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.

0096-3518/80/0400-0137$00.75 © 1980 IEEE

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

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

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

where

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

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 produce the speech estimate

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

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

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

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 deterministic waveform of unknown amplitude and phase. In 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

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

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

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

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

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

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

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

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 $|x|$ ($\geqslant 3$), which represents a constraint on the signal-to-noise ratios,

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

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

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

Maximizing this function with respect to A leads to the estimator

$$\hat{A} = \frac{1}{2}\left[|y_n| + \sqrt{|y_n|^2 - \lambda_w(n)}\right]. \tag{19}$$

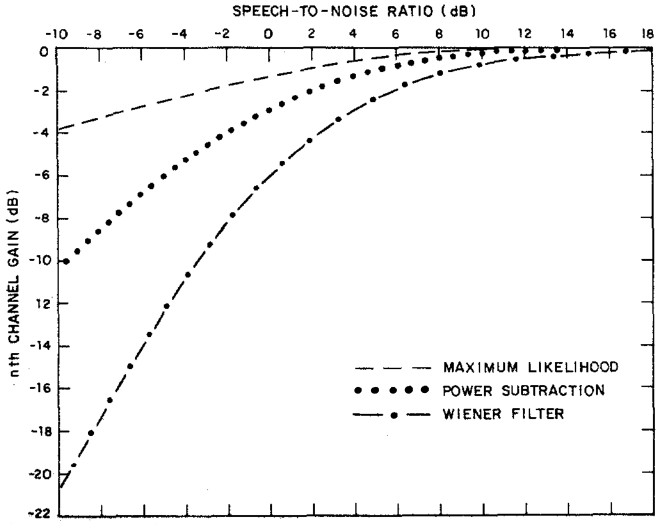


Fig. 1. Power subtraction, Wiener filter, and maximum likelihood suppression rules.

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

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

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

$$\begin{aligned} &H_0\text{: speech absent:} \quad |y_n| = |w_n| \\ &H_1\text{: speech present:} \quad |y_n| = |Ae^{j\theta} + w_n|. \end{aligned} \tag{21}$$

Only the measured envelope is used in this measurement model since it has already been shown that the measured phase provides no useful information in the suppression of the noise.

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

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

and substituting (26), (27), and (28) into (25) results in the following expression for the *a posteriori* probability for the presence of speech:

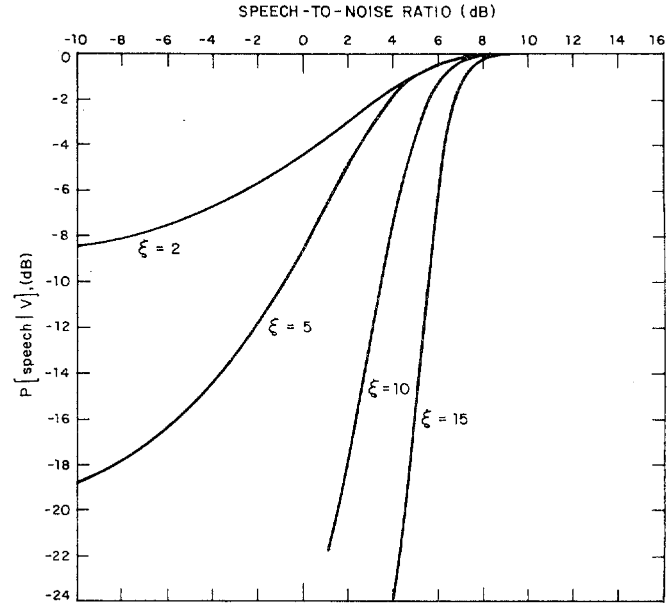


Fig. 2. *A posteriori* probability for the speech state.

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

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

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

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.

In order to interpret the role of the parameter ξ, it is noted that in a radar or communications context (from which the

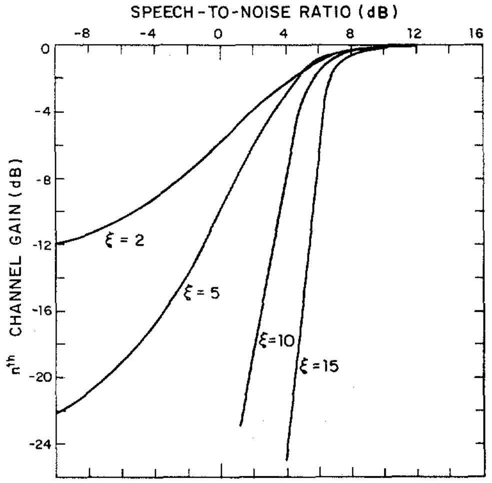


Fig. 4. The channel vocoder filter bank.

