SNR Estimation of Speech Signals Using Subbands and Fourth-Order Statistics

Elias Nemer, *Student Member, IEEE,* Rafik Goubran, *Member, IEEE,* and Samy Mahmoud, *Senior Member, IEEE*

***Abstract*— This letter addresses the problem of instantaneous signal-to-noise ratio (SNR) estimation during speech activity for the purpose of improving the performance of speech enhancement algorithms. It is shown that the kurtosis of noisy speech may be used to individually estimate speech and noise energies when speech is divided into narrow bands. Based on this concept, a novel method is proposed to continuously estimate the SNR across the frequency bands without the need for a speech detector. The derivations are based on a sinusoidal model for speech and a Gaussian assumption about the noise. Experimental results using recorded speech and noise show that the model and the derivations are valid, though not entirely accurate across the whole spectrum; it is also found that many noise types encountered in mobile telephony are not far from Gaussianity as far as higher statistics are concerned, making this scheme quite effective.**

***Index Terms*— Higher-order statistics, signal-to-noise ratio, speech processing.**

I. Introduction

SPEECH enhancement algorithms based on spectral decomposition, such as Wiener filtering and maximum likelihood [1], rely on an accurate estimation of the background noise energy and the signal-to-noise ratio (SNR) in the various frequency bands. Traditionally, the noise spectrum is estimated during segments of nonspeech activity and used for SNR computations. In practice, this is seldom sufficient as the noise spectrum changes during speech. The resulting poor SNR estimation limits the effectiveness of the suppression filters and often results in noise artifacts.

A number of approaches have been proposed to estimate the SNR without the need for a speech detector, but the problem is far from being closed: The iterative estimation proposed in [2] works well in most situations, but being based on relative energy levels, cannot distinguish between rising noise energy and the presence of speech. The spectral analysis method in [3] requires a long segment of speech and a good frequency resolution to work effectively and overcome the discrete Fourier transform (DFT) windowing errors.

Higher-order statistics (HOS) have shown promising results in a number of applications and are of particular value when dealing with a mixture of Gaussian and non-Gaussian processes [4]. In [5], we explored some of the peculiarities of the third-order statistics of speech and showed how it may be used to estimate the pitch and detect voicing.

The idea of using HOS for SNR estimation hinges on being able to separate signal and noise energies based on these statistics. In this letter, we show that the fourth-order statistics, along with a subbanding scheme, allows us to do just that when speech is divided into narrow bands, such that at most one harmonic falls in a given band. The expression for the kurtosis of subbanded speech is derived assuming a sinusoidal model [6], and is shown to be a function of the speech energy. This, coupled with the kurtosis of noise being zero, allows one to individually estimate speech and noise energies from the kurtosis and variance of noisy speech. The resulting scheme is attractive in allowing a continuous estimation of the SNR during both speech and nonspeech segments.

II. A Model for Subbanded Speech

The *zero-phase harmonic representation* proposed in [6] is among the simplest sinusoidal models for speech analysis and synthesis. Its elegance is in the use of the same expression for both voiced and unvoiced speech and allowing for a soft decision whereby a frame may contain both types. A short-term segment of speech is expressed as a sum of sine waves that are coherent (in-phase) during steady voiced speech and incoherent during unvoiced speech, as follows:

$$s(n) = \sum_{m=1}^{M} A_m . \cos[(n - n_0)\omega_m + \psi_m + \phi_m] \qquad (1)$$

where n_0 is the voice onset time, M the number of sinusoids, A_m the amplitude, and ω_m the excitation frequency of the mth sine wave. The first phase term is due to the onset time n_0 of the pitch pulse; the second depends on a frequency cutoff ω_c and a voicing probability, P_v, so the higher the voicing probability the more sine waves are declared voiced with zero phase. The third phase component is the system phase ϕ_m along frequency track m, often assumed zero or a linear function of the frequency.

Given that speech is divided in narrow bands such that at most one harmonic falls in each band, then in light of the model, voiced speech in a given band is modeled as a single sinusoid with deterministic phase, whereas unvoiced speech is modeled as a sinusoid with random phase.

III. Higher Moments of Subbanded Speech

A. Definitions

If $x(n)$, $n = 0, \pm 1, \pm 2, \pm 3, \cdots$, is a real stationary discrete-time signal and its moments up to order p exist, then

Manuscript received November 4, 1998. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. J. H. L. Hansen.

E. Nemer is with Nortel Networks, Verdun, P.Q. H3E 1H6, Canada (e-mail: enemer@ieee.org).

R. Goubran and S. Mahmoud are with the Department of Systems and Computer Engineering, Carleton University, Ottawa, Ont. K1S 5B6, Canada (e-mail: goubran@sce.carleton.ca; mahmoud@sce.carleton.ca).

