SPEECH ENHANCEMENT USING A TERNARY-DECISION BASED FILTER

T.S. Sun, S. Nandkumar, J. Carmody, J. Rothweiler, A. Goldschen, N. Russell, S. Mpasi, P. Green

Martin Marietta Laboratories
1450 South Rolling Road
Baltimore, MD 21227

ABSTRACT

A new speech enhancement scheme based on a generalized Wiener filter formulation is proposed. A ternary-valued parameter is derived empirically based on the likelihood of the input signal vector being classified as speech. This parameter controls the Wiener filter coefficient in order to obtain an improved speech spectral estimate. This "ternary-decision" concept renders a logical compromise between the simple, hard, "binary speech/noise decision" filtering and the elaborate "soft-decision" filtering approaches in terms of practicality and performance. An important feature in our scheme is that we exploit the interframe spectral relationship to reinforce the assessment of the likelihood of weak speech components. This feature prevents many weak formants from being disproportionally attenuated as in most previous schemes. Other important features of our scheme include a novel speech/noise classifier and a robust noise median amplitude tracker, both of which make the estimate of noise spectrum more reliable. A preliminary evaluation of this new scheme is reported here.

1. INTRODUCTION

Speech enhancement is required to ensure the quality and reliability of voice communication systems under noisy environments. Many speech enhancement schemes have been based on a Wiener filtering framework with varying degrees of success. Representatives of these schemes include the two-state "hard-decision" filtering and "soft-decision" filtering [1], iterative filtering [2], and constrained iterative filtering [3]. Each of these schemes has its strengths and weaknesses. The two-state "hard-decision" filtering approach reduces the residual noise somewhat, but it tends to degrade the naturalness of the speech. The "soft-decision" filtering approach improves the naturalness of speech, but it tends to dull the speech because of a lack of nonlinear characteristics in its filter function. In addition, most schemes tend to attenuate weak speech components disproportionally because the filter function depends only on the estimated noise spectra. The iterative filtering approaches are seen to improve perceptual quality over the above schemes, but they have the disadvantage of increased computational cost. Besides Wiener filter based schemes, there have been schemes based on cepstral and spectral domain noise subtraction. In particular, the INTEL method [7], which uses root cepstral transformation, has been shown to be effective and robust for enhancing speech degraded by a wide range of noise conditions. However, this scheme has the drawback of causing unnatural distortions to the processed speech. For practical purposes, it is desirable to develop a scheme to achieve a well-balanced compromise among the performances of these schemes regarding speech quality, reduction of residual noise, and computational efficiency. This is the primary motivation for our ternary-decision scheme.

2. BASIC CONCEPT

The foundation of our approach is a generalized Wiener filtering model where the predicted speech component at the kth frequency unit, $\hat{\mathbf{s}}_k$, is given by,

$$\hat{\mathbf{s}}_k = \frac{\langle s_k^2 \rangle}{\langle s_k^2 \rangle + \lambda \langle w_k^2 \rangle} \mathbf{y}_k, \tag{1}$$

where $\mathbf{y}_k$ is the input signal assumed to be speech corrupted with additive noise, and $\langle s_k^2 \rangle$ and $\langle w_k^2 \rangle$ are expected power spectra of speech and noise, respectively. The parameter λ is used to control the amount of noise suppression according to the likelihood that the kth frequency peak is speech or noise. In the extreme case where the speech is assumed to be present in every measurement, $\lambda = 1$, and the above expression becomes identical to the classical Wiener filtering. For practical purposes, we consider a ternary realization of λ, which is a compromise between the binary "hard" decision case and the continuous "soft" decision case [1]. We can derive the above filter function explicitly based on a least mean square error criterion. Using a treatment similar to the one used in [1, 2], we consider the noisy speech in the frequency domain to be represented by two independent random processes,

$$\mathbf{x}_k = \mathbf{s}_k + \mathbf{n}_k, \tag{2}$$

where $\mathbf{s}_k$ and $\mathbf{n}_k$ are the speech and noise vectors, respectively. The phase is not important in processing speech and is assumed to be uniformly distributed. The noise is assumed to be a Gaussian process with a Rayleigh distribution for the amplitude. We can decompose the noise amplitude into its mean $\langle n_k \rangle$ and another variable $\mathbf{w}_k$, and define input noisy signal $\mathbf{y}_k$ as below,

$$\mathbf{n}_k = \mathbf{w}_k + \langle n_k \rangle, \text{ and } \mathbf{y}_k = \mathbf{x}_k - \langle n_k \rangle.$$

