Acoustic Noise Reduction For Mobile Telephony

Elias Nemer
Nortel Networks
enemer@ieee.org

1.0 INTRODUCTION

1.1 Context and Motivation

In the context of mobile telephony, speech signals are often corrupted by surrounding acoustic noise such as engine, traffic and wind as well as by system-introduced noise such as quantization, handoff, and channel interference. This in turn has an adverse effect on the perceived quality and intelligibility of speech as well as on the performance of speech processing algorithms throughout the network, such as speech coding and recognition.

Wireless telephony by definition has a lower speech quality than wireline, due to the speech coding process. If, in addition, the cellular system encodes a noisy signal prior to its transmission, then further degradation may occur, since coders rely on a model for the clean signal which is not suitable for the noisy signal. Similarly, speech recognition systems will degrade drastically in noisy environments, due to differences between testing and training conditions.

The aim of acoustic noise reduction is to minimize the effect of noise on the performance of voice communication systems. This means improving the perceived quality to the human listener as well as providing a more appropriate signal for estimating crucial speech parameters such as spectral content, pitch, voicing, and others.

The *quality* of speech signals is a subjective measure which reflects on the way the signal is perceived by listeners. It can be expressed in terms of how pleasant speech sounds to the human ear. *Intelligibility* on the other hand is an objective measure of the amount of information which can be extracted by listeners from the given signal [20]. *Intelligibility* is important in situations -such as military or emergencies- where the content of the message is critical.

1.2 Technical Challenge

Noise reduction is an ancient art, but is still far from a perfect science. Conceptually, it can be viewed as the combination of the classical problem of signal estimation, coupled with psychoacoustic aspects that account for the characteristics of the speech signal and the peculiarities of human hearing. The challenge is that the latter aspect is less understood, thus preventing one from formulating the problem in a way that leads to a globally optimum solution. As a result, a number

of suboptimal solutions based on mathematically tractable distortion measures or on some properties of the auditory system have been proposed.

1.3 Structure of this Paper

The paper is structured as follows: section 2 is the business perspective on NR. Section 3 provides a brief background on speech properties. The common approaches to NR are described in section 4. The system aspects of using NR in a cellular systems are discussed in sections 5 and 6. An overview of some commercial implementations is given in section 7.

2.0 NOISE REDUCTION: A BUSINESS PERSPECTIVE

2.1 Business Rationale

- *Market needs:* speech quality is gaining increasing importance in the context of Personal Communication Services as greater consumer acceptance is being sought by advertising these to be as reliable in service and quality as the wireline counterpart. Analysts predict that service functionality, which includes voice quality, will eventually become more important than price as a differentiator in wireless services.
- *Competitive importance:* telcos and manufacturers seek various proprietary solutions to improve the end-to-end voice quality on their networks in an attempt to differentiate their products from others.
- *Strategic importance:* voice processing technologies, such as speech enhancement or echo cancellation, are an integral part of voice communication network equipment. Ownership of this technology, as opposed to reliance on acquired solutions, is almost a necessity to minimize exposure given the strategic importance of these features.

2.2 Target Markets

Wireless telephony constitutes a large potential market for noise reduction. The technology however is also applicable to other communication contexts where ambient noise needs to be removed. The market thus targeted includes, though not limited to, the following applications:

- *Cellular phone hands free cradles:* to reduce ambient road, engine and background speaker to the far-end listener.
- *International long distance telephony:* to improve the intelligibility on low-quality international voice circuits where old analog technologies cause static and other switching noise on these communications.
- *Voice-activated phone dialing:* to increase the voice recognition hit rates when background noise is present.
- *Internet telephony:* to improve voice quality in noisy office environments, particularly where the input microphone is placed somewhere far from the speaker (ex on the computer terminal).
- *Teleconferencing:* to remove the effect of interfering speakers or background noise (chair movements, etc...) in handsfree teleconferencing.

- *Cockpit Intercom Systems:* to improve aircraft intercom and air-to-ground communications by removing wind and engine noise.
- *Voice Storage systems:* to improve compression vocoder performance in voice storage systems where ambient noise is an issue.
- *Emergency 2-way radio:* to improve voice intelligibility where vehicle noise (siren, etc.) disrupts communications.

2.3 Telephony Application Objectives

In wireless telephony, the objectives of telephone companies are to:

- Provide higher quality services to subscribers, that exceed other telcos'.
- Provide a 'comfortable' call to the end user, thus increasing the call holding patterns, and thus revenues.

The objectives of manufacturers on the other hand are to:

- Provide telcos with equipments and features that delivers consistent and high end-to-end voice quality to all users on that telco's network.
- Provide end-users with hands-free equipment and handsets that have more professional sounding quality, particularly appealing to business users.

3.0 Background on Speech Production and Perception

3.1 Speech Production

Speech is produced when air is forced from the lungs through the vocal cords and along the vocal tract. The vocal tract introduces short-term correlations (of the order of 1 ms) into the signal, and can be thought of as a filter with broad resonances, called *formants*, whose frequencies are controlled by varying the shape of the tract, like moving the position of the tongue. Speech sounds can be broken into three classes depending on their mode of excitation:

1. *Voiced* sounds: are quasi-periodic in the time domain and harmonically structured in the frequency domain. Their short-time spectrum is characterized by its *fine* and *formant* structure. The *fine* harmonic structure is attributed to the vibrating vocal chords. This *pitch* period is typically between 2 and 20 ms. The spectral envelope is characterized by a set of peaks, the *formants.* For the average vocal tract there are 3 to 5 formants below 5 Khz. The amplitudes and locations of the first 3 formants, usually occurring below 3 Khz, are quite important both in speech synthesis and perception.

2. *Unvoiced* sounds: result when the excitation is a noise-like turbulence produced by forcing air at high velocities through a constriction in the vocal tract. Such sounds show little long-term periodicity, although short-term correlations due to the vocal tract are still present. This aperiodic speech constitute about a third of the total speech. Example of unvoiced phonemes include /s/ (as in sort) and /S/ (shore).

