Enhancement of Noise-Corrupted Speech Using Sinusoidal Analysis-Synthesis

by

Alan Seefeldt

B.S., General Engineering

Harvey Mudd College, 1995

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirement for the Degree of
Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 1997
[June 1997]

Signature of Author____________________
Department of Electrical Engineering and Computer Science
May 19, 1997

Certified by____________________
Alan V. Oppenheim
Ford Professor of Electrical Engineering
Thesis Supervisor

Accepted by____________________
Arthur C. Smith
Chairman, Committee on Graduate Students
Department of Electrical Engineering and Computer Science

Enhancement of Noise-Corrupted Speech
Using Sinusoidal-Analysis Synthesis

by

Alan Seefeldt

Submitted to the Department of Electrical Engineering and Computer Science
on May 19, 1997, in partial fulfillment of the
requirements for the Degree of Master of Science in
Electrical Engineering and Computer Science

Abstract

This thesis proposes a method for the single-sensor enhancement of speech that has been corrupted by additive broadband noise. The method is based on a technique known as Sinusoidal Analysis-Synthesis (SAS) and involves two steps. First, sinusoidal tracks relevant to the speech alone are extracted from the short-time spectrum of the corrupted speech. Secondly, extracted tracks are processed to reduce the perceptual level of any remaining noise.

In order to evaluate the potential of this enhancement technique, an upper bound on its performance is examined. The speech-only tracks are extracted from the corrupted speech by using tracks from the corresponding uncorrupted speech as a guide. These extracted tracks are then processed within the single-sensor framework. The resulting enhancement represents an upper limit on performance for whatever type of track processing is performed. Several types of track processing are explored, and the best-case results are compared to traditional spectral subtraction enhancement.

Thesis Supervisor: Alan V. Oppenheim
Title: Distinguished Professor of Electrical Engineering

Acknowledgments

Foremostly, I would like to thank my advisor, Al Oppenheim, for introducing me to the speech enhancement problem and encouraging me to explore the subject without any preconceptions. He provided excellent guidance and support as I discovered for myself the difficulty of this long-standing problem. I would also like to thank Tom Quatieri for the time he took to consult with me about the sinusoidal analysis-synthesis system and listen to my ideas about its use for enhancement. To my fellow members of the Digital Signal Processing Group, I offer many thanks for listening to, commenting on, and sometimes laughing at the processed speech resulting from the work in this thesis.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Speech Enhancement: Definition and Background

Speech enhancement can be defined generally as the perceptual improvement of speech that has undergone some type of distortion. In this thesis we examine the case where speech has been corrupted by additive broadband noise that is independent of the speech. In addition, we assume that enhancement is performed within the *single-sensor* framework, meaning that only a single recording of the corrupted speech can be used by any algorithm; additional simultaneous recordings of the corrupted speech, or the noise by itself, from other sensors are not available. Within these constraints, we focus on the goal of reducing the level of the corrupting noise while maintaining the integrity of the underlying speech such that the overall effect is perceptually more pleasing, though not necessarily more intelligible, to a listener. This is a very subjective goal, and there exists no simple mathematical objective function which can be minimized to achieve the desired enhancement. Therefore, listening comparisons must be made in order to judge the quality of any developed techniques.

The described single-sensor enhancement problem is long-standing and has many applications [1]. It is a very difficult task, and essentially all techniques are faced with the same dilemma. In the enhanced signal, the noise level may be perceptually lower, but unnatural sounding artifacts are introduced. Whether or not these artifacts are more or less objectionable than the original noise is usually unclear, and oftentimes listeners prefer the original noise, because it sounds more natural. The enhancement method examined in this thesis attempts an improvement over past work but does not avoid this dilemma.

Two common approaches to speech enhancement are *noise removal* and *reconstruction.* Noise removal focuses on the noise and tries to suppress it in the corrupted speech. An example is the method of spectral subtraction, which subtracts the expected value of the noise spectrum from the short-time spectral magnitude of the corrupted speech. No direct knowledge of the underlying speech is utilized. With reconstruction, on the other hand, parameters for a model of speech production are estimated from the corrupted speech, and an approximation of the uncorrupted speech is synthesized from these parameters. The quality of the enhanced speech is therefore limited by the parametric model. As an example of reconstruction, speech can be modeled as a time-varying all-pole filter that is excited by a periodic pulse train during voiced speech and by white noise during unvoiced speech. The filter coefficients, the voiced/unvoiced decision, and the pitch-period are then estimated from the corrupted speech.

1.2 Sinusoidal Analysis-Synthesis for Enhancement

The enhancement scheme investigated in this thesis can be considered a combination of the reconstruction and noise removal concepts. Sinusoidal analysis-synthesis (SAS), a technique developed by MacAulay and Quatieri [2], is the parametric representation of speech associated with the reconstruction component. SAS approximates speech as a sum of finite-duration, time-varying sinusoids, each referred to as a *track.* These tracks are formed by matching peaks frame-to-frame in the short-time discrete Fourier transform (STDFT) magnitude of a speech signal. For enhancement, the goal is to estimate from the STDFT of corrupted speech only those tracks that are relevant to the underlying speech, thereby eliminating the broadband noise lying in between these tracks. This speech-only track estimation problem is separated into two steps, the first of which can be considered reconstruction and the second noise removal. In the first step, speech-only tracks are *extracted* from the corrupted STDFT magnitude by matching appropriately selected peaks, and in the second step, the corrupted parameters of the resulting tracks are processed to reduce the level of any remaining noise.

The idea of using SAS for enhancement was first proposed in [3], but their approach is based on a modified SAS system with a very small parameter set that produces a more synthetic sounding speech estimate in comparison to the system discussed above. By using a larger parameter set, the SAS system being investigated in this thesis avoids restrictions on the potential quality of the enhanced speech. However, parameter estimation from the corrupted speech (*i.e.* track extraction) is made much more difficult. The system in [3] must also be trained on the uncorrupted speech of each individual user, whereas our approach assumes speaker independence.

1.3 Thesis Outline

This thesis focuses on determining and evaluating an upper bound for the performance of the proposed SAS enhancement technique. In the framework that is developed, corrupted speech is generated by adding white Gaussian noise to uncorrupted speech. The track extraction step is performed with guidance from the uncorrupted speech, thereby violating the single-sensor assumption. Track processing is then explored within the boundaries of single-sensor enhancement. Because the extraction step is, in a sense, optimized, the resulting speech estimate represents an upper limit on performance for whatever type of processing is performed. If this upper bound affords convincing enhancement, justification is provided for the development of a technique to perform track extraction without precise knowledge of the uncorrupted speech.

Chapters 2 and 3 provide background information on SAS and spectral subtraction. In Chapter 4, SAS enhancement is proposed, and development of the evaluation framework is initiated by defining a procedure for extracting speech-only tracks. Chapter 5 explores techniques for track amplitude processing, and in Chapter 6, the best-case SAS enhancement results are compared to spectral subtraction through a listening test. Lastly, Chapter 7 presents a summary and offers directions for future work.

Chapter 2

Sinusoidal Analysis-Synthesis

Sinusoidal analysis-synthesis (SAS) is the basis of the speech enhancement technique explored in this thesis. The SAS model approximates speech as a finite sum of sinusoids, each with time-varying amplitude and phase. The goal for enhancement is to estimate only speech-relevant sinusoids from corrupted speech. SAS processing, as developed by MacAulay and Quatieri in [2], generates these sinusoids from uncorrupted speech and serves as a starting point for our enhancement system. This chapter describes the procedure and prepares us for the application of SAS to speech enhancement.

2.1 Sinusoidal Analysis-Synthesis Framework

Under the SAS framework, a digitized speech signal, *s*[*n*], is approximated as

$$\tilde{s}[n] = \sum_{k=1}^{Q} a_k(n)\cos[\theta_k(n)]\ , \quad \omega_k(n) = \dot{\theta}_k(n) \tag{2.1}$$

where $a_k(n)$, $\theta_k(n)$, and $\omega_k(n)$ are the time-varying amplitudes, phases, and instantaneous frequencies of each sinusoid. These three sets of functions are generated from the short-time discrete Fourier transform (STDFT) of *s*[*n*] given by

$$S[f,d] = \sum_{n=0}^{M-1} w[n]s[n+Td]e^{-j\frac{2\pi fn}{N}}\ , \quad 0 \le f < N/2, \tag{2.2}$$

where *w*[*n*] is a windowing function, *M* is the frame length, *T* is the frame interval, *N* is the DFT length, *d* is the frame number, and *f* is the DFT bin number. In the SAS system, peaks from the STDFT magnitude are selected and then matched frame-to-frame to form what are called *tracks*. The amplitudes, phases, and frequencies associated with the peaks of each track serve as sample points of the functions $a_k(n)$, $\theta_k(n)$, and $\omega_k(n)$. In order to completely specify these functions, amplitude and phase must be interpolated along the tracks, after which the speech estimate is calculated according to (2.1).

The formation of tracks begins by selecting peaks in the STDFT magnitude of the speech signal. Specifically, a peak is found at bin value p if the following condition holds:

$$|S[p,d]| > |S[p-1,d]| \text{ and } |S[p,d]| > |S[p+1,d]| . \tag{2.3}$$

Let p be the bin value of the kth peak in the STDFT magnitude at frame d. Then the frequency, amplitude, and phase associated with the kth peak are

$$\Omega_{k,d} = \frac{2\pi p}{N}, \tag{2.4a}$$

$$A_{k,d} = |S[p,d]|, \tag{2.4b}$$

$$\text{and } \Theta_{k,d} = ARG\{S[p,d]\} + \frac{M\Omega_{k,d}}{2} , \tag{2.4c}$$

where $\Theta_{k,d}$ is referenced to the center of the STDFT frame. With all peaks selected at each frame, tracks are formed by pairing peaks between successive frames using a nearest-neighbor frequency matching algorithm. For each peak k in frame d the algorithm tries to find a matching peak l in frame d+1 such that the quantity $|\Omega_{l,d+1} - \Omega_{k,d}|$ is minimized. In addition, $|\Omega_{l,d+1} - \Omega_{k,d}|$ is required to be less than a pre-defined matching tolerance Δ. If no peak l is found to satisfy this constraint, peak k is left unmatched. Following this procedure, two or more peaks in frame d may be paired to the same peak l. In such a case, the peak in frame d whose frequency has the smallest absolute difference with $\Omega_{l,d+1}$ remains paired. Peak l is then removed from the list of possible matching candidates, and the process starts again for the other peaks in frame d that had been paired to peak l. The result of the algorithm is a set of tracks, each composed of a series of matched peaks spanning one or more frames. The beginning/end of track is defined by a peak in a given frame that is not matched to a peak in the previous/successive frame. The structure of the tracks is illustrated in Figure 2.1, where each vertical line represents the STDFT at a given frame, and the X's represent peaks in the STDFT magnitude. The dashed lines connecting the X's represent the tracks that are formed.

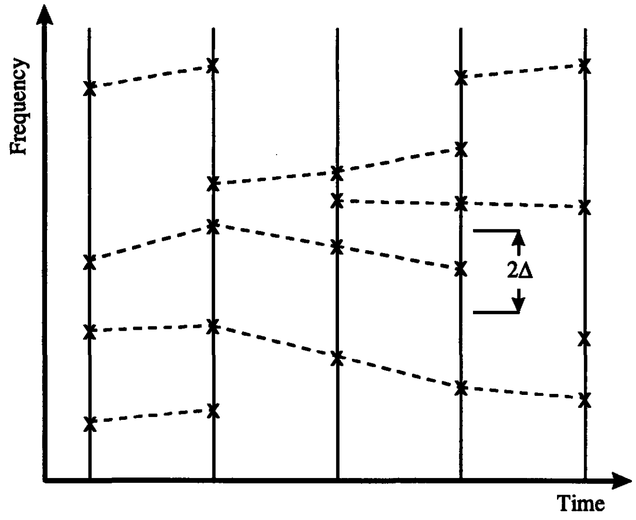


Fig. 2.1 Track formation by matching peaks frame-to-frame in the STDFT magnitude (note the matching tolerance Δ)

Interpolation of amplitude and phase between the peaks of each track is the next step in SAS. The process is performed piece-wise from frame-to-frame. To explain this in more detail, we rewrite (2.1) as

$$\tilde{s}[n+Td]=\frac{1}{N}\sum_{k=1}^{Q(d)} a_{k,d}(n)\cos[\theta_{k,d}(n)]\ ,\quad \omega_{k,d}(n)=\dot{\theta}_{k,d}(n),\quad 0\le n<T, \tag{2.5}$$

where $a_{k,d}(n)$, $\theta_{k,d}(n)$, and $\omega_{k,d}(n)$ are functions existing over only a single frame interval, and where the number of sinusoids, $Q(d)$, varies with the frame number. The functions $a_{k,d}(n)$ and $\theta_{k,d}(n)$ describe the interpolation of amplitude and phase between a peak k in frame d that is paired to a peak l in frame d+1.

