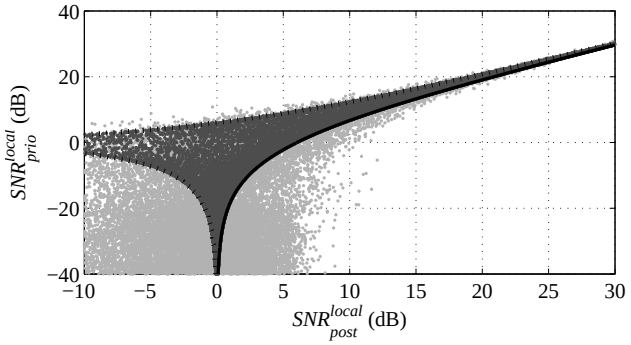

Figure 1: SNR^{local}_{prio} versus SNR^{local}_{post}. Dark gray features: clean speech and noise amplitudes are known in (5) and (6). Light gray features: clean speech amplitude is known but estimated noise PSD is used in (5) and (6).

the clean speech and the noise amplitudes are known, the features lie between two curves : the solid one (resp. dashed) corresponds to the limit case where $\alpha(p,k) = 0$ (π) in (7), *i.e.* noise and clean speech spectral components are added in phase (phase opposition). These two limits define an area where the feature repartition depends on the true phase difference $\alpha(p,k)$. When an estimated noise PSD is used in (5) and (6) instead of the local noise, the estimation errors lead to an important dispersion of the features outside of the limit area for low SNR values and decrease the quality of the enhanced speech.

4. DECISION-DIRECTED APPROACH

Using a given estimation of the noise PSD, the *a posteriori* and *a priori* SNRs are estimated as follows

$$\hat{SNR}_{post}(p,k) = \frac{|X(p,k)|^2}{\hat{\gamma}_n(p,k)}, \quad (8)$$

$$\hat{SNR}_{prio}(p,k) = \beta\frac{|\hat{S}(p-1,k)|^2}{\hat{\gamma}_n(p,k)} + (1-\beta)\mathrm{P}[\hat{SNR}_{post}(p,k) - 1] \quad (9)$$

where P[.] denotes the half-wave rectification and $\hat{S}(p-1,k)$ is the estimated speech spectrum at previous frame. This *a priori* SNR estimator corresponds to the so-called decision-directed (DD) approach [2, 3] whose behavior is controlled by the parameter β (typically $\beta = 0.98$). The approaches based on (8) and (9) to compute the spectral gain will be referred to the DD algorithm.

4.1 Analysis

We can emphasize two effects of the DD algorithm which have been interpreted by Cappé in [3]:

- When the *a posteriori* SNR is much larger than 0dB, $\hat{SNR}_{prio}(p,k)$ corresponds to a frame delayed version of $\hat{SNR}_{post}(p,k) - 1 = \hat{SNR}_{inst}(p,k)$.
- When the *a posteriori* SNR is lower or close to 0dB, $\hat{SNR}_{prio}(p,k)$ corresponds to a highly smoothed and delayed version of $\hat{SNR}_{inst}(p,k)$. The direct consequence for the enhanced speech is the reduction of the musical noise effect due to a lower variance.

This behavior is illustrated in Fig. 2 where we consider the case of speech corrupted by additive car noise at a 12 dB global SNR. The time varying *instantaneous* and *a priori* SNR are represented for the frequency band centered on 467 Hz. The 20 first and the 17 last frames contain only noise whereas the 19 frames in the middle contain noisy speech including speech onset and offset. Notice that in this experiment, we have chosen the Wiener filter [1], without loss of generality, to compute the spectral gain of the DD approach :

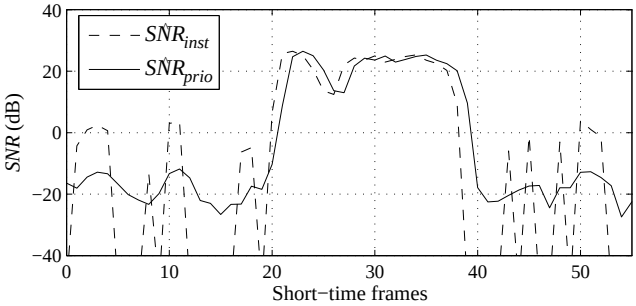

Figure 2: SNR evolution over short-time frames ($f = 467$ *Hz*). Solid line: *instantaneous* SNR; dashed line: *a priori* SNR.

$$G_{DD}(p,k) = \frac{\hat{SNR}_{prio}(p,k)}{1 + \hat{SNR}_{prio}(p,k)}. \quad (10)$$

The smoothing effect and the delay introduced by the DD algorithm are clearly visible on Fig. 2. We can emphasize that this delay is a drawback especially for speech non-stationarities, *e.g.* speech onset and offset. Furthermore, it introduces a permanent bias in gain estimation which limits its noise reduction performance and generates an annoying reverberation effect.