3. *Plosive* sounds result when a complete closure is made in the vocal tract, and air pressure is built up behind this closure and released suddenly. Examples include /p/ (port) and /b/ (bay).

Some sounds are considered to be a mixture of two types: voiced fricatives (/z/) result when both vocal cord vibration and a constriction in the vocal tract are present.

3.2 Speech Perception

There is a limit to the sensitivity of the human ear, known as the threshold of hearing. This threshold varies with frequency: we are able to hear a much softer sound at 4 kHz than at 50 Hz or 15 kHz. At 25 kHz, this threshold is off the scale: no matter how loud the sound is, we can't hear it.

Another aspect of hearing is the phenomenon of *masking* in which the perception of one sound is obscured by the presence of another [19]. Simultaneous sounds cause frequency masking, where a lower frequency sound generally masks a higher-frequency one; sounds delayed with respect to one another can cause temporal masking of one or both sounds. Masking is due to the non-linearity of human hearing, that prevents treating the perception of many sounds as a summation of responses to their tone and bandlimited noise components.

The auditory system is more sensitive to the presence of *energy* than the absence of it, and tends to ignore aspects of phase.Voiced speech has a high amplitude and a concentration of energy at low frequency and is therefore more perceptually important than unvoiced speech. For this reason, most enhancement algorithms tend to concentrate on improving the periodic portions of speech. A proper representation of spectral amplitudes at harmonic frequencies particularly in the first 3 or 4 formant regions is of crucial important for high speech quality.

4.0 NOISE REDUCTION TECHNIQUES

4.1 Comb Filtering

Since the important audible component of speech is periodic, its harmonic frequencies may be identified for the purpose of either preservation or suppression. One basic method involves comb filtering, in which a dynamic filter is designed to *comb* through the spectrum, modifying energy at equally spaced frequencies to attenuate or enhance them. The frequency response of the filter resembles a *comb*, with large values at a specified F_O and its multiples, and low values between these harmonics. The impulse response over one period is given by [13]:

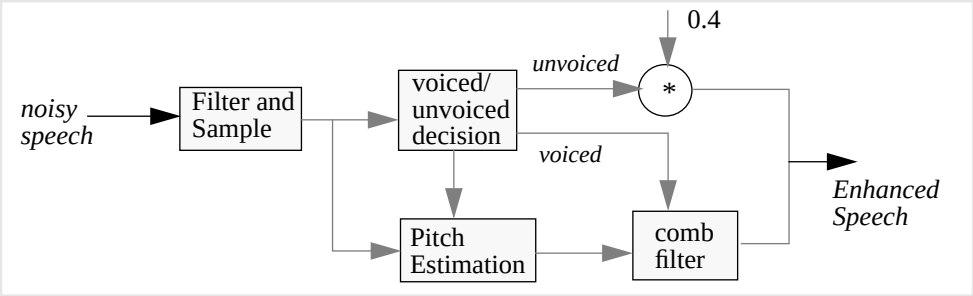
$$h(n) = \sum_{k=-L}^{L} a_k \cdot \delta(n - N_k) \qquad \textbf{(Eq 1)}$$

where $\delta(n)$ is a unit sample function, a_k are the filter coefficients, the length of the filter is $(2L + 1)$ pitch periods, and the pitch period estimate N_k is computed based on the pitch information of the speech frame. The filter coefficients are unchanged and only N_k is updated once every pitch period based on the pitch information T_k of the speech waveform being processed.

To the extent that the speech waveform is periodic over the $2L + 1$ pitch periods that the filter is applied, the speech samples will add constructively while the noise samples sum to zero. The operation depends on an accurate estimate of the desired signal's period, and its performance is best when this signal has stationary traits. F_O estimation is not always easy and speech signals very often change from one period to the next, in terms of harmonics and spectral envelope.

In the case of a spectral change but constant F_O during the comb's window, the filter spreads the change out over the duration of the window, thus smearing it in time. On the other hand, if F_O changes within the window, a reduction in the reinforcement of the periods in *y(n)* will result. Thus comb filters work best only during sections of speech where F_O is not changing rapidly. Moreover, they can only be applied to the voiced segments, thus requiring a voiced/unvoiced detector, in addition to a pitch estimator. Figure 1 shows the system diagram of the comb filter used in [13].

FIGURE 1. Block diagram of Lim's comb filtering



The results in [13] indicated that even with perfect estimates of the fundamental frequency, the adaptive comb filter does not achieve a significant increase in intelligibility at any S/N ratio. A substantial decrease of intelligibility was observed when the length of the filter was between 7 and 13 pitch periods. At a length of 3 periods, intelligibility does not decrease, but noticeable increases in S/N ratio were achieved. When the filter length was 3, 7, and 13 pitch periods, the increases were 3.5 dB, 7 dB and 10 dB respectively.

4.1.1 Variations of the basic method

Some noteworthy variations are proposed to the basic comb filter in order to deal with smearing spectral changes and segment discontinuity. In [26], the coefficient set is chosen to adapt to the data, as opposed to being constant. The argument is that pitch periods may not be similar and thus choosing a constant set does not account for such transitions between periods. The coefficients a_k are chosen such that the samples that are *k* periods away best predict the current sample, thus making the results less dependent on the precision of the pitch estimate.

The method in [14] uses a class of windows that has a variable weight in each pitch period. The rationale is that when comb filtering is done in a pitch synchronous way, segment discontinuity results when the pitch period changes. This class of windows is claimed to minimize this effect.