Simple linear interpolation is adequate for the amplitude function, in which case it is given by

$$a_{k,d}(t)=A_{k,d}+\left(\frac{A_{l,d+1}-A_{k,d}}{T}\right)t,\qquad 0\le t<T, \tag{2.6}$$

where the time variable t is used instead of n to indicate that the function is continuous. By itself, (2.6) results in a track whose amplitude begins and ends abruptly. To prevent this, the amplitude is additionally required to ramp up from zero at the beginning and down to zero at the end, a process that is referred to as the "birth" and "death" of a track. More specifically, suppose that peak k in frame d represents the beginning of a track. Then, over the frame interval preceding d, amplitude is interpolated from zero up to $A_{k,d}$ with frequency held constant at $\Omega_{k,d}$. An analogous procedure is followed at the end of a track.

Generating the phase functions, $\theta_{k,d}(t)$, is more complicated, because the phase of a peak is not unique, and because the proper relationship between frequency and phase must be maintained. The end-points of the phase function are constrained to be

$$\theta_{k,d}(0) = \Theta_{k,d} \tag{2.7a}$$

$$\text{and } \theta_{k,d}(T) = \Theta_{l,d+1} + 2\pi P, \tag{2.7b}$$

where P is an integer, and the term $2\pi P$ is necessary because the $\Theta_{k,d}$ are measured modulo-2π. Since the time derivative of the phase function is the instantaneous frequency, two more end-point constraints are introduced:

$$\dot{\theta}_{k,d}(0) = \Omega_{k,d} \tag{2.7c}$$

$$\text{and } \dot{\theta}_{k,d}(T) = \Omega_{l,d+1}. \tag{2.7d}$$

The four constraints listed in (2.7a-d) can all be satisfied if the phase function is represented with a cubic polynomial:

$$\theta_{k,d}(t) = \alpha + \beta t + \chi t^2 + \delta t^3. \tag{2.8}$$

Combining (2.7a-d) and (2.8) yields the polynomial coefficients:

$$\alpha = \Theta_{k,d}, \tag{2.9a}$$

$$\beta = \Omega_{k,d}, \tag{2.9b}$$

$$\chi = \frac{3}{T^2}(\Theta_{l,d+1} - \Theta_{k,d} - \Omega_{k,d}T + 2\pi P) - \frac{1}{T}(\Omega_{l,d+1} - \Omega_{k,d}), \tag{2.9c}$$

$$\text{and } \delta = \frac{-2}{T^3}(\Theta_{l,d+1} - \Theta_{k,d} - \Omega_{k,d}T + 2\pi P) + \frac{1}{T^2}(\Omega_{l,d+1} - \Omega_{k,d}). \tag{2.9d}$$

Both χ and δ are dependent on P, and therefore different values of P will yield different phase functions. Intuitively, P should be chosen so that $\theta_{k,d}(t)$ is "maximally smooth," a notion which is quantified by minimizing the variation of the phase function over the frame interval. Specifically, the function

$$f(P) = \int_0^R [\ddot{\theta}(t;P)]^2 \, dt \tag{2.10}$$

is chosen to represent smoothness. Minimizing $f(P)$ with respect to P yields the optimum value

$$P_{opt} = \text{round}\left\{\frac{1}{2\pi}[(\Theta_{k,d} + \Omega_{k,d}T - \Theta_{l,d+1}) + \frac{T}{2}(\Omega_{l,d+1} - \Omega_{k,d})]\right\}, \tag{2.11}$$

where round$\{x\}$ chooses the integer closest to x. The expression for P_{opt} is substituted into (2.10c-d), and $\theta_{k,d}(t)$ is then completely specified.

Once $a_{k,d}(t)$ and $\theta_{k,d}(t)$ are generated for all peaks in all frames, the speech estimate can be computed according to (2.5). While (2.5) indicates that the $a_{k,d}(t)$ and $\theta_{k,d}(t)$ are sampled at integer time values, these functions can be re-sampled using any time interval since they are continuous. As one consequence, arbitrary time expansion and contraction of the original speech signal, without a change in pitch, is easily achieved under the SAS framework. In fact, numerous other transformations, such as pitch shifting and tone shaping, can be implemented in the context of SAS [4]. These applications are not relevant to the enhancement work at hand but demonstrate a broader application of SAS in speech processing.

2.2 SAS Implementation

In selecting the parameters for the SAS process, the goal is to create a speech estimate that sounds as close to the original speech as possible. In our implementation we focus specifically on digital speech signals sampled at a rate of 10kHz. Our first consideration is the set of parameters used in the STDFT: the windowing function $w[n]$, the frame length M, the frame interval T, and the DFT length N. The Hamming window provides adequate suppression of side-lobe leakage between peaks in the STDFT magnitude. Given the relationship between window length and main-lobe width for the Hamming window, it is shown in [2] that the frame length must be at least 2.5 times the voiced pitch period in order to resolve frequencies separated by the pitch fundamental of a voiced speech segment. Assuming that the lowest pitch fundamental to be encountered is 100Hz, this translates to a frame length of 25ms. A frame interval of 10ms provides adequate time-resolution to follow transitions within the speech, and with a 10kHz sampling rate, we then have M=250 and T=100. The DFT length, N, must be large enough so that peaks in the underlying discrete-time Fourier Transform (DTFT) are well represented by the DFT samples. The value N=1024 works well, providing a bin-width of 9.77Hz.

Our next consideration is the formation of SAS tracks. In the peak matching algorithm, the matching tolerance, Δ, should reflect the largest frame-to-frame variation expected in the fundamental pitch of a voiced speech segment; the algorithm should be able to form a single track that follows the pitch fundamental over time. On the other hand, the matching tolerance should not be so large that tracks vary wildly in frequency during unvoiced sections of speech for which the frequencies of the STDFT peaks exhibit little ordered structure. A value of Δ=100Hz is adequate in meeting both these requirements.

The last parameters to be examined are those associated with selecting peaks in the STDFT. If every peak generated according to condition (2.4) is used to create tracks, the resulting speech estimate sounds almost perceptually indistinguishable from the original, but many more peaks than necessary are being used to achieve this quality. Because of low-level recording noise, areas of silence in the speech signal are not precisely zero, and therefore very small amplitude peaks are found in the STDFT of these regions. Consequently, it is more appropriate to impose a minimum peak amplitude that is referenced to the average energy of the entire speech signal. Toward this end we define a peak-to-signal-ratio (PSR), and the minimum peak amplitude is derived from this ratio. Let σ_s^2 denote the sample variance of the speech signal given by

$$\sigma_s^2 = \frac{1}{L}\left[\left(\sum_{n=0}^{L-1} s^2[n]\right) - \left(\sum_{n=0}^{L-1} s[n]\right)\right], \tag{2.12}$$

where L is the total length of the signal. Then consider a white zero-mean Gaussian random process with variance σ_s^2. The expected value of the STDFT squared magnitude of this process is

$$\psi_s = \sigma_s^2 \sum_{n=0}^{M} w^2[n] \ , \tag{2.13}$$

independent of time and frequency. The minimum peak amplitude is defined to be

$$A_{\min} = \sqrt{\psi_s \left(10^{\frac{PSR}{10}} \right)}, \tag{2.14}$$

where the PSR is in decibels. In addition to (2.3), we then require that $|S[p,d]| > A_{\min}$. Empirically, -30dB is approximately the largest PSR threshold for which high quality synthetic speech is still obtained. At this value, peaks within the silence regions are not selected, but regions of the synthetic signal containing speech are perceptually indistinguishable from those when all possible peaks are selected. As the PSR is increased beyond -30dB, some peaks relevant to the speech are ignored, and artifacts are heard in the synthetic speech.

To illustrate the SAS implementation, Figure 2.2 depicts spectrograms for two speech utterances, one female and one male. The spectrogram is a plot of $|S[p,d]|$ in decibels, with time (frame number) on the horizontal axis and frequency (bin number) on the vertical axis. The STDFT magnitude is represented by gray-scale, with a lighter shade corresponding to a larger value. Below each spectrogram image is a spectrogram-like plot displaying the frequency contours of the corresponding SAS tracks. During voiced sections of the speech, the track lines follow the smoothly-varying, harmonic ridges of the spectrogram very well, and during unvoiced sections, the track lines are shorter and more erratic in frequency since the underlying spectrogram is noise-like. For both utterances, the speech estimate produced from the depicted tracks is nearly perceptually indistinguishable from the original, aside from the absence of low-level recording noise during regions of non-speech activity.

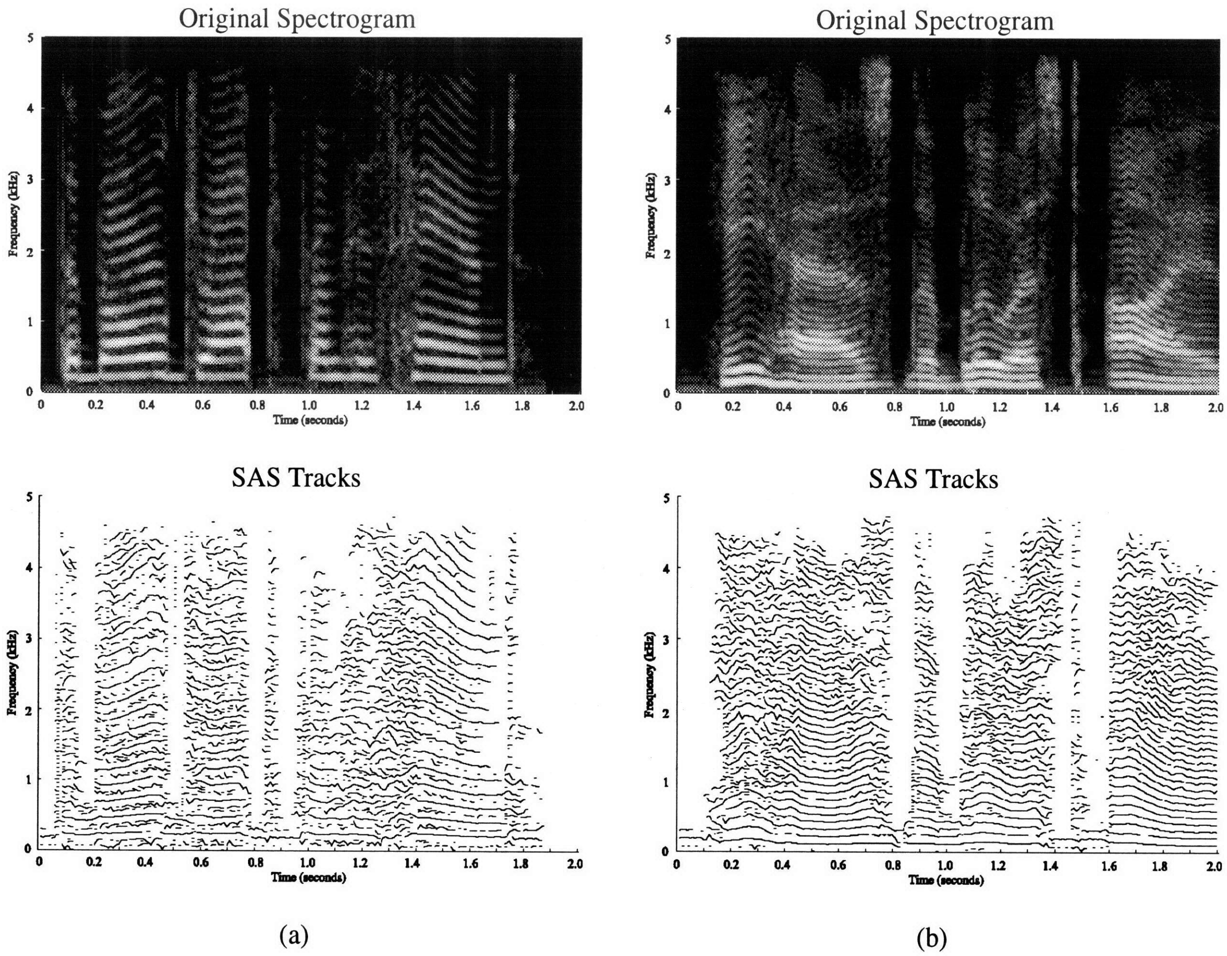


Fig. 2.2 Spectrograms and SAS tracks: (a) female speech, "The bowl dropped from his hand." and (b) male speech, "He has the bluest eyes."