In order to describe more precisely the behavior of the DD approach, the 2-tuple $(\hat{SNR}_{post}, \hat{SNR}_{prio})$ is represented in Fig. 3 where the *a posteriori* and *a priori* SNRs are estimated using (8) and (9), respectively. To analyze this figure, the case where SNRs are computed using known clean speech amplitude and estimated noise PSD (*cf.* Fig. 1) is used as reference. Note that in such a case, the estimated *a posteriori* SNR are the same in Fig. 1 and 3 since equation (5) resumes to (8). In Fig. 3 a large part of the *a priori* SNR features (approximately 60%) is underestimated which illustrates the effect of the DD bias on SNR estimation.

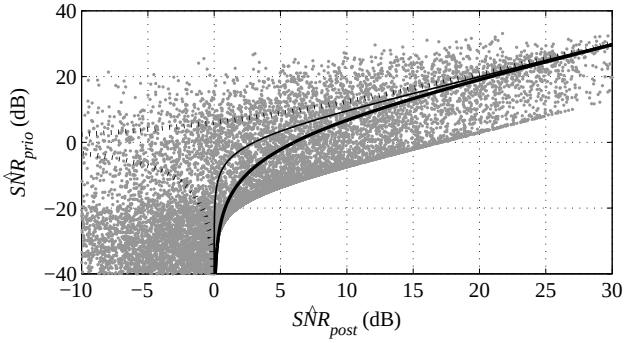

Figure 3: $\hat{SNR}_{prio}$ versus $\hat{SNR}_{post}$ for the DD approach. The three lines illustrate equation (7) where $\alpha(p,k) = 0$ (bold solid line), $\alpha(p,k) = \pi$ (dashed line) and $\alpha(p,k) = \frac{\pi}{2}$ (thin solid line).

If we consider the case where a speech component appears abruptly at frame p, assuming that $\hat{S}(p-1,k) = 0$ in equation (9), then for the current frame we have

$$\hat{SNR}_{prio}(p,k) = (1-\beta)\mathrm{P}[\hat{SNR}_{post}(p,k) - 1]. \qquad (11)$$

Actually, the estimated *a priori* SNR will be a version of the *a posteriori* SNR attenuated by $(1-\beta)$. If $\beta = 0.98$ (typical value), this attenuation is around 17dB. Note that if $\alpha(p,k) = \frac{\pi}{2}$, equation (7) becomes

$$SNR^{local}_{prio}(p,k) = SNR^{local}_{post}(p,k) - 1. \qquad (12)$$

This relationship is illustrated in Fig. 3 by the thin solid line. Thus, the attenuation introduced by $1-\beta$ in equation (11) is materialized by a high concentration of features around a shifted version (by -17dB) of this thin line curve. This offset corresponds to the maximum bias and it is consistent with the degradation introduced by the DD approach during speech onsets and more generally when speech amplitude increases rapidly.

We can also observe in Fig. 3 that some *a priori* SNR features are overestimated. This case occurs when a speech component disappears abruptly, *i.e.* $\mathrm{P}[\hat{SNR}_{post}(p,k) - 1] = 0$ leading to

$$\hat{SNR}_{prio}(p,k) = \beta \frac{|\hat{S}(p-1,k)|^2}{\hat{\gamma}_n(p,k)} \qquad (13)$$

whereas a null value would be the best estimate. This overestimation is related to the speech spectrum of the previous frame. The reverberation effect characteristic of the DD approach is explained by both underestimation and overestimation of the *a priori* SNR features.

4.2 Comparison between *a posteriori* and *a priori* SNRs

It is interesting to underline the behavior of the *a posteriori* and *a priori* SNR estimators. It is well known that using only the *a posteriori* SNR to enhance the noisy speech results in a very high amount of musical noise, leading to a poor signal quality. However, this technique leads to the lowest degradation level for the speech components themselves. The *a priori* SNR, estimated using the DD approach, is widely used instead of the *a posteriori* SNR because the musical noise is reduced to an acceptable level. However, this estimated SNR is biased and then the performance is reduced during speech activity.

In order to measure the performance of SNR estimators, it is useful to compare the estimated SNR values to the true (or local, *cf.* equations (5) and (6) with $|N(p,k)|^2$ known) ones as shown in Fig. 4. The SNRs are plotted for 50 frames of speech activity to focus the analysis on the behavior of the SNR estimators for speech components.