4.2 Wiener Filtering and Related Methods

Wiener filters are an instance of the general class of optimum filters [24], where it is desired to find an estimate $\hat{S}_t$ for a signal S_t given a number of observations assumed to be the sum of the desired signal plus unwanted noise: $X_\alpha = S_\alpha + N_\alpha \qquad \alpha \in I \quad I = \{t-a, \ldots, t+b\}$. The estimate $\hat{S}_t$ is obtained by a linear filter acting on the set of observations. The filter is *optimum* with respect to minimizing the *mean square error*: $E[e_t^2] = E[(S_t - \hat{S}_t)^2]$. In the case where the signal and noise are independent processes, it can be shown that the filter that minimizes this error must satisfy the equations:

$$R_S(m) = \sum_{\beta=0}^{p} h_\beta \{ R_S(m-\beta) + R_N(m-\beta) \} \qquad m \in \{0, 1, \ldots p\}. \qquad \textbf{(Eq 2)}$$

That is the filter coefficients (h_β) could be found by solving the p+1 simultaneous equations (Eq 2). Here, $R_S(m)$ and $R_N(m)$ are the autocorrelation functions of the clean signal and the noise

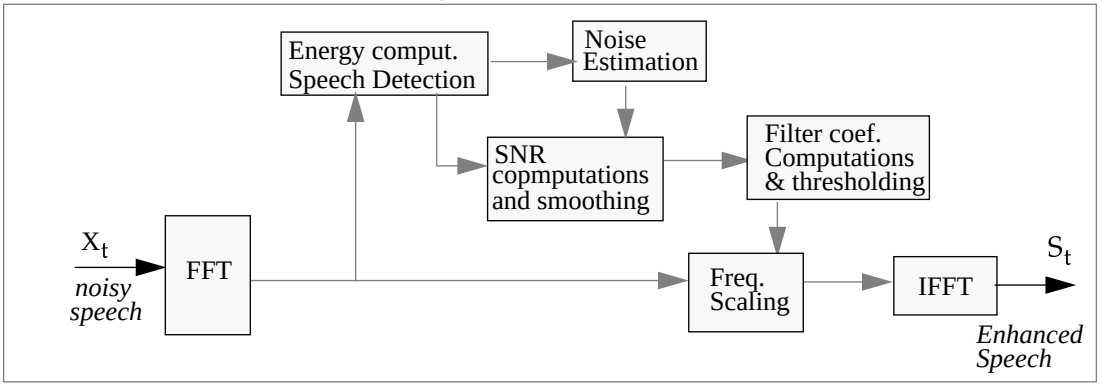
4.2.1 Frequency-domain filters

If S_t is to be estimated using the entire realization of X_t, that is $I = (-\infty,\infty)$, (this is the case when the entire signal is recorded, and then played back) and the processes are independent and zero-mean, then the *optimum* filter is given in the frequency domain by:

$$H(f) = \frac{P_S(f)}{P_S(f) + P_N(f)} = \frac{SNR(f)}{SNR(f) + 1}. \qquad \textbf{(Eq 3)}$$

Thus assuming the power spectrum of the speech and noise could somehow be estimated, a speech enhancement system may be achieved by spectral decomposition and appropriate scaling of the various frequency coefficients. A generic Wiener-based enhancement system is shown in Figure 2.

FIGURE 2. Generic Wiener-based speech enhancement



Power subtraction filters

Consider a filter whose response is the square root of a Wiener filter, thus:

$$|H(f)|_P = \sqrt{\frac{|S(f)|^2}{|S(f)|^2 + |N(f)|^2}}. \qquad \textbf{(Eq 4)}$$

The power spectrum of the output of this filter is:

$$P_{out} = P_{in}|H(f)|^2 = (|S(f)|^2 + |N(f)|^2) \bullet \frac{|S(f)|^2}{|S(f)|^2 + |N(f)|^2} = |S(f)|^2 \qquad \textbf{(Eq 5)}$$

This filtering operation is thus equivalent to *subtracting* an estimate of the noise power from the spectrum of the noisy speech. Therefore, it is a *power spectral subtraction* system.

Magnitude subtraction filters

Similarly, by considering a filter whose response is: $|H(f)|_M = 1 - \sqrt{\frac{|N(f)|^2}{|S(f)|^2 + |N(f)|^2}}$,

a *magnitude* spectral subtraction is obtained.

4.2.2 Pros and Cons

The advantages of the Wiener-based approach is its intuitive basis, its implementation simplicity and its practical effectiveness. Its limitations on the other hand are:

- Noticeable residual noise which consists of narrow-band signals with time-varying frequencies and amplitudes. This is referred to as musical noise, and is mostly perceived in weak bands and weak speech frames.
- The dependence on a good noise and SNR estimation, which is a problem in itself. Noise is often estimated during speech pauses, which in turn relies on the assumption that the background noise environment remains locally stationary.
- Ignoring the effects on phase. In experiments where only the complex phase was modified, it was found that adding random noise above a threshold causes roughness in the reconstructed speech.

4.2.3 Some Reported systems

In [2], a power subtraction system with the concept of noise overestimation is introduced. The variable oversubtraction factor is based on the estimated SNR in each band. The rationale is that strong SNR bands are indicative of a strong speech component and there is no need for aggressive subtraction. The suppression schemes in [3] and [25] are classical examples of magnitude subtraction systems where noise is estimated by averaging the signal magnitude spectrum during *non-speech* activity. In [1], a parametrized Wiener filter with noise oversubtraction is used. Some of the classical concepts in [2] are extended to the new form of subtraction. For instance, the concept of 'spectral floor' is implemented by setting a lower bound on the filter (i.e., preventing the filter gain from going below a -10 dB limit). In addition to producing a 'broadband noise' effect, this prevents coefficient randomness around small values of the gain which would result in more annoying musicality of the noise.

4.3 ML and MMSE Spectral Estimation Methods

While Wiener-based approaches did not make assumptions about the distributions of speech and noise, the following ones are estimation-oriented techniques based on assumed *apriori* distributions of the signals. The solution proposed is an estimator of the spectral components; it is therefore the distribution of this spectral component (DFT coefficient), rather than the time samples, that is of importance.

