SPEECH ENHANCEMENT USING FOURTH-ORDER CUMULANTS AND TIME-DOMAIN OPTIMAL FILTERS

Elias Nemer¹, Rafik Goubran² and Samy Mahmoud²

¹Nortel Networks
16 Place Du Commerce
Verdun, Quebec
Canada, H3E 1H6
enemer@ieee.org

²Systems & Computer Eng’g
Carleton University
Ottawa, Ontario
Canada, K1S 5B6
[goubran, mahmoud]@sce.carleton.ca

ABSTRACT

A new method for speech enhancement based on optimal filtering, subbands, and higher-order cumulants is proposed in this paper. The key idea is to use the 4th cumulant to estimate the parameters required for the enhancement filters. It is shown that the kurtosis of noisy speech may be used to estimate the SNR and the probability of speech presence when speech is divided in narrow bands and modeled as a sinusoidal signal. The resulting algorithm is tested in typical mobile noise conditions and proves effective under such types as street, office and fan noises. Compared to the TIA IS-127 standard for noise reduction, the proposed algorithm is better at preserving the harmonic structure of the speech and results in overall more noise reduction in Gaussian-like conditions. However, this comes at the cost of slightly more noise artifact, mostly at very low SNR and non-Gaussian conditions.

1. INTRODUCTION

Speech enhancement by spectral decomposing and filtering [1] [2][9] remains a common and effective approach for enhancing speech degraded by acoustic additive noise when only the noisy speech is available. This general class is based on optimal filters and encompasses such methods as Wiener filtering, spectral subtraction, and maximum likelihood (ML) estimations. A common set of requirements in this class includes:

- An appropriate suppression rule that is based on some criteria of optimality [1] [2].
- An estimation of the speech and noise power spectral densities, or their respective autocorrelation.
- A quantification of the probability of speech presence in a given band, to further attenuate non-speech bands [9].

While much of the published work has focused on appropriate suppression rules, little has been done in the other aspects, not the least being the estimation of the 2nd statistics of the speech and noise, such as the SNR, which remains a crucial aspect for effective enhancement, or the quantification of the uncertainty of speech presence which is shown to improve the suppression of noise residuals [9] under a number of suppression rules. Higher-order statistics (HOS) have shown promising potential in a number of signal processing applications, and are of particular value when dealing with a mixture of Gaussian and non-Gaussian processes [8]. Their application to speech processing has been primarily motivated by their inherent Gaussian suppression and phase preservation properties. However, the traditional approaches have ignored the peculiarities of the HOS of speech signals or have neglected to show that they are indeed different from those of Gaussian processes. In [7], we explored some of these peculiarities by deriving the expression for the 3rd-order cumulant of the LPC residual of voiced speech and showed how it may be used to estimate the pitch. In this paper we show that using the 4th statistic along with a subbanding scheme allows us to build a speech enhancement system based on optimal filters. The idea is to use the kurtosis of the noisy speech to estimate such parameters as the SNR and the probability of speech presence in a band.

2. A SINUSOIDAL MODEL FOR SPEECH

The zero-phase harmonic representation proposed in [5] 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:

\[
x(n) = \sum_{m=1}^{M} a_{m} \cos \left( (n - n_{o}) w_{m} + \psi_{m} + \theta_{m} \right)
\]

where \(n_{o}\) is the voice onset time, \(M\) is the number of sinusoids, \(a_{m}\) the amplitude of the \(m^{th}\) sine wave and \(w_{m}\) the excitation frequencies. The first phase term is due to the onset time \(n_{o}\) of the pitch pulse. The second phase component depends on a frequency cutoff \(w_{c}\) and a voicing probability, \(P_{v}\), so that the higher the voicing probability the more sine waves are declared voiced with zero phase. The third phase component is the system phase \(\theta_{m}\) along frequency track \(m\), often assumed to be zero or a linear function of frequency.

Given that speech is divided in narrow bands such that at most two harmonics fall in each, then in light of the model, voiced speech in a given band is modeled as the sum of two sinusoids with deterministic phases, and unvoiced speech as two sinusoids with random phases.