Chapter 3

Spectral Subtraction

Spectral subtraction is a popular single-sensor enhancement technique that pre-dates SAS. It focuses on enhancing only the short-time spectral amplitude of corrupted speech and defines a class of estimators for doing so. In this chapter, background on the technique is presented in anticipation of its later use for the development and evaluation of SAS enhancement. Various forms of spectral subtraction, found mainly in [1], are reviewed and compared, and then a well known implementation is examined.

3.1 Formulations of Spectral Subtraction

For the enhancement problem being considered, let $s[n]$, $z[n]$, and $y[n]$ be the uncorrupted speech signal, the additive noise, and the corrupted speech signal, respectively, and let $S[f,d]$, $Y[f,d]$, and $Z[f,d]$ be their corresponding STDFT's. We have

$$y[n]=s[n]+z[n], \tag{3.1}$$

from which we then know

$$|Y[f,d]|^2=|S[f,d]|^2+|Z[f,d]|^2+S[f,d]Z^*[f,d]+S^*[f,d]Z[f,d]. \tag{3.2}$$

Assuming that the noise is zero-mean and uncorrelated with the speech, we can take the expected value of (3.2) with $S[f,d]$ treated as a known value and obtain

$$E\left\{|Y[f,d]|^2\right\}=|S[f,d]|^2+E\left\{|Z[f,d]|^2\right\}, \tag{3.3}$$

since the expected values of $Z[f,d]$ and $Z^*[f,d]$ are both zero. For the single-sensor enhancement problem, we can compute $|Y[f,d]|^2$, and $E\{|Z[f,d]|^2\}$ may be known or can be estimated by averaging adjacent frames of $|Y[f,d]|^2$ during areas of non-speech activity. With these two available quantities, an estimate of $|S[f,d]|^2$ suggested by (3.3) is

$$\left|\hat{S}[f,d]\right|^2 = \left|Y[f,d]\right|^2 - E\left\{\left|Z[f,d]\right|^2\right\}. \tag{3.4}$$

Estimation of the short-time spectral amplitude of the uncorrupted speech using (3.4) is referred to specifically as *power spectral subtraction.* A more general formulation estimates $|S[f,d]|$ as

$$\left|\hat{S}[f,d]\right|^a = \left|Y[f,d]\right|^a - kE\left\{\left|Z[f,d]\right|^a\right\}, \tag{3.5}$$

where the parameters a and k can be varied to achieve different attenuation characteristics. Use of a=1 and k=1 has received considerable attention, and with these values we refer to (3.5) as *amplitude spectral subtraction.* The estimate $|\hat{S}[f,d]|^a$ in (3.5) is not guaranteed to be positive, and if the right hand side does become negative, setting $|\hat{S}[f,d]|^a$ equal to zero is the most widely accepted practice.

In order to construct an estimate of $s[n]$ using any form of spectral subtraction, an estimate of the complex-valued function $S[f,d]$, rather than its magnitude, is required. The most common solution is to approximate the phase of $S[f,d]$ as that of $Y[f,d]$. Generating $|\hat{S}[f,d]|$ from (3.5), we have

$$\hat{S}[f,d] = \left|\hat{S}[f,d]\right| \exp\left(j\angle Y[f,d]\right). \tag{3.6}$$

The entire spectral subtraction process can then be written as a time-varying, zero-phase frequency response $S[f,d]$ applied to $Y[f,d]$:

$$H[f,d] = \frac{\hat{S}[f,d]}{Y[f,d]} = \left(\frac{\left|Y[f,d]\right|^a - kE\left\{\left|Z[f,d]\right|^a\right\}}{\left|Y[f,d]\right|^a} \right)^{1/a} \tag{3.7}$$

With this frequency response representation of spectral subtraction, a connection to Wiener filtering can be drawn, and as a result, the formulation in (3.5) is both reinforced and expanded.

If, for $y[n]$=$s[n]$+$z[n]$, both $s[n]$ and $z[n]$ can be represented by uncorrelated stationary random processes, then the minimum mean-square linear estimator of $s[n]$ given $y[n]$ is the non-causal Wiener filter with frequency response

$$H(\omega) = \frac{P_s(\omega)}{P_s(\omega) + P_z(\omega)}, \tag{3.8}$$

where $P_s(\omega)$ and $P_z(\omega)$ are the power spectral densities of $s[n]$ and $z[n]$. The speech signal is not stationary, so one possible approximation to (3.8) is a time-varying Wiener filter that is applied to $Y[f,d]$:

$$H[f,d] = \frac{E\left\{\left|S[f,d]\right|^2\right\}}{E\left\{\left|S[f,d]\right|^2\right\} + E\left\{\left|Z[f,d]\right|^2\right\}}. \tag{3.9}$$

If $E\{|S[f,d]|^2\}$ is estimated as the right-hand side of (3.4), we obtain *Wiener spectral subtraction*:

$$H[f,d] = \frac{\left|Y[f,d]\right|^2 - E\left\{\left|Z[f,d]\right|^2\right\}}{\left|Y[f,d]\right|^2}. \tag{3.10}$$

Comparing to (3.7), we see that (3.10) is just the square of the suppression filter for power spectral subtraction. A generalized form of (3.9), known as a parametric Wiener filter, can also be considered:

$$H[f,d] = \left[\frac{E\left\{\left|S[f,d]\right|^2\right\}}{E\left\{\left|S[f,d]\right|^2\right\} + \alpha E\left\{\left|Z[f,d]\right|^2\right\}}\right]^{\beta}, \tag{3.11}$$

where α and β are varied to obtain different characteristics. Using the filter $H[f,d]$, our estimate of $S[f,d]$ is given by

$$\hat{S}[f,d] = H[f,d]Y[f,d]. \tag{3.12}$$

If we estimate $E\{|S[f,d]|^2\}$ as $\hat{S}[f,d]$ and combine (3.11) with (3.12), we obtain the implicit relationship

$$\left|\hat{S}[f,d]\right| = \left[\frac{\left|\hat{S}[f,d]\right|^2}{\left|\hat{S}[f,d]\right|^2 + \alpha E\left\{\left|Z[f,d]\right|^2\right\}}\right]^{\beta} \left|Y[f,d]\right|, \tag{3.13}$$

which can be solved for $|\hat{S}[f,d]|$. If α=1 and β=1/2, the solution is exactly the method of power spectral subtraction in (3.4), and for α=1/4 and β=1 a solution to (3.13) is

$$\left|\hat{S}[f,d]\right| = \tfrac{1}{2}\left|Y[f,d]\right| + \tfrac{1}{2}\left(\left|Y[f,d]\right|^2 - E\left\{\left|Z[f,d]\right|^2\right\}\right)^{1/2}. \tag{3.14}$$

Coincidentally, this same estimator is found by solving for the maximum likelihood estimate of $S[f,d]$, assuming that the noise is Gaussian at each frequency [5]. We therefore refer to (3.14) as *maximum likelihood spectral subtraction.*

3.2 Comparison of Formulations

If the corrupting noise, $z[n]$, is zero-mean white Gaussian with variance σ_n^2, a quantitative comparison can be made between the four specific forms of spectral subtraction that have been discussed: power, amplitude, Wiener, and maximum likelihood. All of these formulations, except amplitude, involve the term $E\{|Z[f,d]|^2\}$ given by

$$\psi_n = \sigma_n^2 \sum_{n=0}^{M} w^2[n] \ , \tag{3.15}$$

independent of time and frequency. The quantity $E\{|S[f,d]|\}$, necessary for amplitude spectral subtraction, can be computed by considering the probability density of $|Z[f,d]|$, which is Rayleigh [5]:

$$p(|Z[f,d]|) = \frac{2|Z[f,d]|}{\psi_n} e^{\frac{-|Z[f,d]|^2}{\psi_n}} \ , \ |Z[f,d]| \geq 0 \, , \tag{3.16}$$

from which the expected value of $|Z[f,d]|$ is found to be

$$E\{|Z[f,d]|\} = \frac{\sqrt{\pi\psi_n}}{2} \, . \tag{3.17}$$

Now, for any f and d, we define

$$\rho = \frac{|Y[f,d]|^2}{\psi_n} \, , \tag{3.18}$$

which can be considered a (speech-plus-noise)-to-noise ratio. The spectral subtraction estimate of $S[f,d]$ is then given by

$$\hat{S}[f,d] = H(\rho)Y[f,d] \ , \tag{3.19}$$

where $H(\rho)$ is a gain function whose form is dependent on the specific type of spectral subtraction. Through manipulation of (3.7), (3.10), and (3.13) we have:

$$\textit{Power}:\ H(\rho)=\left(\frac{\rho-1}{\rho}\right)^{1/2} \tag{3.20a}$$

$$\textit{Amplitude}:\ H(\rho)=\frac{\sqrt{\rho}-\sqrt{\pi}/2}{\sqrt{\rho}} \tag{3.20b}$$

$$\textit{Wiener}:\ H(\rho)=\frac{\rho-1}{\rho} \tag{3.20c}$$

$$\textit{Maximum Likelihood}:\ H(\rho)=\frac{1}{2}+\frac{1}{2}\left(\frac{\rho-1}{\rho}\right)^{1/2} \tag{3.20d}$$

Figure 3.1 shows a plot of $H(\rho)$ versus ρ for each type of spectral subtraction, and in all cases, the attenuation of $H(\rho)$ increases as ρ decreases. This can be viewed as an attempt to increase the overall SNR of the speech estimate by attenuating spectral areas with a relatively low SNR while leaving those with a relatively high SNR unaffected. On the whole, maximum likelihood attenuates the least, while amplitude spectral subtraction attenuates the most.

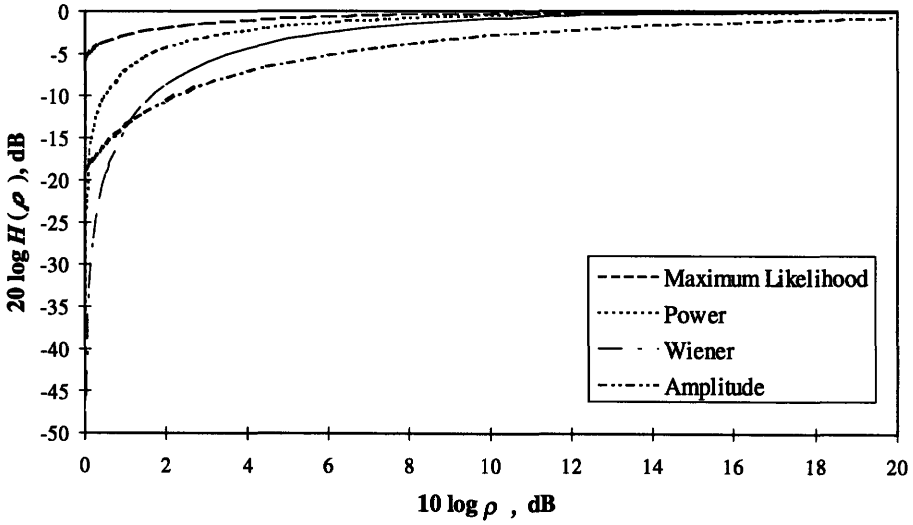


Fig 3.1 Gain function for various forms of spectral subtraction

Further information is obtained by examining the expected value of $|\hat{S}[f,d]|$ given $|S[f,d]|$. To simplify the notation, let A=$|S[f,d]|$, B=$|Y[f,d]|$, and C=$|\hat{S}[f,d]|$. The probability density of B given A is Rician [5]:

$$p(B|A)=\frac{2B}{\psi_n}e^{-\frac{A^2+B^2}{\psi_n}}I_0\left(\frac{2AB}{\psi_n}\right),\ B>0, \tag{3.21}$$

where $I_0(x)$ is the modified Bessel function of order zero. From this distribution, $E\{B|A\}$ can be computed numerically, and by combining (3.19), (3.20a-d), and (3.21), $E\{C|A\}$ can be computed for each type of spectral subtraction. Figure 3.2 is a plot of these curves for $\psi_n=1$. As anticipated, $E\{B|A\}$ is always larger than A, but as A increases, $E\{B|A\}$ asymptotically approaches A. We also note that all the spectral subtraction estimates of A are biased, *i.e.* $E\{C|A\} \neq A$. With maximum likelihood, a relatively large upward bias is observed for smaller values of A, while the bias is relatively small for larger values of A. With power, Wiener, and amplitude, a much smaller upward bias is achieved for smaller values of A at the expense of a large downward bias for larger values of A.

