# ABSTRACT

The ability of listeners with hearing loss to understand speech in noise is severely degraded compared to that of normal-hearing listeners. While several physiological and psychophysical models exist to explain how normal-hearing listeners detect various stimuli in noise, few of these models have been used as the basis of any sort of corrective algorithms for use in hearing-aids or communication devices. Here, a signal processing implementation of a physiological model was developed for the detection of tones in noise. The performance of the detector was evaluated for conditions under which speech in noise commonly occurs, with the goal of using the detector as part of a larger noise-reduction system. The potential benefits of the general class of noise-reduction (NR) algorithms that perform time-frequency gain manipulation were examined through the use of the ideal binary mask, a tool used in automatic speech recognition. By degrading the ideal binary mask, the detection parameters necessary to provide benefit were derived. Lastly, a phase-opponent noise-reduction (PONR) algorithm was developed and tested. Testing consisted of both an indirect measurement of intelligibility, as well as overall preference testing. The PONR algorithm resulted in up to 10 dB improvement in the signal-to-noise ratio of the speech in noise; this improvement, however, did not result in an improvement in intelligibility. This finding is consistent with the results from the binary mask experiments, as the PONR system was unable to detect the 90-95% of the speech energy necessary to show improvement in intelligibility.

# Time-Frequency Gain Manipulation for Noise-Reduction in Hearing Aids: Ideal and Phase-Opponent Detectors

By

Michael C. Anzalone
B.S. Boston University, 1999
M.S. Syracuse University, 2001

DISSERTATION

Submitted in partial fulfillment of the requirements for the
degree of Doctoral of Philosophy in Bioengineering
in the Graduate School of Syracuse University

May 2005

Approved: _____
Dr. Laurel H. Carney

Date: _____

# Table of Contents

# List of Figures

# Acknowledgements

All of this work would not have been possible without the help and support of my advisor, Dr. Laurel Carney. Her generous offer to help with my qualifying exam has turned into four years of advising, for which I am thankful. She constantly provided the necessary encouragement that I needed to continue my education and development as both an engineer and a researcher. She is truly an asset to the Syracuse University community and the Institute for Sensory Research.

I would also like thank the various professors within the department that have contributed to my development as a graduate student and a person. Dr. Robert Smith provided the opportunity to come to Syracuse University as a teaching assistant, and then allowed me to join his laboratory as a research assistant. He and Dr. Evan Relkin were instrumental in convincing me to continue on for my PhD after receiving my Master's degree.

My work has encompassed many different fields, and I have had the pleasure of working with people from each of these fields. Manny Ares provided critique and many insights into detection that helped shape Chapter 2. Dr. Carol Espy-Wilson and Om Deshmukh, both electrical engineers, provided valuable insight into speech recognition and detection, as well as suggesting the allpass filter that was used in Chapter 4. Much of the work would not have been possible without the aid of Dr. Karen Doherty, who attempted to teach an engineer the field of audiology. Her advice, as well as the use of her lab, contributed to the experiments of Chapters 3 and 4.

My time at Syracuse University would not have been the same with the great group of students that I have known over the last six years. Both Sean Davidson and Paul Nelson have provided many hours of commiseration over classwork, homework, experiments, and the general experience of being a graduate student. I will miss the many spontaneous conversations that have tended to occur in my office. Lauren Calandruccio has been an invaluable friend over the last two years. She provided the necessary audiological support for Chapters 3 and 4. I would also like to thank Felicia (Yan) Gai and Satish Iyengar, who have provided aid in one form or another throughout the years.

Throughout my six years, I have received funding from a variety of sources. The Department of Bioengineering and Neuroscience and the Graduate School provided me with support through both a teaching assistantship and a University Fellowship. I have been a research assistant, funded by grants from the National Institutes of Health, as well as from the Gerber endowment. I am grateful to all of these sources.

My family has always been close to my heart, and without them I would not be where I am today. Thanks PJ and Chrissy! I would also like to thank my mom, Kathleen Clark-Anzalone, for the never-ending encouragement and support, and for all of the work that has gone into making me the person I am today.

Lastly, I'd like the acknowledge my fiancé, Nicole Sanpetrino. We met while working at ISR, and as my work has progressed, so has my love and devotion to her. It is because of her constant support that I am finally finished!

# Chapter 1

# Introduction

## 1.1 Background

### 1.1.1 Speech Intelligibility and Listeners with Hearing Loss

One of the greatest problems facing listeners with hearing-loss is the loss of the ability to understand speech-in-noise. While this loss is known to be the result of the impairment of the listener's auditory system, the exact physiological causes are unknown. Much of the current effort in hearing-aid research has been to develop new algorithms that can help with speech-in-noise; it is these algorithms that currently differentiate the various hearing-aids, which generally use the same hardware components.

The ability of a listener to understand speech in noise can be measured through the use of a reception threshold for speech (RTS). This threshold measures the signal-to-noise ratio (SNR) that is necessary for the listener to correctly identify a given percentage of the speech, usually 50 or 100%. Under certain conditions, the RTS of normal-hearing listeners can be as low as -5 dB; this ability to understand speech-in-noise, even when the level of the noise is greater than the speech, exceeds the capabilities of any computer system or algorithm that currently exists. When the auditory system is impaired, however, the RTS is increased. When the noise has the same long-term spectral shape as speech, the increase is on the order of 2-5 dB (Plomp, 1994). When the noise is amplitude-modulated (Eisenberg et al., 1995; Takahashi and Bacon, 1992), or is a

competing speaker (Carhart and Tillman, 1970), the RTS can increase by an even greater amount, 7-15 dB, for listeners with hearing loss.