4.3.1 ML Estimation of the Speech Envelope

Speech enhancement is formulated in [15] as a maximum likelihood estimation of the speech spectral envelope. In this model, the noise is assumed to be a *Gaussian* random process and the speech a deterministic waveform of unknown amplitude and phase. Thus, each DFT channel value is given by: $y_k = s_k + w_k$, where the speech is given by: $s_k = A_k e^{j\theta_k}$, A_k representing the speech envelope and θ_k its phase. The estimator is based on maximizing the *posteriori* probability of the channel measurement $\overline{p(y_k|A_k)}$. The estimator of the speech spectrum envelope s_k, given the noisy speech spectrum value y_k is shown to be:

$$\hat{s_k} = \left[\frac{1}{2} + \frac{1}{2}\sqrt{\frac{|y_k|^2 - \lambda_w(k)}{|y_k|^2}}\right] \cdot y_k . \qquad \textbf{(Eq 6)}$$

where $|y(k)|^2$ is the measured envelope energy and $\lambda_w(k)$ the estimate of the noise energy. The estimation is further refined to account for the probability of speech presence in channel *k*. By assuming the probability of the noise and speech states equally likely, and the *a priori* signal to noise ratio known, the probability of speech presence in band *k* may be quantified based on $|y(k)|^2$, $\lambda_w(k)$ and the *apriori* signal to noise ratio: $SNR_{prior} = A_k^2 / \lambda_w(k)$. The resulting estimator is simply the product of the estimator for the clean signal and the probability of speech presence.

4.3.2 MMSE-based Estimation of the Speech Amplitude

The system proposed in [6] is closely related to and builds on the same principle as [15], except that the DFT coefficients of the noisy observations are assumed to have a Gaussian distribution. The enhancement problem is formulated as an MMSE estimator of the speech spectral amplitude. Given the same definition of speech, noise and noisy channel measurements as before, the MMSE estimator of the speech amplitude A_k is given by the conditional expectation, that is:

$\hat{A}_k = E\{A_k | Y_0, Y_1, \ldots\}$, where $\{Y_0, Y_1, \ldots\}$ is the set of spectral observations over a few frames. By making Gaussian assumptions about these spectral values, the estimator can be expressed as a spectral gain $G(p, k)$ that is applied to the spectral amplitudes $Y_k(p)$ (p denotes the frame index) and is a function of the local and the true SNR. This estimator may be shown [4] to be a smooth transition between the classical Wiener and power spectral subtraction filters.

An important aspect of this system is the estimation of the true SNR at each frequency: first, a local SNR is computed: $SNR_{post}(p, k) = [|Y(p, k)|^2 / \lambda_w(k)] - 1$ at frame p and frequency k, using the total energy $|Y(p, k)|^2$ and the estimate of the noise energy $\lambda_w(k)$. The so-called *apriori* SNR (i.e., the true SNR) is found by using information from the previous frame, after filtering:

$$SNR_{prior}(p, k) = (1 - \alpha) P[SNR_{post}(p, k)] + \alpha \frac{[G(p-1, k)]^2 |Y(p-1, k)|^2}{\lambda_w(k)}, \quad \textbf{(Eq 7)}$$

where $P[x] = x$ when $x > 0$ and 0 otherwise. As in [15], the estimator is conditioned on the probability of speech presence at frequency *k*, which may be expressed in terms of the two SNR terms defined above. It is shown in [4] that the proposed SNR smoothing yields an elimination of any noise artifacts without bringing distortion to the speech signal.

4.4 Kalman Filters

The motivation for studying a Kalman filter based noise suppression system is that it can handle colored noise and has a reasonable numerical complexity. Its use was proposed in [21] for where experimental results reveal an advantage over Wiener filtering, for the ideal cases where the speech parameters are available. The method was first proposed for white Gaussian noise, then extended [12] by incorporating a colored noise model.

A key issue in Kalman filtering is its reliance on a set of unknown parameters that have to be estimated from a noisy signal. These include the model parameters for the speech and noise. It is often assumed that the noise is long-time stationary and consequently its parameters are estimated during speech pauses. Speech is viewed as a short-time stationary process, e.g., 10-40 ms. Thus an instantaneous model of the speech has to be obtained from a short segment of noisy measurements.

The noisy speech is modeled ([21], [12]) as the sum of an AR process and a noise process: $x(n) = s(n) + v(n)$, where *x(n)* denotes the measured signal, *s(n)* the speech, and *v(n)* a zero-mean white noise process with variance σ_v^2. Furthermore: $s(n) = \sum_{i=1}^{p} a_i s(n-i) + w(n)$,

where *w(n)* is a zero-mean white Gaussian process with variance σ_w^2. The above two equations can be written in state-space form (bold letters refer to matrices):

$$\begin{aligned} \boldsymbol{s}(\boldsymbol{n}) &= \boldsymbol{F}\boldsymbol{s}(\boldsymbol{n}-\boldsymbol{1}) + \boldsymbol{g}w(n) \\ x(n) &= \boldsymbol{h}^T\boldsymbol{s}(\boldsymbol{n}) + v(n) \end{aligned} \qquad \textbf{(Eq 8)}$$

From standard Kalman filtering theory, the state vector estimate may be found and from it, the sample estimate at time instant *n* is then obtained by: $s(n) = \boldsymbol{h}^T\hat{\boldsymbol{s}}(\boldsymbol{n})$.

