SINGLE-SENSOR SPEECH ENHANCEMENT USING A SOFT-DECISION/VARIABLE ATTENUATION ALGORITHM

E. Bryan George

Signal Processing Center of Technology
Lockheed Sanders Inc.
Nashua, NH 03061

ABSTRACT

This paper presents an algorithm for single-sensor enhancment of speech corrupted by additive random noise, based on soft-decision and statistical signal processing concepts and incorporating fully automatic noise estimation/tracking algorithms. The *Soft-Decision/Variable Attenuation* (SDVA) algorithm uses a compressive noise reduction model within the framework of short-time Fourier processing. The SDVA algorithm is fast, effective and robust, and has been applied in realistic RF and telephone environments.

1. INTRODUCTION

The most popular algorithms for single-sensor enhancement of speech degraded by additive noise are based on processing short-term spectral information derived from noisy speech [1, 2]. An important reason for this popularity is computational efficiency. Spectral subtraction, for instance, operates on the *discrete short-time Fourier transform* (DSTFT) magnitude of input speech [1]. Since the DSTFT is computed using the FFT, this algorithm is implementable using available DSP hardware. While efficient, these algorithms tend to produce speech with "musical" artifacts that are often more objectionable than the original noise. In addition, single-sensor speech enhancement algorithms proposed to date require speech activity detection and have difficulty dealing with the effects of nonstationary noise signals.

This paper presents a novel approach to the problem of speech enhancement by DSTFT processing. This new approach, referred to as the *soft-decision/variable attenuation* (SDVA) algorithm, combines the computational efficiency of other DSTFT-based enhancement algorithms with a compressive, rather than subtractive, noise suppression model. While achieving processing gains comparable to spectral subtraction, this model has the distinct advantage of producing residual noise without musical artifacts. In addition, the SDVA algorithm is formulated such that no speech activity detection is required, and has been designed to track noise level variations caused by automatic gain control (AGC) [3].

2. SPECTRAL SUBTRACTION ANALYSIS

Spectral subtraction operates on the assumption that if a noise signal $e[n]$ is added to a desired signal $x[n]$ to produce a noisy signal $s[n]$, and the added signals correspond to uncorrelated random variables, then the *power spectral density* (PSD) of $s[n]$, $\mathcal{P}_s(e^{j\omega})$, is given in terms of the PSD's of $x[n]$ and $e[n]$ by

$$\mathcal{P}_s(e^{j\omega}) = \mathcal{P}_x(e^{j\omega}) + \mathcal{P}_e(e^{j\omega}), \tag{1}$$

implying that if $\mathcal{P}_e(e^{j\omega})$ is known, then $\mathcal{P}_x(e^{j\omega})$ can be recovered by subtracting $\mathcal{P}_e(e^{j\omega})$ from $\mathcal{P}_s(e^{j\omega})$. The phase information associated with $x[n]$ is assumed to be relatively unaffected by the noise addition, and is thus obtained from the complex spectrum of $s[n]$.

Unfortunately, while the assumption of additive power spectra is true in a broad statistical sense, it fails to be accurate when applied to isolated windowed segments of $s[n]$ used to calculate the DSTFT. An example is illustrated in Figure 1. In this figure, the discrete Fourier transform (DFT) magnitude of one frame of Gaussian white noise is plotted, along with the noise PSD estimate $\hat{\mathcal{P}}_e[n,k]$ (dotted line) derived from the mean spectral magnitude of a number of frames. Clearly, the DFT magnitude of a single frame of noise bears little resemblance to the nominally smooth noise PSD. Because the spectral magnitude of a single noise frame varies both above and below the mean, the magnitude-squared spectrum after subtraction, $|\hat{X}[n,k]|^2$ will have negative values. Since there is no provision in the spectral subtraction model for negative values, such values are typically "rectified" by assigning them a zero or arbitrarily small value.