While the most obvious symptom of hearing-loss is the increase in a listener's threshold, the problems associated with hearing-loss are much greater, making the identification of the exact cause of increased reception thresholds difficult. Amplification of the input signal to account for the increased thresholds can improve intelligibility, but does not completely restore the SRT (Peters et al., 1998; Bentler and Duve, 2000). In addition to increased thresholds, listeners with hearing-loss also suffer from reduced spectral resolution (Glasberg and Moore, 1986), reduced temporal resolution, as well as abnormal growth-of-loudness (Fowler, 1936; Steinberg and Gardner, 1937). All of these issues can potentially contribute to the increased SRTs.

Whether the loss of temporal resolution associated with hearing-loss is involved in the degraded SRTs is unknown. Speech perception in noise has been shown to correlate with temporal gap detection in listeners with hearing-loss (Dreschler and Plomp, 1985; Noordhoek et al., 2001), suggesting a link between the two. However, no relationship has been found between speech intelligibility and two other measures of temporal resolution: differences in masking of amplitude-modulated noise and forward/backward masking (Festen and Plomp, 1983) and temporal distortion introduced by wavelet transformations (van Schijndel, 2001).

Evidence suggests that reduced spectral resolution plays a larger role in the increased SRTs than does reduced temporal resolution. Speech intelligibility has been shown to be correlated with spectral resolution (van Schijndel et al., 2001; Leek and Summers, 1996; Dreshcler and Plomp, 1985). Decreased spectral resolution is the result of wider auditory

filters in listeners with hearing-loss; it is believed that these wider filters result in a decrease in the SNR of the internal auditory representation of the speech for listeners with hearing loss (Leek and Summers, 1996).

### 1.1.2 Noise-Reduction Techniques – Single Channel

Because of the complex, non-linear nature of the auditory system, the introduction of any sort of correction to account for the loss of the cochlear amplifier is difficult. The main attempt to correct the speech-in-noise problem has been the development of noise-reduction (NR) algorithms that attempt to increase the SNR of the incoming signal. Unfortunately, while many NR algorithms exist, most are not able to show an improvement in intelligibility when tested on listeners with hearing loss (Dillon and Lovegrove, 1993; Levitt et al., 1993; Levitt, 2001; Moore, 2003).

One of the classical methods of reducing noise in a system is through the application of a Wiener filter (Wiener, 1949). The Wiener filter is the optimal filter for a given noise and signal; to use this type of filtering, the statistics of both the signal and noise must be known, and both must be stationary. Unfortunately, the statistics of the speech and noise are generally unknown, and both are non-stationary, making the use of the Wiener filter difficult. Nonetheless, when Wiener filtering is used on speech corrupted by a known noise, intelligibility does not always improve (Levitt et al., 1993). A more generalized form of the Wiener filter is the Kalman filter (Kalman, 1960), which uses an adaptive system to allow for non-stationary signals and noise.

Most current NR algorithms employ spectral subtraction (Boll, 1979). Classically, the short-term power spectrum of the noise, which is assumed to be additive and stationary, is estimated and removed from the short-term power spectrum of the incoming

signal by subtraction. While the resulting signal has an increase in the SNR, the residual noise left by spectral subtraction has a "musical" quality. This musical noise is the result of mismatches between the short-term estimates of the noise's power spectrum and the instantaneous power spectrum of the noise. When subtracted, the differences between the estimated and actual power spectrum results in sinusoids of short duration that are randomly distributed across frequency and time.

Much of the current research into NR algorithms focuses on the enhancement of the spectral-subtraction technique and the removal of the musical noise that results from its application. To this end, NR algorithms have begun to incorporate perceptual properties of the human auditory system to aid in the removal of the musical noise (Tsoukalas et al., 1997; Virage, 1999; Arehart et al., 2003). These algorithms use auditory masking thresholds (AMT) to determine the parameters of the spectral subtraction. The musical noise is reduced by attempting to mask it with the energy contained within the speech signal. Of these studies, only Arehart et al. (2003) actually tested the NR algorithm on listeners, and obtained a small (2-8%) improvement in intelligibility.

Current hearing-aid technology allows for simple NR algorithms to be used. Most perform a simple version of spectral subtraction, wherein the gains of the separate frequency bands of the hearing-aid are varied in time to attenuate bands that do not have speech present. The absence or presence of speech is determined by examining each band for the amount of modulation present (Kuk et al., 2002; Edwards et al., 1998) or by attempting to measure a SNR for that band (Phonak, 2000). This type of NR is termed time-frequency gain manipulation, as the gains for each frequency band are changed with time. Research has shown mixed results with these sorts of NR algorithms, with some

showing improvement (Stein and Dempesy-Hart, 1984; Rankovic et al, 1992) and others none (Klein, 1989; Fabry and Van Tasell, 1990).

## 1.2 Objectives

The problem of understanding speech in noise is one that has received much attention, with very little progress made in improving the performance of listeners with hearing loss. The introduction of perceptual ideas to a NR algorithm poses many interesting opportunities for the enhancement of the algorithms. Current NR algorithms have begun to use the AMT to improve performance, but still rely on various statistical models for the separation of speech and noise. Many models exist, both psychophysical and physiological, that attempt to describe how the normal human auditory system detects various stimuli in noise. The overall objective of this thesis was to develop a NR algorithm using one of these models, the phase-opponency (PO) model (Carney et al., 2002), as the basis for separation of speech from background noise. The algorithm was then tested to determine whether it improved listeners' ability to understand speech-in-noise.