Publisher Item Identifier S 1070-9908(99)04965-2.

its pth-order moment function is given by

$$m_{px}(\tau_1, \tau_2, \cdots, \tau_{p-1}) \equiv E[x(n)x(n+\tau_1) \cdots x(n+\tau_{p-1})]$$

and depends only on the time differences τ_i. For a deterministic signal, the statistical expectation $E\{.\}$ is replaced by a time summation (or averaging for power signals). The pth moment of $x(n)$ is found by setting all lags to zero:

$$M_{px} \equiv m_{px}(0, 0, \cdots, 0) = E[x^p(n)]. \tag{2}$$

The cumulant functions of $x(n)$ may be written in terms of the moment functions, m_{px} [4]. The kurtosis is obtained by setting all lags to zero in the fourth cumulant:

$$C_{4x} \equiv M_{4x} - 3\{M_{2x}\}^2. \tag{3}$$

B. Voiced Speech—Deterministic Phase

Given a signal consisting of a single sinusoid of deterministic amplitude, frequency and phase: $s(n) = A\cos(nT\omega + \theta)$, where T is the sampling period. Its higher-order moments computed over a length-N segment are

$$M_{ps} = E[s^p(n)] = \frac{1}{N}\sum_{n=0}^{N-1}\{A\cos(nT\omega+\theta)\}^p$$

for integer values of p. The above is evaluated for $p = 2$ and 4 by first noting the identity

$$\sum_{n=0}^{N-1}\cos(2knT\omega) = \frac{\sin(Tk\omega N)\cos[Tk\omega(N-1)]}{\sin(Tk\omega)}$$

for real values of k and M. Thus, the second moment is

$$\begin{aligned} M_{2s} &= \frac{A^2}{2N}\sum_{n=0}^{N-1}\{\cos[2(nT\omega+\theta)]+1\} \\ &= \frac{A^2}{2}\left\{\frac{\sin(T\omega N)\cos(T\omega[N-1]+2\theta)}{N\sin(T\omega)}+1\right\} \end{aligned} \tag{4}$$

and the fourth moment[1]

$$\begin{aligned} M_{4s} &= \frac{A^4}{8N}\sum_{n=0}^{N-1}\{\cos[4(nT\omega+\theta)]+4\cos[2(nT\omega+\theta)]+3\} \\ &= \frac{A^4}{8}\left\{\frac{\sin(2T\omega N)\cos[2(T\omega[N-1]+2\theta)]}{N\sin(2T\omega)}\right. \\ &\quad \left. + \frac{\sin(T\omega N)\cos(T\omega[N-1]+2\theta)}{N\sin(T\omega)}+3\right\}. \end{aligned} \tag{5}$$

In the above expressions, the trigonometric terms vanish whenever N is a multiple of the signal period ($N = 2k\pi/Tw$) or is large enough. The bias error due to these terms is bounded by $(1/N)$. If the effect of these terms is removed, the expressions for the moments simplify to

$$M_{2s} = A^2/2, \quad \text{and} \quad M_{4s} = 3A^4/8. \tag{6}$$

Thus, the kurtosis may be written in terms of the second moment M_{2s}, or signal energy, using (3) and (6), as follows:

$$C_{4s} = M_{4s} - 3(M_{2s})^2 = -3A^4/8 = -1.5(E_s)^2 \tag{7}$$

[1]We use the identity $\cos^4(A) = \frac{1}{8}\cos 4A + \frac{1}{2}\cos(2A) + \frac{3}{8}$.

where E_s denotes speech energy: $E_s \equiv M_{2s}$. It is to note here that another way to eliminate the bias terms is to compute the higher moments for few overlapping frames then average the results prior to computing the kurtosis (7). This is equivalent to integrating the results over the entire range of the phase θ (assumed uniform over $[-\pi, \pi]$) in (4) and (5). It is easy to see from the equations above that this integration will result in a zero bias term.

C. Unvoiced Speech—Random Phase

Given a signal consisting of a single sinusoid of deterministic amplitude and frequency and uniformly distributed random phase, $s(n) = A\cos(nT\omega + \theta)$, where T is the sampling period. Its higher-order moments are

$$M_{ps} = E[s^p(n)] = A^p E[\cos^p(nT\omega+\theta)]$$

for integer values of p. We first note that

$$E[\cos k(nT\omega+\theta)] = \frac{1}{2\pi}\int_{-\pi}^{\pi}\cos k(nT\omega+x)\,dx = 0$$

for integer values of k. Therefore, the second moment is

$$M_{2s} = \frac{A^2}{2}E[\cos 2(nT\omega+\theta)+1] = A^2/2 \tag{8}$$

and the fourth moment is

$$\begin{aligned} M_{4s} &= \frac{A^4}{8}E[\cos 4(nT\omega+\theta)+4\cos 2(nT\omega+\theta)+3] \\ &= 3A^4/8. \end{aligned} \tag{9}$$

The kurtosis is therefore the same as in the deterministic case given by (7).