Hence,

$$\mathbf{y}_k = \mathbf{s}_k + \mathbf{w}_k, \text{ with } \langle n_k \rangle = 0.$$

0-7803-2431-5/95 $4.00 © 1995 IEEE

Next, we consider the following two-state prediction model,

$$\begin{aligned} \hat{\mathbf{s}}_k &= \alpha \mathbf{y}_k \text{ if } \mathbf{s}_k > 0 & (3) \\ \hat{\mathbf{s}}_k &= \beta \mathbf{w}_k \text{ if } \mathbf{s}_k = 0. & (4) \end{aligned}$$

The prediction error $\mathbf{d}_k$ is then

$$\begin{aligned} \mathbf{d}_k &= \hat{\mathbf{s}}_k - \mathbf{s}_k \\ &= (\alpha - 1)\mathbf{s}_k + \alpha \mathbf{w}_k \text{ if } \mathbf{s}_k > 0 \\ &= \beta \mathbf{w}_k \text{ if } \mathbf{s}_k = 0. \end{aligned}$$

The mean square error can be expressed in terms of the following probabilities: p (probability of $\mathbf{y}_k$ being in the speech-plus-noise state); q (conditional probability of $\mathbf{y}_k$ being correctly classified given that $\mathbf{y}_k$ is in the speech-plus-noise state); and r (conditional probability of $\mathbf{y}_k$ being correctly classified given that $\mathbf{y}_k$ is in the noise-only state). The mean square error can now be expanded as,

$$\begin{aligned} \langle d_k^2 \rangle &= pq\langle[(\alpha - 1)s_k + \alpha w_k]^2\rangle + \\ &\quad p(1-q)\langle[(\beta - 1)s_k + \beta w_k]^2\rangle + \\ &\quad (1-p)r\langle(\beta w_k)^2\rangle + \\ &\quad (1-p)(1-r)\langle(\alpha w_k)^2\rangle \\ \langle d_k^2 \rangle &= [pq(\alpha - 1)^2 + p(1-q)(\beta - 1)^2]\langle s_k^2\rangle + \\ &\quad \{[pq + (1-p)(1-r)]\alpha^2 + [p(1-q) + \\ &\quad (1-p)r]\beta^2\}\langle w_k^2\rangle. \end{aligned}$$

Minimization of $\langle d_k^2 \rangle$ with respect to α and β, respectively, leads to

$$\alpha = \frac{\langle s_k^2\rangle}{\langle s_k^2\rangle + [1 + \frac{(1-p)(1-r)}{pq}]\langle w_k^2\rangle}, \qquad (5)$$

$$\beta = \frac{\langle s_k^2\rangle}{\langle s_k^2\rangle + [1 + \frac{(1-p)r}{p(1-q}]\langle w_k^2\rangle}. \qquad (6)$$

Note that α is related to the amount of distortion whereas β is related to the amount of residual noise in the filtered output. Combining Eqns (3),(4),(5), and (6) leads to Eqn (1), where

$$\lambda = 1 + \frac{(1-p)(1-r)}{pq},$$

if $\mathbf{y}_k$ is classified as a speech-plus-noise measurement, or

$$\lambda = 1 + \frac{(1-p)r}{p(1-q)},$$

if $\mathbf{y}_k$ is classified as a noise-only measurement. In Eqn (1), the noise and speech spectra are estimated using the FFT. The noise spectra are updated continuously over sections classified as "noise-only". The speech spectra are obtained using $\langle s^2\rangle = \langle y^2\rangle - \langle w^2\rangle$.