The PO model attempts to explain human listeners' ability to detect tones in noise; it uses the pattern of temporal information present in auditory neurons to perform this task. The second chapter of this thesis develops a simplified version of the PO model that captures the essential qualities of the model without use of the computationally expensive auditory-nerve model. The resulting PO detector is then compared to several classical detectors to determine its overall performance, as well as to demonstrate its ability to detect tones in conditions that are detrimental to other detectors.

The PO detector developed in Chapter 2 was deemed a good fit with the time-frequency gain manipulation strategy of NR. However, previous results have shown that this NR strategy was not always capable of increasing the performance of listeners. To examine whether this was the result of its implementation in previous studies, or a fundamental flaw of the strategy itself, experiments were performed to assess the maximum potential benefits that could be achieved using time-frequency gain manipulation. Rather than use a real-world detector, the gains were changed based on the speech in quiet, thus representing the ideal condition. Stimuli were processed using this ideal case, and presented to listeners to determine any benefit. How the gains changed with time and frequency was then systematically varied to determine the necessary detection performance of a speech detector to achieve the performance improvements seen with the ideal condition. These results are presented in Chapter 3.

In the fourth chapter, the previously developed PO detector was used in a complete NR algorithm referred to as the Phase-Opponent Noise-Reduction (PONR) algorithm. The results of testing the PONR algorithm on listeners with both normal hearing and hearing loss will be presented.

# Chapter 2

## Evaluation of a Signal-Processing-Based Phase-Opponent Detector

## 2.1 INTRODUCTION

The use of biologically motivated designs has become increasingly popular within the fields of engineering and computer science. Advances in computational power have allowed the creation of a wide variety of biologically based designs in areas such as neural networks, sensor design, and signal-processing algorithms. While detailed models of various biological systems exist, it is often possible to simplify these models to obtain a signal-processing algorithm that has many of the desired features of the original biological model. However, the detection of signals in noise is an area in which biological systems generally outperform current detection algorithms in adverse conditions. In this study, we implemented a signal-processing algorithm based on the mammalian auditory system.

Optimal solutions for the detection of a narrowband signal in noise have been known for some time (Kay, 1993; Van Trees, 1971; Whalen, 1971; Wiener, 1949; Hippenstiel, 2002). However, due to the difficulties in implementing these optimal detectors, non-optimal detectors are often used in practical applications. Non-optimal detectors are generally less complex than optimal detectors; the mathematical complexity of an optimal detector can require large amounts of computation for accurate detection [e.g., Kalman filtering requires several matrix inversions (Kalman, 1960)]. Non-optimal

detectors generally require less computational power yet often perform at levels comparable to those attained using optimal detectors.

To use an optimal detector under conditions of non-stationary noise, one must have either an accurate model or *a priori* knowledge of the noise statistics. In practical applications, *a priori* knowledge is not available, requiring the use of models and estimates of the noise. To obtain accurate noise estimates, however, the signal must be sampled for long periods of time, resulting in a system that can be sluggish. The requirement for noise estimation also increases the computational load for optimal detectors.

To contend with the problem of noise with fluctuating amplitudes or noise with unknown statistics, one can examine a system that handles it remarkably well: the human auditory system. Human listeners are able to detect narrowband signals in a vast array of noise conditions, including fluctuating noises (Kidd et al., 1989), and across a large dynamic range (Moore, 1997). While the exact physiological mechanisms behind the remarkable performance of the auditory system are unknown, many existing models can be used as the starting point for novel detectors that share some of the qualities of the auditory system. One such model is the phase-opponency (PO) model (Carney et al., 2002), which has recently been shown to accurately predict human performance for detection of tones in wideband noise of unknown amplitudes. This model forms the basis of the PO detector proposed here.

The PO model (Carney et al., 2002) is a physiological model for the detection of tones in the presence of additive white noise. Unlike traditional models of detection in the auditory system, the PO model relies on the temporal information contained in the

discharge patterns of auditory neurons. The model consists of two auditory neurons that are tuned to overlapping regions of the auditory spectrum, with the frequency to be detected falling in this overlapping region. Because the tone is within both neurons' frequency regions, both neurons synchronize their response to the tone. However, the two neurons are chosen such that the synchronization of their responses is 180° out-of-phase with each other, hence the name phase-opponency. When only noise is present, the outputs of the two neurons are partially correlated; the addition of a tonal signal results in the outputs becoming negatively correlated, as the two neurons synchronize to the signal. The PO model, because it relies on temporal information, is able to explain the robust performance of humans in detecting tones in noise of unknown amplitude.

The focus of previous work on the PO model was to explain human psychophysical performance. This study developed a signal-processing version of a PO detector and compared it to other detectors. The general form of a PO detector is shown in Fig. 2-1. The PO model is based on the phase-locking of auditory neurons, which rolls off at high frequencies (Johnson, 1980); the PO detector, however, is not limited in the frequencies it can handle.