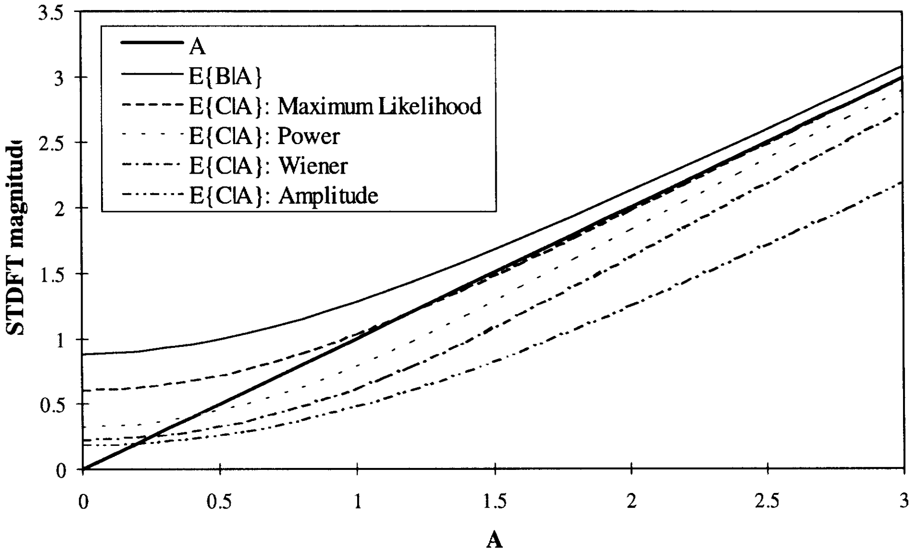


Fig 3.2 $E\{C|A\}$ vs. A for various forms of spectral subtraction

3.3 Implementation of Spectral Subtraction

In this section we discuss the specific implementation of spectral subtraction which will eventually be used as a basis for comparison with the SAS enhancement scheme developed in this thesis. The implementation is documented in detail by Boll [6], and it incorporates several steps beyond basic spectral subtraction. In developing this implementation, we consider speech corrupted by additive white Gaussian noise to achieve SNR's from 18dB down to 0dB, and we assume that the variance of the noise is known so that $E\{|Z[f,d]|^2\}$ can be computed exactly according to (3.15).

As a first consideration, $\hat{S}[f,d]$ must be transformed to a time-domain signal in order to produce an estimate of the uncorrupted speech. To make this process simple, several modifications are made to the STDFT parameters that were selected in Section 2.2 for SAS. The frame length remains 25ms, but the frame rate is increased from 10ms to 12.5ms so that adjacent frames overlap by exactly one-half. Also, a rectangular window is used instead of a Hamming window. To transform the STDFT to a time-domain signal, the inverse DFT of each frame is calculated, multiplied by a Hanning window, and overlapped and added with the adjacent frames. Since a series of Hanning windows adds to one when the windows are overlapped by half their length, the original time domain signal is reconstructed perfectly if the STDFT is not modified.

In Boll's system, $\hat{S}[f,d]$ is calculated by applying amplitude spectral subtraction to $Y[f,d]$. If $\hat{S}[f,d]$ is transformed to the time domain after only this first step, the perceptual level of the noise is noticeably reduced, and the underlying speech remains intact. The noise, however, no longer sounds white. Instead, the residual manifests itself as randomly spaced spikes in the STDFT magnitude and sounds like the sum of tone generators with random fundamental frequencies turned on and off at the frame rate – some have described it as "musical noise" or "doodly-doos." To further reduce the level of this noise residual, Boll applies a filter along each bin of $|\hat{S}[f,d]|$ that exploits the frame-to-frame randomness of the offending spikes. Operation of the filter can be stated as follows:

If $|\hat{S}[f,d]| < \kappa$, then

$$\bar{S}[f,d] = \left| \min_{d-1 \le c \le d+1} \left\{ \left| \hat{S}[f,c] \right| \right\} \right| \exp\left(j \angle \hat{S}[f,d] \right),$$

otherwise

$$\bar{S}[f,d] = \hat{S}[f,d],$$

where κ is some threshold and $\bar{S}[f,d]$ is the new estimate of $S[f,d]$. This adaptive order-statistic filter targets only low-energy, randomly time-varying spectral components for attenuation and leaves high-energy components unchanged. Empirically, a threshold of $\kappa = 1.5\psi_n$

provides substantial reduction of the noise residual at the expense of slightly muffled sounding speech.

As a final processing step, the speech estimate is attenuated in areas of non-speech activity. This has the effect of "evening out" the perceptual level of the noise residual, since the residual is masked somewhat by speech activity. The non-speech areas are detected by calculating the value

$$Q(d) = 20\log_{10}\left[\frac{1}{N/2}\sum_{f=0}^{N/2-1}\left|\bar{S}[f,d]\right| \Big/ E\left\{\left|Z[f,d]\right|\right\}\right] \qquad (3.22)$$

at each frame. If $Q(d)$ is less than some threshold, then frame d is classified as non-speech, and $|\bar{S}[f,d]|$ is attenuated by some constant factor for every f. Empirically, a threshold of -18dB proves satisfactory for detecting the non-speech frames, and attenuating these frames by -25dB results in a residual level that sounds perceptually in balance with its level during speech activity. Having the attenuation turn on and off instantly, however, makes the speech estimate sound "jerky," but forcing the attenuation level to fade in and out over three frames solves this problem.

Chapter 4

SAS for Speech Enhancement

In Chapter 2 we saw that SAS creates its speech estimate from a subset of the STDFT values of a speech signal, completely ignoring spectral areas surrounding the tracks that are generated. Thus, SAS inherently lends itself to speech compression, and in practice it has proven very effective for low-rate speech coding [7]. Viewed differently, this same property suggests a method for enhancing speech corrupted by broadband noise. If tracks associated only with the original speech can be estimated from the STDFT of the corrupted speech, then the spectral components of the noise lying in between these tracks will be eliminated in the synthesized signal. In our enhancement system we separate this speech-only track extraction problem into two parts. First, tracks are *extracted* from the corrupted speech by appropriately selecting and matching peaks from its STDFT. Then, the corrupted frequencies, amplitudes, and phases associated with the peaks of these extracted tracks are processed to reduce any remaining noise. After this, interpolation is performed along the speech-only tracks, and the resulting sinusoids are summed to produce an estimate of the uncorrupted speech.

This thesis attempts to evaluate the performance of the SAS enhancement system when track extraction is performed by employing specific knowledge of tracks from the uncorrupted speech. If tracks extracted in this manner can be processed within the single-sensor framework to yield significant enhancement, we conclude that the use of SAS for speech enhancement has definite potential, and justification is provided for developing a method to blindly extract speech-only tracks. If convincing enhancement is not obtained, we know that a blind track extraction procedure would perform no better, and time has not been wasted developing such a technique in vain.

Figure 2.3 depicts a block diagram of the overall evaluation framework for the thesis. An uncorrupted speech signal is deliberately corrupted with additive white Gaussian noise, after which the STDFT's of both the uncorrupted and corrupted signals are computed. SAS tracks are generated in the uncorrupted STDFT according to the procedure outlined in Chapter 2, and these tracks are then used to extract speech-only tracks from the corrupted STDFT. In this chapter, the formation and synthesis of the extracted tracks is examined. Chapter 5 then explores various types of track processing within the limits of single-sensor enhancement. In Chapter 6, the best results obtained from synthesizing the speech-only track estimates are compared to results obtained through spectral subtraction. If the development of blind track extraction is to be justified, SAS enhancement based on tracks extracted with knowledge of the uncorrupted speech should provide a *significant* improvement over spectral subtraction. If it does not, then deterioration of the SAS enhanced speech resulting from blind track extraction would likely result in an enhancement that is worse than spectral subtraction.

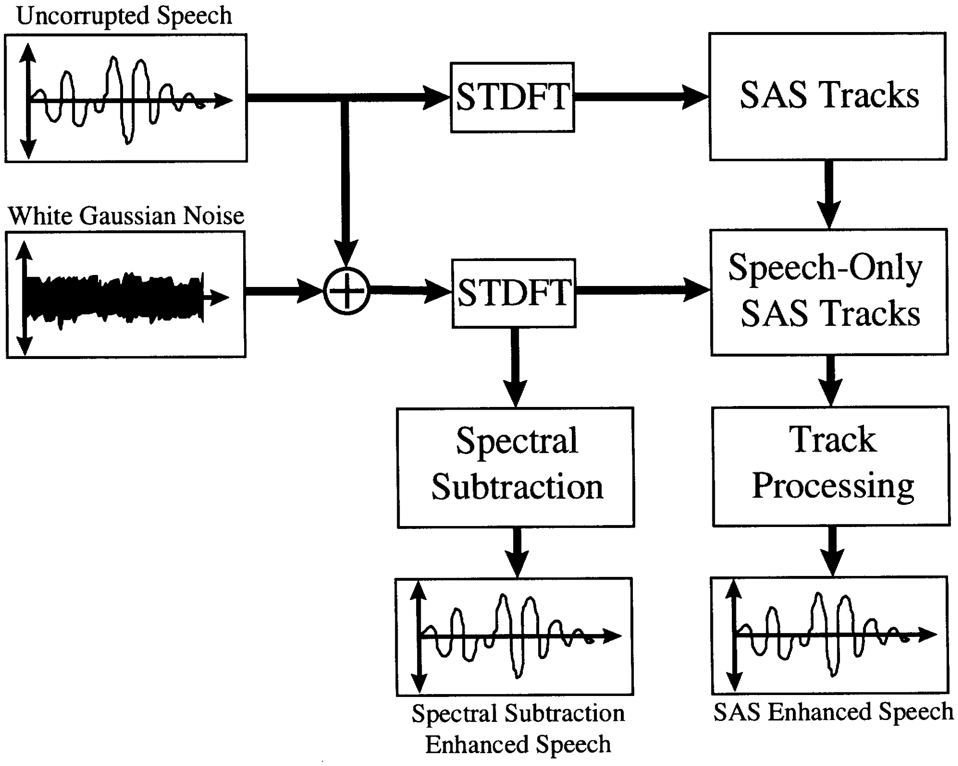


Fig. 2.3 SAS enhancement evaluation framework

4.1 Speech-Only Track Extraction

Track extraction was defined as the process of selecting and matching peaks from the corrupted STDFT to form tracks relevant to the speech. By using tracks from the uncorrupted speech as a guide in this process, we provide an upper bound on the performance of any track extraction algorithm operating within the limits of single-sensor enhancement, where no specific knowledge of the uncorrupted speech is available. The location of a track within the STDFT of the uncorrupted speech is given by the frequencies (bin values) of its matched peaks, and for each of these tracks, a corresponding track with approximately the same location is formed in the corrupted STDFT. The frequency, amplitude, and phase of each peak defining any such track are taken from the corrupted STDFT, but the manner in which the peaks are selected and matched from frame-to-frame is dictated by a track from the uncorrupted speech.

Specifically, track extraction proceeds as follows. For each peak along a track from the uncorrupted speech, a peak at approximately the same frequency is sought in the corresponding frame of the corrupted STDFT magnitude. The corrupted peak whose frequency is closest to that of the uncorrupted peak is selected, and if the absolute difference between the frequencies of the two peaks is below some threshold, the corrupted peak is included in the extracted track. If not, the corrupted peak is considered "lost," and its parameters are interpolated after all other peaks from the associated track are extracted from the corrupted STDFT. The SAS amplitude and phase interpolation formulas given in (2.6) and (2.8) are utilized, but instead of applying these formulas over a single frame interval, they are allowed to span any consecutive number of frames containing lost peaks. The parameters of the lost peaks are then found by sampling the resulting amplitude and phase functions at each frame interval lying within the interpolation region.

Figure 4.1 depicts the track extraction procedure by plotting together peaks from the STDFT's of the uncorrupted and corrupted speech. The tracks from the uncorrupted speech (dashed lines) are the same as those seen in Figure 2.1. The extracted tracks, represented by the dark solid lines, are seen to connect peaks from the corrupted STDFT (the O's) that are close to peaks from the uncorrupted speech (the X's). In several places, peaks are lost in the corrupted STDFT, and interpolation is required.

