A TWO-STEP NOISE REDUCTION TECHNIQUE

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

[1] France Telecom R&D - DIH/IPS, 2 Avenue Pierre Marzin, 22307 Lannion Cedex, France
[2] ENSSAT - LASTI, 6 Rue de Kerampont, B.P. 447, 22305 Lannion Cedex, France
E-mail: cyril.plapous,claude.marro,laurent.mauuary@francetelecom.com; pascal.scalart@enssat.fr

ABSTRACT

This paper addresses the problem of single microphone speech enhancement in noisy environments. Common short-time noise reduction techniques proposed in the art are expressed as a spectral gain depending on the *a priori* SNR. In the well-known decision-directed approach, the *a priori* SNR depends on the speech spectrum estimation in the previous frame. As a consequence the gain function matches the previous frame rather than the current one which degrades the noise reduction performance. We propose a new method called Two-Step Noise Reduction (TSNR) technique which solves this problem while maintaining the benefits of the decision-directed approach. This method is analyzed and results in voice communication and speech recognition context are given.

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 where efficient noise reduction techniques are required. Scalart and Vieira Filho presented in [1] an unified view of the main single microphone noise reduction techniques where the noise reduction process relies on the estimation of a short-time suppression 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 thanks to the decision-directed approach proposed by Ephraim 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 delay of one frame. This bias is due to the use of the speech spectrum estimated at the previous frame to compute the current *a priori* SNR. In fact, 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 suppression system. We propose a new method, called Two-Step Noise Reduction (TSNR) technique, to refine the estimation of the *a priori* SNR which suppresses these drawbacks while maintaining the advantages of the decision-directed approach, like the highly reduced musical noise effect. An analysis of the TSNR technique behavior is proposed and some results are given in the context of voice communication and speech recognition using one of the databases that were used for the competitive selection of the ETSI/STQ/AURORA/WI008 standardization [4].

2. CLASSICAL NOISE REDUCTION RULE

In the classical additive noise model, the noisy speech is given by $x(t) = s(t) + b(t)$ where $s(t)$ and $b(t)$ denote the speech and the noise signal, respectively. Let $S(p, \omega_k)$, $B(p, \omega_k)$ and $X(p, \omega_k)$ designate the ω_k spectral component of short-time frame p of the speech $s(t)$, the noise $b(t)$ and the noisy speech $x(t)$, respectively. The quasi-stationarity of the speech is assumed over the duration of the analysis frame. The noise reduction process consists in the application of a spectral gain $G(p, \omega_k)$ to each short-time spectrum value $X(p, \omega_k)$. In practice, the spectral gain requires the evaluation of two parameters. The *a posteriori* SNR is the first parameter given by

$$SNR_{post}(p, \omega_k) = \frac{|X(p, \omega_k)|^2}{E\{|B(p, \omega_k)|^2\}} \quad (1)$$

where E is the expectation operator. The *a priori* SNR, which is the second parameter of the noise suppression rule is expressed as

$$SNR_{prio}(p, \omega_k) = \frac{E\{|S(p, \omega_k)|^2\}}{E\{|B(p, \omega_k)|^2\}} \quad (2)$$

and requires the unknown information of the speech spectrum. Let us define a new parameter, the *instantaneous* SNR,

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

This parameter can be interpreted as an estimation of the local *a priori* SNR in a way equivalent to the spectral subtraction. So, to evaluate the accuracy of the *a priori* SNR estimator, it is better to compare it to the *instantaneous* SNR instead of the *a posteriori* SNR . Both the gain function and the *a priori* SNR, described in the literature as functions of the *a posteriori* SNR, can be easily redefined as functions of the *instantaneous* SNR. Consequently, in the following we will only refer to the *instantaneous* SNR and to the *a priori* SNR. In practical implementations of speech enhancement systems, the power spectrum density of the speech $|S(p, \omega_k)|^2$ and the noise $|B(p, \omega_k)|^2$ are unknown as only the noisy speech is available. Then, both the *instantaneous* SNR and the *a priori* SNR have to be estimated. The noise power spectral density is estimated during speech pauses using the classical recursive relation

$$\hat{\gamma}_{bb}(p, \omega_k) = \lambda \hat{\gamma}_{bb}(p-1, \omega_k) + (1-\lambda)|X(p, \omega_k)|^2 \quad (4)$$

where $0 < \lambda < 1$ is the smoothing factor. Then the two estimated SNRs can be computed as follow

$$\hat{SNR}_{inst}(p, \omega_k) = \frac{|X(p, \omega_k)|^2}{\hat{\gamma}_{bb}(p, \omega_k)} - 1 \quad (5)$$

