SPEECH ENHANCEMENT USING HARMONIC REGENERATION

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

[1] France Télécom - TECH/SSTP, 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@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 introduce harmonic distortion in enhanced speech because of the non reliability of estimators for small signal-to-noise ratios. We propose a new method called Harmonic Regeneration Noise Reduction technique which solves this problem. A fully harmonic signal is calculated based on the distorted signal using a non-linearity to regenerate harmonics in an efficient way. This artificial signal is then used to compute a suppression gain able to preserve the speech harmonics. This method is theoretically analyzed, then objective and formal subjective results are given and show a significant improvement compared to classical noise reduction techniques.

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. In [1] is presented 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 Signal-to-Noise Ratio (SNR) for each frequency bin. As a consequence, the performance (trade-off between distortions and noise reduction) of the noise reduction technique depends on the quality of the SNR estimator. One major limitation that exists in common short-time suppression techniques is that they are unable to enhance speech harmonics where the SNR is small. Notice that in most of spoken languages, voiced sounds represent a large amount of the pronounced sounds. Then it is very interesting to overcome this limitation. For each frequency bin, an harmonic with an unfavorable local SNR is suppressed by the noise reduction algorithm because it considers that no speech signal is present. This phenomenon appears because the suppression gain is computed only based on *a posteriori* measures. Finally, common noise reduction algorithms suppress some harmonics existing in the original signal and then the enhanced signal sounds degraded.

To overcome this problem, we propose a new method, called Harmonic Regeneration Noise Reduction (HRNR) technique, that takes into account the harmonic characteristic of speech. In this approach, the output signal of a common noise reduction technique (with missing or degraded harmonics) is further processed to create an artificial signal where the missing harmonics have been automatically regenerated. Then this artificial signal is used to compute a suppression gain that will preserve all the harmonics. An analysis of HRNR technique behavior is proposed and results are given in terms of objective and subjective measures in the context of voice communication.

2. CLASSICAL NOISE REDUCTION RULE

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 $S(p, \omega_k)$, $N(p, \omega_k)$ and $X(p, \omega_k)$ designate the ω_k spectral component of short-time frame p of the speech $s(t)$, the noise $n(t)$ 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

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

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

$$SNR_{prio}(p, \omega_k) = \frac{E\{|S(p, \omega_k)|^2\}}{E\{|N(p, \omega_k)|^2\}}. \tag{2}$$

In practical implementations of speech enhancement systems, the power spectrum density of the speech $|S(p, \omega_k)|^2$ and the noise $|N(p, \omega_k)|^2$ are unknown as only the noisy speech is available. Then, both the *a posteriori* 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

$$\hat{\gamma}_{nn}(p, \omega_k) = \lambda \hat{\gamma}_{nn}(p-1, \omega_k) + (1-\lambda)|X(p, \omega_k)|^2 \tag{3}$$

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

$$S\hat{N}R_{post}(p, \omega_k) = \frac{|X(p, \omega_k)|^2}{\hat{\gamma}_{nn}(p, \omega_k)}, \tag{4}$$

$$\begin{aligned} S\hat{N}R_{prio}(p, \omega_k) = {} & \beta \frac{|\hat{S}(p-1, \omega_k)|^2}{\hat{\gamma}_{nn}(p, \omega_k)} \\ & + (1-\beta) P[S\hat{N}R_{post}(p, \omega_k) - 1] \end{aligned} \tag{5}$$

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 (5) corresponds to the so-called decision-directed approach [2, 3] with a behavior controlled by the parameter β (typically $\beta = 0.98$). The multiplicative gain function $G(p, \omega_k)$ is obtained by

$$G(p, \omega_k) = g(S\hat{N}R_{prio}(p, \omega_k), S\hat{N}R_{post}(p, \omega_k)) \qquad (6)$$

and the resulting speech spectrum is estimated as follows

$$\hat{S}(p, \omega_k) = G(p, \omega_k)X(p, \omega_k). \qquad (7)$$

The function g can be the different gain functions proposed in the literature (*e.g.* amplitude and power spectral subtraction, Wiener filtering, *etc.*) [1, 2, 4] and especially can be replaced by the Two-Step Noise Reduction approach proposed in [5]. For the following, the function g corresponds to the Wiener filter, and then

$$G(p, \omega_k) = \frac{S\hat{N}R_{prio}(p, \omega_k)}{1 + S\hat{N}R_{prio}(p, \omega_k)}. \qquad (8)$$