E. Bryan George is now with the Systems and Information Science Laboratory, Texas Instruments, Inc., Dallas TX 75265

0-7803-2431-5/95 $4.00 © 1995 IEEE

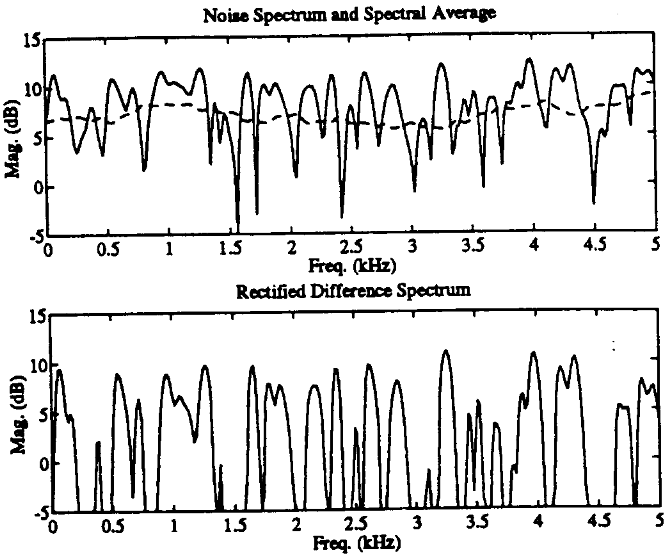

Figure 1: Illustration of spectral subtraction applied to an isolated noise frame spectrum.

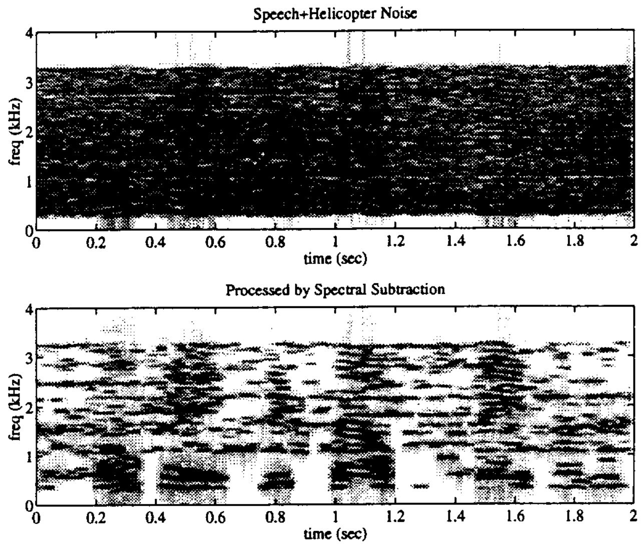

Figure 2: Narrowband spectrograms of noisy speech and speech processed using spectral subtraction.

The spectral artifacts resulting from this processing are shown in Figure 2 for the case of speech corrupted by additive noise in a helicopter environment. Although the result of spectral subtraction on isolated frames is an overall reduction in noise power, this figure demonstrates that the character of the noise residual is that of a collection of random tone bursts that account for the "musical noise" artifact noted in spectral subtraction (and in subtractive speech enhancement algorithms in general). Although additional processing steps may be taken to reduce the artifacts caused by spectral subtraction [1], such steps are only partially effective.

3. SDVA NOISE SUPPRESSION

The observations presented so far suggest that an effective alternative to spectral subtraction should meet two important criteria: First, it must be able to differentiate between any speech components present in the spectrum and the effects of additive noise, and be capable of suppressing noise while leaving speech information undistorted. Second, the algorithm must process noisy speech such that the character of any residual noise is no more objectionable than the original noise.