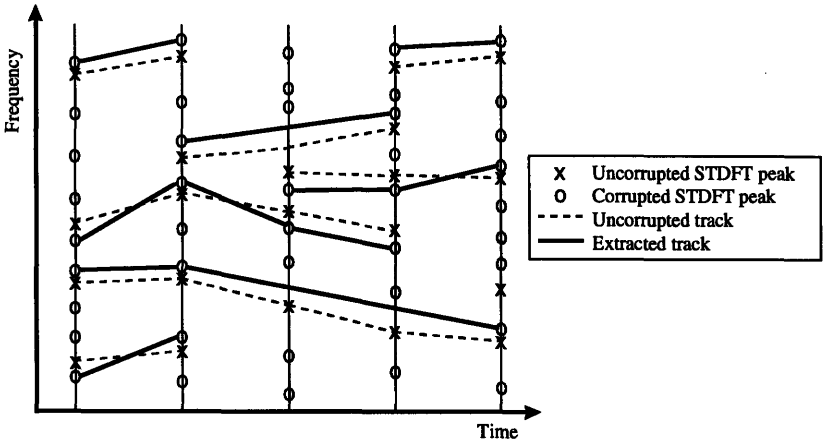


Fig 4.1 Illustration of the track extraction procedure

4.2 Synthesis of Unprocessed Extracted Tracks

By synthesizing extracted tracks before any processing is applied to their corrupted parameters, we examine the enhancement properties of speech-only track extraction by itself. In experiments designed toward this end, speech utterances were corrupted with additive white Gaussian noise to achieve SNR's ranging from 18dB down to 0dB. Speech-only tracks were extracted using a threshold of 40Hz to match peaks defining the uncorrupted tracks to peaks in the corrupted STDFT. The extracted tracks were then interpolated and summed to produce an estimate of the uncorrupted speech. For all SNR's, the underlying speech from the corrupted signal was perceptually intact in the estimate, but a large amount of noise remained. Instead of sounding broadband, the noise now sounded highly correlated with the speech. The fact that the estimate was completely silenced during regions of non-speech activity was a partial cause, and in addition, the noise that remained during speech activity possessed a speech-like quality. This quality did not change with SNR; the residual simply became louder as the SNR decreased. Several informal listeners found the new speech-like noise residual to be very unnatural sounding and preferred the original corrupted speech, even though the noise level was perceived to be higher. Enhancement was definitely not achieved, and the need for track processing was made clear.

Allowing only one of the three extracted track parameters (frequency, amplitude, or phase) to be taken from the corrupted spectrum, while the other two are retained from the uncorrupted tracks, isolates each parameter's contribution to the overall noise residual. When this modified extraction procedure was utilized in our experiments, corrupted track amplitudes were quickly identified as the largest perceptual component of the noise residual. With corrupted amplitudes alone, the synthesized speech sounded almost identical to the speech synthesized from extracted tracks with all three parameters corrupted. When only frequency or phase were corrupted, the resulting synthetic speech sounded very close to speech synthesized from the uncorrupted tracks. It was concluded that track processing should focus on the extracted track amplitudes, while frequency and phase can essentially be ignored.

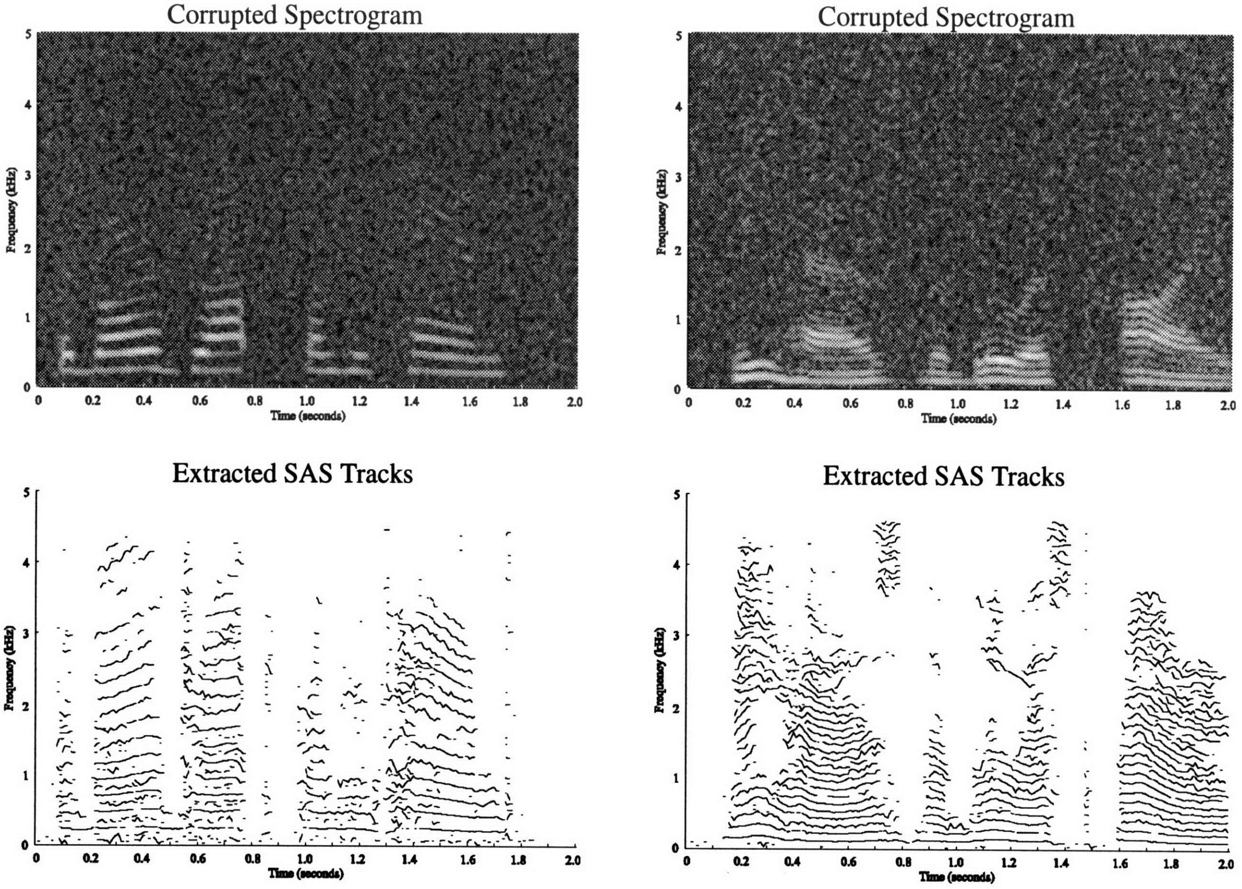
With the above conclusion, characterizing the quality of any amplitude extracted from the corrupted STDFT is appropriate. For this purpose, we define a peak-to-noise-ratio (PNR) in the same manner that a peak-to-signal-ratio (PSR) was defined in Section 2.2. Specifically, for any peak *k* in frame *d* of the STDFT of the uncorrupted speech, the PNR in decibels is defined to be

$$PNR_{k,d} = 10\log_{10}\left(\frac{A_{k,d}^2}{\psi_n}\right), \tag{4.1}$$

where ψ_n is the expected value of the Gaussian noise STDFT squared magnitude given in (3.15). Intuitively, we expect that as the PNR of a peak from an uncorrupted track decreases, the corresponding amplitude that is extracted from the noisy speech will be more corrupt relative to the original amplitude.

If peaks with smaller PNR's are excluded from the track extraction procees, we find that the perceptual level of the residual noise in the speech estimate can be reduced while maintaining the integrity of the underlying speech. This is achieved by referencing the minimum peak amplitude threshold used during track formation in the uncorrupted speech to a particular PNR rather than PSR, so that the amplitude threshold increases with the noise power. A portion of the speech-only track information resulting from the use of a -30dB PSR, as specified in Section 2.2, is discarded in exchange for a reduction of the noise residual. For all SNR's ranging from 18dB down to 0dB, a minimum PNR of -12dB was judged to provide the best compromise between noise reduction and preservation of the underlying speech. With this threshold, any artifacts due to the lost speech-only track information seemed to be perceptually masked by the noise residual. Figure 4.2 illustrates the implementation of track extraction with a -12dB PNR minimum amplitude threshold. We see spectrograms of the two speech utterances from Figure 2.2 after white Gaussian noise has been added to achieve an SNR of 6dB. Below each spectrogram is a spectrogram-like plot of the extracted tracks. Since a -12dB PNR threshold is used, we see fewer tracks than in Figure 2.2, where a -30dB PSR threshold is used.

In summary, the track extraction procedure developed in this chapter is not considered a successful enhancement technique by itself. Corrupted track amplitudes were identified as the most significant cause of the perceptually objectionable noise residual resulting from the synthesis of extracted tracks. By excluding uncorrupted STDFT peaks with smaller PNR's from the extraction process, the audible effects of the noise residual in the speech estimate are somewhat reduced. Based on these findings, the exploration of track processing in Chapter 5 focuses on the amplitudes of tracks that are extracted using a -12dB PNR threshold.



(a)

(b)

Fig. 4.2 Corrupted spectrograms and extracted SAS tracks: 6 dB SNR and -12dB PNR. (a) female speech, "The bowl dropped from his hand." and (b) male speech, "He has the bluest eyes."

Chapter 5

Extracted Track Amplitude Processing

In Chapter 4, we developed a procedure to extract speech-only tracks from the STDFT of corrupted speech by using tracks from the corresponding uncorrupted speech as a guide. Assuming the performance of the extraction step in our SAS enhancement system to be thusly optimized, we now focus on processing the corrupted amplitudes of extracted tracks to reduce the noise residual in the speech estimate. Processing is performed within the limits of single-sensor enhancement, utilizing no specific knowledge of the uncorrupted speech. In this chapter we consider two methods: spectral subtraction and track amplitude smoothing.

5.1 Bias Removal Through Spectral Subtraction

During the discussion of spectral subtraction in Chapter 3, we saw in (3.3) and in Figure 3.2 that the STDFT magnitude of the corrupted speech will be, on average, larger than the STDFT magnitude of the uncorrupted speech at any particular time-frequency location. SAS track amplitudes are taken directly from the STDFT magnitude, but in our enhancement system, an extracted track amplitude is not necessarily taken from the same frequency location as the corresponding uncorrupted track amplitude. The two locations are constrained to be close, however, and we therefore observe that an extracted amplitude is biased above its corresponding uncorrupted amplitude in a manner approximated by (3.3). As an illustration, Figure 5.1 plots the amplitudes of an uncorrupted track and its corresponding extracted track over time. The tracks are from the female utterance that appears in Figures 2.2 and 4.2. With an overall SNR of 6dB for the corrupted speech, the average PNR of the uncorrupted track is 1.12dB. At every frame, except one, the extracted amplitude is larger than the uncorrupted amplitude.

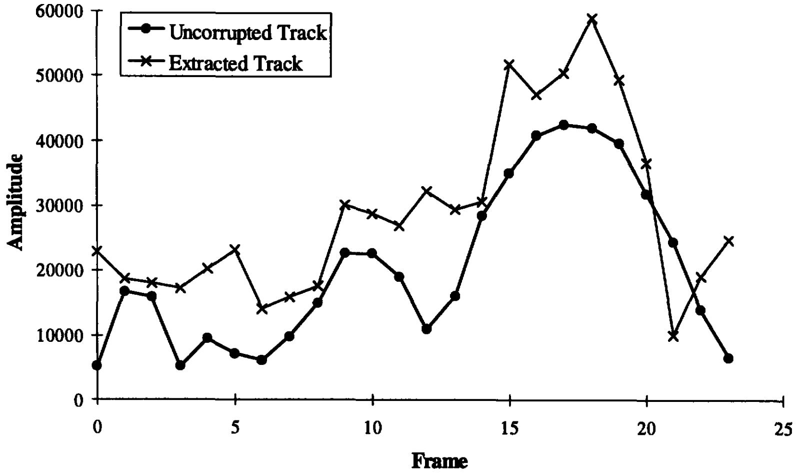


Fig 5.1 Uncorrupted track amplitudes and corresponding extracted track amplitudes taken from female speech with a 6dB SNR

Although spectral subtraction was originally developed to remove the bias from the entire STDFT magnitude of corrupted speech, its application at a particular STDFT bin number is independent of information at any other bin number. Therefore, spectral subtraction can be applied without modification to any track amplitude if the expected value of the noise spectrum at the corresponding frequency is known, as is the case for the Gaussian noise being considered. Stated more formally, let A be the amplitude of an uncorrupted track peak, let B be the corresponding extracted amplitude, and let ψ_n be the expected value of the Gaussian noise STDFT sqaured magnitude. Redefining the (speech-plus-noise)-to-noise ratio of (3.18) to be

$$\rho = \frac{B^2}{\psi_n}, \tag{5.1}$$

an estimate of A is

$$\hat{A} = H(\rho)B, \tag{5.2}$$

where $H(\rho)$ is any one of the gain functions given in (3.20a-d) for the discussed forms of spectral subtraction: maximum likelihood, power, Wiener, and amplitude.

In our experiments, each type of spectral subtraction was applied to tracks extracted according to Section 4.2. In all four cases, the noise residual in the resulting speech estimate was at a lower perceptual level and sounded less "harsh" in comparison to the residual described in

Section 4.3. Stated more precisely, it sounded as though a low-pass filter had been applied to the residual that existed before spectral subtraction was utilized. This makes sense because spectral subtraction applies greater attenuation to extracted tracks with smaller amplitudes, as demonstrated by Figure 3.1, and these smaller amplitude tracks tend to lie at higher frequencies. The degree to which each form of spectral subtraction perceptually reduced the noise residual is also in agreement with Figure 3.1 – maximum likelihood was the least effective and amplitude the most, with power and Wiener lying in between.