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 noise suppression systems [11]. The filters in the channel vocoder were originally chosen to provide a good compromise 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.

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

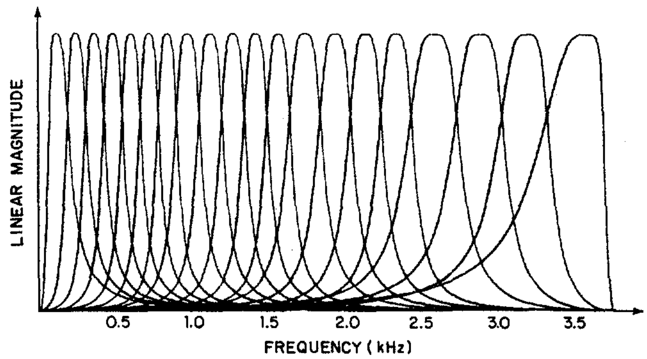


Fig. 3. Suppression rules for maximum likelihood with soft suppression.

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 knowledge of whether or not a particular frame contains speech. One approach to making this determination has been devel-

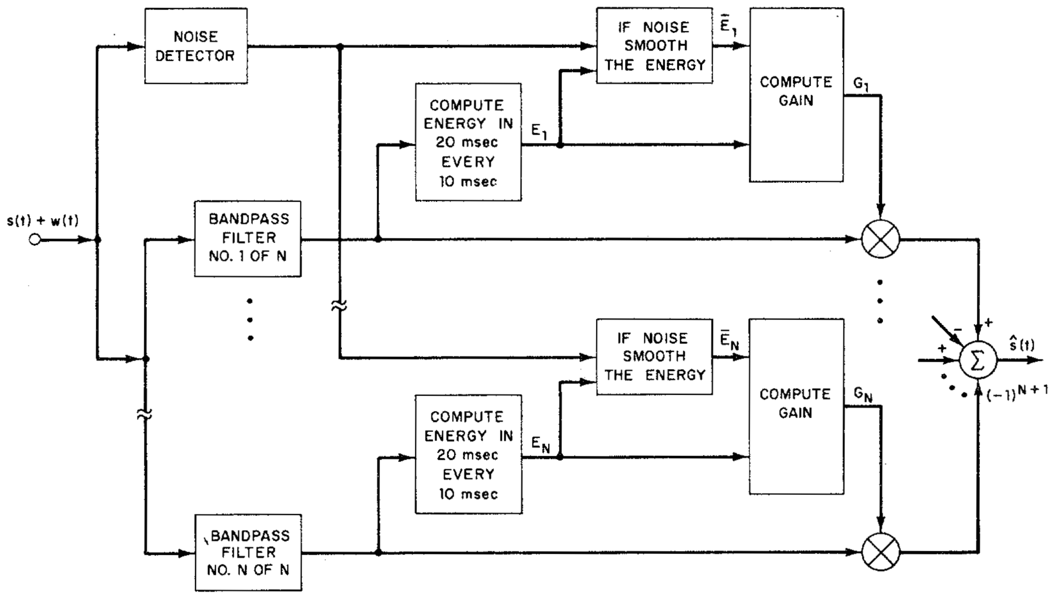

NOISE DETECTOR
IF NOISE SMOOTH THE ENERGY
$\bar{E}_1$
COMPUTE GAIN
G_1
COMPUTE ENERGY IN 20 msec EVERY 10 msec
E_1
s(t) + w(t)
BANDPASS FILTER NO. 1 OF N
IF NOISE SMOOTH THE ENERGY
$\bar{E}_N$
COMPUTE GAIN
G_N
Σ
$\hat{s}(t)$
+
$(-1)^{N+1}$
COMPUTE ENERGY IN 20 msec EVERY 10 msec
E_N
BANDPASS FILTER NO. N OF N

