Speech Enhancement Using a Soft-Decision Noise Suppression Filter

ROBERT J. McAULAY, MEMBER, IEEE, AND MARILYN L. MALPASS

Abstract—One way of enhancing speech in an additive acoustic noise environment is to perform a spectral decomposition of a frame of noisy speech and to attenuate a particular spectral line depending on how much the measured speech plus noise power exceeds an estimate of the background noise. Using a two-state model for the speech event (speech absent or speech present) and using the maximum likelihood estimator of the magnitude of the speech spectrum results in a new class of suppression curves which permits a tradeoff of noise suppression against speech distortion. The algorithm has been implemented in real time in the time domain, exploiting the structure of the channel vocoder. Extensive testing has shown that the noise can be made imperceptible by proper choice of the suppression factor.

I. Introduction

THE need for secure military voice communication has led to the consideration of narrow-band digital voice terminals. A preferred algorithm for this task is linear-predictive coding (LPC) which has demonstrated the ability to produce very intelligible speech with diagnostic rhyme test (DRT) scores in excess of 90 percent at data rates as low as 2400 bits/s [1]. Unfortunately, these results have been achieved only for clean speech, whereas many of the practical environments in which these terminals would be deployed, such as the airborne command post or the cockpits of jet fighter aircraft and helicopters, are characterized by a high ambient noise level, which in many cases causes the vocoded speech to suffer a significant degradation in intelligibility [2]. This has stimulated research into the problem of extracting the speech parameters (pitch, buzz–hiss, and spectrum) from noisy speech in the hope that more robust algorithms could be found [3]–[5].

Another approach to the noisy speech problem is to develop a prefilter that would enhance the speech prior to encoding so that the existing LPC vocoder could be applied in tandem without modification. Two general classes of algorithms have emerged: noise canceling and noise suppression prefilters. In the first case, the coefficients of a tapped delay line are adapted to produce a minimum mean-squared error estimate of the noise signal which is then subtracted from the noisy speech waveform to effect the noise cancellation [6]. In order to train the coefficients of the noise-canceling filter, it is usually necessary to use a second microphone to provide a speech-free measurement of the background noise. Application of this technique to the cancellation of E4A advanced airborne command post noise has shown that although significant improvement in signal-to-noise ratio (SNR) can be obtained, the improvement in intelligibility, as measured by the diagnostic rhyme test (DRT), is marginal [7]. Recent work by Sambur [8] has attempted to exploit the periodicity of voiced speech to eliminate the requirement for a second microphone. Thorough evaluation of this algorithm has not yet been published.

Considerably more work has been expended on the development of noise suppression prefilters. In this approach, a spectral decomposition of a frame of noisy speech is performed, and a particular spectral line is attenuated depending on how much the measured speech plus noise power exceeds an estimate of the background noise power [9]–[13]. Algorithms using the FFT have been tested against wide-band noise and improvements in intelligibility have been indicated, although no quantitative results have been given [11]. To date, the attenuation curves have been proposed on more or less an ad hoc basis; hence, it is of interest to determine whether or not a more fundamental theoretical analysis could lead to a new suppression curve with substantially different properties. In the next section, an analytical model is proposed and used to determine the conditions under which the existing suppression curves can be justified. Having established a common basis, a new suppression curve is derived, recognizing the fact that the degree of suppression should be weighted by the probability that a given measurement corresponds to speech plus noise or to noise alone. It is shown that a class of curves is obtained by varying the value of a suppression factor. This is a parameter that can be chosen to trade off noise suppression against speech distortion. The algorithm has been implemented in real time in the time domain, exploiting the structure of the channel vocoder to perform the spectral decomposition. Extensive testing has shown that the noise can be made imperceptible by proper choice of the suppression factor.

II. Analysis

The prefilter design problem arises because a speech signal $s(t)$ has been corrupted by acoustically coupled background noise $w(t)$ to form the measurement $y(t) = s(t) + w(t)$. In speech, it is not easy to specify a criterion which would lead to a "best" estimate of $s(t)$; hence, a variety of algorithms are often proposed and evaluated by listening to the processed results. In order to provide a common theoretical basis for relating some of these algorithms, it has been found useful to

Manuscript received July 13, 1979; revised November 26, 1979. This work was supported by the Department of the Air Force. The views and conclusions contained in this document are those of the contractor and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the United States Government.