and

$$S\hat{N}R_{prio}(p,\omega_k) = \beta\frac{|\hat{S}(p-1,\omega_k)|^2}{\hat{\gamma}_{bb}(p,\omega_k)} + (1-\beta)P[S\hat{N}R_{inst}(p,\omega_k)] \quad (6)$$

where P denotes the half-wave rectification and $\hat{S}(p-1,\omega_k)$ is the estimated speech spectrum at previous frame. The estimator of the *a priori* SNR described by (6) corresponds to the so-called decision-directed approach [2] which has a behavior controlled by the parameter β (typically equal to 0.98). The multiplicative gain function $G(p,\omega_k)$ is obtained by

$$G(p,\omega_k) = f(S\hat{N}R_{prio}(p,\omega_k), S\hat{N}R_{inst}(p,\omega_k)) \quad (7)$$

and the resulting speech spectrum is estimated as follows

$$\hat{S}(p,\omega_k) = G(p,\omega_k)X(p,\omega_k). \quad (8)$$

The function f depends on *a priori* SNR and/or *instantaneous* SNR. Then the analysis proposed below is valid with the different gain functions proposed in the literature (*e.g.* amplitude and power spectral subtraction, Wiener filtering, etc.) [1, 2, 5].

3. TWO-STEP NOISE REDUCTION TECHNIQUE (TSNR)

3.1. Principle of the two-step procedure

In order to enhance the performance of the noise reduction process, we propose to estimate the multiplicative gain $G(p,\omega_k)$ in a two-step procedure. This method will be referred to as the Two-Step Noise Reduction (TSNR) algorithm in the following. In the first step we compute the multiplicative gain $G_{dd}(p,\omega_k)$ function of the parameter $S\hat{N}R_{prio\_dd}(p,\omega_k)$ and/or $S\hat{N}R_{inst}(p,\omega_k)$ as described in section 2. This method will be referred to as the decision-directed (DD) algorithm. The multiplicative gain obtained in the first step will then be used to refine the *a priori* SNR estimation using the following equation

$$S\hat{N}R_{prio\_2step}(p,\omega_k) = \frac{|G_{dd}(p,\omega_k)X(p,\omega_k)|^2}{\hat{\gamma}_{bb}(p,\omega_k)}. \quad (9)$$

The numerator of (9) gives a more accurate estimation of the power spectrum density of speech.

Finally, we compute the multiplicative gain

$$G_{2step}(p,\omega_k) = h(S\hat{N}R_{prio\_2step}(p,\omega_k), S\hat{N}R_{inst}(p,\omega_k)) \quad (10)$$

which is used to enhance the noisy speech

$$\hat{S}(p,\omega_k) = G_{2step}(p,\omega_k)X(p,\omega_k). \quad (11)$$

Note that h may be different from the function f defined in (7). Furthermore, this approach can be extended to multiple steps in an iterative procedure, however we observed that the major improvement is due to the first two steps.

3.2. Analysis of the two-step procedure

Figure 1 shows the behavior of the DD algorithm and the TSNR algorithm. We consider the case of speech corrupted by additive car noise at a 12 dB global SNR. Only the estimates at frequency 372 Hz are displayed. Note that this case illustrates the typical behavior of the represented SNR estimators. The first 23 short-time frames consist in noise and the last 27 short-time frames consist in speech including a transient between noise and speech around frame 23.

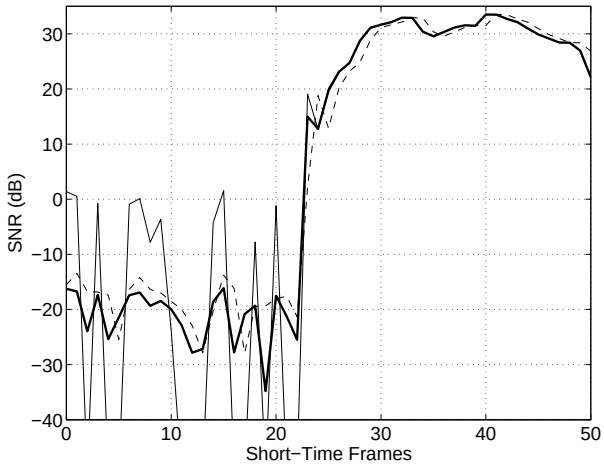

Fig. 1. SNR evolution over short-time frames ($f = 372\ Hz$). Solid line: *instantaneous* SNR; dashed line: *a priori* SNR of the DD algorithm; Bold line: *a priori* SNR of the TSNR algorithm.