The first stage of the detector is a pair of band-pass filters, with one filter tuned higher than the frequency band of interest, and the other tuned lower. Similar to the PO model, the pass-bands of the two filters overlap in the frequency region of interest. The exact magnitude responses of the band-pass filters are less critical than their phase responses, which differ by 180° (i.e., are phase-opponent) in the narrowband frequency

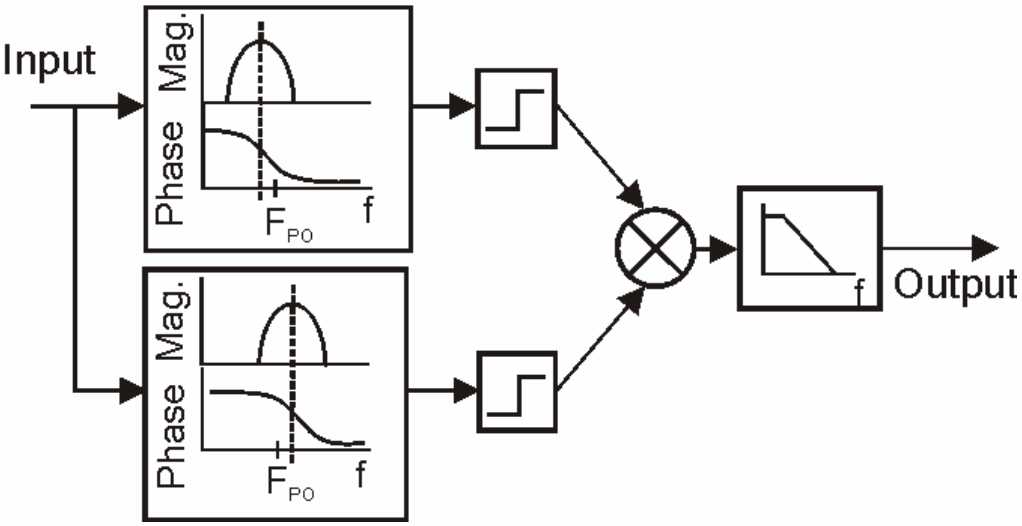


**Figure 2-1: General Phase-Opponent Detector**

The detector consists of two band-pass filters; one is tuned below the frequency of interest ($F_{PO}$), while the other is tuned above it. The exact spacing of the filters is such that the two filters have a phase difference of 180º at $PO_f$. The outputs of the two filters are subject to a hard-saturating non-linearity (a signum function) to remove all effects of level, leaving only the temporal information contained in the zero-crossings of the filtered signal. A running cross-correlation is performed on these saturated filter outputs by multiplying them together and low-pass filtering the resulting waveform. When a narrowband signal is present at $PO_f$, the output of the system will be driven towards -1; the output will fluctuate between -1 and +1 when only noise is present.