Fig. 5. Block diagram of the noise suppression prefilter.

oped by Roberts [17] who noted that a 4 s histogram of the frame energies of the input signal was bimodal. He found that by setting a detection threshold between the modes, correct speech and noise classification could be made most of the time. A modification of this algorithm, which is described in detail in the Appendix, was used to determine with high confidence the frames that were absent of speech. For those frames, the average noise power in each channel was estimated by smoothing the measurements in (32) using a 1 s time constant according to the recursion

$$\lambda_n(m) = \lambda_n(m-1) + \alpha[V_n^2(m) - \lambda_n(m-1)] \tag{33}$$

where $V_n^2(m)$, $\lambda_n(m)$ represents the measured power and the average noise power computed for the mth frame. The major drawback of this algorithm is the relatively long adaptation time needed to determine the detection threshold and then the additional training period required to learn the channel noise powers.

Using the measurement of $V_n^2(m)$ and the estimated average value of the noise energy $\lambda_n(m-1)$, the gain factor

$$g_n(m) = \frac{V_n^2(m) - \lambda_n(m-1)}{V_n^2(m)} \tag{34}$$

is computed for each channel.[1] Since $V_n^2(m)/\lambda_n(m-1)$ can be expressed in terms of $g(m)$, then the noise suppression rule, (30) and (31), can be written as

$$G_n(m) = \frac{\hat{s}_n}{y_n} = \frac{1}{2}\left(1 + \sqrt{g_n(m)}\right) \cdot \frac{\exp(-\xi) I_0\left(2\sqrt{\frac{\xi}{1-g_n(m)}}\right)}{1 + \exp(-\xi) I_0\left(2\sqrt{\frac{\xi}{1-g_n(m)}}\right)}. \tag{35}$$

The advantage in using $g(m)$ as the independent variable is the fact that $0 \leqslant g(m) \leqslant 1$ which permits the use of a simple software divide routine in forming the normalization. For a given value of the suppression factor ξ, the measured gain $g(m)$ is used as a pointer for a table lookup to determine the attenuation prescribed by (35). Fifteen tables corresponding to values of $\xi = 1, 2, 3, \cdots, 15$ have been included in the prefilter, with each table consisting of 50 values of the suppression rule computed for equal increments of $g(m)$ from 0 to 1. No attempt was made to optimize the design of these tables. All of the coding was done in machine language on the LDVT [19], which has the ability to key in a new value of the suppression factor in real time. This meant that the prefilter could easily be adjusted to accomodate a wide class of operational environments. This turned out to be a significant capability for effective noise suppression. Since the algorithm was designed to operate in real time, a 10 ms delay had to be incurred between the time the energies were measured and the time the corresponding gains could be computed and applied to the channel waveforms. This was done by computing the energies (block floating point) in 10 ms segments and adding consecutive segments together to produce the desired 20 ms energy measurement. This permitted computation of the raw gains $G(m)$ every 10 ms. In order to avoid the introduction of discontinuities in the output waveform, the final output is a smoothed gain $\bar{G}(m)$ obtained according to

$$\bar{G}(m) = \bar{G}(m-1) + \beta(m)[G(m) - \bar{G}(m-1)]. \tag{36}$$

Since the introduction of smoothing can cause the prefilter to be slow to respond to a leading edge transition, which could result in speech distortion, the weighting factor in (36) is chosen adaptively according to the rule

$$\beta(m) = \begin{cases} 1 & \text{if } g(m) \geq \bar{G}(m-1) \\ \frac{1}{2} & \text{if } g(m) < \bar{G}(m-1). \end{cases} \tag{37}$$

In this way, the prefilter responds immediately to an increase in the SNR which should minimize the potential for leading edge distortion. During a trailing edge, in which the gain will be decreasing, the smoothed gain will be used which will tend

[1] This is where the normalization by N in (32) disappears.

to maintain the speech signal even though the noise becomes dominant. It is the gain $\overline{G}(m)$ in (37) that is applied to the waveform at the output of each of the channel filters. These waveforms were then added together alternately 180° out of phase to produce the prefilter output waveform $\hat{s}(t)$.