The authors are with the M.I.T. Lincoln Laboratory, Lexington, MA 02173.

analyze the prefilter for a frame of data of length $T(T \sim 20$ ms). A further simplification occurs by expanding $y(t)$ in terms of a set of basis functions $\{\phi_n(t)\}$ in such a way that the expansion coefficients are uncorrelated random variables. If the covariance function of $y(t)$ is $R_y(t,u)$, then a suitable set of basis functions are obtained from the Karhunen-Loève expansion

$$\lambda(n)\,\phi_n(t) = \int_0^T R_y(t,u)\,\phi_n(u)\,du \qquad 0 \leqslant t \leqslant T. \tag{1}$$

Then on $(0, T)$

$$y(t) = \sum_{n=1}^{\infty} y_n \phi_n(t) \tag{2a}$$

$$y_n = \int_0^T y(t)\,\phi_n(t)\,dt = s_n + w_n. \tag{2b}$$

Van Trees [14] shows that if the correlation time of $y(t)$ is less than the frame interval T, then an appropriate set of eigenfunctions and eigenvalues are

$$\phi_n(t) = \frac{1}{\sqrt{T}} \exp\left(j\frac{2\pi nt}{T}\right) \tag{3a}$$

$$\lambda(n) = S_y\left(\frac{n}{T}\right) \tag{3b}$$

where

$$S_y(f) = \int_0^T R_y(\tau)\,e^{-j2\pi ft}\,dt \tag{4}$$

is the power spectrum of the observed process. Since a narrowband vocoder usually operates over a bandwidth less than 4 kHz, only a finite number of expansion coefficients are needed to characterize $y(t)$. The prefilter design problem then reduces to the problem of optimally extracting the speech random variable s_n from the noisy observation $y_n = s_n + w_n$. If the speech and the noise are modeled as independent Gaussian random processes, then the expansion coefficients are independent Gaussian random variables with variances

$$\sigma_y^2(n) = \lambda_s(n) + \lambda_w(n) \tag{5}$$

where

$$\lambda_s(n) = S_s\left(\frac{n}{T}\right) \tag{6a}$$

$$\lambda_w(n) = S_w\left(\frac{n}{T}\right) \tag{6b}$$

represent the power in the nth harmonic line of the speech and noise spectra.

A. Power Subtraction

Since it is well known that the perception of speech is phase insensitive, a reasonable criterion for a prefilter design is to

$$\hat{s}(t) = \sum_{n=1}^{N} \hat{s}_n \phi_n(t) \qquad 0 \leqslant t \leqslant T \tag{7}$$

where $\hat{s} = \sqrt{\lambda_s(n)}$ since if $\lambda_s(n)$ were known, the spectrum of $\hat{s}(t)$ would be identical to the spectrum of $s(t)$. Of course, it is not known and provision must be made for estimating its value from an observation of y_n and knowledge of $\lambda_w(n)$. Since y_n is a complex Gaussian variate with variance $\sigma_y^2(n)$, its real and imaginary parts are Gaussian with variance $\sigma_y^2(n)/2$. Hence, the probability density function for y_n is

$$p(y_n) = \frac{1}{\pi\,[\lambda_s(n) + \lambda_w(n)]} \exp\left\{-\frac{|y_n|^2}{[\lambda_s(n) + \lambda_w(n)]}\right\}; \tag{8}$$

then by maximizing $p(y_n)$ with respect to $\lambda_s(n)$, the maximum likelihood estimate of $\lambda_s(n)$ can be found to be

$$\hat{\lambda}_s(n) = |y_n|^2 - \lambda_w(n). \tag{9}$$

In order to maintain an identity system in the absence of noise, the input phase can be appended to the prefilter output by taking

$$\hat{s}_n = \sqrt{\hat{\lambda}_s(n)}\,\frac{y_n}{|y_n|} = \left[\frac{|y_n|^2 - \lambda_w(n)}{|y_n|^2}\right]^{1/2} \cdot y_n \tag{10}$$

which is known as the method of power subtraction. Modifications of this algorithm have been studied extensively by Boll [10], Preuss [12], and Berouti *et al.* [13].