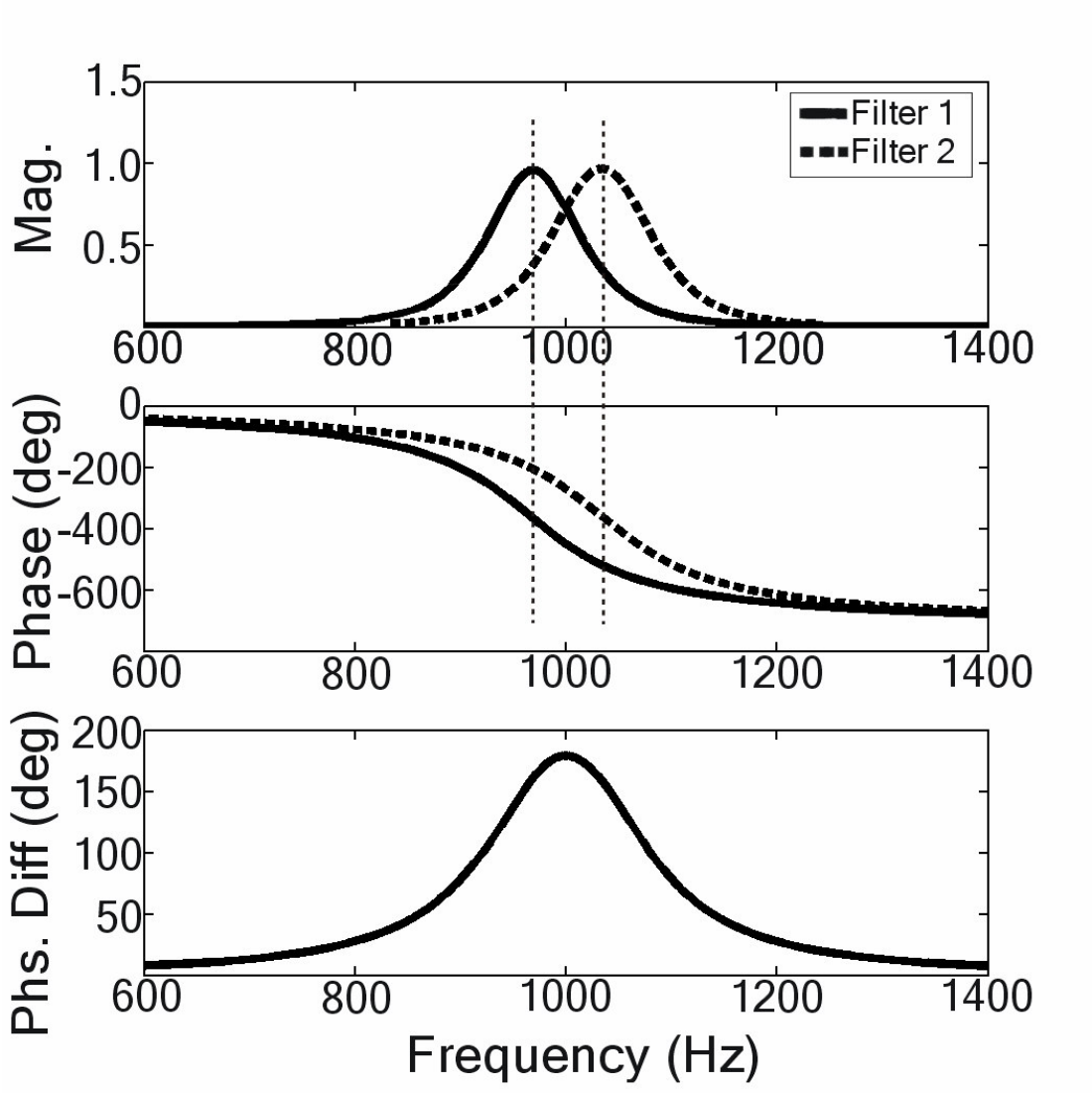


**Figure 2-2: Phase-Opponent Relation of Filters**

The two gammatone filters in the PO detector are arranged such that the difference in their phase functions is 180° at 1 kHz. The two filters' magnitudes are shown in the top panel, phases are shown in the middle panel, and the difference between them is shown in the lower panel.

region of interest (Fig. 2-2). As long as the zero-crossings of the outputs of the two band-pass filters are synchronized to the narrowband signal to be detected, the only important parameter of the two filters is their phase relationship at the signal frequency.

The output of each band-pass filter is subject to a hard-saturating non-linearity (Fig. 2-1). The non-linearity minimizes the magnitude information in the filter outputs, forcing subsequent components of the PO detector to rely solely on temporal information. The band-pass filter followed by the saturating non-linearity is a simplified version of the auditory neuron in the PO model.

The saturated outputs of the filters are multiplied together and low-pass filtered. The multiplication and low-pass filtering represent a running cross-correlation between the saturated outputs of the two band-pass filters. When only noise is present, the outputs are partially correlated because of the overlap between the two filters' pass-bands. When a narrowband signal is present, however, the two outputs are negatively correlated.

The PO detector was compared to two other detectors (Fig. 2-3). The non-coherent quadrature detector is an optimal detector for sinusoids of unknown phase (Robertson, 1967; Van Trees, 1971; Whalen, 1971; Kay, 1993). This detector is a generalization of a matched filter that attempts to detect the signal on the basis of the envelope of the sinusoid signal.

The energy detector is a simple detector that isolates a signal from noise by filtering the majority of the noise energy from the signal (Fletcher, 1940). The energy of the filter output is then used as a decision variable. To provide some robustness to detection in noise with unknown amplitude, the basic energy detector can be modified to include a mechanism for the estimation of the energy contained in the noise, which is then

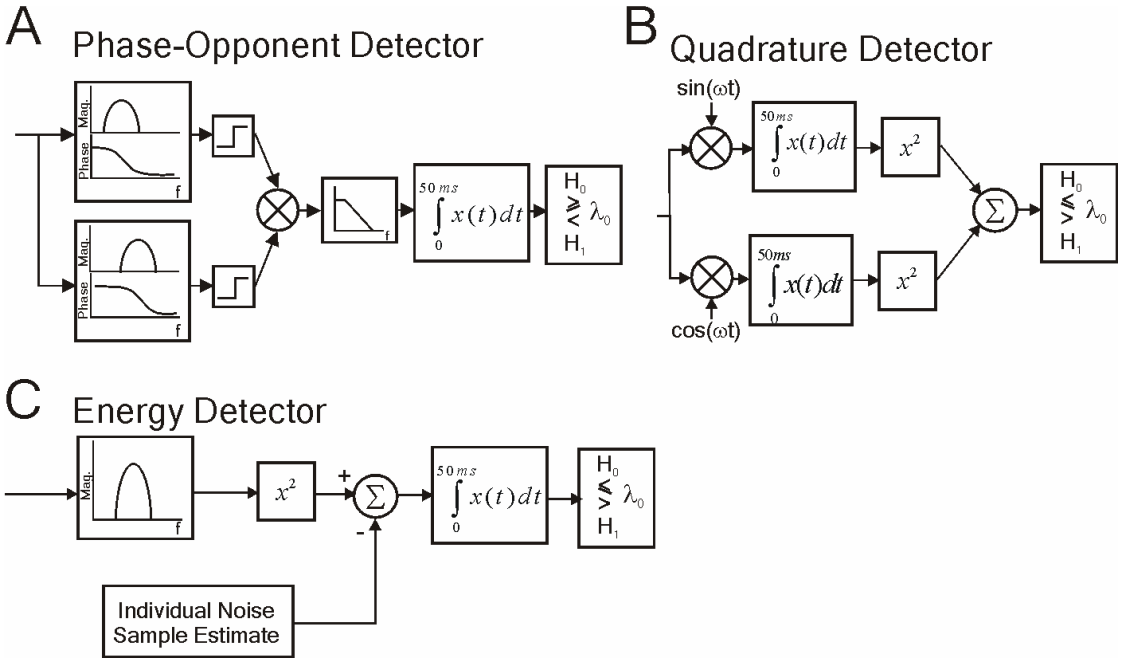


**Figure 2-3: Schematics of Proposed PO Detector and Common Detectors**

Schematic diagrams of the 3 detectors compared in this study. All detectors were designed to detect a 1-kHz tone pulse embedded in white Gaussian noise. The PO detector (A) consisted of two 4$^{th}$-order gammatone filters with bandwidths of 80 Hz. The low-pass filter had a cut-off of 1 kHz and was implemented as a 4$^{th}$-order Butterworth filter. The quadrature detector (B) was the optimal detector to which the PO detector was compared. Shown in (C) is an energy detector with constant-false-alarm rate noise estimation. It consisted of a 4$^{th}$-order gammatone filter with a bandwidth of 20 Hz, from which an estimate of the energy contained in the noise was subtracted. The estimate of the noise was obtained by taking the average energy of an independent noise that was matched to the input noise's spectrum level and duration. The noise estimation was performed separately for each noise token used.

subtracted from the energy at the output of the filter (Boll, 1979). This modification assumes that an estimate can be made from *a priori* information about the noise or from time periods of the input when it is assumed there is no signal present (Boll, 1979).