The formulation of λ cannot be used to construct a filter directly because the probabilities involved in the derivation are unknown. For practical implementation, we simplify this model by assuming that each input frame is characterized by one of three possible states, based on its likelihood of containing a speech component. According to the above defined likelihood, the three states can be labelled as "most likely," "equally likely," and "least likely," and each is assigned a λ value. Using Eqns (5) and (6) as a starting guideline, the three λ values are determined empirically. Adjustment of these parameters is guided by the tolerance to distortion and residual noise.

3. ALGORITHM FEATURES

Figure 1 shows the top level flow diagram of our algorithm. Most of the steps in this diagram involve common signal processing routines, except for the following features, which are unique to our algorithm.

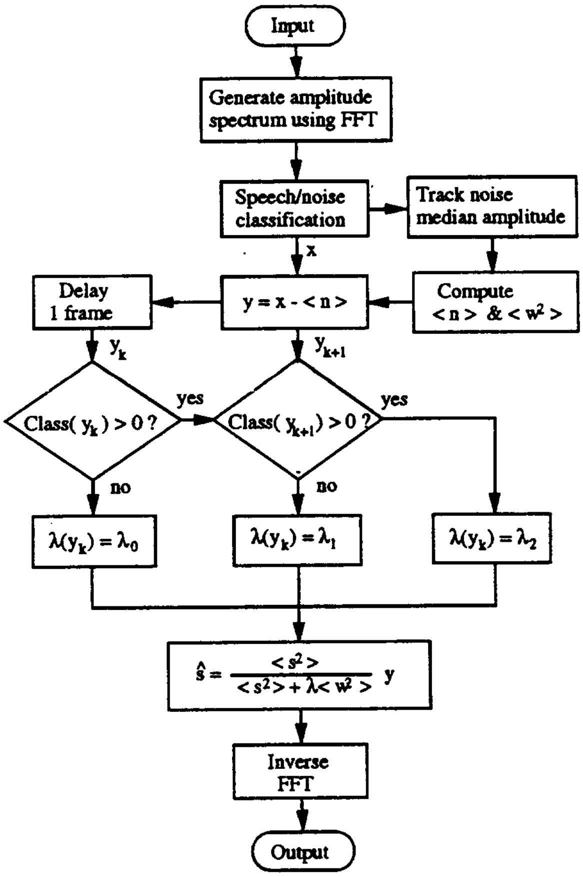

Figure 1: The proposed Ternary-Decision Based Filtering (TDBF) scheme.

3.1. Frame-by-Frame Speech/Noise Classification

This feature enables us to prevent speech-containing frames from biasing the noise estimation process. Our speech-noise classifier uses a four-parameter decision rule, which is based primarily on the difference of energy distribution between speech and noise in the frequency and time domains. In the frequency domain, the energy of a speech tends to be localized in a smaller number of frequency bins compared to a wideband noise. The first parameter quantifies this property in terms of the fourth root of the fourth moment of spectral amplitudes. In the time domain, the total energy of a speech-plus-noise frame is generally greater than that of a noise-only frame and this is represented by the second parameter. The third parameter, defined as the average energy of noise-only frames, is used as a normalization reference. This parameter must be initialized once in the beginning and is updated thereafter using a sliding window averaging routine. The fourth parameter is an empirical constant, which can be adjusted to match various noise conditions, such as white and colored noises. The decision rule is such that a speech frame is declared if the product of the first two parameters is greater than the product of the third and fourth parameters. This new classification scheme results in improved performance compared to two established schemes, VOICEN [4] and APID [5], as indicated in Table 1. The classification results in Table 1 are based on a total of 464 frames contained in the same five speech files used for enhancement evaluation (Section 4). The boundaries between voice and noise only (or silence) regions are marked manually using the clean speech files. Because the ratio between the numbers of speech frames and noise frames is about 2 to 1, the scores for correct classification of each voice frame and each noise frame are weighted so that a random classifier would score 50 on the average out of a perfect score of 100. The new classification scheme is also seen to result in robust performance as SNR decreases.