B. Wiener Filtering

Whereas the power subtraction algorithm arises from an attempt to obtain the best estimate of the speech spectrum, the Wiener filter corresponds to the criterion of minimizing the mean-squared error of best time domain fit to the speech waveform. Van Trees [14, pp. 198-206] has shown that this can be done by choosing the channel coefficients to be

$$\hat{s}_n = \frac{\lambda_s(n)}{\lambda_s(n) + \lambda_w(n)} \cdot y_n. \tag{11}$$

Since the speech eigenvalues are unknown *a priori*, the maximum likelihood estimate developed in (8) can be used in (11) to result in the suppression rule

$$\hat{s}_n = \left[\frac{|y_n|^2 - \lambda_w(n)}{|y_n|^2}\right] \cdot y_n \tag{12}$$

which is simply the square of the suppression rule for the method of power subtraction.

C. Maximum Likelihood Envelope Estimation

The previous results were obtained assuming that the speech and the noise were independent Gaussian random processes. In the interest of exploring the importance of this assumption, an alternative model is proposed in which the noise is a Gaussian random process, while the speech is characterized by a

this case, the channel measurement is $y_n = s_n + w_n$ where now $s_n = A \exp(j\theta)$ where A determines the speech envelope and θ its phase. For the perception of speech, an optimum estimate of its envelope is desired since this would represent an estimate of the speech spectrum in the nth channel. For Gaussian noise, the probability density function of the channel measurement y_n is

$$p(y_n|A,\theta) = \frac{1}{\pi\lambda_w(n)} \cdot \exp\left[-\frac{|y_n|^2 - 2A\,\mathrm{Re}\,(e^{-j\theta}y_n) + A^2}{\lambda_w(n)}\right]. \quad (13)$$

To obtain the maximum likelihood estimate of A, a maximum of $p(y_n|A,\theta)$ is sought. However, the speech phase θ shows up as a nuisance parameter. Its effect can be eliminated by maximizing the average likelihood function

$$\overline{p(y_n|A)} = \int_0^{2\pi} p(y_n|A,\theta)\,p(\theta)\,d\theta \quad (14)$$

where $p(\theta)$ is the probability density function for the phase. Since it is reasonable to assume a uniform distribution on $(0, 2\pi)$, then the likelihood function for the spectral envelope becomes

$$\overline{p(y_n|A)} = \frac{1}{\pi\lambda_w(n)} \cdot \exp\left[-\frac{|y_n|^2 + A^2}{\lambda_w(n)}\right] \cdot \frac{1}{2\pi}\int_0^{2\pi} \exp\left[\frac{2A\,\mathrm{Re}\,(e^{-j\theta}y_n)}{\lambda_w(n)}\right] d\theta. \quad (15)$$

The integral appearing in (15) is known as the modified Bessel function of the first kind and is labeled

$$I_0(|x|) = \frac{1}{2\pi}\int_0^{2\pi} \exp\,[\mathrm{Re}\,(e^{-j\theta}x)]\,d\theta \quad (16)$$

where $x = 2Ay_n/\lambda_w(n)$ depends on the *a priori* signal-to-noise ratio $A^2/\lambda_w(n)$ and the *a posteriori* signal-to-noise ratio $|y_n|^2/\lambda_w(n)$. For large values of $|x|$ (≥ 3), which represents a constraint on the signal-to-noise ratios,

$$I_0(|x|) \sim \frac{1}{\sqrt{2\pi|x|}} \exp(|x|). \quad (17)$$

For this condition, the likelihood function for the spectral envelope becomes

$$\overline{p(y_n|A)} = \frac{1}{\pi\lambda_w(n)} \cdot \frac{1}{\sqrt{2\pi\dfrac{2A|y_n|}{\lambda_w(n)}}} \cdot \exp\left[-\frac{|y_n|^2 - 2A|y_n| + A^2}{\lambda_w(n)}\right]. \quad (18)$$