3. SPEECH HARMONIC REGENERATION

The output signal $\hat{S}(p, \omega_k)$, or $\hat{s}(t)$ in the time domain, obtained by the noise reduction technique presented in the previous section suffers from distortions. In fact some harmonics have been considered as noise only components and then have been suppressed. We propose to process this signal to create a fully harmonic signal where all the missing harmonics are regenerated. This will be called the speech harmonic regeneration step.

3.1. Principle of harmonic regeneration

There exists a simple and efficient way to restore signal harmonics, it consists in applying a non-linear function NL (absolute value, minimum or maximum relative to a threshold, *etc.*) to the time signal. Then the artificially restored signal $s_{harmo}(t)$ is obtained by

$$s_{harmo}(t) = NL(\hat{s}(t)). \qquad (9)$$

Notice that the restored harmonics of $s_{harmo}(t)$ are created at the same positions as the clean speech ones. This very interesting and important characteristic is implicitly assured because a non-linearity in the time domain is used to restore them. For illustration, Fig. 1 shows the typical behavior and the interest of the non-linearity. Figure 1.(a) represents a reference frame of voiced clean speech. Figure 1.(b) represents the same frame after being corrupted by noise and enhanced by the noise reduction technique presented in section 2. It appears clearly that some harmonics have been completely suppressed or severely degraded. Figure 1.(c) represents the artificially restored frame obtained using (9). It can be shown that the non-linearity applied to the signal $\hat{s}(t)$ has successfully restored the suppressed or degraded harmonics.

The signal $s_{harmo}(t)$ possesses very useful information that can be exploited to compute a new suppression gain which will be able to preserve all the harmonics of the speech signal. The new suppression gain, $G_{harmo}(p, \omega_k)$, which preserves the harmonics is computed as follows

$$G_{harmo}(p, \omega_k) = h(S\hat{N}R_{harmo}(p, \omega_k), S\hat{N}R_{post}(p, \omega_k)), \qquad (10)$$

where

$$S\hat{N}R_{harmo}(p, \omega_k) = \frac{\rho|\hat{S}(p, \omega_k)|^2 + (1 - \rho)|S_{harmo}(p, \omega_k)|^2}{\hat{\gamma}_{nn}(p, \omega_k)}. \qquad (11)$$

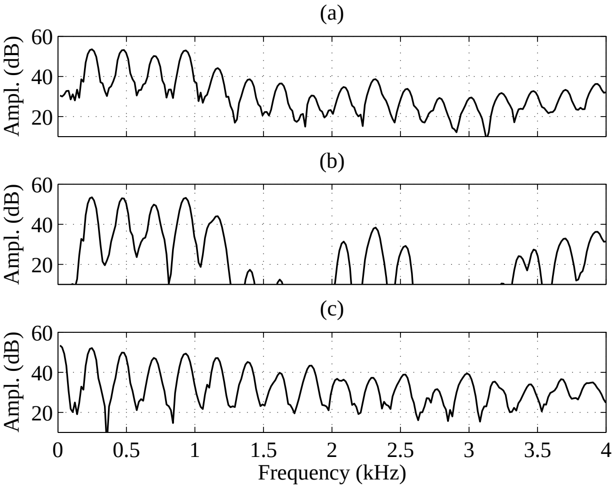

Fig. 1: Effect of the non-linearity on a voiced frame. (a) Clean speech spectrum; (b) Enhanced speech spectrum after classical suppression rule; (c) Restored speech spectrum after harmonic regeneration.