Despite the relatively large reduction of the noise residual provided by amplitude spectral subtraction in our experiments, the underlying speech in the estimate retained nearly all the clarity that was present when extracted tracks were synthesized without modification, suggesting that even more attenuation might be applied without severely damaging the quality of the speech. By increasing the parameter *k* in the spectral subtraction formulation of (3.7), more attenuation is provided. With *a*=1 and *k* allowed to vary, the gain function for amplitude spectral subtraction becomes

$$H(\rho) = \frac{\sqrt{\rho} - k\sqrt{\pi}/2}{\sqrt{\rho}}, \tag{5.3}$$

and Figure 5.2 is a plot of this function for various values of *k*. A further reduction of the noise residual was heard as *k* increased, but the underlying speech started to sound muffled, especially for *k*>1.5. A value of *k*=1.25 was judged to provide the best compromise, and with this specific value, (5.3) will be referred to as *modified* amplitude spectral subtraction. After its application, the level of the noise residual was low enough so that a preference among informal listeners between the speech estimate and the corrupted speech leaned towards the estimate, especially for very low SNRs. The residual still sounded very correlated to the speech, but was now much more tonal – during longer voiced sections of an utterance, it might be described as a "warble." Figure 5.3 demonstrates the application of modified amplitude spectral subtraction to the extracted track from Figure 5.1. The extracted track amplitudes are no longer biased above the corresponding uncorrupted amplitudes, and, in fact, the majority of them lie below the uncorrupted amplitudes, as predicted by Figure 3.2.

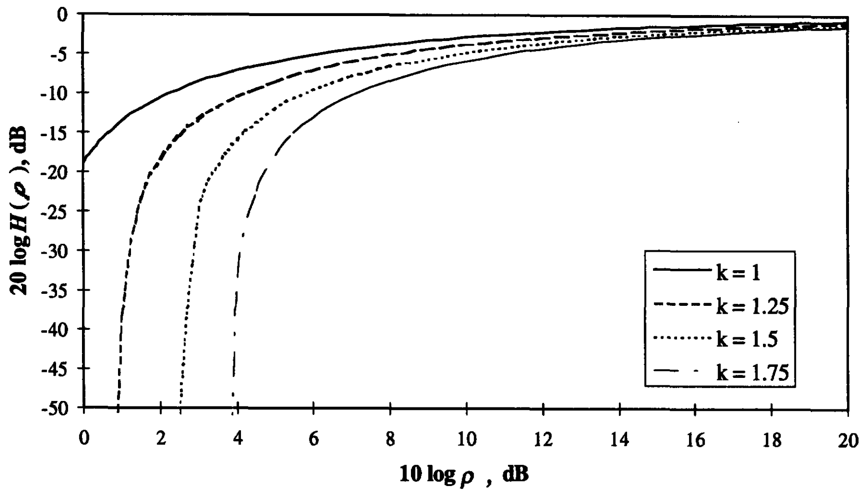


Fig 5.2 Gain function for amplitude spectral subtraction with varying *k*

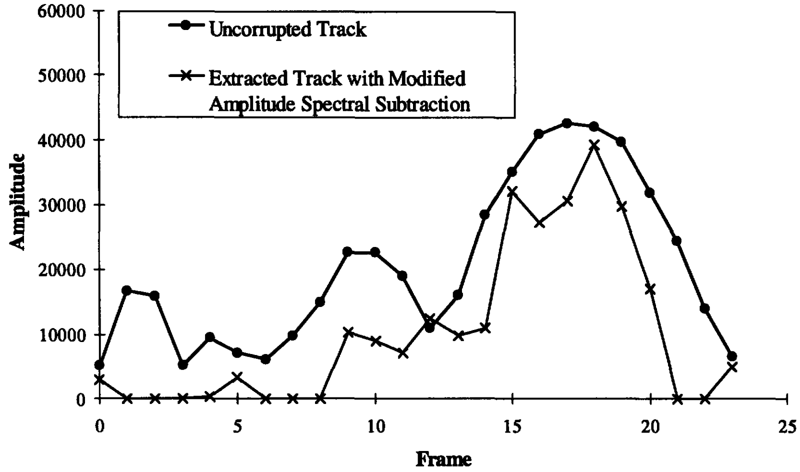


Fig 5.3 Uncorrupted track amplitudes and corresponding extracted track amplitudes after the application of modified amplitude spectral subtraction

Using modified amplitude spectral subtraction as a first step in track processing is certainly reasonable, but it possesses a short-coming in the context of SAS. By operating on each track amplitude independently, the highly structured manner in which these amplitudes are connected over time is completely ignored. Exploiting this structure to further reduce the noise residual is the next step in our development as we consider smoothing the series of amplitudes that define each track.

5.2 Track Amplitude Smoothing

A large number of tracks formed in an uncorrupted speech utterance, especially those from a voiced section, exhibit amplitudes that change quite slowly and predictably from frame-to-frame. A series of extracted track amplitudes, on the other hand, tends to vary much more randomly because of the corrupting noise. These observations are demonstrated in Figure 5.4 where we see pairs of uncorrupted and extracted tracks from the male and female speech of Figures 2.2 and 4.2. Each extracted track is from a corrupted utterance with an overall SNR of 6dB, and modified amplitude spectral subtraction has been applied. The average PNR of the corresponding uncorrupted track is printed at the top of each plot. We see that the uncorrupted tracks (dark lines) tend to vary smoothly over time while the extracted tracks (lighter lines) look much more jagged. From these plots it is does not appear that the exact, or even approximate, shape of the uncorrupted tracks is recoverable from the extracted tracks, but it does seem plausible that smoothing the extracted track amplitudes over time might further reduce the noise residual in the speech estimate.

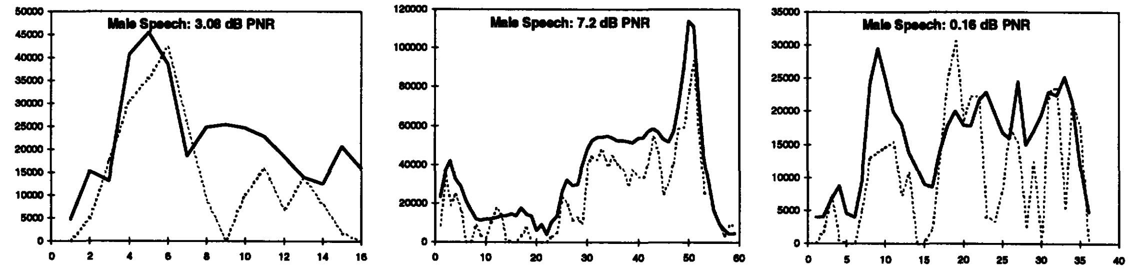


Fig 5.4 Uncorrupted track amplitudes (solid lines) and the corresponding extracted track amplitudes (dashed lines) after modified amplitude spectral subtraction. Top row: female speech. Bottom row: male speech.

5.2.1 Moving Average Filters

The first type of smoother that we have considered is a weighted moving average, which can be represented as a linear time-invariant (LTI) filter with finite-length impulse response $h[n]$ satisfying

$$h[n] = h[-n], \quad |n| \leq L \qquad (4.20a)$$

$$h[n] = 0, \quad |n| > L,$$

and

$$\sum_{n=-L}^{L} h[n] = 1, \qquad (4.20b)$$

where the length of the filter is $2L+1$, and the time index n represents frame number. We do not have any strict constraints in the frequency domain for the filter other than a desire for it to exhibit a low-pass characteristic. Several popular symmetric windowing functions, such as the rectangular, Bartlett, Hanning, and Hamming, all meet this requirement, and they are easily normalized to comply with (4.20b). By experimenting with different windows, various tradeoffs between low-pass bandwidth and high-pass attenuation are achieved [8], and for each window, the amount of smoothing increases with the filter length.

The track amplitude sequences being filtered are finite length, so the manner in which the filter is applied at the endpoints of each sequence must be considered. Let $a[n]$ be a sequence of track amplitudes having length T. We want to perform the convolution $\tilde{a}[n] = h[n] * a[n]$ and then replace the original track amplitudes with $\tilde{a}[n]$ for $0 \leq n < T$. Because $h[n]$ is symmetric about $n = 0$, $a[n]$ must be extended so that it is defined for $-L \leq n < T + L$. One option is to set $a[n]$=0 at the added beginning and end points, but this causes $\tilde{a}[n]$ to taper down towards zero at the beginning and end of the track. For very short tracks, where $T \leq L$, the entire track is severely attenuated. A better solution is to extend the values $a[0]$ and $a[T$-1$]$ into the added beginning and end points, *i.e.* $a[n]$=$a[0]$ for $-L \leq n < 0$ and $a[n]$=$a[T$-1$]$ for $T \leq n < T + L$. This way, any filtered track can still begin or end with a large amplitude, and very short tracks are not attenuated.

In our experiments, filtering extracted tracks with a moving average of the proper length resulted in a small perceptual reduction of the noise that remained after the application of spectral subtraction. The warble that was heard during longer voiced sections was transformed into a noise that sounded quieter and less time-varying. With a rectangular window defining the shape of the impulse response, best results seemed to be obtained using L=2, and with any of the tapered windows (Blackmann, Hanning, or Hamming), L=3 seemed best. Improvements afforded by the rectangular and tapered windows seemed perceptually equivalent, so the rectangular window appeared the better choice since it had a shorter length. After filtering the track amplitudes, transition areas in the speech estimate became slightly slurred, but for low SNR's these degradations were judged to be less perceptible and unnatural sounding than the warbles that were eliminated. For SNR's greater than 12 dB, however, the tradeoff was debatable. As L was increased beyond three, the slurring in transition areas became worse, and no additional reduction of the noise residual was heard during longer voiced segments.

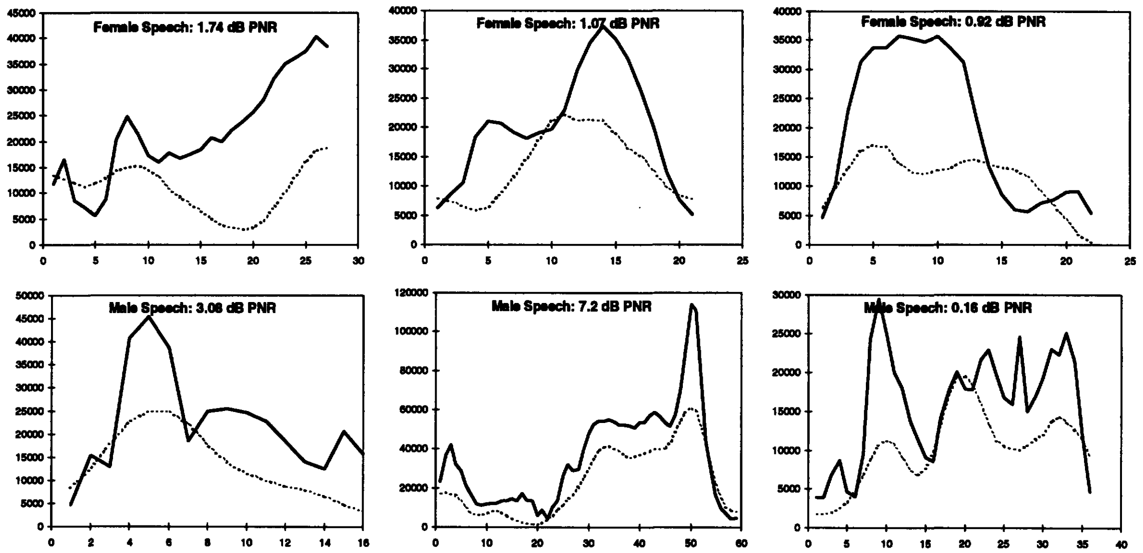


Fig 5.5 Uncorrupted track amplitudes (solid lines) and the corresponding extracted track amplitudes (dashed lines) after modified amplitude spectral subtraction and a moving average filter with a rectangular window and L=2.

Our experiments suggest that, among the specified class of weighted moving average filters, a filter with a rectangular window and a length of five frames (50ms) provides the best improvement to the speech estimate. In comparison to the improvement heard after the application of spectral subtraction, however, the added benefit of the moving average amplitude smoother was fairly small. Figure 5.5 depicts the same tracks from Figure 5.4, but now a moving average filter with a rectangular window and L=2 has been applied to the extracted tracks.

5.2.2 Median Filters

Another type of smoothing filter that we have considered is the running median which, unlike the moving average, is non-linear. A running median replaces each value of a sequence with the median of a surrounding sub-sequence, and its application to a sequence of track amplitudes $a[n]$ with length T is analogous to the moving average filter. First, $a[n]$ is extended so that it is defined over the range $-L \leq n < T + L$. Then, for $0 \leq n < T$, $\tilde{a}[n]$ is equal to the median of the sub-sequence $a[n-L]\ldots a[n+L]$, where the length of the filter is $2L+1$.