Maximizing this function with respect to A leads to the estimator

$$\hat{A} = \frac{1}{\quad}$$

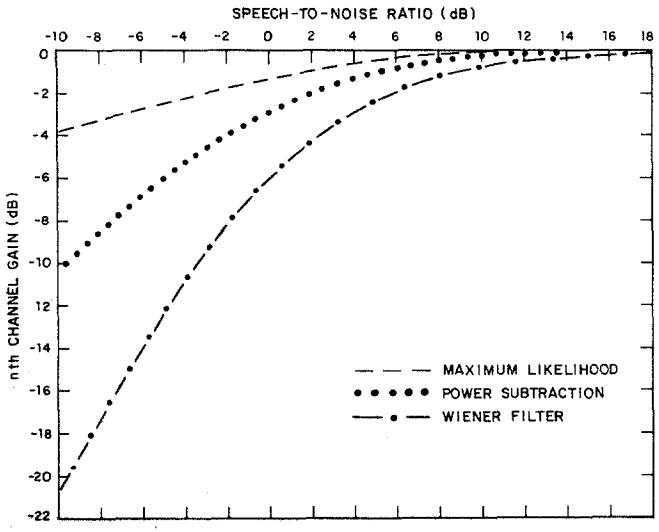

Fig. 1. Power subtraction, Wiener filter, and maximum likelihood suppression rules.

As before, the input phase can be appended to this estimate of the envelope to produce the maximum likelihood estimate of the speech waveform:

$$\hat{s}_n = \hat{A}\frac{y_n}{|y_n|} = \left[\frac{1}{2} + \frac{1}{2}\sqrt{\frac{|y_n|^2 - \lambda_w(n)}{|y_n|^2}}\right] \cdot y_n. \quad (20)$$

D. Two-State Soft Decision Maximum Likelihood Envelope Estimation

The suppression rules for the power subtraction, Wiener filtering, and maximum likelihood algorithms are illustrated in Fig. 1. Their suppression capabilities were evaluated for speech in airborne command post noise using a real-time implementation of the prefilter (to be described in detail in Section III). While it was difficult to determine which algorithm did the best job of extracting the speech when speech was present, it was apparent that none of the algorithms adequately suppressed the background noise when speech was absent. This is hardly surprising in view of the fact that the suppression rules were derived on the assumption that speech was always present in the measured data. Had a detector been used to determine that a given frame of data consisted of noise alone, then obviously a better suppression rule would have been to apply greater attenuation than indicated by the curves in Fig. 1. From this point of view, it follows that a better suppression curve might evolve if a two-state model for the speech event is considered at the outset, that is, either speech is present or it is not. Mathematically, this leads to the binary hypothesis model

$$H_0: \text{ speech absent: } |y_n| = |w_n|$$
$$H_1: \text{ speech present: } |y_n| = |Ae^{j\theta} + w_n|. \quad (21)$$

Only the measured envelope is used in this measurement model since it has already been shown that the measured phase pro-

A useful criterion for estimating the spectral envelope A is to choose $\hat{A}$ to minimize the mean-squared spectral error $E(\hat{A} - A)^2$. It is well known [14] that the resulting estimator is the conditional mean $\hat{A} = E(A|V)$ where $V = |y_n|$ is used for notational convenience to represent the measured envelope. Reference to the nth channel will be implied. In this formulation, the expectation operator is used to indicate averaging over the ensemble of noise sample functions, speech envelopes and phases, and the ensemble of speech events. The averaging for the latter case is carried out explicitly and results in the estimator

$$\hat{A} = E(A|V,H_1)P(H_1|V) + E(A|V,H_0)P(H_0|V) \tag{22}$$

where $P(H_k|V)$ is the probability that the speech is in state H_k given that the measured envelope has the value V. Since $E(A|V,H_0)$ represents the average value of A given an observation V *and* the fact that speech is absent, then obviously this value must be zero; hence, (22) reduces to

$$\hat{A} = E(A|V,H_1)P(H_1|V). \tag{23}$$

Since $E(A|V,H_1)$ represents the minimum variance estimate of A when speech is present, and since the maximum likelihood estimator is asymptotically efficient for large SNR, it suffices to replace $E(A|V,H_1)$ by the estimator derived in (19); hence,

$$\hat{A} \sim \frac{1}{2}[V + \sqrt{V^2 - \lambda_w}\,]P(H_1|V). \tag{24}$$

Application of Bayes rule gives

$$P(H_1|V) = \frac{p(V|H_1)P(H_1)}{p(V|H_1)P(H_1) + p(V|H_0)P(H_0)} \tag{25}$$

where $p(V|H_k)$ is the *a priori* probability density function for the measured envelope given the speech state H_k. Assuming that the speech and noise states are equally likely (a worst case assumption),

$$P(H_1) = P(H_0) = \frac{1}{2}. \tag{26}$$

Under hypothesis H_0, $V = |w|$, and since the noise is complex Gaussian with mean zero and variance λ_w, it follows that the envelope has the Rayleigh pdf

$$p(V|H_0) = \frac{2V}{\lambda_w}\exp\left(-\frac{V^2}{\lambda_w}\right). \tag{27}$$