These criteria can be formalized by a statistical analysis of the short-time spectrum of noise signals. For instance, it can be shown [4] that the DSTFT $E[n, k]$ of zero-mean colored Gaussian noise is itself a complex Gaussian random variable for each frequency bin k. The significance of this statistical characterization of noise spectra becomes clear after considering the cumulative density function (CDF) of $|E[n, k]|^2$, which has the particularly simple form

$$C(\epsilon, k] = \text{Prob}(|E[n, k]|^2 < \epsilon) = 1 - e^{-\epsilon/\mathcal{P}_e[n,k]}. \quad (2)$$

A plot of this CDF is shown in Figure 3, with the variable $\epsilon/\mathcal{P}_e[n, k]$ plotted on a logarithmic scale. A qualitative interpretation of $C(\epsilon, k]$ is as the probability that noise spectral magnitudes are upper-bounded by ϵ. By this interpretation, Equation 3 indicates a dichotomy between magnitude values well below the mean, which are more likely associated with the noise process, and magnitude values well above the mean, which are very likely to correspond to the desired signal.

This observation suggests a simple noise suppression strategy: Given an estimate of the noise floor, $\hat{\mathcal{P}}_e[n, k]$, evaluate $C(\epsilon, k]$ for each value of the "noisy" DSTFT magnitude $|S[n, k]|^2$ to determine the probability that the spectral value is associated with the desired signal, *i.e.*

$$\text{P}_{\text{sig}}[n, k] = 1 - e^{-|S[n,k]|^2/\hat{\mathcal{P}}_e[n,k]}. \quad (3)$$

If the "signal probability" $\text{P}_{\text{sig}}[n, k]$ is low, attenuate the corresponding value of $|S[n, k]|$; otherwise, do not attenuate. Finally, combine the processed magnitude spectrum with the phase[1] of $S[n, k]$ and resynthesize the processed speech signal.

This "compressive" noise suppression strategy has a distinct advantage over subtractive approaches. Since

[1] As in spectral subtraction, the phase is unaltered.

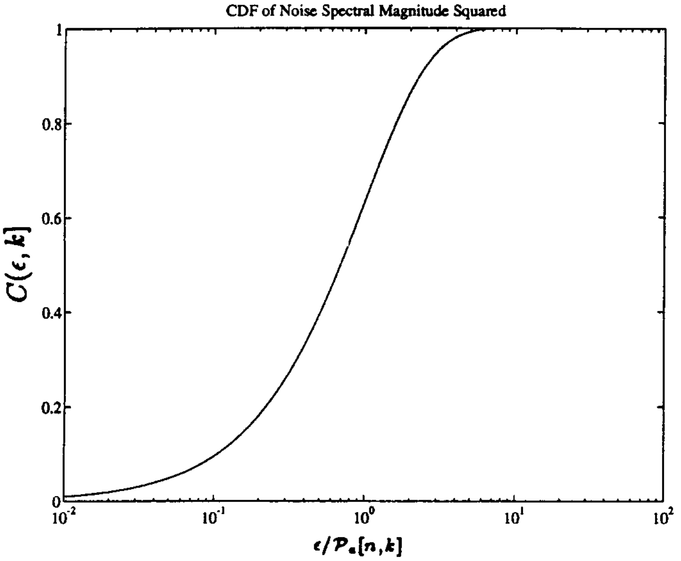

Figure 3: "Semilog" plot of noise spectrum CDF.

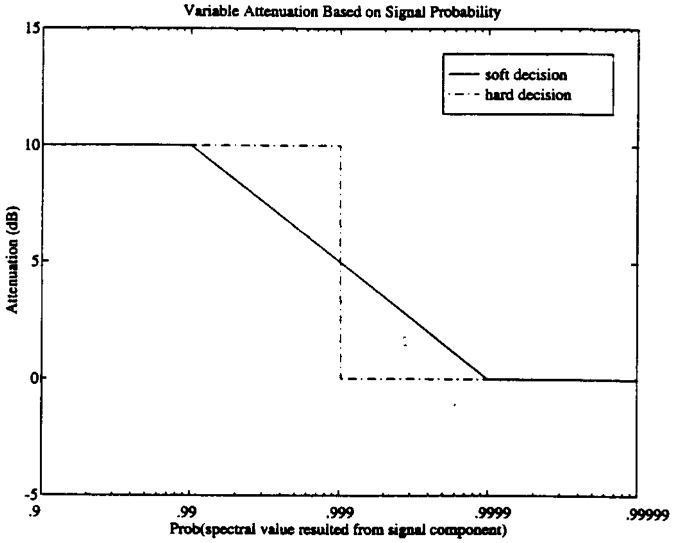

Figure 4: SDVA spectrum attenuation rule.