3. ENHANCEMENT BY OPTIMAL FILTERS

The problem of extracting a signal from noise entails estimating a desired process \(S(n)\) from the \((p+1)\) most recent noisy observations:

10.21437/Eurospeech.1999-578
\( X_\alpha = S_\alpha + N_\alpha \quad \alpha \in \{ n - p, \ldots, n \} \)

If \( S_\alpha \) and \( N_\alpha \) are independent random processes, then it may be shown [4] that the optimal linear that will minimize the mean-square estimation error can be found by solving a set of \((p+1)\) linear equations involving the autocorrelation of speech: \( R_S[\tau] \) and that of the noise: \( R_N[\tau] \). In the simplest case where a single-tap filters are used (i.e., \( p = 0 \)), then the filter coefficients are given by:

\[
h_0 = \frac{R_S[0]}{R_S[0] + R_N[0]} = \frac{\text{SNR}}{\text{SNR} + 1} \tag{2}\]

Thus enhancement requires the knowledge of the SNR in each band. Moreover, the likelihood of speech presence in that band will allow further attenuation of non-speech bands. In the following sections, these two entities are estimated using the kurtosis of noisy speech.

4. HIGHER STATISTICS OF SUBBANDED SPEECH

4.1 Definitions

If \( x(n) \), \( n = 0, \pm 1, \pm 2, \pm 3, \ldots \) is a real stationary discrete-time signal and its moments up to order \( p \) exist, then its \( p^{th} \) order moment function is given by:

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

and depends only on the time differences \( \tau_i \) for all \( i \). The statistical expectation \( E\{.\} \) is replaced by a time summation (or averaging) for deterministic signals. The \( p^{th} \) moment of \( x(n) \) is found by setting all lags to zero:

\[
M_{px} = m_{px}(0,0,\ldots,0) = E\{|x|^p\}
\]

The cumulant functions of \( x(n) \) may be written in terms of the moments functions, \( m_{px}[8] \). The kurtosis is obtained by setting all lags to zero in the 4th cumulant:

\[
C_4x = M_{4x} - 3 \{M_{2x}\}^2 \tag{3}
\]

When estimating HOS from finite data record, the variance of the estimators is reduced by normalizing these statistics by the variance of \( x(n) \); thus the normalized kurtosis is:

\[
\gamma_4x = C_4x / [M_{2x}]^2 = [M_{4x} / [M_{2x}]]^2 - 3 \tag{4}
\]

4.2 The Kurtosis of Subbanded Speech

- **Proposition 1:** According to the sinusoidal model, the kurtosis of subbanded speech (both voiced and unvoiced) may be written in terms of the 4th powers of the harmonic amplitudes,

\[
C_4s = -3 \left[ a_1^4 + a_2^4 \right] / 8 \tag{5}
\]

and is upper and lower bounded by scale factors of the signal energy:

\[
-1.5 E_s^2 \leq C_4s \leq -0.75 E_s^2 \tag{6}
\]

- **Proof:** The signal is modeled as:

\[
s(t) = a_1 \cos (w_1 t + \phi_1) + a_2 \cos (w_2 t + \phi_2) \tag{7}
\]

where the phases are deterministic for voiced speech and uniformly distributed for unvoiced speech. The proof of the proposition is a special case of the more general expression for the cumulant slice derived in [10]. Moreover, it may be shown that the same results hold whether the phases are random or deterministic. Since the number of sinusoids is less than 4, then the expression is the same whether the frequencies are harmonically related or not.

To determine the bounds on the kurtosis, the two extreme cases on the amplitudes are considered:

**Case 1:** \( a_1 = a_2 \). The signal energy is: \( E_s = a_1^2 \). The kurtosis becomes: \( C_4s \simeq -3a_1^4 / 4 = -0.75 E_s^2 \).

**Case 2:** \( a_1 \neq a_2 \). The signal energy is: \( E_s \simeq a_1^2 / 2 \). The kurtosis becomes: \( C_4s \simeq -3a_1^4 / 8 = -1.5 E_s^2 \).

Therefore, for any values of the two amplitudes, the kurtosis is bounded by the signal energy as given in Eq 6. As a result, the normalized kurtosis is bounded by:

\[
-1.5 \leq \gamma_4s \leq -0.75 \tag{8}
\]

5. FOC-BASED ENHANCEMENT ALGORITHM

5.1 Probability of Speech Presence

The probability of speech presence in a band is the complement of the probability of noise-only being present:

\[
\text{Prob} [\text{Speech Presence}] = 1 - \text{Prob} [\text{Noise Only}] \tag{9}
\]

- **Proposition 2:** The probability of noise-only bands is determined by the estimate of the kurtosis, scaled by the variance of the estimator expressed in terms of the noise energy. Denoting this scaled values by ‘b’, then a reasonable quantification of this probability is:

\[
\text{Prob} [\text{Noise Only}] = \text{erfc} (|b|) \tag{10}
\]