IV. SNR Estimation from the Kurtosis

The energy and kurtosis in each band k are computed using N-point segments

$$E_x(k) = M_{2x} = \frac{1}{N}\sum_{n=0}^{N-1}x_k^2(n) \tag{10}$$

$$C_{4x}(k) = \frac{1}{N}\sum_{n=0}^{N-1}x_k^4(n) - 3\left(\frac{1}{N}\sum_{n=0}^{N-1}x_k^2(n)\right)^2. \tag{11}$$

Since cumulants are cumulative and since the kurtosis of Gaussian noise is zero [7], then when the signal consists of both speech and noise, the second moment is the sum of the two moments (or energies), whereas the kurtosis is simply that of speech and, from (7), is expressed in terms of speech energy:

$$E_x(k) = E_s(k) + E_n(k) \tag{12}$$

$$C_{4x}(k) = C_{4s}(k) = -1.5[E_s(k)]^2. \tag{13}$$

Therefore, in a given band k, speech energy is estimated using $C_{4x}(k)$ and (13)

$$\tilde{E}_s(k) = \sqrt{P\left[\frac{C_{4x}(k)}{-1.5}\right]}; \quad P[x] = \begin{cases} x & x > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

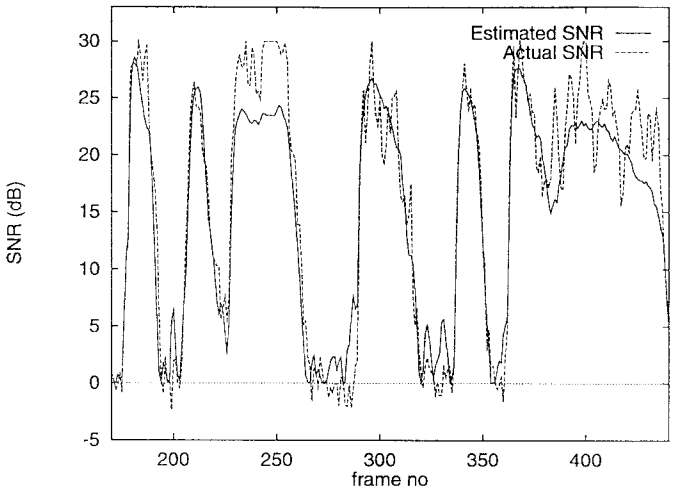

Fig. 1. Actual versus estimated SNR in a lower band (Gaussian).

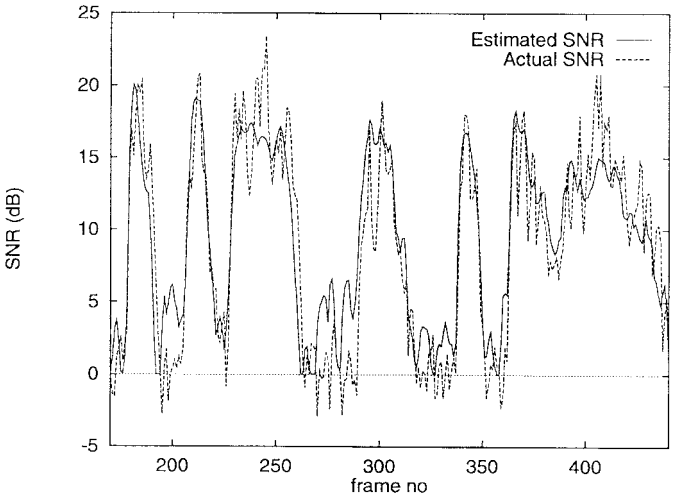

Fig. 2. Actual versus estimated SNR in a lower band (Street).