it does not introduce negative values, the processed spectrum does not require magnitude rectification, thus avoiding the musical artifact associated with subtractive enhancement algorithms. Unfortunately, the "hard decision" boundary described above can cause spectral discontinuities that are perceived as annoying "pops" in the processed speech. An alternative approach is to define a soft-decision boundary to match the soft signal/noise probability boundary indicated by Equation 2.

One such approach is illustrated in Figure 4. Based on empirical studies of noisy speech, it was found that a gating attenuation based on logarithmic values of $P_{\text{sig}}[n, k]$ is sufficient to suppress most noise, pass most speech information, and provide a smooth transition between the two cases. Equivalently (ref. Equation 2), this approach may be viewed as a linearly gated subtraction of log magnitude values. SDVA noise suppression is similar in form and motivation to approaches due to McAulay and Malpass [5] and Ephraim and Malah [6]. However, the SDVA algorithm has no subtractive component, and the evaluation of $\text{P}_{\text{sig}}[n, k]$ is less computationally intense than suppression rules requiring the evaluation of Bessel functions.

4. NOISE ESTIMATION/TRACKING

The effect of nonstationary noise represents a significant obstacle to practical speech enhancement. A particularly troublesome problem is that of fast AGC in pre-recorded speech. In fast AGC, the speech signal is compressed to have a constant envelope, meaning that the noise level of the signal varies inversely with the speech energy. Since many speech enhancement algorithms rely on noise floor estimates computed during non-speech activity and assume that the noise level remains relatively stationary, AGC has an obvious negative impact on their performance.

To deal with this problem, the SDVA algorithm has been developed with a noise tracking capability that operates even during speech activity. The basis of this noise tracking is the generation of a noise floor estimate in each frame of the DSTFT. This is done by estimating and removing the effects of speech components using an analysis-by-synthesis sinusoidal model [7, 3]. At this point a noise floor estimate exists for each frame that contains information about the spectral shape of the noise floor (which is relatively stable) and the noise level (which is not).

Unfortunately, the instantaneous noise floor is not completely reliable for use in the SDVA algorithm, since it may incorporate the effects of fricative sounds that are indistinguishable from noise in the short term spectrum. Therefore, the SDVA algorithm incorporates a "gain-normalized" long-term spectral averaging function to smooth noise floor estimates given by sinusoidal analysis, while at the same time preserving the effects of short-term noise level variations.

Figure 5 is a block diagram of the SDVA algorithm that illustrates the structure of the noise suppression and tracking algorithms described. The algorithm is structurally similar to spectral subtraction; the primary differences are the nature of noise suppression as discussed in Section 3 and the noise floor tracking algorithm described above. While noise floor tracking represents a computational overhead in comparison to spectral subtraction, the amount of computation required is comparable to speech activity detection, which is not required in the SDVA algorithm.

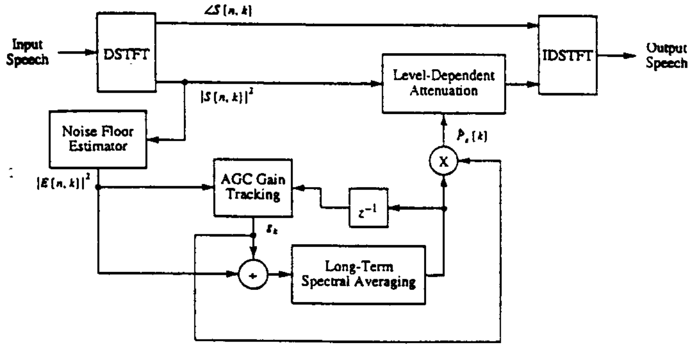

Figure 5: Block diagram of the SDVA algorithm.