3.2. Median Noise Amplitude Tracking

Instead of computing the sliding average of the spectral envelops in the noise frames, we compute the median value of the envelop over a number of frames. This can reduce the risk of bias caused by large amplitudes, which occur occasionally due to misclassifications or accidental spikes in the noise. For computational efficiency, the entire spectrum is divided into 16 overlapping sections and the median value in each section is determined over a number of frames. The mean noise spectral amplitude is constructed by means of linear interpolation among those points.

3.3. Two-Step Ternary Decision Scheme

The fraction of input amplitude that can be retained as speech is controlled by the ternary parameter λ in the filter coefficient. Each input amplitude is classified into one of three states based on its likelihood to be speech. In the first step, the three states are classified using two adaptive thresholds, which are set to be proportional to the mean value of the noise amplitude. The two proportional constants are predetermined empirically. In the second step, the continuity between spectral peaks in the adjacent frames is tested. Measurements classified initially to be in the second state can be reclassified as speech or noise if they appear continuously in adjacent frames. In this way, many weak speech components with equal probabilities in either the speech-plus-noise or noise-only case can be enhanced effectively.

4. EVALUATION

Evaluation of the proposed algorithm consists of two parts: (a) an objective evaluation based on two SNR measures and the Itakura-Saito distance measure [6] and (b) a subjective evaluation consisting of five listeners. In the objective evaluation, five speech files randomly selected from the Resource Management Database are degraded with additive white Gaussian noise to simulate test data. The ratio of average amplitude squares between the entire original signal and the added noise sequence is termed reference SNR. The two SNR measures comprise the segmental SNR [6] and the global SNR as defined below,

$$\text{Global SNR} = 10log_{10}\left(\frac{1}{N}\sum_{n=1}^{N}\frac{\langle x_{sp}^2\rangle - \langle x_{ns}^2\rangle}{\langle x_{ns}^2\rangle}\right),$$

where $\langle x_{sp}^2\rangle$ and $\langle x_{ns}^2\rangle$ denote the average amplitude squares of samples taken from the voice and noise regions of the processed speech files, respectively, and N (=5) denotes the number of speech files; and

$$\text{Segmental SNR} = \frac{10}{M}\sum_{m=0}^{M-1} log_{10}\left(\frac{\sum_{k=Km+1}^{Km+K} s_k^2}{\sum_{k=Km+1}^{Km+K}[\gamma x_k - s_k]^2}\right),$$

where s_k and x_k are signal amplitudes taken from the voice regions of the clean and processed speech files, respectively, M is the total number of segments in the five speech files, K (=160) is the segment length, and γ is a scaling constant adjusted to maximize the segmental SNR value in each case. The voice and noise-only regions in the above SNR formulations are obtained using manually marked boundaries. Based on the above definition, the global SNR can be regarded as a performance measure with respect to the residual noise, whereas the segmental SNR can be regarded as a performance measure with respect to signal distortion. The scheme used for comparative evaluation is based on the INTEL [7] method with a square root cepstral transformation, which has been a well established technology in our laboratory. Although higher roots have been frequently used in the cepstral compression, the square root produces lower distortion and better perceptual quality. To obtain a fair comparison, the new speech/noise classifier has been used in both schemes (note that the INTEL and APID combination has a poorer performance). The Global SNR measures shown in Table 2 indicate superior noise reduction, with the exception of the clean speech case, for the proposed ternary decision based filter (TDBF) compared to the original input SNR and INTEL processing. The segmental SNR measure related to signal distortion also shows significant improvement for the proposed scheme over INTEL and the original noisy utterance (Table 3). Although the INTEL has the best global SNR in the clean speech case, it has the worst segmental SNR among the three groups. The gain-optimized Itakura-Saito distance measure, which is a measure of spectral distortion and is correlated to perceptual quality, also shows the lowest distances from the noise-free original for the proposed scheme (Table 4).