The comparison to the quadrature detector and energy detector with spectral subtraction was valid, because the PO detector, similar to these two detectors, does not require detection of phase information related to the signal.

# 2.2 METHODS

## 2.2.1 Detector Implementation

All of the detectors were digitally implemented on a PC in MATLAB with a sampling rate of 100 kHz. Block diagrams of the three detectors are shown in Fig. 2-3. In all detectors, the band-pass filters were based on 4$^{th}$-order gammatone filters (Patterson et al., 1988) of the form:

$$h(t) = At^3 e^{-\frac{t}{\tau}} \cos(\omega_0 t), t \geq 0$$

$$G(\omega) = 0.5A\tau^4 \left[ \frac{1}{[1 + j\tau(\omega - \omega_0)]^4} + \frac{1}{[1 + j\tau(\omega + \omega_0)]^4} \right]$$

where $\omega_0$ is the center frequency of the filter and $\tau$ determines the bandwidth of the filter. The filters were normalized to have 0 dB gain at $\omega_0$. This general form was chosen because the filter's response closely matches that of the auditory neuron, and the phase response of the filter is simply:

$$\angle(\omega) = 4\tan^{-1}[-\tau(\omega - \omega_0)]$$

All bandwidths specified for a given detector represent the bandwidth 3 dB down from the peak of the gammatone filter (located at $\omega_0$.)

For detection of a 1-kHz tone, the center frequencies of the two band-pass filters in the PO detector were 966.2 Hz and 1033.8 Hz, which resulted in the two filters having 180° of phase-difference at 1 kHz. The filter bandwidths were 80 Hz ($\tau$ = 0.002 sec), which produced the best detection performance for a 50-ms tone. The hard saturating non-linearity was a signum function; the low-pass filter was a $4^{th}$-order Butterworth filter with a cut-off frequency of 1 kHz. The 1 kHz cut-off was chosen to remove the 2 kHz component that results from the multiplication of the two band-pass filter outputs while still allowing the output of the multiplication to change rapidly. The output of the low-pass filter was integrated over 50 ms, which was the entire duration of the signal to be detected.

The energy detector with spectral subtraction consisted of a $4^{th}$-order gammatone filter centered at 1 kHz. The bandwidth was 20 Hz ($\tau$ = 0.0078 sec). This bandwidth was chosen to capture the energy contained in the main lobe of a 50-ms tone burst (the signal that was most often used for analyzing performance). The energy detector with spectral subtraction simply squared the output of the band-pass filter and subtracted from it an estimate of the noise energy. This estimate was based on the mean energy in an independent sample of noise that was matched to the spectrum level and duration of the input noise. The estimation was performed for each token individually; a global estimate of the noise was not used.

The quadrature detector multiplied the input by sine and cosine waves of amplitude 1 and frequency 1 kHz. The two separate branches were then integrated over 50 ms (the

duration of the signal to be detected), squared, and summed (Robertson, 1967; Whalen, 1971).

For all three detectors, the decision variable was the last sample of the output of each detector. Because each detector integrated its output over the entire duration of the signal (50 ms), the last sample was all that was necessary. The criterion value was chosen based on the overall distributions of decision variables for each detector, and was different for each detector. While the energy detector with spectral subtraction estimated the noise for each noise token, the criterion value to which the output was compared remained fixed for all noise tokens.

The signal to be detected was a 1000-Hz sinewave, the amplitude of which was varied with respect to the noise amplitude to provide a given SNR. Noise spectrum levels ranged from 0 to 100 dB re 1 V, depending on the simulation. The duration for both signal and noise was 50 ms, gated with a rectangular window. Noise was generated using the Gaussian random number generator provided in MATLAB.

### 2.2.2 ROC Curve Generation

To analyze the performance of the detectors, receiver-operator characteristic (ROC) curves were generated by simulation. While analytical expressions for the energy and quadrature detectors are known, the non-linearity in the PO detector hampers the determination of an analytical expression for its performance.

The responses of the detectors were simulated in response to 10,000 noise-only trials and 10,000 signal-plus-noise trials. The SNR for the signal-plus-noise trials was kept constant, and the same noise tokens were used for all three detectors. The probability of false alarm and probability of detection were determined as a function of the criterion for

detection. The probability of detection for a given criterion was the proportion of responses to the signal-plus-noise that were above the criterion. Similarly, the probability of false alarm was the proportion of responses to the noise alone that were above the criterion. By systematically varying the criterion across the entire distribution of detector outputs, a complete ROC curve was generated. Under each simulation condition (described in the next section), ROC curves were generated for SNRs ranging from -5 dB to 25 dB in 1 dB steps. SNR was measured as the ratio of the energy of the signal to the spectrum level of the noise sample.