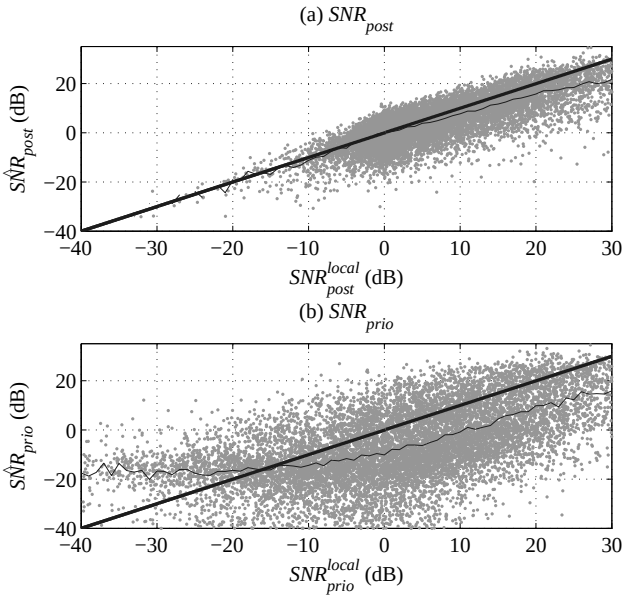

Figure 4: Estimated SNRs versus true SNRs (*i.e.* local SNRs) in case of (a) *a posteriori* SNR and (b) *a priori* SNR. The bold line represents a perfect estimator and the thin line represents the mean of the estimated SNR versus the true SNR.

In the two cases depicted in Fig. 4.(a) (*a posteriori* SNR) and Fig. 4.(b) (*a priori* SNR), the bold line corresponds to a perfect SNR estimator ($\hat{SNR} = SNR^{local}$) that can be used as a reference to evaluate the performance of the real estimators. It is obvious that the features corresponding to the *a posteriori* SNR estimator are closer to the reference bold line and less dispersed than the *a priori* SNR estimator ones. The thin line represents the mean of the estimated SNR knowing the true SNR. This mean is closer to the perfect estimator

for the *a posteriori* SNR estimator. It is slightly underestimated for high SNR whereas for the *a priori* SNR the underestimation (*cf.* equation (11)) is large for SNR greater than -17dB. However, since the dispersion is high for the *a priori* SNR features, even if the mean is largely underestimated, the case where SNR features are overestimated exists (*cf.* equation (13)). Finally, these results confirm that the *a posteriori* SNR estimator is more reliable than the *a priori* one for speech components.

5. RELIABLE SNR FEATURES SELECTION

Since the *a posteriori* SNR estimator is better for speech components than the *a priori* SNR estimator of the DD approach, a judicious strategy would be to determine when it is possible to use it and when it will lead to musical noise. In order to select only the reliable *a posteriori* SNR components, we propose to separate the SNR features in the space defined by the 2-tuple $(\hat{SNR}_{post}, \hat{SNR}_{prio})$ using two thresholds. Given the threshold η for the *a priori* SNR, it is possible to compute the threshold δ for the *a posteriori* SNR using (7) which depends on the phase parameter $\alpha(p,k)$. As displayed in Fig. 5, these SNR features will be then separated into four quadrants. By processing output signals using

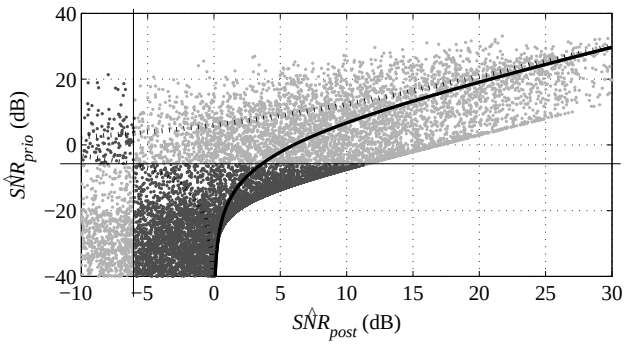

Figure 5: Separation of the features defined by the 2-tuple $(\hat{SNR}_{post}, \hat{SNR}_{prio})$ into 4 quadrants using 2 thresholds on $\hat{SNR}_{post}$ and $\hat{SNR}_{prio}$.

the *a posteriori* SNR values of each quadrant, informal listening tests confirm that a classification can be made. Based on these experiments, we propose to choose $\eta = -6$dB for the *a priori* SNR threshold, which leads to $\delta \approx -6$dB for the *a posteriori* SNR threshold. In order to compute δ, we chose $\alpha(p,k) = \pi$ in (7) because it corresponds to the smallest resulting threshold δ and then preserves SNR values corresponding to speech whatever the phase difference between speech and noise is (*cf.* algorithm proposed in the following). This choice is natural because we cannot estimate this phase difference and consequently it leads to the less speech component suppression. However, any other choice can be made for $\alpha(p,k)$.