The solid curve represents the time varying *instantaneous* SNR. The dashed curve and the bold curve represent the *a priori* SNR evolution for the DD algorithm and for the TSNR algorithm, respectively. Notice that in this experiment, we have chosen the multiplicative Wiener gain, without loss of generality, to compute both gains $G_{dd}(p,\omega_k)$ and $G_{2step}(p,\omega_k)$. Thus the generic gain expression is

$$G_{generic}(p,\omega_k) = \frac{S\hat{N}R_{prio\_generic}(p,\omega_k)}{1+S\hat{N}R_{prio\_generic}(p,\omega_k)} \quad (12)$$

where the subscript $generic$ must be replaced by dd and $2step$, respectively. We can emphasize two effects of the DD algorithm which have been interpreted by Cappé in [3]:

- When the *instantaneous* SNR is much larger than 0 dB, the $S\hat{N}R_{prio\_dd}(p,\omega_k)$ corresponds to a delayed version of the *instantaneous* SNR. This delay is equal to the frame duration.
- When the *instantaneous* SNR is lower or close to 0 dB, the $S\hat{N}R_{prio\_dd}(p,\omega_k)$ corresponds to a highly smoothed and delayed version of the *instantaneous* SNR. Thus the variance of the *a priori* SNR is reduced compared to the *instantaneous* SNR. The direct consequence for the enhanced speech is the reduction of the musical noise effect.

The delay introduced by the DD algorithm is a drawback especially when the speech signal is non-stationary like during onset or ending of speech. Furthermore, this delay introduces a bias in the gain estimation and thus limits the noise reduction performance. The analysis proposed below shows that the TSNR algorithm is able to suppress the delay while maintaining the benefits of the DD algorithm.

The conclusions of Cappé [3] concerning the DD algorithm directly apply to the first step of the TSNR algorithm and furthermore can be used to analyze the second step:

- When the *instantaneous* SNR is much larger than 0 dB, we can make from (6) the following approximation [3]

$$S\hat{N}R_{prio\_dd}(p,\omega_k) \approx \beta S\hat{N}R_{inst}(p-1,\omega_k). \quad (13)$$

So, the multiplicative gain obtained after the first step can be approximated by

$$G_{dd}(p,\omega_k) \approx \frac{\beta S\hat{N}R_{inst}(p-1,\omega_k)}{1+\beta S\hat{N}R_{inst}(p-1,\omega_k)}. \quad (14)$$

Furthermore, by considering that $S\hat{N}R_{inst}(p-1,\omega_k) \gg 1$ and that β is very close to 1, (14) reduces to $G_{dd}(p,\omega_k) \approx 1$. If we introduce this approximation in equation (9), this leads to

$$S\hat{N}R_{prio\_2step}(p,\omega_k) \approx \frac{|X(p,\omega_k)|^2}{\hat{\gamma}_{bb}(p,\omega_k)}. \quad (15)$$

Finally, by applying $S\hat{N}R_{inst}(p,\omega_k) \gg 1$ in (5), the following relation can be derived

$$S\hat{N}R_{prio\_2step}(p,\omega_k) \approx S\hat{N}R_{inst}(p,\omega_k). \quad (16)$$

This result shows that the TSNR algorithm succeeds in suppressing the delay introduced by the DD algorithm. This result is illustrated by Fig. 1. When the signal is composed of a mixture of speech and noise (right-part of Fig. 1), the bold curve is superimposed on the solid curve, then the TSNR algorithm efficiently suppresses the delay introduced by the DD algorithm and its negative consequences on the multiplicative gain.

- When the *instantaneous* SNR is lower or close to 0 dB, the $S\hat{N}R_{prio\_2step}(p,\omega_k)$ is further reduced compared to $S\hat{N}R_{prio\_dd}(p,\omega_k)$ which is equivalent to further reduce the noise when speech components are absent, even during speech activity. This is illustrated in left-part of Fig. 1. Furthermore, it appears that the second step helps in reducing the delay introduced by the smoothing effect even when the SNR is small while keeping the smoothing effect provided by the DD algorithm.

To summarize, the TSNR algorithm improves the performance of the noise reduction since the gain is well adapted to the current frame to enhance, whatever the *instantaneous* SNR may be. Notice that when more than two steps are used, the behavior is similar to the TSNR algorithm but without noticeable improvement.