IV. Experimental Results and Conclusions

Since the prefiltering algorithm operated in real time, it was possible to perform extensive listening tests on a large speech and noise database. It was of particular interest to determine the operational performance of the prefilter in conjunction with a 2400 bits/s vocoder operating in a background of E4A advanced airborne command post noise (ACPN). Source tapes were available for this environment, consisting of lists spoken by six male speakers for which a DRT score and a diagnostic acceptability measure (DAM) could be computed. The recordings were made using both a high-quality Altec microphone and a noise-canceling microphone.

The first experiment consisted of listening to the output of the prefilter for various values of the suppression factor. It was always possible to select a suppression factor which would render the background noise imperceptible, although, for cases in which the SNR was low enough, the cost in doing this was the introduction of various degrees of speech distortion. In these cases, if the suppression factor was subsequently reduced, the speech distortion was reduced at the expense of introducing a perceptible level of background noise.

In the next experiment, the prefilter was connected in tandem with the 2400 bit/s LPC vocoder which used the Gold-Rabiner pitch estimator [19], [20]. An unexpected result was obtained. If the suppression factor was set to remove the residual noise at the output of the prefilter, then the speech quality at the vocoder was poor due to both buzz-hiss errors and spectral distortion. If, however, the suppression factor was chosen so that the noise at the vocoder output was negligible, then a significantly lower value of the suppression factor was needed and the speech quality was quite good, although the Gold-Rabiner algorithm continued to make buzz-hiss errors, but at a lower rate. In other words, LPC itself has some suppression capabilities against weak noise which can usefully be exploited in the tandem connection. It was the flexibility in selecting the prefilter suppression factor which made this result possible.

Since the deployment of the LPC vocoder does allow for flexibility in the specification of the pitch extractor, it was of interest to determine whether or not algorithms that were specially designed to operate in noise would operate more effectively in the tandem connection. Such an algorithm, based on maximum likelihood estimation techniques, has been under development for some time [21] and was chosen to be tested against the Gold-Rabiner algorithm. In the subjective listening tests it was found that, indeed, smoother pitch tracks could be obtained with a lower rate of buzz-hiss errors.

Although the results of using the prefilter always produced subjectively more pleasant sounding speech to the ear since the annoying and tiresome background noise was removed, it was important to determine whether or not there was a corresponding quantitative improvement in intelligibility. To do this, DRT scores are being obtained for the prefiltered speech and the speech out of the LPC tandem for both the Gold-Rabiner and the maximum likelihood pitch extraction algorithms. Results are currently being obtained for both the Altec dynamic microphone and the confidencer noise-canceling microphone and will be reported once all of the data have been collected and analyzed.

So far the focus has been on the 19-channel prefilter based on the principles of channel vocoder design. This was strictly a pragmatic choice which was made to facilitate the development of a real-time testbed. Questions relating to the number of filters, the bandwidths, and the choice of center frequencies remain to be addressed. Although the time-domain structure of the channel prefilter is well suited to an analog implementation using CCD technology, it is of interest to determine the tradeoffs with respect to a frequency-domain approach using the FFT. Whatever candidate system is chosen for evaluation, using the class of suppression rules developed in this study allows the overall design to be optimized with respect to the noise suppression/speech distortion tradeoff by choosing an appropriate suppression factor. In this way, performance differences can be attributed to the system design parameters independent of a particular suppression rule which may have represented a poor choice for the particular signal and noise conditions used in the evaluation.

Appendix

Modified Roberts Noise Detection Algorithm

In order to estimate the statistics of the background noise, it is desirable to inspect only those frames of data which have a high probability of containing no speech. To accomplish this, an adaptive energy threshold marking the probable boundary between noise and noise plus speech is established by monitoring the energy on a frame-by-frame basis and maintaining energy histograms which reflect the bimodal distribution of the energy. The flowchart for the algorithm, shown in Fig. 6, is described in the following paragraphs.

For each frame, the sum of the squares of the input samples is computed. If this energy does not exceed 16 bits (i.e., does not strongly imply the presence of speech), the adaptive threshold algorithm is exercised. First, a decay factor of 0.995 is applied to a 128-bin histogram of uniform ranges of energy, causing exponential decay of the histogram values with a time constant of 4 s. The value of the bin which encompasses the energy of the current frame is incremented by 160. A typical energy histogram after adaptation is complete is shown in Fig. 7(a).