This particular choice is illustrated in Fig. 5. The two thresholds separate the SNR features into four quadrants (two in dark gray dots and two in light gray). The interest of this separation is the possibility to classify the features into different categories. The right dark gray features lead to high level musical noise only and the ones in the two left quadrants lead to very low and inaudible components that are consequently useless. Finally, the right light gray features can be classified as SNR components leading to speech only, without musical noise. We can emphasize that a reliable classification is obtained because the behaviors of the *a posteriori* and *a priori* SNR estimators are complementary. Actually, the *a posteriori* SNR estimator is efficient for speech components but poor for musical noise and the *a priori* SNR estimator of the DD approach is efficient for musical noise but biased for speech components. As a consequence, an efficient separation of the SNR features can be done in the space defined by the 2-tuple $(\hat{SNR}_{post}, \hat{SNR}_{prio})$.

Based on this classification, we propose to re-estimate the *a posteriori* SNR using only the reliable features and to use it to compute the spectral gain. This algorithm called RFSNR (Reliable Features Selection Noise Reduction) is described as follows

step 1: The *a posteriori* and *a priori* SNRs are computed using relations (8) and (9), respectively.

step 2: The *a posteriori* SNR is re-estimated as follows

$$\hat{SNR}^{thr}_{post}(p,k) = \begin{cases} \hat{SNR}_{post}(p,k) & \text{if } \hat{SNR}_{post}(p,k) \geq \delta \\ & \text{and} \\ & \hat{SNR}_{prio}(p,k) \geq \eta, \\ 1 & \text{else,} \end{cases} \tag{14}$$

where *thr* indicates that the *a posteriori* SNR is processed using thresholds.

step 3: This re-estimated and unbiased SNR, $\hat{SNR}^{thr}_{post}(p,k)$, is directly used to compute the spectral gain, the Wiener filter [1] for example. This gain is then applied to the noisy speech to obtain the enhanced signal. We can emphasize that the *a priori* SNR is used only to select the reliable *a posteriori* SNR features, and will not be used to compute the spectral gain as in [2] since it is biased.

step 4: Another spectral gain is computed based on *a posteriori* and *a priori* SNRs of step 1 and will be used to obtain $\hat{S}(p,k)$ needed in step 1 for the next frame. Actually, this is what is done in the classical DD approach.

Notice that the two right quadrants in Fig. 5 correspond to the case where a threshold is applied only to the *a posteriori* SNR values in a way close to the generalized spectral subtraction [5]. In that case, a threshold of 10dB is required to suppress all the musical noise (dark gray features) but then all the speech components corresponding to light gray dots lying between -6 and 10dB (abscissa axis) are suppressed too. Finally, using two thresholds (14) avoids this problem and allows to preserve the features corresponding to speech components while suppressing the musical noise.

6. RESULTS

In this section, the Wiener filter, *cf.* equation (10), is chosen for the DD and RFSNR approaches. Figure 6 shows three spectrograms. Figure 6.(a) represents the noisy speech corrupted by car noise (SNR=12dB), Fig. 6.(b) shows the enhanced speech, free of musical noise, obtained with the RFSNR technique and Fig. 6.(c) represents the musical noise successfully removed. This musical noise corresponds only to the right dark gray and to the left features of Fig. 5 which confirms that the proposed features selection based on equation (14) is powerful to remove it. Notice that this very high level of musical noise is the one present in enhanced speech using only unprocessed *a posteriori* SNR (8). Furthermore,

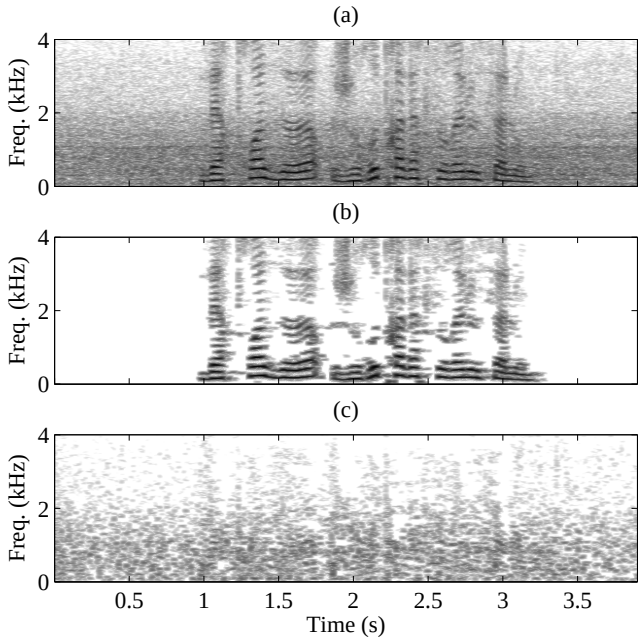

Figure 6: Speech spectrograms. (a) Noisy speech; (b) Noisy speech enhanced by RFSNR technique; (c) Musical noise successfully removed using the RFSNR technique.