The function h can be the different gain functions proposed in the literature (*e.g.* amplitude and power spectral subtraction, Wiener filtering, *etc.*) [1, 2, 4]. The ρ parameter is a constant used to control the mixing level of $|\hat{S}(p, \omega_k)|^2$ and $|S_{harmo}(p, \omega_k)|^2$ depending on the chosen non-linear function (typically $0 < \rho < 1$). This mixing is necessary because the harmonic function is able to restore harmonics at the desired frequencies, but with biased amplitudes. Finally, the resulting speech spectrum is estimated as follows

$$\hat{S}(p, \omega_k) = G_{harmo}(p, \omega_k)X(p, \omega_k). \qquad (12)$$

The suppression gain $G_{harmo}(p, \omega_k)$ has the ability to preserve the harmonics suppressed by most of the common algorithms, and then avoids distortions.

3.2. Analysis of harmonic regeneration

To analyze the harmonic regeneration step, we will focus on a particular non-linearity, without loss of generality, the maximum (Max) relative to zero. Replacing the non-linear function NL by the Max function in (9), it follows

$$s_{harmo}(t) = Max(\hat{s}(t), 0) = \hat{s}(t)p(\hat{s}(t)) \qquad (13)$$

where $p(t)$ is defined as

$$p(t) = \begin{cases} 1 & \text{if } t > 0 \\ 0 & \text{if } t < 0. \end{cases} \qquad (14)$$

Figure 2 represents a frame of the voiced speech signal $\hat{s}(t)$ (dotted line) and the corresponding $p(\hat{s}(t))$ signal (dashed line). Notice that this signal is scaled to make the figure clearer. If we analyze this figure, we notice that the $p(\hat{s}(t))$ signal resumes to a repetition of an elementary waveform (solid line) with a periodicity T, corresponding to the voiced speech pitch. Since the quasi-stationarity of the speech is assumed over the frame duration, the Fourier transform (FT) of $p(\hat{s}(t))$ reduces to the sampling, by step of $\frac{1}{T}$, of the

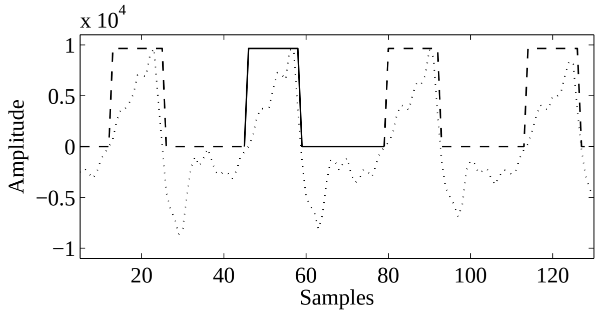

Fig. 2: Voiced speech frame $\hat{s}(t)$ (dotted line) and associated scaled $p(\hat{s}(t))$ signal (dashed line). Repeated elementary waveform (solid line).

FT of the elementary waveform:

$$FT(p(\hat{s}(t))) = \frac{1}{T}\sum_{m=-\infty}^{\infty} R\left(\frac{m}{T}\right)\delta\left(f-\frac{m}{T}\right) \tag{15}$$

where f denotes the continuous frequency and $R(\frac{m}{T})$ is the Fourier transform (FT) of the elementary waveform taken at discrete frequency $\frac{m}{T}$. Finally, using (13), the FT of $s_{harmo}(t)$ can be written as

$$FT(s_{harmo}(t)) = FT(\hat{s}(t)) * \frac{e^{-j\theta}}{T}\sum_{m=-\infty}^{\infty} R\left(\frac{m}{T}\right)\delta\left(f-\frac{m}{T}\right) \tag{16}$$

where θ is the phase at origin. So the spectrum of the restored signal, $s_{harmo}(t)$, is a convolution between the spectrum of $\hat{s}(t)$, signal enhanced by the classical rule, and an harmonic comb. This one has the same fundamental frequency as the frame of voiced speech signal processed which explains the phenomenon of harmonic regeneration. The main advantage of this method is its simplicity to restore speech harmonics at desired positions.

4. EXPERIMENTAL RESULTS

In the following, the technique presented in section 2 will be referred as Classical Noise Reduction (CNR) technique and the associated suppression gain is the Wiener filter (g in section 2), expressed by (8). For the proposed HRNR technique, the two associated suppression gains are also expressed by (8) ($g = h$, *cf.* section 3). The chosen non-linear function is the maximum (Max) relative to zero, *cf.* (13). The mixing parameter ρ used in (11) is set to 0.5.