Under hypothesis H_1, $V = |Ae^{j\theta} + w|$ and the envelope has the Rician pdf

$$p(V|H_1) = \frac{2V}{\lambda_w}\exp\left(-\frac{V^2 + A^2}{\lambda_w}\right) I_0\left(\frac{2AV}{\lambda_w}\right). \tag{28}$$

Defining the *a priori* signal-to-noise ratio ξ to be

$$\xi = \frac{A^2}{\lambda_w} \tag{29}$$

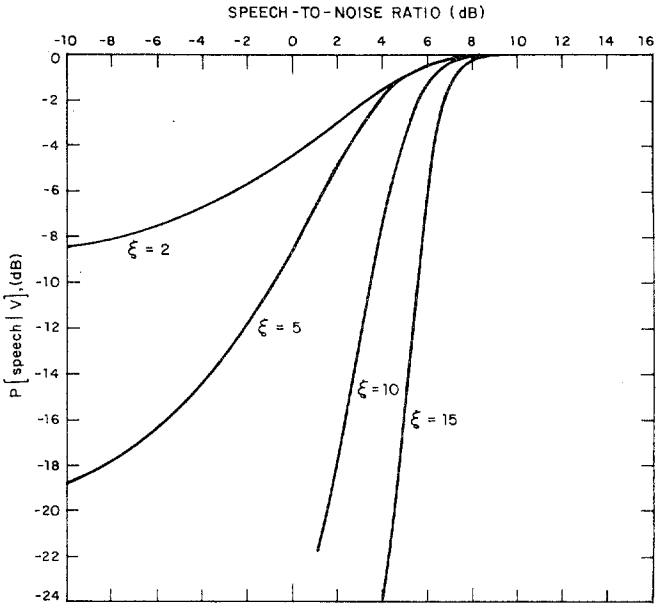

Fig. 2. *A posteriori* probability for the speech state.

presence of speech:

$$P(H_1|V) = \frac{\exp(-\xi)I_0\left[2\sqrt{\xi\left(\frac{V^2}{\lambda_w}\right)}\right]}{1 + \exp(-\xi)I_0\left[2\sqrt{\xi\left(\frac{V^2}{\lambda_w}\right)}\right]}. \tag{30}$$

It is this term which contributes the "soft-decision" aspect to the maximum likelihood envelope estimator in contradistinction with "hard decision" for which the speech plus noise is either passed as is or is suppressed completely. Appending the measured phase to the estimated envelope in order to preserve the identity system in the absence of noise, the final suppression rule is then

$$\hat{s} = \hat{A}\frac{y}{|y|} = \left[\frac{1}{2} + \frac{1}{2}\sqrt{\frac{V^2 - \lambda_w}{V^2}}\right]\cdot P(H_1|V)\cdot y. \tag{31}$$

In Fig. 2 several curves for the *a posteriori* probability for the speech state $P(H_1|V)$ are plotted as a function of the *a posteriori* speech-to-noise ratio V^2/λ_w (i.e., the measured SNR) for various values of the *a priori* signal-to-noise ratio ξ. The channel gains obtained when these *a posteriori* probabilities are appended to the maximum likelihood suppression rule are shown in Fig. 3. The two-state soft-decision maximum likelihood algorithm applies considerably more suppression when the measurement corresponds to low speech SNR. Since this case "most likely" corresponds to noise alone, it is seen that the effect of the residual noise (false alarms) should be considerably reduced. When the speech SNR is large, the measured SNR (i.e., the *a posteriori* SNR V^2/λ_w) will be large and it "most likely" means that speech is present, in which case the original maximum likelihood algorithm is the correct rule for extracting the speech envelope.