The running median filter is markedly different from the moving average in two ways. First, if the sequence being filtered is composed of constant sub-sequences, each at least $L+1$ samples long, then the sequence is unchanged by the median filter. Therefore, a median filter is capable of preserving edges within a sequence, whereas a moving average filter smears such discontinuities. Secondly, if every sample of a sequence equals some constant except for one outlying sample, then every sample of the median filtered sequence is equal to this constant, no matter how large the outlier. The median filter removes the outlier without affecting the rest of the sequence, while a moving average filter results in a hump around the outlier. Figure 5.6 illustrates these differences between the running median and moving average filters, both with L=2, as they are applied to two artificially constructed sequences.

The edge preservation and outlier suppressing properties of the running median filter have made it a popular tool in image processing. For example, images that are corrupted with impulse-like noise can often be enhanced with a 2-dimensional median filter [9]. More related to the work at hand, median filtering has proven useful in smoothing the contours of estimated short-time speech features, such as zero-crossing rate and pitch period, which contain large-scale discontinuities that need to be preserved and small-scale estimation noise that should be suppressed [10]. In addition, a series combination of a running median and moving average has been used for this same application with reported improvements over standard median filtering. It is not clear from Figure 5.5 that the sequences of extracted track amplitudes exhibit any properties that make them obvious candidates for median filtering, but one might hypothesize that the edge preservation property could reduce the slurring problem associated with the moving average filter.

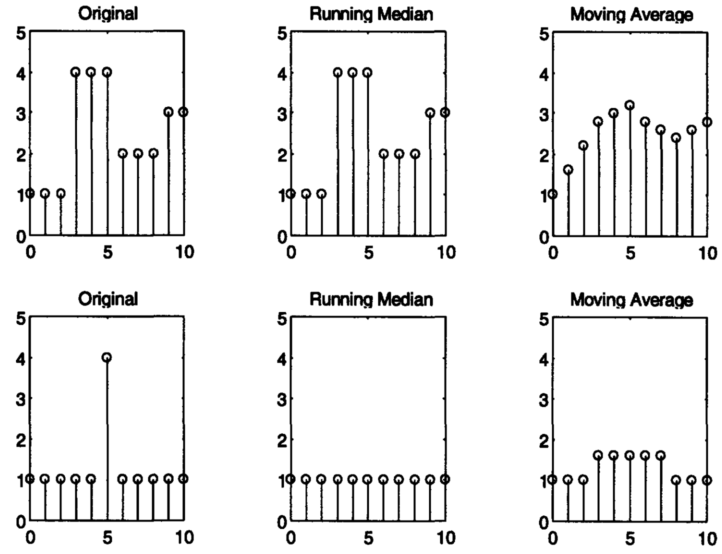


Fig 5.6 Demonstration of the differences between a moving average and running median filter. Top row: edge preservation. Bottom row: outlier suppression.

In our experiments, median filtering of the extracted track amplitudes with L=1 had almost no perceptual effect on the speech estimate. The warbling was reduced somewhat, perhaps, but any improvements were negligible in comparison to those resulting from a moving average filter of the same length. At the same time, no noticeable artifacts were introduced into the speech estimate. With L=2, median filtering reduced the warbling by an amount comparable to a moving average filter, but an equivalent amount of slurring was heard. As L increased beyond two, the slurring became worse, and no further reduction of the residual was perceived. When a series combination of a running median and moving average was applied with combinations of L=1 and

L=2 for the two filters, no additional improvements over the standard median filter were heard in the speech estimate.

In summary, the best results obtained in our experiments from median filtering seemed perceptually equivalent to those obtained from the moving average. Figure 5.7 depicts the tracks from Figure 5.4, but now a running median filter with *L*=2 has been applied to the extracted tracks. Comparison with Figure 5.5 reveals few visual differences between the general shape of each median and moving average filtered track, an observation that coincides with the perceptual equivalence of the resulting speech estimates. Apparently, sequences of extracted track amplitudes possess very few of the features that highlight differences between the application of a running median and a moving average.

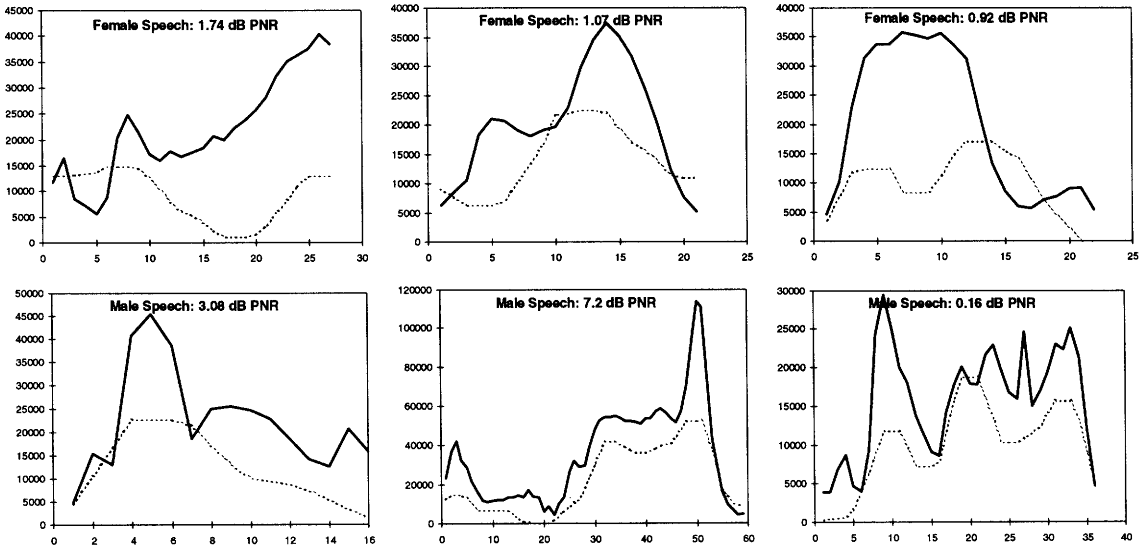


Fig 5.7 Uncorrupted track amplitudes (solid lines) and the corresponding extracted track amplitudes (dashed lines) after modified amplitude spectral subtraction and a running median filter with *L*=2.

5.3 Conclusions

In this chapter, two types of track amplitude processing, spectral subtraction and smoothing, were found effective for reducing the noise residual in the speech estimate produced from extracted SAS tracks. Removing the upward bias of extracted track amplitudes by applying modified amplitude spectral subtraction resulted in a substantial reduction of the noise residual.

Spectral subtraction, however, does not take advantage of the track structure in which amplitudes from the uncorrupted speech tend to vary smoothly over time. Based on this observation, the amplitudes defining each extracted track were processed with a smoothing filter. Both a moving average and running median filter resulted in an equivalent perceptual reduction of the noise residual at the expense of some minor slurring in the underlying speech. The improvements due to smoothing, however, were small in comparison to the improvements heard after spectral subtraction was applied.

Examining these results, we note that spectral subtraction can be utilized outside of the proposed SAS enhancement framework, while track amplitude smoothing is a procedure that is entirely unique to SAS. The use of SAS for enhancement would appear more credible if the SAS-specific track amplitude processing had a larger perceptual impact. However, benefits resulting from the very fact that the speech estimate is constructed from extracted SAS tracks have not yet been evaluated through comparison with a non-SAS enhancement procedure. By comparing traditional spectral subtraction with the best-case SAS enhancement, the overall effect of the SAS-specific processing can be isolated and judged. This is the focus of the next chapter.

Chapter 6

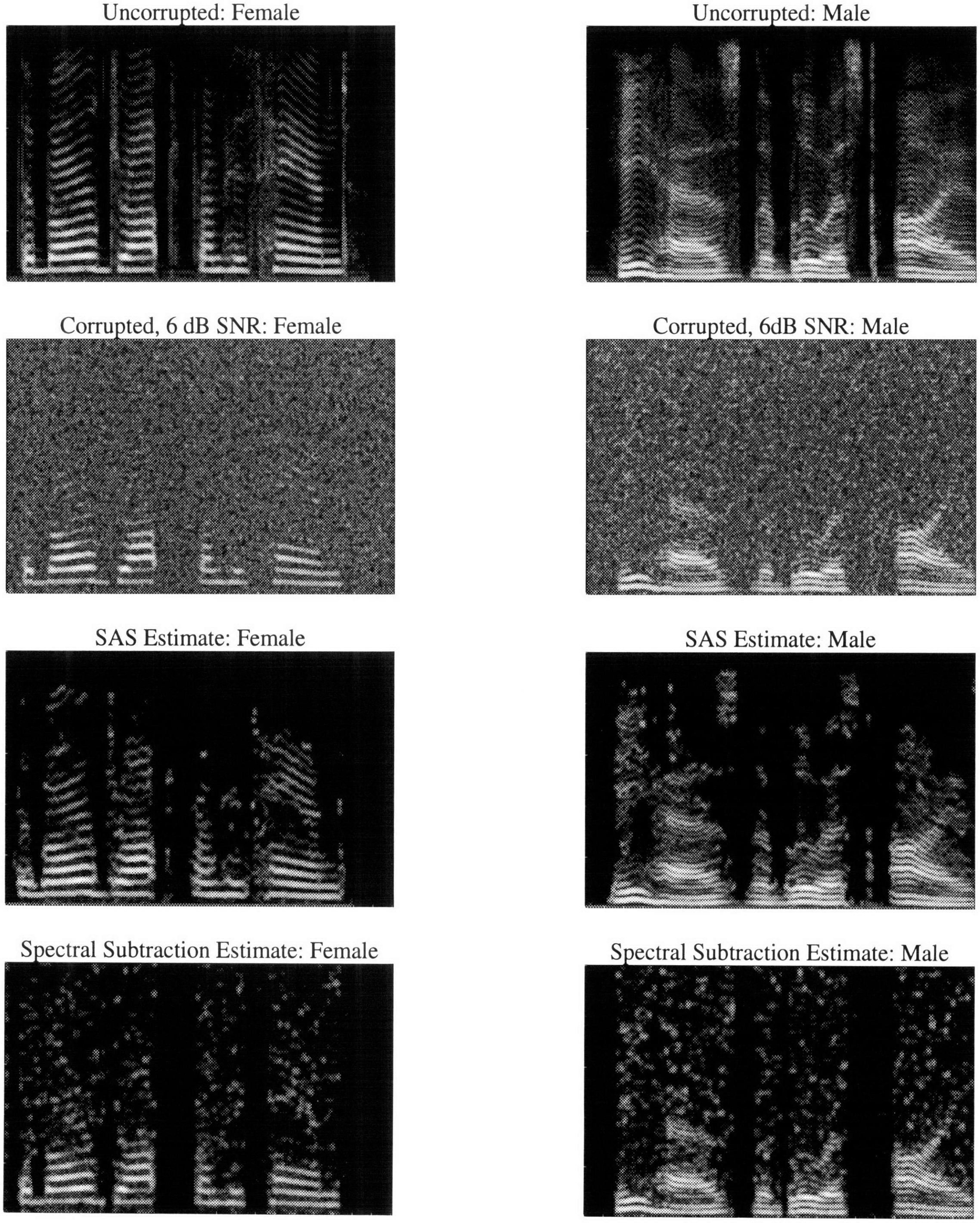
Comparison of SAS and Spectral Subtraction

In order to determine the relative benefit of the SAS-specific processing in the enhancement procedure discussed in Chapter 5, a comparison is made with traditional spectral subtraction, as implemented in Section 3.3. For the development of a blind track extraction algorithm to be justified, SAS enhancement based on tracks extracted with specific knowledge of the uncorrupted speech should demonstrate a clear and significant improvement over traditional spectral subtraction. The existence of such an improvement was evaluated through an informal listening test involving multiple subjects and utterances, and the results are presented in this chapter.

6.1 Qualitative Comparison of SAS and Spectral Subtraction Enhancement

Before the listening test is discussed, some general qualitative comparisons can be made between the SAS enhancement technique and traditional spectral subtraction. Figure 6.1 shows spectrograms of the uncorrupted and corrupted speech from Figures 2.2 and 4.2. In addition, we see spectrograms of both the SAS and spectral subtraction speech estimates. The most obvious visual difference between the two estimates is that SAS preserves some of the higher-frequency/lower-energy harmonic lines, while the spectral subtraction spectrograms look very "spotted" is these areas. The fact that tracks are being used to create the SAS estimate allows most of these noise spots to be eliminated, and track amplitude smoothing interpolates between the speech-relevant spots to more faithfully reproduce the harmonic lines. In both cases, however, very low energy portions of the original speech are completely lost.