4.1. Illustration of HRNR behavior

Figure 3 shows three spectrograms, Fig. 3.(a) represents the noisy speech corrupted by car noise (SNR=12dB) and Fig. 3.(b) and Fig. 3.(c) represent the noisy speech enhanced using CNR technique and using the proposed HRNR technique, respectively. Notice that no threshold is used to constraint the noise reduction filter in these two cases to make the spectrograms clearer. It appears that many harmonics can be preserved using HRNR technique whereas they are suppressed with CNR technique. So, to take into account the voiced characteristic of speech makes it possible to enhance harmonics completely degraded by noise.

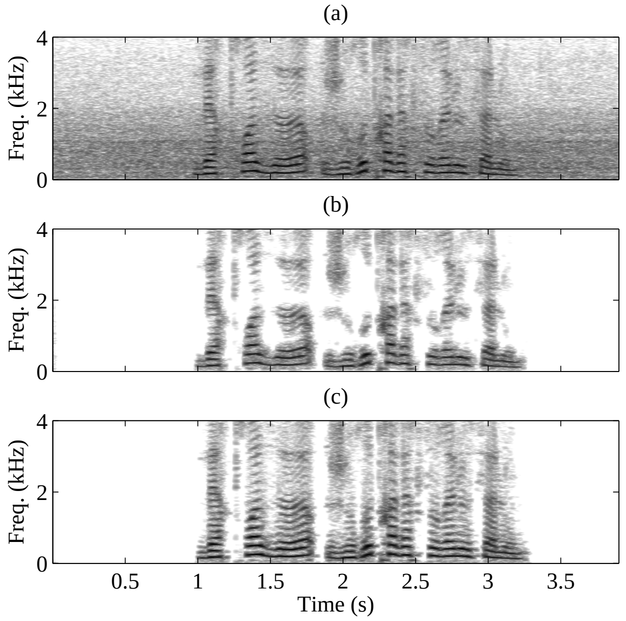

Fig. 3: Speech spectrograms. (a) Noisy speech corrupted by car noise at 12dB SNR; (b) Noisy speech enhanced by CNR technique; (c) Noisy speech enhanced by HRNR technique.

4.2. Objective results

Figure 4 shows the cepstral distance (CD) between clean speech and speech enhanced by CNR technique, Fig. 4.(a), and speech enhanced by HRNR technique, Fig. 4.(b), respectively. The noisy speech to enhance is the same as in Fig. 3.(a). The CD is a degradation measure correlated with subjective tests. The CD for HRNR technique is much smaller than for CNR technique, therefore the HRNR technique introduces less distortions than the CNR one. Notice that in Fig. 4, high peaks are located in low energy zones and then correspond to low perceptually important zones (*cf.* Fig. 3). Finally, this results in a better quality of the enhanced speech.

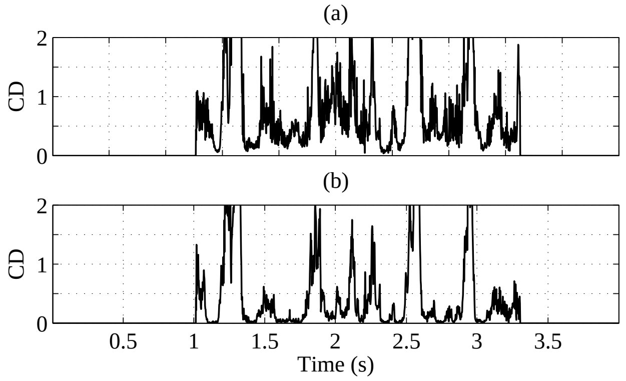

Fig. 4: Cepstral distances (CD) between clean speech and (a) speech enhanced by CNR technique and (b) speech enhanced by HRNR technique, respectively.

The input SNRs of noisy speech and the corresponding segmental SNR improvements obtained by CNR and HRNR 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 input SNRs are computed in two ways, using the segmental SNR and using the ITU-T recommendation P.56 speech voltmeter (SVP56). In fact, the SVP56 measure is used to create the noisy speech sentences at given SNRs. The segmental SNR measure takes into account both residual noise level and speech degradation. The proposed HRNR technique achieves the best results (bold values) under all noise conditions. Furthermore, the residual noise has the same structure and level (the noise reduction is limited to 19dB using a threshold applied to the suppression gain) for the two compared techniques. As a consequence, the differences between the segmental SNR improvement of CNR and HRNR techniques is mainly due to the very small level of speech degradation of HRNR technique.