4. EXPERIMENTAL RESULTS

4.1. Voice communication

Figure 2 shows the efficiency of the TSNR algorithm when the noisy signal is mainly noise, like during speech pauses or during speech activity in frequency areas with no speech component. The solid curve is the amplitude of noise without processing. The bold curve corresponds to the residual noise with the TSNR algorithm. Compared to the dashed curve which corresponds to the residual noise delivered by the DD algorithm, the TSNR algorithm exhibits an extra reduction of 10 dB on average. This is an interesting property since spectral valleys between speech harmonics are well enhanced and more generally the level of the residual musical noise is reduced.

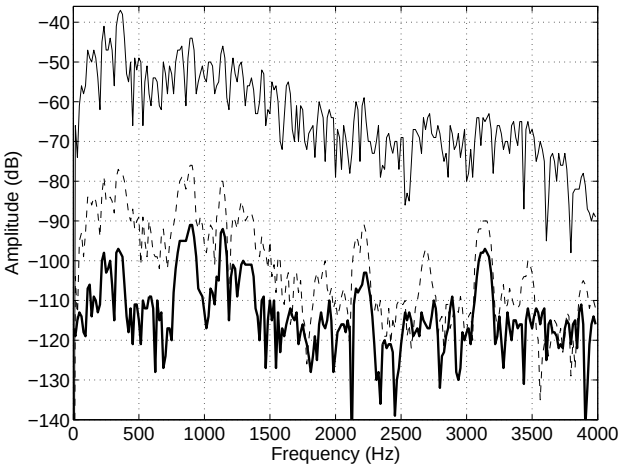

Fig. 2. Amplitude of the signal in small *instantaneous* SNR areas. Solid line: noise; dashed line: residual noise of DD algorithm; bold line: residual noise of TSNR algorithm

In Fig. 3 a silence to noise transient is isolated in order to show the improvement obtained by suppressing the bias in the *a priori* SNR estimation. The solid curve is the amplitude of clean speech and will be considered as the reference for the two other curves. The dashed curve corresponds to the enhanced speech using the DD algorithm. The bold curve corresponds to the enhanced speech using the TSNR algorithm. It can be observed that there is a significant improvement of about 1 to 5 dB on most of the harmonics. This property is mainly due to the ability of the TSNR algorithm to update the *a priori* SNR faster than the DD algorithm. For each frequency component, the bias of the multiplicative gain is removed and the non-stationarity of the speech signal can be immediately tracked. Note that this phenomenon occurs not only for onset and ending of speech, but also during speech activity in frequency areas where the SNR exhibits abrupt changes (*e.g.* unvoiced to voiced transitions, etc.).

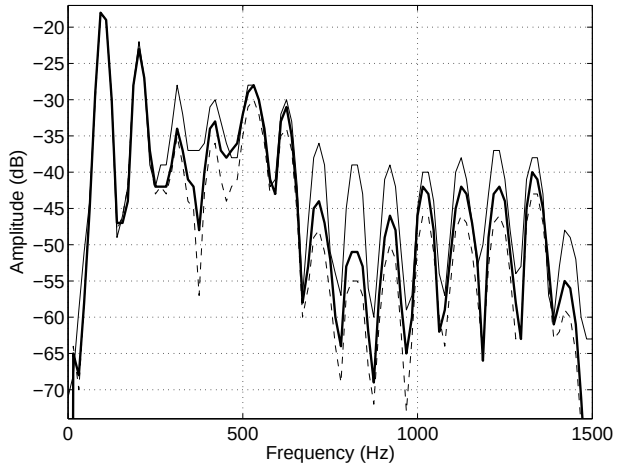

Fig. 3. Amplitude of the signal in high *instantaneous* SNR areas. Solid line: clean speech; dashed line: enhanced speech of DD algorithm; bold line: enhanced speech of TSNR algorithm

4.2. Speech recognition

The TSNR algorithm was included in the ETSI standard Distributed Speech Recognition (DSR) advanced front-end, ETSI 202 050 version 1.1.1 (ES202) [6]. In order to quantify the benefits provided by the TSNR algorithm, speech recognition experiments were carried out with ES202 and with modified version of ES202 where the second step of the TSNR algorithm was removed, which corresponds to the DD algorithm (ES202dd).

Notice that in this ES202 front-end, to compute both gains $G_{dd}(p,\omega_k)$ and $G_{2step}(p,\omega_k)$, we have chosen the following gain

$$G_{generic}(p,\omega_k) = \frac{\sqrt{\hat{SNR}_{prio\_generic}(p,\omega_k)}}{1+\sqrt{\hat{SNR}_{prio\_generic}(p,\omega_k)}} \qquad (17)$$

where the subscript $generic$ must be replaced by dd and $2step$, respectively. This gain, which is smoother than the Wiener gain, is well adapted to speech recognition applications.