4.4.1 Parameter Estimation

In [12], it is assumed that σ_v^2 is known or computed using an extra microphone. The $\boldsymbol{a}$ vector (the AR parameters of speech) and σ_w^2 are calculated iteratively. That is, estimates $\hat{\boldsymbol{a}}_1$ and σ_{w1}^2, are calculated directly from the noisy observations using the Durbin algorithm and then substituted into the appropriate Kalman filter. A new set of coefficients, $\hat{\boldsymbol{a}}_2$ and σ_{w2}^2, are calculated using the filtered output and plugged into the filter. This procedure is iterated until a final set $\hat{\boldsymbol{a}}_n$ and σ_{wn}^2, is obtained. Experimental results reported in [12] found that this approach leads to good parameter estimation during high SNR segments but performs poorly during low SNR frames and causes a distortion at the filter output. In [8], Higher Order Cumulants are used to estimate the speech parameters. Experimental results found that the method is effective at low SNR conditions (5 dB) when 4^{th}-order cumulants are used. The parameter estimation was superior to using 2^{nd}-order statistics and it was also found that the use of third order statistics is limited in effectiveness.

4.5 Psychoacoustic-based Methods

Formant-based Spectral Shaping

The frequency domain postfiltering approach proposed in [28] consists of approximating the noisy speech spectrum by LPC analysis and then modifying this spectrum (through frequency filtering) so that the spectral *formants* are *sharpened* and the valleys deepened. The filter is ‘adapted’ for every new frame of speech (128 samples) and is represented by a set of DFT coefficients *H(k)*. These coefficients are multiplied by a modified version of the speech transform *P(k)* to yield the transform of the enhanced speech. The core operation of the algorithm is to scan the LPC spectrum and detect the 3 or 4 main formants (and valleys) through *peak picking*.

Exploiting the Characteristics of Human Hearing

These methods exploit frequency masking or other properties of the auditory system to gain an advantage in locating or attenuating noise bands or in preventing unnecessary attenuation in speech bands. In [23] and [27], a cross-band masking threshold is used to control the spectral subtraction process. The idea is to find the best compromise between noise reduction and speech distortion in a perceptual sense. By identifying the bands where natural noise masking occurs, there is no need for an aggressive noise suppression.

The algorithm proposed in [5] uses the property of lateral inhibition of the auditory system. It convolves the so-called function of spatial lateral inhibition with the power spectrum of the noisy speech input to yield an estimate of the clean speech spectrum without apriori knowledge of the noise spectrum.

4.6 Higher-Order Statistics

4.6.1 Rationale and Challenges

Higher-order statistics (HOS) have shown promising results in such fields as radar, image processing, seismology, and array processing [18] and are of particular value when dealing with a mixture of Gaussian and non-Gaussian random processes and system nonlinearity. Their key attractive features is their *Gaussian blindness:* The higher order cumulants of a Gaussian process are identically zero. Thus, given a speech signal *s(n)* corrupted by additive Gaussian noise *g(n)*: $x(n) = s(n) + g(n)$, the higher order cumulants of *s(n)* are simply those of *x(n)*:

$$cumulant(s) = cumulant(x) \quad \textbf{(Eq 9)}$$

Analysis in the higher order domain is therefore a way of filtering out all Gaussian noise (both white and colored). In spite of their attractive noise-suppression properties, the application of HOS to speech processing has, in general, lead to mixed results. To make effective use of HOS, some crucial issues need to be addressed:

1. ***The burden of proof and interpretation***: for the statement in Eq 9 to have a meaningful implication, the cumulants of the speech itself have to be non-zero. While this is in general true, one needs to consider specific speech domains to guarantee this non-zero condition. In addition, one needs a framework where the speech HOS are expressed in terms of useful speech attributes, such as energy, frequency, pitch, etc.

2. ***Implementation issues and limitations***: The Gaussian-blindness feature of HOS is true only in a statistical sense. When analysis is done using only one realization and finite data records, issues related to the bias and variance of the HOS estimators need to be accounted for and factored into the algorithm. As an example, the detection of speech and noise can only be made in a probabilistic manner with a confidence interval.

4.6.2 Some Reported Systems

In [7], a speech enhancement system based on 3rd-order statistics is proposed and consists of time averaging the bispectrum then resynthesizing speech using a bispectrum-to-Fourier reconstruction. The bispectrum is computed from the Fourier transform:

$$B(f_1, f_2) = X(f_1) \cdot X(f_2) \cdot X^*(f_1 + f_2) .$$

and averaged over three consecutive segments to approximate statistical averaging. The authors reported an SNR improvement of about 1 to 2 dB when the input SNR was below 6 dB, for both white and color noise. The enhancer was more effective for voiced speech, where the enhancement attained 2.7 dB, than in unvoiced speech, that was typically enhanced by 1 to 1.5 dB.

In [17], an algorithm for noise reduction based on optimal filters, subbands and higher-order cumulants is proposed. A filter bank is used to subdivide the signal into narrow bands, and enhancement is carried out in each of these using causal optimal filters and the *p* most recent samples. The key idea is to use the 4th-order statistics of the noisy speech to estimate the required parameters for the enhancement filters, such as the SNR, the autocorrelation of speech and the

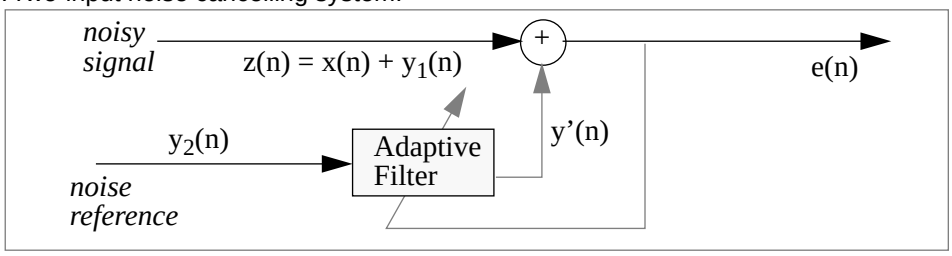
probability of speech presence. The concepts are based on established properties of the higher-order statistics of speech [16]. For instance, it is shown that:

- The normalized autocorrelation of speech may be estimated from the diagonal slice of the 4th-order cumulant normalized by the kurtosis: $Ra[\tau] = C_{4s}[\tau] / (C_{4s}[0])$.
- The probability that a band contains only noise is determined by the estimate of the kurtosis, scaled by the variance of the estimator: $Prob[NoiseOnly] = erfc(|b|)$.
- In bands where speech and noise are present, an upper bound on the speech energy is derived from the kurtosis, from which a lower bound on the noise may be deduced.