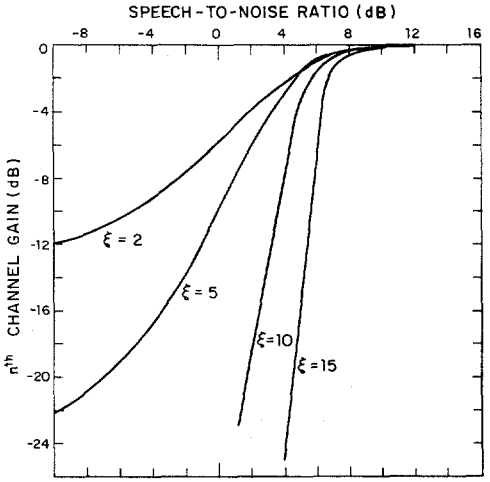

Fig. 4. The channel vocoder filter bank.

preceding theory was extracted), one would choose ξ (the *a priori* SNR A^2/λ_w) in order to guarantee a specified performance in terms of false alarms and missed detections. In speech, however, one must deal with whatever SNR exists as a consequence of the particular acoustic environment in which one is forced to operate; hence, the concept of an *a priori* SNR which can be controlled by the system designer is inappropriate. In terms of a noise suppression prefilter, however, Fig. 3 shows that the parameter $\xi = A^2/\lambda_w$ simply controls the amount of suppression applied to a particular frequency channel; hence, it is convenient to refer to it as the "suppression factor." From this point of view, the theory has simply provided the catalyst for generating a new class of suppression curves.

III. Implementation

All of the noise suppression prefilters that have been reported on to date have been implemented in the frequency domain. This corresponds nicely to the theoretical orthogonal channel decomposition used in Section II and exploits the properties of the FFT for filtering by circular convolution. Since the present work evolved from an attempt to implement a time-domain Kalman filter based on a parallel formant model for speech [15], and since a contemporary implementation of a channel vocoder is being developed using CCD technology to produce a package which operates at rates from 1.2 to 4.8 kbits/s, requires about 50 integrated circuits, occupies 0.22 ft^3, requires 5 W, and weighs 5 lb [16], it seemed appropriate to attempt a time-domain implementation of the prefilter that could exploit this emerging technology. As in the channel vocoder, 19 filters are used to span the frequency range 180–3720 Hz (the sampling rate was 7575 Hz). Each filter in the bank is a result of a bandpass transformation of a second-order Butterworth filter. The center frequencies and the bandwidths for each of the filters in the bank are listed in Table I and a plot of their linear magnitudes is shown in Fig. 4.

Although theory requires that the channels be orthogonal, in practice, overlapping filters provide for spectral smoothing which is known to be an important factor in the design of

TABLE I
Channel Filter Specifications

Channel Number	Center Frequency	3 dB Bandwidth
0	240	120
1	360	120
2	480	120
3	600	120
4	720	120
5	840	120
6	975	150
7	1125	150
8	1275	150
9	1425	150
10	1575	150
11	1750	200
12	1950	200
13	2150	200
14	2350	300
15	2600	300
16	2900	300
17	3200	300
18	3535	370

Sampling Rate = 132 μs

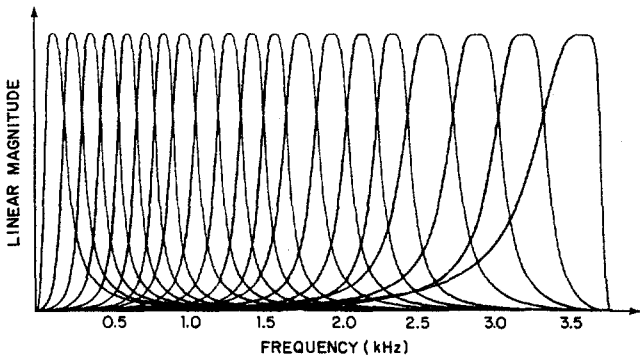

Fig. 3. Suppression rules for maximum likelihood with soft suppression.

for smoothing the envelope of the speech spectrum; hence, their lack of orthogonality turns out to be an asset in this particular case. Since the 19 filters span the frequency range of the speech signal, the front end of the channel vocoder, in the absence of noise, represents an identity system provided the outputs of each of the channels are added alternately out of phase, as shown in the block diagram in Fig. 5.

In order to compute the channel gains, measurements must be made to determine the instantaneous signal power and the average noise power at the output of each of the channel filters. Since the speech parameters change very little in 20 ms, some temporal smoothing can be exploited by computing the signal power in the nth channel from

$$V_n^2 = \frac{1}{N} \sum_{k=1}^{N} y_n^2(k) \tag{32}$$

where $y_n(k)$ represents the signal sample out of the nth channel at time k where there are N such samples in the 20 ms frame (the normalization by N will be unnecessary).

Determination of the background noise power requires

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.