- **Proof:** The kurtosis of Gaussian noise is zero only in a statistical average sense. Since in practice finite length frames are used, the decision that a given band in a frame is noise can only be made in a probabilistic sense with a confidence level that takes into account the variance and distribution of the estimator of the kurtosis. It is possible to show that in the case of a white Gaussian process \( g(n) \), the bias and variance of this estimate can be quantified in terms of the process variance \( \nu_g \) and the frame length \( N \). A new unbiased estimator is thus used and defined as:

\[
\hat{K}_U = \left( 1 + \frac{2}{N} \right) M_{4g} - 3 (M_{2g})^2 \tag{11}
\]

where \( M_{2g} \) and \( M_{4g} \) denote the computed 2nd and 4th moments. The distribution of this estimator is not straightforward, but an approximation is used here and the estimator assumed normally distributed. A unit-variance version of this zero-mean variable is defined as:

\[
\hat{K}_U a = \frac{\hat{K}_U}{\sqrt{\frac{3\nu_g^4}{N} \left( 104 + \frac{452}{N} + \frac{596}{N^2} \right)}} \tag{12}
\]

Thus, given the estimate of the kurtosis in a given band and the corresponding scaled value, ‘b’, the probability that the
band consists only of noise is given by Eq 10.

5.2 Noise Energy and SNR Estimation

Given that the 4th cumulant of noisy speech is simply that of the clean speech, and given that the HOS of Gaussian noise are zero, then the energies are estimated as follows:

• In bands where only noise is present, the noise energy is estimated using the total energy in that band.
• In bands where steady speech is present, an upper bound on the speech energy is derived from the kurtosis, from which a lower bound on the noise is deduced.

Once the energy of the speech and noise are estimated in a given band, the SNR is computed as:

$$\text{SNR}(k) = 0.1 \cdot \text{Prob noise}_b$$

where Pos \(x\) = x when \(x > 0\) and 0 otherwise. \(E_N^j\) is the estimate of the noise energy and \(E_X\) the total band energy. These three ideas are detailed below.

5.3.1 Noise Energy from noise-only bands

The probability that a band contains only noise can be quantified using the value of the kurtosis and the variance of the kurtosis estimator as explained in Proposition 2. In these bands, the estimate of the noise energy is updated using the total signal energy. In addition to the probability of noise, the normalized kurtosis is used to discriminate noise only bands, thus:

If \((\text{Prob}[\text{noise}] > T_{\text{noise}} \text{ and } \gamma_4 < T_{\gamma_4})\)  

$$\hat{E}_N^j(j) = (1 - \beta) \hat{E}_N^j(j - 1) + \beta E_X$$

where \(j\) denotes the iteration index, \(\hat{E}_N^j\) the estimate of the noise energy, \(E_X\) the total band energy and \(\beta\) is an integration constant with \(\beta < 0.1\).

5.4.2 Energies from steady state speech bands

• Proposition 3: In the case where steady speech is present in a band, then a lower bound on the noise energy in this band may be computed from the kurtosis and the total band energy:

$$\hat{E}_N^j = E_X - \sqrt{\frac{C_{4X}[0]}{-0.75}}$$

The detection of this steady state condition can be done using the normalized kurtosis.

• Proof: When a band contains steady speech, the kurtosis is upper and lower bounded by a scale factor of the signal energy and the normalized kurtosis is then contained in the range [-1.5, -0.75]. When noise is present, the normalization may yield a smaller amplitude for the normalized kurtosis. However the normalized kurtosis is always negative and cannot be smaller than -1.5. Since cumulants are cumulative and since the kurtosis of Gaussian noise is zero, then when the signal consists of both speech and noise, the total energy is the sum of the two energies, whereas the kurtosis of noisy speech is simply that of clean speech:

$$E_X = E_S + E_N \text{ and } C_{4X} = C_{4S}$$

From Eq 6, one can compute an upper bound on the speech energy from the kurtosis: \(E_S \leq \frac{C_{4X}[0]}{-0.75}\) and from it a lower bound on the noise is deduced using Eq 15. Thus the noise energy is updated as follows:

If \((\gamma_4 < 0 \text{ and } \gamma_4 \geq -1.5 \text{ and Prob}[\text{noise}] < T_{\text{noise}})\)  

$$\hat{E}_N^j(j) = (1 - \beta) \hat{E}_N^j(j - 1) + \beta [\hat{E}_N^j]$$

where \(j\) denotes the iteration index, \(\hat{E}_N^j\) the estimate of the noise energy, \(E_X\) the total band energy and \(\beta\) an integration constant with \(\beta < 0.1\).