Figure 6 demonstrates the effect of the SDVA algorithm on a segment of noisy speech transmitted in a helicopter environment over an HF channel, with fast AGC at the receiver. As seen in this figure (cf. Figure 2), the SDVA algorithm preserves the noisy character of the original distortion while achieving attenuation comparable to spectral subtraction. Furthermore, despite the presence of significant noise level variations, the noise tracking capability of the SDVA algorithm ensures that the speech is not inadvertantly suppressed by the algorithm despite the fact that its level is comparable to the noise floor.

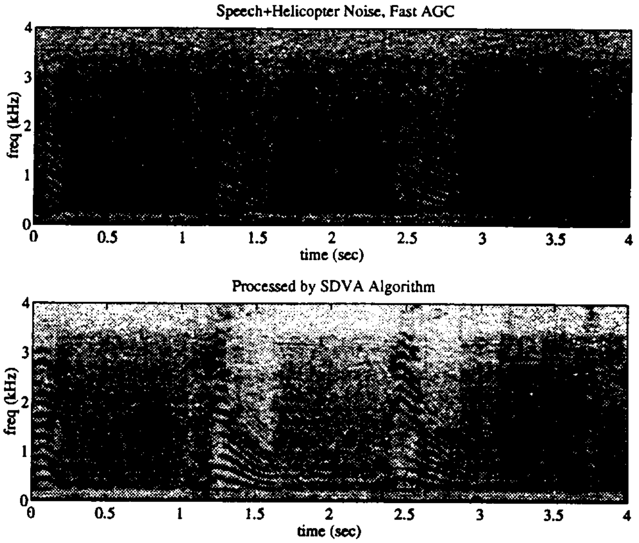

Figure 6: Narrowband spectrograms of noisy speech and speech processed using the SDVA algorithm.

5. CONCLUSION

This paper has introduced a novel approach to the problem of single-sensor speech enhancement known as the *Soft-Decision/Variable Attenuation* (SDVA) algorithm. As with classical speech enhancement algorithms, the SDVA algorithm is a straightforward spectral processing algorithm that is easily implemented in real-time using inexpensive, off-the-shelf DSP processors. In addition, the SDVA algorithm also eliminates the objectionable artifacts associated with classical approaches, and its ability to track and adapt to both rapid noise level variations and long-term changes in the spectral character of additive noise makes it particularly well-suited for operation in realistic communications environments.

6. REFERENCES

[1] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on Acoust., Speech and Signal Processing*, ASSP-27(2):113–120, April 1979.

[2] M. R. Weiss, E. Aschkenasy, and T. W. Parsons. Study and development of the INTEL technique for improving speech intelligibility. Final Rep. NSC-FR/4023, Nicolet Scientific Corp., December 1974.

[3] E. B. George. Improved speech enhancement using refined spectral subtraction and analysis-by-synthesis PSD estimation. In *Proc. Speech Research Symposium XIII*, pages 279–286, June 1993.

[4] S. M. Kay. *Fundamentals of statistical signal processing: Estimation theory.* Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

[5] R. J. McAulay and M.-L. Malpass. Speech Enhancement Using a Soft-Decision Noise Suppression Filter. *IEEE Trans. on Acoust., Speech and Signal Processing*, ASSP-28(2):137–144, April 1980.

[6] Y. Ephraim and D. Malah. Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator. *IEEE Trans. on Acoust., Speech and Signal Processing*, ASSP-32(6):1109–1121, December 1984.

[7] E. B. George and M. J. T. Smith. Analysis-by-Synthesis/Overlap-Add Sinusoidal Modeling Applied to the Analysis and Synthesis of Musical Tones. *J. Audio Eng. Soc.*, 40(6):497–516, June 1992.