A second 128-point cumulative histogram is then formed to represent the area under the first histogram by computing the accumulated scores from the low-energy bin through a high-energy bin. Fig. 7(b) shows the result of accumulating the scores from the histogram in Fig. 7(b). If the 10th point of the second histogram exceeds 25 percent of the total area, it is assumed that there is no noise present.

If noise is present, a search is made through the second histogram for the point which represents 80 percent of the total area. The quantum of energy corresponding to this point becomes the new threshold candidate. If this candidate exceeds the current threshold, the threshold is updated using a decay factor of 0.95, a slow time constant of 400 ms. If the candidate is below the current threshold, the threshold is up-

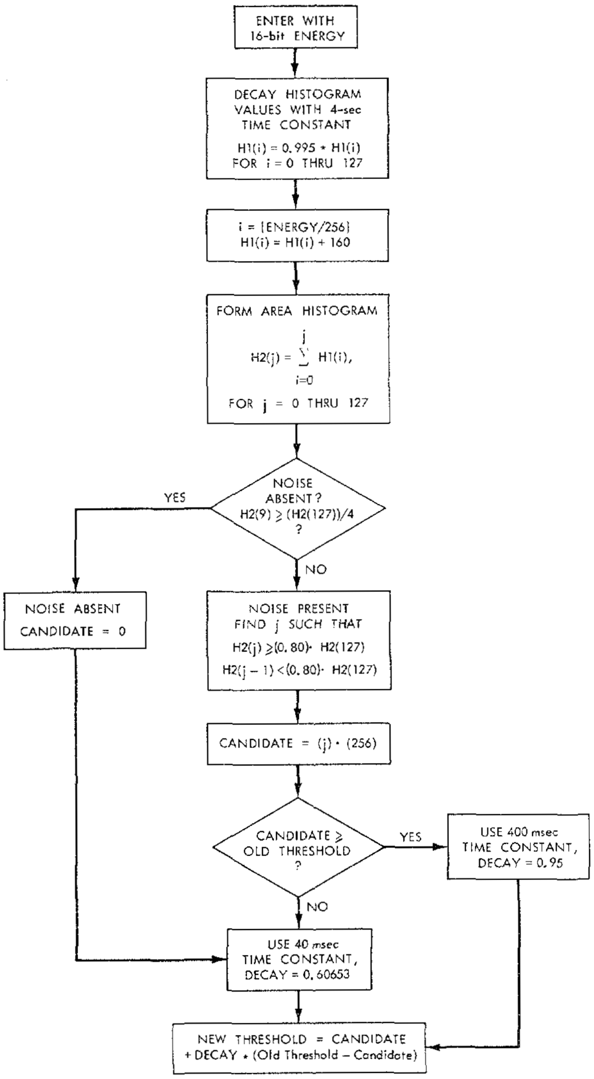


Fig. 6. Modified Roberts noise detection algorithm.

dated with a decay factor of 0.60653, a fast time constant of 40 ms.

If noise is absent, the new threshold candidate is set to zero, and the threshold is updated using a decay factor of 0.60653, a fast time constant of 40 ms. Finally, the threshold is held to a minimum of 1024 to guarantee updating of the estimated noise components when background noise suddenly disappears.

Acknowledgment

The authors appreciate the technical interaction provided by their colleagues J. Tierney and T. Bially. Their intuition regarding the appropriateness of the channel vocoder front end as a preprocessing structure was particularly helpful. The authors would also like to thank A. J. McLaughlin for his support and encouragement throughout the course of this work. Finally, the authors appreciate the comments made by a conscientious and knowledgeable reviewer which resulted in an improved final draft.