5.5 Other Considerations

5.6.1 Inter-Frame SNR Smoothing

The SNR smoothing scheme proposed in [1] is used whereby the variation in SNR between successive frames is reduced significantly by averaging the locally computed SNR (SNR\(_{\text{post}}\)) with the SNR estimated in the previous frame after the filtering operation (SNR\(_{\text{est}}\)):

$$\text{SNR}_{\text{prior}}(k) = (1 - \gamma) \text{SNR}_{\text{post}}(k) + \gamma \text{SNR}_{\text{est}}(k)$$

Where SNR\(_{\text{post}}(k)\) refers to the local SNR from this iteration (Eq 13) and SNR\(_{\text{est}}(k)\) is computed in the previous iteration after the filtering operation, i.e. using the energy of the estimated speech instead of the energy of the noisy speech.

5.7.2 Frequency Masking Psycocoacoustics

In computing SNR\(_{\text{post}}(k)\) and SNR\(_{\text{est}}(k)\), the masking effect of the auditory system is taking into account by convolving both total and noise energies with the critical band filter at each frequency. The auditory filters are modeled using the polynomial functions proposed in [6].

5.8.3 Subbanding Filters

A filter bank consisting of \(M=50\) cosine-modulated filter pairs is used for analysis and synthesis. The expression for the analysis and synthesis filters is the one proposed in [3]:

$$H(l) = 2P(l) \cos \left( \frac{(2k + 1)(2l - L + 1 + M)\pi}{4M} \right)$$

$$F(l) = 2P(l) \cos \left( \frac{(2k + 1)(2l - L + 1 - M)\pi}{4M} \right)$$

where \(M\) is the number of filters (bands), \(L\) is the length of the filter and \(P(l)\) is the prototype filter.

5.9.4 Bounds on Filter Gains

In practice, a lower bound is used for the filter coefficients to prevent too much audible variations in the noise levels:

If \((\text{SNR}_{\text{prior}} < T_{\text{SNR2}} \text{ and Prob}[\text{noise}] > T_{\text{noise2}})\)  

$$h_0 = \text{MinGain}$$

If \((\text{SNR}_{\text{prior}} \geq T_{\text{SNR1}} \text{ and Prob}[\text{noise}] < T_{\text{Noise}})\)  

$$h_0 = 1$$
6. EXPERIMENTAL RESULTS

A block diagram of the algorithm is given in Fig 1. In addition to comparing the estimated and actual SNR in each band, subjective listening is used to evaluate the overall effectiveness of the algorithm. The deciding criteria include the degree of noise reduction, the degree of distortion of the speech and the degree of distortion -if any- of the residual noise. To provide a frame of reference, the results are compared to the TIA noise reduction algorithm [11]. This algorithm is based on spectral subtraction, using heuristic rules for gain computations and using SNR and stationarity measures for noise estimation.

At the range of the SNR considered [10, 14 dB], the degree of noise reduction is overall higher for the HOS-based algorithm than the TIA’s IS-127. This is particularly true for the Gaussian-like noises such as street and fan noise. The effectiveness however diminishes in the case of non-Gaussian noise, such as office noise where dominant conversations and impulsive machine noise are not detected as noise given their non-zero HOS. In general, it is found that the noise effectiveness comes at the cost of a slightly more audible noise artifacts in the case of the HOS algorithm, and likely due to the high variance of the HOS estimator that results in SNR fluctuation during periods of non-speech. The case of office noise may be heard in the attached wave file (available on the CD ROM only).

7. CONCLUSION

This paper presented an algorithm for enhancing speech corrupted by Gaussian noise, using optimal filtering in the time domain, subbands and fourth order cumulants. The idea is to use the kurtosis to estimate the required parameters for the enhancements filters, namely the SNR and the probability of speech presence.

The rationale for using a subband approach is that speech can be easily modeled analytically and the expression for the kurtosis may be quantified in terms of speech amplitudes from which a bound on the energy may be inferred. The resulting algorithm is shown to be effective on typical noises encountered in mobile telephony such as street, office and fan noise. It does not however eliminate the non-Gaussian components in these noises, such as dominant conversations, as these processes do not have zero HOS.

The computational complexity of the algorithm is mostly due to the analysis / synthesis stages which require a relatively large number of filters (40 -> 60). Using efficient implementations of these filters will significantly reduce the operations required.

Compared to the TIA IS-127 standard for noise reduction, the HOS algorithm is better at preserving the harmonic structure of the speech and results in less speech distortion. It also results in overall more reduction of the noise in Gaussian conditions, but that comes at the cost of more noise artifact, particularly at very low SNR and non-Gaussian noise conditions.

8. REFERENCES