### 2.2.3 Simulations

To compare the performance of the PO detector to that of the energy and quadrature detectors, the performance of all three were measured under three noise conditions. The three conditions were known-level Gaussian noise, unknown-level Gaussian noise, and amplitude-modulated Gaussian noise. The detectors were compared, as a function of SNR, at a constant false-alarm rate ($P_{FA} = 0.01$). The noise spectrum level was varied from 0 to 100 dB re 1 V in 10-dB steps to examine the effects of noise spectrum level on the criterion value needed to provide a constant $P_{FA}$ of 0.01 as a function of noise spectrum level.

For the known-level case, the noise was Gaussian white noise with a constant spectrum level. The spectrum level of the noise was fixed at 50 dB re 1 V, and the amplitude of the 1-kHz tone was varied to achieve SNRs of -5 to +20 dB. The criterion values used for each detector were chosen based on the distribution of the noise-only detector outputs to give a $P_{FA}$ of 0.01. The criterion values chosen were specific to each detector.

After baseline performance was established for the known-level case, the level of the noise was varied within the 10,000 tokens used to generate the ROC curve for the detectors. The level of each noise token was varied by randomly selecting a spectrum level for that particular token, unlike the known-level condition, which had a fixed spectrum level. The spectrum level was altered by generating a Gaussian white noise with unity variance, and scaling it to achieve the desired spectrum level, which is related to the overall level. Because the spectrum level (and RMS) changed for each token, this represents the unknown-level condition.

The amount of variation in the noise amplitude was determined by how the spectrum level was chosen for each noise-token. The spectrum level of each noise token was chosen from a uniform distribution centered around 50 dB re 1V. By changing the width of the uniform distribution, the amount of variation across tokens in the noise level could be varied . Three distribution widths were used: 15 dB, 30 dB, and 45 dB. As an example, for the 30-dB width condition, the spectrum level of any individual noise token could vary between 35 and 65 dB re 1 V. In all cases, the SNR for the 10,000 tokens was kept constant (*i.e.*, once the noise spectrum level was chosen, the corresponding tone burst's amplitude was adjusted to keep the same SNR). This unknown–level–with-constant-SNR condition, though uncommon in detection literature, is a common psychophysical task known as a roving-level condition (Kidd et al., 1989). As in the known-level condition, the criterion value for each detector was specific to that detector, and held constant for that particular condition (i.e. each detector had a separate criterion for the 15, 30 and 45 dB level-range conditions). In each case, the criterion value was chosen based on the noise-only distribution to achieve a $P_{FA}$ of 0.01.

The last condition involved a known-level Gaussian white noise that was amplitude modulated by a low-frequency sine wave. Again, families of ROC curves were generated for each detector with a 50 dB re 1 V spectrum level Gaussian noise that was fully modulated. Modulation frequency ranged from 0 Hz (no modulation) to 100 Hz in 5-Hz increments. Detectors were again compared for a constant $P_{FA} = 0.01$.

## 2.3 RESULTS

### 2.3.1 PO Detector

An example of the time-varying output of a PO detector before integration is shown in Fig. 2-4. To demonstrate the response of the detector to both noise and signal plus noise, the illustration reflects an input consisting of a 100-ms, 1-kHz tone burst centered in 300 ms of Gaussian noise. When noise alone was present, the output of the detector fluctuated between the limits of the detector (+1 and -1), with both highly positively correlated outputs (when the detector output was near +1) and highly negatively correlated outputs (when the detector output was near -1). While the output of the detector fluctuates during the noise-only periods, the expected value of the output during these periods was close to zero. The expected value was determined by the correlation between the two filters used in the detector. For the gammatone filters implemented here, the expected value was computationally estimated to be 0.0048 V.

When the 1-kHz tone burst was present, it dominated the temporal properties of the two phase-opponent filters' outputs. This resulted in a high negative correlation between the filter outputs, which resulted in a decrease in the average value of the PO detector's output. This decrease can be seen in Fig. 2-4 as a saturation toward -1. Once the tone