4.7 Methods that Exploit Spatial Noise Correlation

These techniques make use of one or more noise reference inputs (microphones) and attempt to subtract the noise component from the noisy speech signal. The primary microphone picks up the noisy speech signal while a set of secondary microphones measure a signal consisting mainly of noise. The signals from the secondary mics are fed to an adaptive filter which estimates the noise on the primary mic and simply subtract it from the transmitted speech (Figure 3).

FIGURE 3. A Two-input noise cancelling system.



A number of filter structures and adaptation algorithms have been evaluated in the literature. The technique however requires that the noise component in the corrupted signal and the reference noise have a high coherence. Inside a passenger car, it was found [11] that if two microphones are located at a distance greater than 50 cm, the only coherent noise component is that of the engine. In order to cancel 90% of the noise energy, the two mics cannot be more than 5 cm apart, which makes it almost impossible to prevent speech from entering the noise reference. Furthermore, when using a single noise reference it was shown [9] that the noise cancellation has limited effectiveness due to the nonstationarity of the noise process in the automobile. A single noise reference will reduce the noise energy in the dominant band below 1000Hz where the noise in both mics are strongly correlated but will enhance the noise in the band above 1000Hz where the correlation of the noise as well as its energy is much lower.

In case where no speech contamination is present in the secondary reference, the filter continuously adapts the coefficient set in a way to minimize the power of the error *e(n)*. It can be shown [29] that minimizing this error (even during speech) is equivalent to minimizing the power of the noise, defined as $E[\{y_1(n) - y'(n)\}^2]$. The adaptation algorithm may be based on any gradient method, for example: $W_{j+1} = W_j - (\mu \cdot \Delta_j)$, where Δ_j is the gradient of the mean square of the output with respect to the coefficient vector W_j.

The main limitations of the multi-sensor may be practical realization. Most hand held portable phones are designed to be very small in dimension and placing multiple microphones with enough distance to have enough noise -but no speech- correlation between the sensors may prove difficult.

5.0 System Aspects of Noise Reduction

The NR operation, in general, alters the spectral characteristics of the signal and may result in modifying the speech harmonic or its formant structure. Whenever NR is placed prior to encoding or speech recognition, this may have an impact -positive or negative- on the operation of these applications, which rely on spectral analysis. On the other hand, whenever NR is placed at the landline side after the speech decoder, it may be affected, positively or negatively, by the way the mobile encodes the various speech and non-speech segments, by any additional operation that the user terminal may do, and by channel interference, handoffs and other network operations of the cellular network.

5.1 Interaction With Speech Coders

- *Discontinuous Transmission (DTX)*: In some TDMA and GSM systems, a discontinuous transmission feature is implemented in order to save battery life and minimize interference. When no speech activity is detected at the mobile, transmission ceases and the landline decoder generates a so-called comfort noise to keep a seamless interruption in the output signal. While the comfort noise is based on the actual noise present, its spectral characteristics may not be the same. As a result, an NR placed after the decoder at the landline end will be using a wrong reference to update its noise estimates.
- *Variable Rate Coders*: in CDMA systems, variable rate speech coders are used in order to reduce interference by lowering the average transmission rate. During non-speech and unvoiced segments, the coding rate is dropped to 1/4 or 1/8. As a result, a similar problem as in the DTX case occurs when NR is placed at the landline side after the decoder, since the coding of speech and non-speech segments results in different spectral noise characteristics. Moreover, if NR were placed at the mobile prior to encoding, then the reduction of the noise will affect the decision of the rate determination algorithm (RDA). In the good cases, it will result in a lower average rate for transmission, which is always desirable.

5.2 Interaction With the Cellular Network Operations

- *Handoff recovery*: handoffs in cellular networks cause a short interruption in the transmitted signal. If NR is placed at the landline side, then its input signal will temporarily drop to zero, causing an effective resetting of the internal memory of the algorithm. After handoff, the algorithm restarts from a zero initial condition and thus it is desirable that its convergence rate be fast enough in order not to generate a very audible effect.
- *Channel interference and errors*: on analog cellular systems, channel interference may be viewed as an added source of acoustic noise. In digital cellular systems, channel errors cause the speech decoder to repeat or interpolate speech frames, adding sometimes a reverberant effect and sometimes adding a somewhat static noise effect. If NR is placed at the

mobile end, then NR will have no effect on this added noise source. Placing NR at the land-line side however, may have the added advantage of removing this added noise.

5.3 Placement of Noise Reduction in the Network

5.3.1 At the Mobile/wireless Terminal:

This results in removing the noise at the source, prior to encoding or recognition. In this scenario, DSP complexity, battery requirements, and terminal cost are important issues that may impose constraints on the type of algorithm to use. In a hands-free application, NR may be implemented in the external hardware, thus easing the complexity constraints, but also increasing the cost of the accessories. Other points to note here:

- There is a better chance at estimating the noise, thus improving the effectiveness of Wiener-like algorithms. The noise conditions are the same during speech and silence segments.
- There is no adverse effect from DTX or variable rate coding.
- This scheme cannot eliminate any channel or quantization noise from the network.
- NR may negatively impact the coder or speech recognition algorithms.
- There is the possibility of using multi-sensor techniques.
- In variable rate codecs, it will result in a lower overall rate, which is desirable for capacity.

When removing the noise at the source, it is necessary to do so on the reverse link only. In a mobile-to-mobile call, each terminal removes the noise at its end and the far end receives a clean signal.