In the subjective evaluation, noisy speech was recorded digitally from two radio broadcast messages. One recording was a speech segment from a radio station and the other was wideband noise generated by tuning the receiver in between two stations. The proposed TDBF scheme was compared with the noisy original and the INTEL based scheme. Three of five evaluators preferred the proposed scheme because of its naturalness, the other two prefered the INTEL based scheme because of their preference to a more uniform background.

5. CONCLUSIONS

A new generalized Wiener filtering scheme which uses a ternary-decision based parameter is proposed. The ternary-decision based parameter which controls the filter function is experimentally determined using the guidelines of minimum residual noise and spectral distortion. A new speech/noise classifier is also proposed, which provides the enhancement algorithm with boundary points and is robust at low SNRs. The proposed Ternary-Decision Based Filtering (TDBF) scheme is shown to perform well in both areas of successful speech enhancement, that is, reducing residual background noise and producing artifact-free natural speech. Moreover, the proposed scheme is computationally less intensive than INTEL based and iterative speech enhancement approaches.

Classification Performance (Percent Correct)			
Reference	VOICEN	APID	Proposed Classifer
Clean	62	84	89
10 db	59	81	90
5 db	60	82	88
3 db	62	81	87
0 db	55	71	84

Table 1: Comparison of classification performances of VOICEN, APID and the proposed scheme.

Global SNR Measure			
Reference	Original	INTEL	Proposed TDBF
Clean	28.7	36.9	32.3
10	11.8	25.4	27.6
5	6.8	19.8	24.7
3	4.7	17.3	23.3
0	1.8	13.7	21.0

Table 2: Comparison of global SNR values (in dB) for speech processed by INTEL, and the proposed scheme (T-DBF), with respect to original noisy input.

Segmental SNR Measure			
Reference	Original	INTEL	Proposed TDBF
Clean	∞	17.2	33.2
10	6.5	9.5	10.9
5	2.4	6.8	7.6
3	1.2	5.6	6.4
0	0.4	3.7	4.7

Table 3: Comparison of segmental SNR values (in dB) for speech processed by INTEL, and the proposed scheme (T-DBF), with respect to original noisy input.

Gain-Optimized Itakura-Saito Measure			
SNR	Original	INTEL	Proposed TDBF
Clean	0.00	0.34	0.26
10	1.79	1.41	1.40
5	2.05	1.72	1.69
3	2.15	1.85	1.85
0	2.28	2.02	1.99

Table 4: Comparison of the Itakura-Saito Distance Measure for degraded speech (original), processed by INTEL and the proposed scheme (TDBF), with respect to noise-free (clean) speech.

REFERENCES

1. R.J. McAulay and M.L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp.137-145, 1980.

2. J. Lim, A. Oppenheim, "All-Pole Modeling of Degraded Speech," *IEEE Trans. Acous., Speech, Sig. Proc.*, vol. ASSP-26, pp. 197-210, June 1978.

3. S. Nandkumar and J.H.L. Hansen, "Speech enhancement based on a new set of auditory constrained parameters," *Proc. IEEE ICASSP*, pp. 001-004, April 1994.

4. T. Tremain, "The Government Standard Linear Predictive Coding Algorithm: LPC-10," *Speech Technology*, April 1982, pp. 40-49.

5. C.S. Wu, V. Nguyen, V. Goncharoff, W. Kushner, and J. Damoulakis, "Adaptive pitch detection algorithm for noisy signals," *Proc. IEEE ICASSP*, pp 576-579, 1990.

6. S.R. Quackenbush, T.P. Barnwell III, and M.A. Clements, *Objective Measures of Speech Quality*, Englewood Cliffs, N-J: Prentice Hall, 1988.

7. 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*, Griffiss Air Force Base, Rome, NY, Apr 1975.