Table 1: Segmental SNR improvement with HRNR technique compared to CNR technique in various noise and SNR conditions.

Noise	Input SNR (dB)		Seg. SNR improv. (dB)	
type	SVP56	Segmental	CNR	HRNR
Street	12	8.72	10.58	**13.46**
	18	13.14	13.46	**15.62**
	24	18.18	15.80	**17.20**
Car	12	8.95	11.75	**13.84**
	18	13.33	14.43	**15.96**
	24	18.33	16.40	**17.41**
Babble	12	9.74	12.08	**14.70**
	18	14.41	14.50	**16.36**
	24	19.60	16.38	**17.58**

4.3. Subjective test

Since the segmental SNR lacks indication about the subjective preference of listeners, a formal subjective test has been conducted. This formal subjective test is a Comparative Category Rating (CCR) test and follows the UIT-T P.800 recommendation. For each algorithm, the parameters are tuned to obtain optimal tradeoff between noise reduction and speech distortion. This test was conducted with 24 listeners, 4 speakers (2 females and 2 males), 9 sentences per speaker, 3 SNR conditions (12, 18 and 24dB) and 3 noise types (Street, Car and Babble). The listeners had to listen the sentences by pairs (classical technique - proposed technique or in reverse order, the order being random) and then rate the second sentence in contrast to the first one. The scale is -5 to 5 by steps of 1. The listeners use this scale to give global preference that take into account both level of residual noise and level of distortions. The results obtained are displayed in Fig. 5. The CMOS (Comparative Mean Opinion Score) score and the associated confidence interval are function of the SNR and the noise type. A positive value indicates that the HRNR technique is preferred over the CNR one. We can notice that the HRNR technique is always preferred, with significant mean scores, to the CNR technique which is in agreement with the segmental SNR results presented in Table 1. We can observe that there is less improvement for the babble noise (speech-like noise) than for street and car noises. We can also notice that the improvement increases with the SNR.

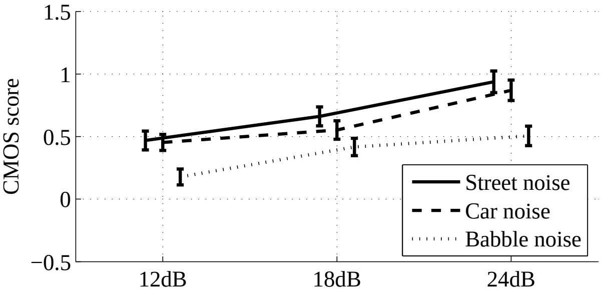

Fig. 5: Formal subjective test results. CMOS scores and confidence intervals for three SNR (12, 18 and 24dB) and three noises.

5. CONCLUSION

In this paper, we proposed a new noise reduction technique based on the principle of harmonic regeneration. Common techniques suffer from harmonic distortions when the SNR is too small. Then we proposed to create a fully harmonic signal using a non-linear function of the distorted signal. This is the speech harmonic regeneration step. Then, this signal is used to compute a suppression gain that is able to preserve speech harmonics, and hence avoids distortions. The role of the non-linearity and the principle of harmonic regeneration are detailed and analyzed. Results are given in terms of spectrum distortions and cepstral distance to illustrate the efficiency of the HRNR technique. Furthermore, results are given in terms of segmental SNR based on a large corpus of signals. All these results exhibit the good performance of the HRNR technique in terms of objective results. To be more complete, results of a formal subjective test are given and confirm the significant efficiency of the proposed technique.

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., Vol. 2, pp. 629–632, 7–10 May 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] 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.

[5] C. Plapous, C. Marro, P. Scalart, and L. Mauuary, “A Two-Step Noise Reduction Technique,” IEEE Int. Conf. on Acoustics, Speech and Signal Proc., Vol. 1, pp. 289–292, 17–21 May 2004.