The ES202 and ES202dd front-ends were evaluated on the *SpeechDatCar German* of the Aurora 3 databases. Aurora 3 is a set of multi-language SpeechDat-Car databases recorded in-car under different driving conditions with close-talking and hands-free microphones.

Three train and test configurations were defined: the well-matched condition (WM), the medium mismatched condition (MM) and the highly mismatched condition (HM). In the WM case, 70% of the entire data is used for training and 30% for testing. The training set contains all the variability that appears in the test set. In the MM case, only far microphone data is used for both training and testing. For the HM case, training data consists of close microphone recordings only while testing is done on far microphone data.

Recognition experiments were carried out using perfect end-pointing. Aurora 3 databases are connected digit tasks. Hence different types of error may occur: substitution error (one word uttered, another word recognized), deletion error (one word uttered, no word recognized) and insertion error (no word uttered, one word recognized). Most of the insertion errors are due to the silence/noise between the words.

We tested the ES202 and ES202dd front-ends by using the back-end configuration as defined by the ETSI Aurora group [4]. The digit models have 16 states with 3 Gaussians per state. The silence model has 3 states with 6 Gaussians per state. Also, a one-state short pause model is used and is tied with the middle state of the silence model.

Figure 4 shows the relative degradation for deletion, substitution and insertion errors when the second step of the TSNR algorithm is removed from the ES202 front-end. For the three types of test (WM,MM and HM), it appears that the TSNR algorithm mainly reduces the substitution and insertion errors. The reduction of insertion errors when applying TSNR algorithm is explained by its benefits in small *instantaneous* SNR (*cf.* Fig. 1 and Fig. 2). Indeed, less noise between words results in less insertion errors.

As already mentioned, SNR for a given frequency exhibits abrupt changes (*e.g.* unvoiced to voiced transitions, etc.) during speech activity. Thus the better behavior of the TSNR algorithm for transients results in an improved noise reduction during speech activity (*cf.* Fig. 1 and Fig. 3). This explains the reduction of the substitution errors.

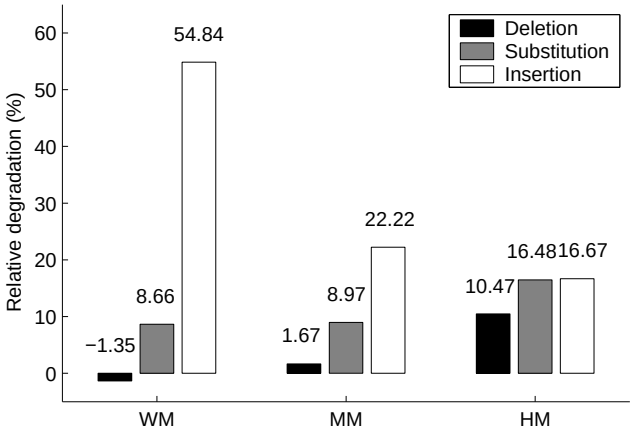

Fig. 4. Relative degradation when the second step of the TSNR algorithm is removed in ES202 front-end.

5. CONCLUSION

In this paper, we proposed a new noise reduction technique based on the estimation of the *a priori* SNR in two steps. The *a priori* SNR estimated in the first step provides interesting properties but suffers from a delay of one frame which is removed by the second step of the TSNR algorithm. So, this technique has the ability to immediately track the non-stationarity of the speech signal without introducing musical noise effects which is illustrated in the context of voice communication. In addition, in automatic speech recognition application, the TSNR algorithm exhibits a significant reduction of substitution and insertion errors leading to a substantial relative recognition performance improvement.

6. REFERENCES

[1] P. Scalart, and J. Vieira Filho, "Speech Enhancement Based on a Priori Signal to Noise Estimation," IEEE Int. Conf. on Acoustics, Speech and Signal Proc., pp. 629–632, 1996.

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

[3] O. Cappé, "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor," IEEE Trans. on Speech and Audio Proc., Vol. 2, No. 2, pp. 345–349, April 1994.

[4] H.G. Hirsch, and D. Pearce, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions," Proc. of the ISCA ITRW ASR2000, pp. 181–188, 2000.

[5] J.S. Lim, and A.V. Oppenheim, "Enhancement and Bandwith Compression of Noisy Speech," IEEE Proc., Vol. 67, No. 12, pp. 1586–1604, December 1979.

[6] "ETSI ES 202 050 v1.1.1 STQ; distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," 2002.