5.3.2 At the Wireless / Wireline Infrastructure

In this scenario, noise reduction is placed in the switching center, thus shared by many individual users, both wireline and wireless. Given the more available computational power, more sophisticated methods may be employed in this scheme. Worth noting points include:

- The scheme allows operators to provide noise reduction for those users that are currently using cellular phones that do have any built-in noise reduction such as AMPS phones, TDMA and GSM based phones.
- This methods poses more challenges since the speech is now encoded at the noisy end (terminal) along with the noise and, given the non-linearity of speech coders, irreversible speech distortion may have resulted.
- NR is directly affected by the RDA and DTX and some global optimization may be needed.
- It allows elimination of the noise introduced by channel errors, or any other network impairments.

5.3.3 The Business Argument

There are three business reasons [10] for placing the noise reduction functionality at the network side, instead of the terminal:

- *Handset Cost:* the marketplace is driving the price points for cellular phones down to virtually zero, thus putting severe constraints on the amount of sophisticated technology that can be included in handsets.

- *Who should pay:* the main beneficiary of noise reduction is not the handset owner but the far-end party (as is the case for echo cancellation); therefore the owner has little immediate justification to spend more for a handset with more expensive noise cancellation included.
- *Quality Consistency:* the wireless operator can achieve a better overall service by providing noise reduction to all subscribers in the network (wireless, wireline, near and far-end) and thus a network-wide, network-consistent solution seems to be the proper way.

6.0 TEN CRITERIA IN CHOOSING A NOISE REDUCTION TECHNOLOGY

1. *Tradoff of complexity and performance*: can the algorithm be configured to operate with less computational power, and thus less precision and performance.
2. *Psychoacoustics*: regardless of what theoretical concept the algorithm is built upon, it should factor in the effect of frequency and, ideally, time domain masking.
3. *Performance*: how does the algorithm fair in ideal conditions, such as white stationary noise, and worse-case conditions, such as bable noise. What are the noise and the speech distortions, and are the latter function of noise characteristics (noise modulation).
4. *Fixed-point performance*: is the difference between the fixed and floating point performance significant. Does the algorithm conceptually require high dynamic range to ensure good performance or stability.
5. *Interaction with codecs*: if placed prior to encoding, how does the algorithm affect the quality of the coded speech or the operation of the rate determination algorithm in a variable rate coder. If placed after the decoder at a centralized place in the network, how is it affected by any DTX operation or the variable rate coder or any other operation performed at the mobile set or in the network.
6. *Convergence speed*: how quickly does the algorithm converge when the noise characteristics abruptly change or when any operation in the network (e.g., handoff) cause temporary interruption of the signal.
7. *Quality vs. intelligibility*: does the algorithm make a good compromise between the two. What is the impact on intelligibility in best and worst case scenarios. What phonemes are affected the most.
8. *Transparency*: does the algorithm disengange itself when the situation is such that it is not required or is not effective (in very clean speech, or in extremely poor signal conditions).
9. *Multiple concepts*: is the algorithm based on classical or new theories, or a mixture of both.
10. *Uplink vs. downlink*: is the performance highly contingent on the placement in the network, or does the algorithm have an adaptive way to adjust its operations.

7.0 SOME COMMERCIAL ALGORITHMS BASED ON SPECTRAL SUBTRACTION

7.1 Proprietary Aspects

While spectral subtraction as a *concept* is a public domain technique, there are a number of variations and implementation aspects of this methods that have been patented. For example, the methods for estimating the noise or computing the filter coefficients, or smoothing the SNR or spliting the signal into bands, etc... The following are some examples of proprietary techniques.

7.2 Patents Based on Spectral Subtraction/Scaling

How subtraction / scaling is done

- In Motorola's [31], spectral scaling with subbanding is used. The scaling is applied in the frequency domain, thus involving an FFT/IFFT of 128 points. FFT bins are mapped to 16 non-homogenous bands roughly following a Bark scale.
- In Nokia's [32] [33], an FFT is used for spectral analysis. In addition, the formant locations are estimated so that speech within the formants is attenuated less that the part in between.
- In Ericsson's [30], attenuation is applied in the time domain on the entire frame (no sub-banding) and is function of the estimated noise level.

Filter gains computation

- In Motorola's, the amount of band attenuation is a non-linear function of the estimated SNR in this band and the total noise energy in the frame. The scaling function does not resemble any of the classical/published optimal filters and is a proprietary feature of the algorithm.
- In Ericsson's, the attenuation function is a logarithmic function of the noise level (not the SNR) relative to a preset threshold, below which no attenuation is deemed required. The attenuation function however is different whether speech is detected in that frame or whether it is purely a noise frame.

Noise estimation

- In Motorola's, noise estimation is done during silence/stationary segments. Silence is detected by summing the SNRs across the 16 bands into a so-called voice metric, and comparing it to a threshold. Prior to summing, the SNRs are first quantized according to a roughly exponential mapping. The other instance for updating the noise is during stationary conditions, detected whenever the short term energy of each band within the frame is close to the long-term energy average.
- In Nokia's, noise is estimated during speech. Each of the QMF filter passbands is further split in two sub-bands using a special filter. The filters passbands are such that only one of them will capture speech harmonic, while the other contains only the noise (or whatever is in between the two consecutive harmonic peaks). The method thus claims to be superior since it can better keep track of noise changes during speech periods.

Artifacts removal

- In Motorola's this is done by detecting very weak frames, and scaling them by the minimum gain (0.17). In addition, sudden noise bursts are detected by counting the number of bands

where the SNR exceeds a given threshold; The argument is based on the fact that in voiced speech, a large number of bands have a high SNR whereas sudden noise burst frames are characterized by a large SNR in only a few bands.

- In Nokia's, the random flutter effect is avoided by not updating the filter coefficients during speech periods. This is effective but likely to result in a poor convergence of the filter gains during changing noise and speech conditions.