In terms of aural perception, the residual in the SAS estimate sounds correlated to the speech since the corrupting noise is effectively filtered by narrow band-pass filters that follow the harmonic lines of the speech. The residual in the spectral subtraction estimate, though perceptually louder, sounds relatively uncorrelated to the speech. One might argue that an uncorrelated residual sounds more natural, but the fact that it is louder might outweigh this advantage.



Uncorrupted: Female

Uncorrupted: Male

Corrupted, 6 dB SNR: Female

Corrupted, 6dB SNR: Male

SAS Estimate: Female

SAS Estimate: Male

Spectral Subtraction Estimate: Female

Spectral Subtraction Estimate: Male

Fig. 5.1 Comparison of SAS and spectral subtraction enhancement for male and female speech

6.2 Listening Test

A very informal listening test comparing SAS enhancement and spectral subtraction was conducted with the assumption that if a significant preference for SAS enhancement existed, it would be evident in such a test. During the test, subjects listened to several sets of speech utterances through headphones. Each set consisted of four signals: 1) an uncorrupted utterance, 2) the same utterance corrupted with white Gaussian noise to achieve an SNR of 6dB, 3) the corresponding SAS speech estimate, and 4) the corresponding spectral subtraction speech estimate. A 6dB SNR was chosen because it was felt that at this noise level either of the two speech estimates would be preferred over the corrupted speech. The estimates were labeled “A” and “B”, and the letters’ correspondence to SAS or spectral subtraction was randomized from set to set. For each set, subjects were allowed to listen to each of the four utterances as many times as they wanted and were simply asked if they preferred listening to “A” or “B.”

In creating the speech estimates for the listening test, spectral subtraction was implemented to the specifications of Section 3.3. With SAS enhancement, however, track amplitude processing was not performed precisely as outlined in Chapter 5. Rather than using k=1.25 for amplitude spectral subtraction, k=1 was used. Then, before smoothing, the adaptive order-statistic filter presented in Section 3.3 was applied to each sequence of track amplitudes, rather than along STDFT bins as Boll originally intended. Lastly, a moving average filter with a rectangular window and L=2 was applied to the track amplitudes, just as discussed in Section 5.2.1. Applying Boll’s filter along the tracks did result in a further reduction of the noise residual after amplitude spectral subtraction with k=1 had been applied, but after the listening test it was recognized that the filter had been designed to be most effective along portions of the STDFT that did not contain speech. Applying it along the tracks completely violated this logic, and upon closer examination it was found that the filter merely attenuated the lower amplitude tracks further. The same perceptual result was obtained much more directly by setting k=1.25 in the spectral subtraction equation. Therefore, Chapter 5 presented the use of amplitude spectral subtraction with k=1.25 as proper procedure and did not mention Boll’s filter for the sake of maintaining coherency. The filter is mentioned here, however, so that the conditions of the listening test are documented with complete accuracy.

6.3 Results and Conclusions

Six different sets of utterances were presented to each subject during the test:

1. Female - “The bowl dropped from his hands.”
2. Male - “He has the bluest eyes.”
3. Female - “We made some fine brownies.”
4. Male - “The chef made lots of stew.”
5. Female - “That shirt seems much too small.”
6. Male - “Stuff those with soft feathers.”

Nine subjects participated, and the results are depicted in Figure 5.2, where the number of times SAS and spectral subtraction were chosen is plotted against both listener and utterance. Altogether, SAS enhancement was chosen 60% of the time and spectral subtraction 40%. Among the 9 listeners, 4 chose SAS more often, 4 chose spectral subtraction more often, and 1 chose both an equal number of times. Among the 6 utterances, the SAS enhancement was selected more often except for one utterance.

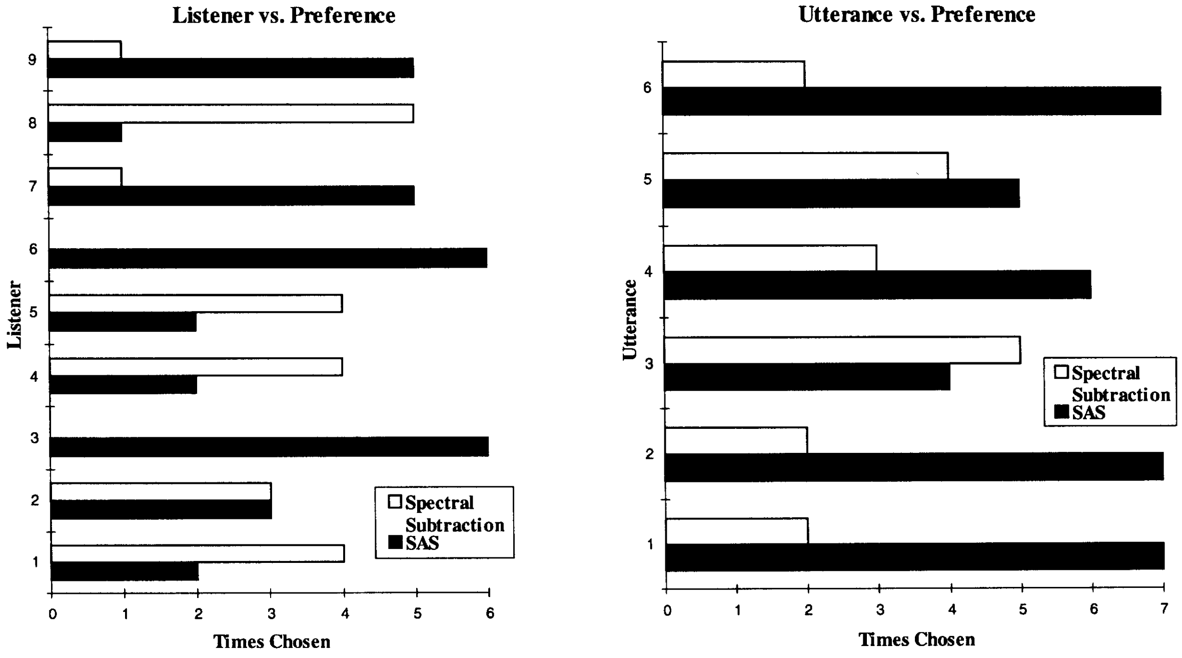


Fig. 5.2 Listening test results

The statistical results of the listening test might be used to indicate a slight preference for SAS enhancement, but they must be qualified by a description of the subjects' reactions and comments. Upon first comparing the two speech estimates, some of the subjects were perplexed, complaining that they could hear no significant difference between the two. They then proceeded to switch back and forth between the estimates, listening more intently for smaller differences. By the end of the test, most of the subjects indicated that they were able to hear a difference between the two estimates in each set, but that selecting one as perceptually more pleasing was oftentimes arbitrary. Two of the subjects, however, were able to quickly perceive two distinct types of enhancement and then consistently pick the type that they liked. Both of these subjects chose SAS, but one other subject was almost as consistent in selecting spectral subtraction. Whatever the subject's preference, no one made any comment about one enhancement being obviously better than the other.

In light of the statistics and especially the subjects' reactions, a strong and immediate preference for SAS enhancement over spectral subtraction was clearly not exhibited in the listening test. The SAS-specific features of our enhancement system did not appear to add any substantial perceptual improvements to spectral subtraction. We therefore conclude that developing an algorithm to blindly extract speech-only tracks from corrupted speech is not yet justified. The upper bound on the performance of SAS enhancement that was examined needs to promise much more convincing gains before such a pursuit makes sense.

Chapter 7

Summary and Future Work

7.1 Summary and Conclusions

This thesis investigated the use of Sinusoidal Analysis-Synthesis (SAS) for the single-sensor enhancement of speech corrupted by additive broadband noise. In the proposed enhancement scheme, speech-only tracks are first extracted from the corrupted speech by appropriately selecting and matching peaks in its STDFT mangitude. Then, the corrupted parameters of these tracks are processed in order to reduce the perceptual level of any remaining noise. To evaluate the potential of this technique, a framework was developed for examining an upper bound on performance. Track extraction was performed with specific knowledge of the uncorrupted speech so that the procedure was, in a sense, optimized. The extracted tracks were then processed within the single-sensor constraints, and the resulting speech estimate represented an upper bound on the performance of SAS enhancement for whatever type of processing was performed.

After experimenting with the synthesis of unproccessed extracted tracks, it was concluded that the extraction procedure, by itself, does not afford convincing enhancement. Corrupted track amplitudes were identified as the largest contributor to the resulting speech-correlated noise residual and therefore became the focus of track processing. Two single-sensor processing techniques were considered. First, spectral subtraction was applied to remove the upward bias in extracted track amplitudes. Of the several forms tested, modified amplitude spectral subtraction seemed to provide the best compromise between noise reduction and speech preservation. Next, various smoothing filters were applied to the extracted track amplitudes in an attempt to take advantage of the SAS track structure. Both moving average and running median filters provided some additional reduction of the noise residual, but the improvements were small in comparison to those resulting from spectral subtraction. A comparison was then made between SAS enhancement and traditional spectral subtraction in order to isolate effects of the SAS-specific processing in our system. An informal listening test indicated no significant preference for SAS enhancement, leading us to conclude that the development of a blind track extraction algorithm is not yet justified.

The results of the listening test suggest that the SAS-specific features of our enhancement system add no substantial improvements to traditional spectral subtraction. This is somewhat surprising given that the noise lying in between the tracks is eliminated. A possible explanation is that the resulting noise residual sounds less natural in comparison to that of spectral subtraction, thereby canceling the benefits of its lower perceptual level. This may be an inherent problem with using SAS for enhancement. While the benefits of eliminating inter-track noise now seem questionable, SAS tracks clearly introduce a large amount of structure to the enhancement problem. In this thesis, efforts made to exploit the track structure produced little improvement in the speech estimate. However, the examined filters were selected based on a simple empirical observation that the track amplitudes should vary smoothly, in some sense, over time. New track processing techniques could conceivably take advantage of the track structure in a more optimal manner, and if such work is continued, the framework developed in this thesis will prove useful.

7.2 Future Work

In light of our conclusions, continuation of the work in this thesis should focus on the development of better track amplitude processing techniques. One course of action is to incorporate more accurate knowledge of speech into the processing. This is an idea that has already demonstrated merit in an enhancement procedure developed by Ephraim [11]. In this system, the probability distributions of both the speech and noise STDFT magnitudes are jointly estimated by using a hidden Markov model (HMM) that is trained on a corpus of representative uncorrupted speech utterances. With these two distributions, estimation of the STDFT magnitude of the uncorrupted speech is formulated in the Bayesian sense, and an improvement over traditional spectral subtraction is reported. Incorporating speech knowledge into the estimation of amplitudes along extracted SAS tracks would clearly have a different specific form, but the general idea is the same. For example, a parametric or statistical model describing the manner in which uncorrupted track amplitudes vary over time could be developed. With a parametric model, a class of curves might be fit to the extracted track amplitudes after spectral subtraction has been applied, and with a statistical model, a minimum-mean-square-error filter might be developed to jointly perform bias removal and smoothing.

References

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proceedings of the IEEE*, vol. 67, no. 12, December 1979.

[2] R. J. MacAulay and T. F. Quatieri, "Speech Analysis-Synthesis Based on a Sinusoidal Representation," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-34, no. 4, August 1986.

[3] R. J. McAulay and T. F. Quatieri, "Noise Reduction Using a Soft-Decision Sine-Wave Vector Quantizer," *Proc. IEEE ICASSP90*, Albuquerque, New Mexico, pp. 821-824, 1990.

[4] R. J. MacAulay and T. F. Quatieri, "Speech Transformations Based on a Sinusoidal Representation," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-34, no. 6, December 1986.

[5] R. J. McAulay and M. L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-28, no. 2, April 1980.

[6] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-29, pp. 113-120, April 1979.

[7] R. J. MacAulay and T. F. Quatieri, "Low-rate speech coding based on the sinusoidal model," in *Advances in Acoustics and Speech Processing*, M. Sondhi and S. Furui, Eds. New York: Marcel Deckker, pp. 165-207, 1992.

[8] A. V. Oppenheim and R. W. Schafer, Chapter 7 of *Discrete-Time Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, 1989.

[9] R. C. Gonzalez and R. E. Woods, Chapter 4 of *Digital Image Processing*, Addison-Wesley Publishing Co., New York, 1993.

[10] L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, "Applications of a Nonlinear Smoothing Algorithm to Speech Processing," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. ASSP-23, no. 6, December 1975.

[11] Y. Ephraim, D. Malah, and B. Juang, "On the Application of Hidden Markov Models for Enhancing Noisy Speech," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. 37, no. 12, December 1989.