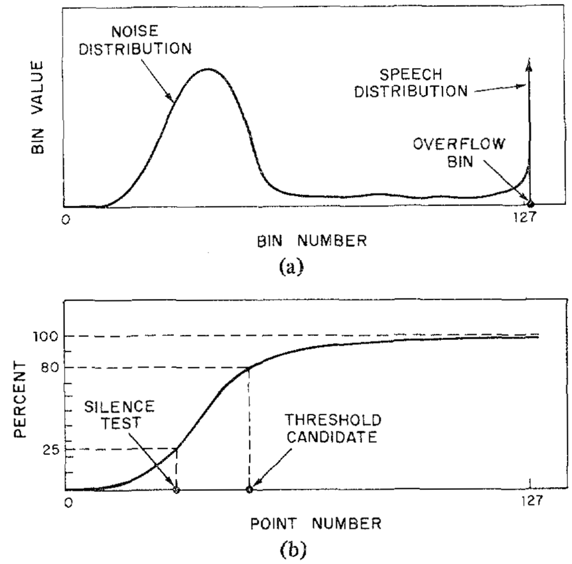


Fig. 7. (a) Typical energy histogram (H_1). (b) Typical cumulative histogram (H_2).

References

[1] T. E. Tremain, J. W. Fussell, R. A. Dean, B. M. Abzug, M. D. Cowing, and P. W. Boudra, Jr., "Implementation of two real time narrowband speech algorithms," in *EASCON '78 Rec.*, Sept. 1978, pp. 698-708.

[2] C. Teacher and H. Watkins, "ANDVT microphone and audio system study," Ketron Final Rep. for the Commander, Naval Electron. Syst. Command, Washington, DC, Aug. 1978.

[3] R. J. McAulay, "Maximum likelihood spectral estimation using state variable techniques," in *Proc. RADC Spectrum Estimation Workshop*, May 1978, pp. 63-68.

[4] D. Paul, "A robust vocoder with pitch-adaptive spectral envelope estimation and an integrated maximum-likelihood pitch estimator," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1979), pp. 64-68.

[5] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, p. 197, 1978.

[6] B. Widrow *et al.*, "Adaptive noise canceling: Principles and applications," *Proc. IEEE*, vol. 63, p. 1692, 1978.

[7] R. C. Rohlfs, "Speech enhancement using the Widrow-Hoff adaptive noise-canceling tapped delay-line filter; A CSP30 implementation," MITRE Tech. Rep. MTR-3626, Mar. 1979.

[8] M. R. Sambur, "Adaptive noise canceling for speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, p. 419, 1978.

[9] M. R. Weiss, E. Aschkenasy, and T. W. Parsons, "Study and development of the INTEL technique for improving speech intelligibility," Tech. Rep. RADC-TR-75-108, RADC, Griffiss Air Force Base, Rome, NY, Apr. 1975.

[10] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, p. 113, 1978.

[11] R. A. Curtis and R. J. Niederjohn, "An investigation of several frequency-domain processing methods for enhancing the intelligibility of speech in wideband random noise," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1978, pp. 602-605.

[12] R. D. Preuss, "A frequency domain noise cancelling preprocessor for narrowband speech communications systems," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1979, pp. 212-215.

[13] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1979, pp. 208-211.

[14] H. L. Van Trees, *Detection, Estimation and Modulation Theory, Part I.* New York: Wiley, 1968, pp. 54-56, 198-206, 205-207.

[15] R. J. McAulay, "A structure for robust speech processing," in *Proc. Int. Conf. Commun.*, vol. 1, June 1978, pp. 8.5.1-8.5.3.

[16] P. E. Blankenship, "An NMOS LSI channel vocoder implementation," in *Proc. EASCON '78*, Sept. 1978, pp. 684-692, and presented at the SPIE Tech. Symp. East '79, Washington, DC, Apr. 1979.

[17] J. Roberts, "Modification to piecewise LPC," MITRE Working Paper WP-21752, May 1978.

[18] P. E. Blankenship, "LDVT: High performance minicomputer for real-time speech processing," in *EASCON '77 Rec.*, Sept. 1977.

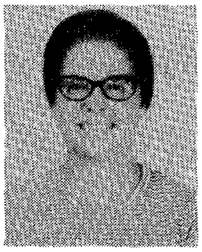
[19] B. Gold and L. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *J. Acoust. Soc. Amer.*, vol. 46, p. 442, 1969.

[20] E. M. Hofstetter, P. E. Blankenship, M. L. Malpass, and S. Seneff, "Vocoder implementations on the Lincoln digital voice terminal," in *EASCON '77 Rec.*, Sept. 1977.