8.0 REFERENCES

[1] L. Arslan, "New methods for adaptive noise suppression", *Proc. ICASSP* 1995, pp. 812 - 815.

[2] M. Beirouti, "Enhancement of speech corrupted by acoustic noise", *Proc. ICASSP* 1979, pp. 208 - 211.

[3] S. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Trans on Acoustics, Speech, and Signal Processing*, Vol. ASSP-27, No 2, April 1979, pp. 113-120.

[4] O. Cappe, "Elimination of the Musical Noise Phenomena with the Ephraim and Malah Noise Suppressor", IEEE trans. on Speech and Audio Processing, Vol. 2, No. 2, April 1994, pp. 345 - 349.

[5] Y. Cheng, D.O'Shaughnessy, "Speech Enhancement Based Conceptually on Auditory Evidence", *IEEE trans. on Signal Processing*, Vol. 39, No. 9, Sept. 1991, pp. 1943-1954.

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

[7] R. Fulchiero and A. Spanias, "Speech Enhancement Using the Bispectrum", *Proc ICASSP* 1993, vol 4, pp. 488 - 491.

[8] S. Gannot, D. Burshtein, E. Weinstein, "Iterative-batch and Sequential Algorithms for Single Microphone Speech Enhancement", Proc. ICASSP 1997, pp. 1215 - 1218.

[9] R. A. Goubran & H.M. Hafez. "Background Acoustic Nlise Reduction in Mobile Telephony". *Proceedings of the 36th IEEE Vehicular Technology Conference*, Dallas, Texas, 1986, p72.

[10] C. Gritton. "Call Quality: A question of noise" in *Telecommunications*, Nov. 96, pp 79.

[11] B.H. Juang. "Speech recognition in adverse environments". *Computer Speech and Language*, 1991, No 5, pp 275-294.

[12] B. Koo, J. Gibson, and S. Gray, "Filtering of Colored Noise for Speech Enhancement and Coding", *Proc. ICASSP* 1989, pp. 349 - 352.

[13] J. Lim, A. Oppenheim, and L. Braida, "Evaluation of an Adaptive Comb Filtering Method for Enhancing Speech Degraded by White Noise Addition", *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. ASSP-26, No. 4, Aug. 1978, pp. 354 - 358.

[14] D. Malah and R. Cox, "A Generalized Comb Filtering Technique for Speech Enhancement", *Proc ICASSP 1982*, pp. 160 - 163.

[15] R. McAulay and M. Malpass, "Speech Enhancement Using a Soft-decision Noise Suppression Filter", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-28, No.2, April 1980, pp. 137 - 145.

[16] E. Nemer. "Speech Analysis and Quality Enhancement Using Higher Order Cumulants", *Ph.D. Thesis*, Carleton University, Ottawa, Canada, 1999.

[17] E. Nemer, R. Goubran, S. Mahmoud, "Speech Enhancement Using HOC and Subband Causal Filters", ICSPAT 1999.

[18] C. Nikias and J. Mendel, “Signal Processing with Higher-Order Statistics”, *IEEE Signal Processing*, July 1993, pp. 10 - 38.
[19] D. O’Shaughnessy, *Speech Communication*, Addison-Wesley. 1987.
[20] D. O’Shaughnessy, “Enhancing Speech Degraded by Additive Noise or Interfering Speakers”, *IEEE Communications Magazine*, Feb 1989, pp. 46-52.
[21] K. Paliwal & A. Basu, “A Speech Enhancement Method Based on Kalman Filtering”, *Proc. ICASSP* 1987, pp. 177 - 180.
[22] S. Seetharaman and M. Jernigan, “Speech Signal Reconstruction Based on Higher Order Spectra”, *Proc ICASSP* 1988, pp. 703-706.
[23] D. Tsoukalas, M. Paraskevas, J. Mourjopoulos, “Speech Enhancement Using Psychoacoustic Criteria”, Proc. ICASSP 1993, pp. 359-362.
[24] H. Van Trees, *Detection, Estimation and Modulation Theory, Part I*. New York: Wiley 1968, pp. 54 - 56, 198 - 206.
[25] P. Vary, “Noise Suppression by Spectral Magnitude Estimation - Mechanism and Theoretical Limits”, *Signal Processing*, Vol.8 No. 4, July 1985, pp 387 - 400.
[26] D. Veeneman and B. Mazor, “A Fully Adaptive Comb Filter for Enhancing Block-coded Speech”, *IEEE trans ASSP*, Jun 1989, pp. 955 - 957.
[27] N. Virag, “Speech enhancement based on masking properties of the auditory system”, *Proc. ICASSP* 1995 pp. 796 - 799.
[28] F. Wang, P. Kabal, and D. O’Shaughnessy, “Frequency Domain Adaptive Postfiltering for Enhancement of Noisy Speech”, *Speech Communication*, Vol. 12, No 1, March 1993, pp. 41-56.
[29] B. Widrow, J.R. Glover, J.M. McCool, J. Kaunitz. “Adaptive Noise Cancelling: Principles and applications”. *Proceeding of the IEEE*, Vol 63, No 12, Dec 1975.

Patents

[30] T. Solve & R. Zack, “System for adaptively reduction noise in speech signals”, *Ericsson* GE Mobile Comm. Inc. Jan 16, 1996. US Patent 5,485,522.
[31] R. Vilmur et al. “Noise Suppression System”, *Motorola*. US Patent 4,811,404. March 7, 1989.
[32] T. Kolehmainen, “A method of and system for noise suppression” European Patent 0 588 526 A1, *Nokia* Mobile Phones Ltd., Salo, Finland, March 23, 1994.
[33] J. Kuusama & A. Makivirta. “System for processing an audio signal so as to reduce the noise contained therein by monitoring the audio signal content within a plurality of frequency bands. US Patent 5,485,524, *Nokia* Technology, GmbH, Germany, Jan 16, 1996.