speech components are enhanced using reliable *a posteriori* SNR estimates and thus do not suffer from the bias introduced by the DD approach.

In order to generalize this result, the input SNRs of noisy speech and the corresponding segmental SNR obtained for DD and RFSNR techniques are presented in Table 1. Each SNR value is a mean over 36 sentences (4 speakers, 2 females and 2 males, and 9 sentences per speaker). The segmental SNR measure takes into account both residual noise level and speech degradation. The proposed RFSNR technique achieves the best results (bold values) under all noise and SNR conditions. Since speech components are enhanced using only reliable *a posteriori* SNR estimates, they do not suffer from the bias introduced by the DD approach which explains the segmental SNR improvement. These remarks are corroborated by informal listening tests. Actually, from a subjective point of view, the annoying reverberation effect of the DD approach is removed. However, some distorsions remains, in particular when SNR is low, since the efficiency of the SNR estimators depends on the quality of the noise PSD estimation.

7. CONCLUSION

In this paper, we proposed and analyzed an SNR estimator based on the selection of the most reliable *a posteriori* SNR features. The *a posteriori* SNR estimator is efficient for speech components but leads to high level musical noise. That is why the DD approach is preferred to compute the *a priori* SNR which efficiently reduces the level of musical noise. However, this estimator is biased for speech components leading to degradation for the enhanced speech and to an annoying reverberation effect. The complementary behaviors of these two estimators precisely allow to classify the features in the space defined by the 2-tuple

Table 1: Segmental SNR obtained for DD and RFSNR in various noise and SNR conditions.

Noise type	Input SNR (dB)	Segmental SNR (dB)	
		DD	RFSNR
Office	24	24.43	**24.68**
	18	19.00	**19.66**
	12	13.59	**14.56**
	6	8.39	**9.46**
	0	3.90	**4.74**
Car	24	23.32	**24.00**
	18	18.07	**18.92**
	12	13.12	**14.11**
	6	8.48	**9.74**
	0	4.39	**5.21**
Street	24	23.20	**23.87**
	18	17.68	**18.61**
	12	12.41	**13.45**
	6	7.41	**8.41**
	0	3.03	**3.70**
Babble	24	24.08	**24.69**
	18	18.13	**19.11**
	12	12.50	**13.48**
	6	7.35	**8.13**
	0	3.16	**3.53**

$(\hat{SNR}_{post}, \hat{SNR}_{prio})$ since reliable and unreliable features are well separated. Finally, the enhanced speech is free of musical noise and does not suffer from the bias mentioned above since only the reliable *a posteriori* SNR features are used to compute the spectral gain. Consequently, the reverberation effect characteristic of the DD approach is also removed.

REFERENCES

[1] P. Scalart, and J. Vieira Filho, "Speech Enhancement Based on a Priori Signal to Noise Estimation," *IEEE ICASSP'96*, Vol. 2, pp. 629–632, 7–10 May 1996.

[2] Y. Ephraïm, and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. ASSP*, Vol. ASSP-32, No. 6, pp. 1109–1121, Dec. 1984.

[3] O. Cappé, "Elimination of the Musical Noise Phenomenon with the Ephraïm and Malah Noise Suppressor," *IEEE Trans. SAP*, Vol. 2, No. 2, pp. 345–349, Apr. 1994.

[4] C. Plapous, C. Marro, and P. Scalart, "Reliable A Posteriori Signal-To-Noise Ratio Features Selection," *IEEE WASPAA*, 16–19 Oct. 2005.

[5] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise," *IEEE ICASSP'79*, Vol. 4, pp. 208–211, Apr. 1979.

[6] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. Speech Audio Processing*, Vol. 9, No. 5, pp. 504–512, Jul. 2001.

[7] P. Renevey, and A. Drygajlo, "Detection of Reliable Features for Speech Recognition in Noisy Conditions Using a Statistical Criterion," *Proc. of Workshop on CRAC*, pp. 71–74, Sept. 2001.