[21] R. J. McAulay, "Design of a robust maximum likelihood pitch estimator for speech in additive noise," Tech. Note 1979-28, M.I.T. Lincoln Laboratory, Lexington, MA, June 1979.

Robert J. McAulay (S'63-M'67) was born in Toronto, Ont., Canada, on October 23, 1939. He received the B.A.Sc. degree in engineering physics with honors from the University of Toronto in 1962, the M.Sc. degree in electrical engineering from the University of Illinois, Urbana, in 1963, and the Ph.D. degree in electrical engineering from the University of California, Berkeley, in 1967.

In 1967 he joined the Radar Signal Processing Group of the M.I.T. Lincoln Laboratory, Lexington, MA, where he worked on problems in estimation theory and signal/filter design using optimal control techniques. From 1970 to 1975 he was a member of the Air Traffic Control Division at Lincoln Laboratory, and worked on the development of aircraft tracking algorithms, optimal MTI digital signal processing, and on the problems of aircraft direction finding for the Discrete Address Beacon System. On a leave of absence from Lincoln Laboratory during the winter and spring of 1974, he was a Visiting Associate Professor at McGill University, Montreal, P.Q., Canada. Since 1975 he has been a member of the Speech Systems Technology Group at Lincoln Laboratory, where he has been involved in the development of robust narrow-band speech vocoders.



Marilyn L. Malpass was born in Cleveland, OH, on July 18, 1938. She received the B.A. degree in physical sciences and math from Radcliffe College, Cambridge, MA, in 1960.

She joined the M.I.T Lincoln Laboratory, Lexington, MA, as a Programmer in June 1960, and later became a member of the Technical Staff. For the past several years she has been involved in real time speech processing and packet network simulations.

Impact of the Ocean Acoustic Transfer Function on the Coherence of Undersea Propagations

ALBERT A. GERLACH

Abstract—**Multipath propagation, in conjunction with source motion, creates a splitting and spreading of the received source-signal energy over the time register time scale-factor (ambiguity) plane. This splitting of the received signal energy is manifested as correlation interactions between both; the set of eigenray signal arrivals (comprising the multipaths) at each receiving sensor, and the set of eigenray signal correlation pairs comprising two (or more) sensors. The more tightly the sets of eigenray signals are clustered in the ambiguity plane, the stronger will be the signal component interactions at each sensor and the less will be the expected correlation degradation between sensors. However, the greater will be the variance. Conversely, the more diffuse the clustering, the weaker will be the signal component interactions and the greater will be the expected intersensor coherence degradation. Also, the less will be the variance. Criteria which define the degree of signal component interactions in terms of the signal and the ocean transfer function parameters have been defined and are displayed in graphical form.**

Introduction

AS A consequence of multipath acoustic propagation, a remotely received signal will consist of the weighted superposition of a number of source signals both slightly compressed (or expanded) and translated in time [1]. The time-variable compression factor is a stochastic process which reflects the fluctuations inherent in both the medium and the source motion. The signal physics are depicted in Fig. 1 for a three eigenray-path model. The diagram is distorted in that the vertical scale is highly exaggerated in comparison with the horizontal scale, and for long ranges (greater than about 200 nmi) the number of path cycles would range from about 7 to greater than 30 (instead of the single cycles shown). The three eigenray paths shown consist of two purely refractive (RR) paths and one refractive surface-reflected (RSR) path. Other paths are also possible [2], [3].

The source signal arriving over each eigenray path differs in relative amplitude A (depicted by the weight of the path), initial time delay τ_0, and time scale-factor shift (or Doppler ratio) δ. For a motional source, the relative amplitudes and the Doppler ratios will be time variable. Signal-wise, the source signal time-scale "t" is transformed into "kt" upon being projected into the medium where $k = 1 + \delta$ is the time scale-factor and δ is the minute scale-factor shift or Doppler ratio. Mathematically, δ is a form of running time average of the scalar or "dot" product between the source velocity and a unit vector along the eigenray path at the source, divided by the sound speed at the source [1]. The Doppler ratios differ as a consequence

Manuscript received January 29, 1979; revised September 6, 1979. This work was supported by the Defense Advanced Research Projects Agency under ARPA Order 3074.

The author is with the Naval Research Laboratory, Washington, DC 20375.