The noise energy is [from (12)]

$$\tilde{E}_n(k) = E_x(k) - \tilde{E}_s(k). \tag{15}$$

To smooth out the variance, both estimators are averaged using the autoregressive scheme

$$\hat{E}_s(k, j) = (1-\alpha)\hat{E}_s(k, j-1) + \alpha\tilde{E}_s(k) \tag{16}$$

$$\hat{E}_n(k, j) = (1-\beta)\hat{E}_n(k, j-1) + \beta\tilde{E}_n(k) \tag{17}$$

where j denotes the iteration index and α and β are integration constants; simulation shows that values of $\alpha \approx 0.5$ and $\beta \leq 0.05$ give best results. The SNR in band k is then estimated using (16) and (17) as

$$\widehat{\mathrm{SNR}}(k, j) = \frac{\hat{E}_s(k, j)}{\hat{E}_n(k, j)} = P\left[\frac{E_x(k, j)}{\hat{E}_n(k, j)} - 1\right]. \tag{18}$$

V. Simulation Results

Clean speech recorded and sampled at 8 kHz is used and mixed with different noise sources. The resulting signal is subbanded using 50 cosine-modulated filters [8]. Analysis in each band is done using 100 points, with 30% overlap. The actual and estimated SNR are computed and limited to the range $[-2, 30]$ dB, which is the main region of operation for most suppression filters used in speech enhancement [1], [9]. Figs. 1 and 2 compare the two quantities in one of the lower bands when Gaussian and street noise are used, respectively. The estimation errors, though not very significant, are partly due to the smoothing effects of the α and β coefficients in (16) and (17), and partly to the estimation errors when computing the HOS of a signal from finite data length. In the context of speech enhancement, however, these errors, which happen mostly at the edge of the SNR range, do not significantly affect the suppression filters, particularly when the SNR smoothing scheme in [9] is used. When the higher bands are examined however, the results have slightly degraded (Fig. 3). A possible explanation is that the sinusoidal model does not hold as well in the higher frequencies making the proposed scheme less effective. As a result, the estimation errors will mostly affect unvoiced speech, where energy is concentrated in the higher bands, but will not be as significant for voiced speech.

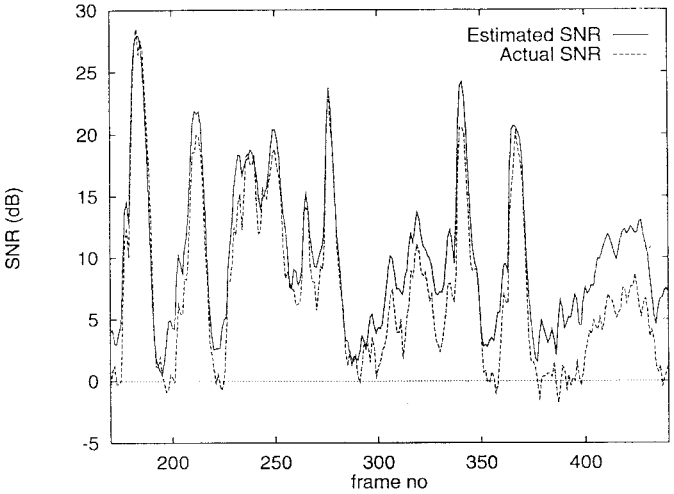

Fig. 3. Actual versus estimated SNR in a high band (Gaussian).

VI. Conclusions

We showed that by using a subbanding scheme, it is possible to separate signal and noise energies using the kurtosis and variance of noisy speech, and provide a continuous estimation of the SNR. We verified that the derivations of the higher moments and the underlying model are valid, though some degradation is noticed in the upper spectrum. Since the proposed scheme relies on higher statistics and not energy, it is effective in conditions where the noise energy changes, as long as the noise remains Gaussian-like. As part of future work, the algorithm is being incorporated in a speech enhancement system based on [9], and will be assessed in mobile telephony environments.

References

[1] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 137–145, Apr. 1980.

[2] H. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Proc. ICASSP 1995*, pp. 153–156.
[3] R. Martin, "An efficient algorithm to estimate the instantaneous SNR of speech signals," in *Proc. Eurospeech 1993*, pp. 1093–1096.
[4] C. Nikias and J. Mendel, "Signal processing with higher-order statistics," *IEEE Signal Processing Mag.*, vol. 10, pp. 10–38, July 1993.
[5] E. Nemer, R. Goubran, and S. Mahmoud, "The third order cumulant of speech signals with application to reliable pitch estimation," in *Proc. IEEE Workshop on Statistical Signal and Array Processing*, Sept. 1998, pp. 427–430.
[6] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, p. 744, Aug. 1986.
[7] D. R. Brilinger, "An introduction to polyspectra," *Ann. Math. Stat.*, vol. 36, pp. 1351–1374, 1965.
[8] R. D. Koilpillai and P. P. Vaidyanathan, "Cosine-modulated filter banks satisfying perfect reconstruction," *IEEE Trans. Signal Processing*, vol. 40, pp. 770–783, Apr. 1992.
[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *Trans. Acoust., Speech, Signal Procesing* vol. ASSP-32, pp. 1109–1121, Dec. 1984.