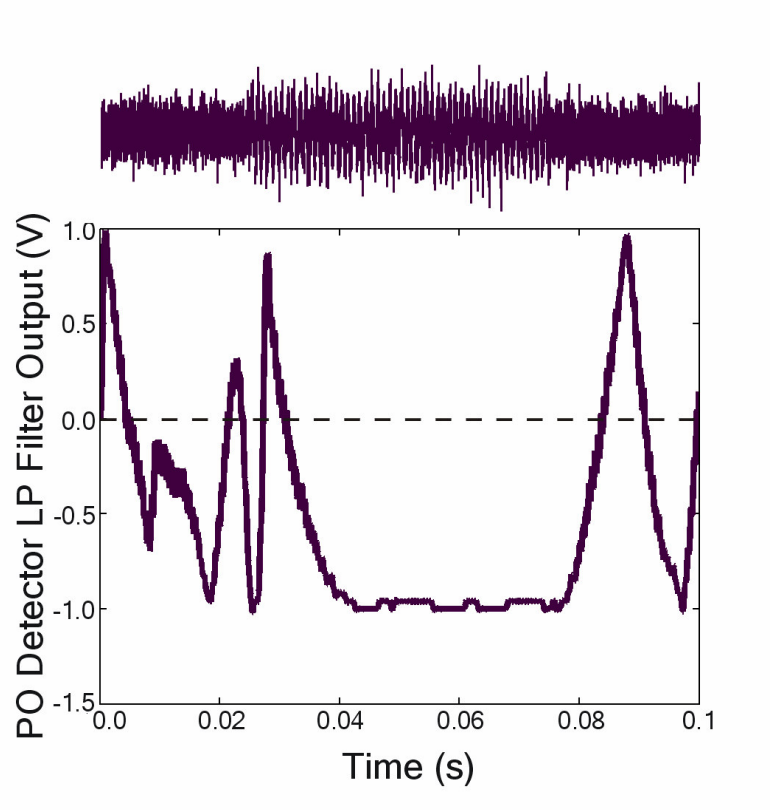


**Figure 2-4: Example Output of the PO Detector Before Integration**

When only noise was present, the output of the detector (before the integration) fluctuated across the entire range of the system (-1 to 1). The presence of the tone resulted in a saturation of the output toward -1, as the output of the two phase-opponent filters was dominated by the tone, leading to a highly negative correlation between the filters. The sluggishness of the system was determined by the cut-off frequency of the low-pass filter at the output of the detector. Decreasing the cutoff frequency would result in a smoother output during the noise-only periods that would be slightly negative (see text).

burst ended, the output of the detector again began to fluctuate from -1 to 1, with an expected value of approximately 0.

From the PO detector responses, ROC curves were generated. Subsets of these curves are shown in Fig. 2-5 for a constant-level white Gaussian noise with varying SNRs. Although the PO detector contains non-linearities, the ROC curves follow the expected shape for a linear detector. The PO detector performed reasonable detection at adverse SNRs over the range of $P_{FA}$ simulated. ROC curves were also derived for the two other detectors, and had the same overall form as the ROC curves of the PO detector shown in Fig. 2-5. It is from ROC curves such as these that the rest of the data shown for all detectors was derived.

### 2.3.2 Known- and Unknown-level Gaussian Noise

The performance of the three detectors is shown for the known-level noise condition in Fig. 2-6(a). The probability of detection is plotted as a function of SNR for a constant $P_{FA} = 0.01$. As expected, the performance of all of the detectors improved with increasing SNR, from chance at -5 dB to perfect detection at 15 dB. The quadrature detector's performance was better than that of the energy detector with spectral subtraction and PO detectors for this condition; this was expected, as the quadrature detector is an optimal detector for a sinusoid of unknown phase in known-level noise. The PO detector, however, performs within 5 dB of the optimal quadrature detector and within 3 dB of the energy detector with spectral subtraction. The general trends of all detectors' performance were similar, with the energy and PO detectors' performance appearing to be a simple shift of the quadrature detector's performance to slightly higher SNRs (3 dB for the energy detector with spectral subtraction, 5 dB for the PO detector). The

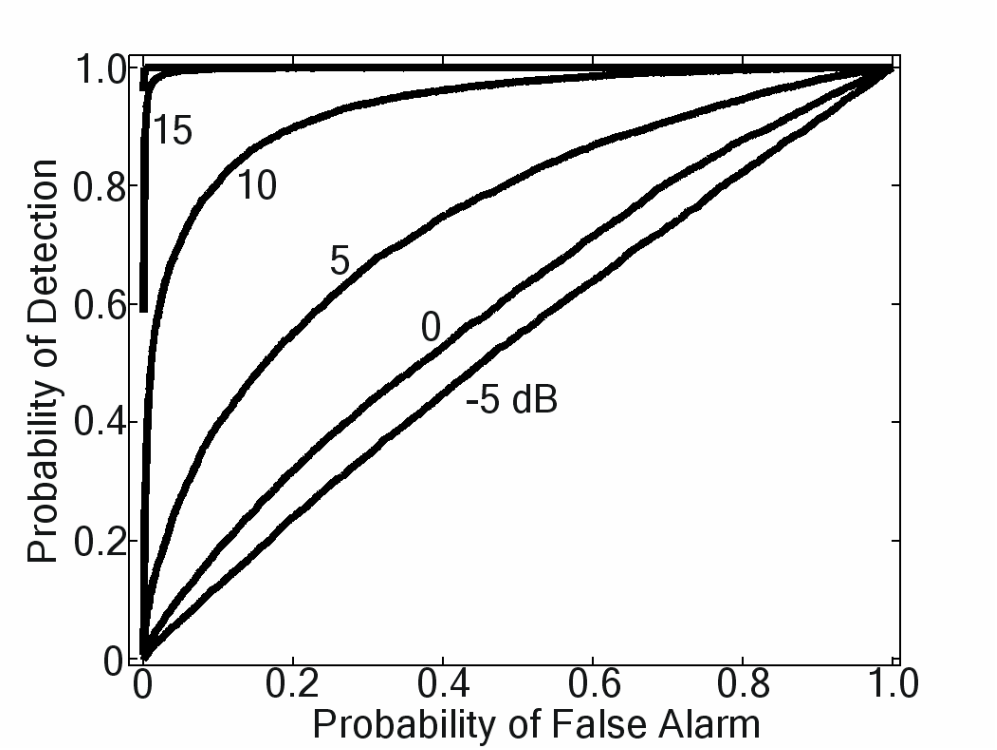


**Figure 2-5: PO Detector Receiver-Operator Characteristic (ROC) Curves**

These ROC curves are the result of MATLAB simulations of the PO detector. The curves range from chance (-5 dB) to perfect detection (above ~10 dB). Although the detector was non-linear, the ROC curves followed the expected shape for a linear detector. Performance in all noise cases was determined by taking a vertical cut at $P_{FA}$ = 0.01 and replotting $P_D$ as a function of SNR.

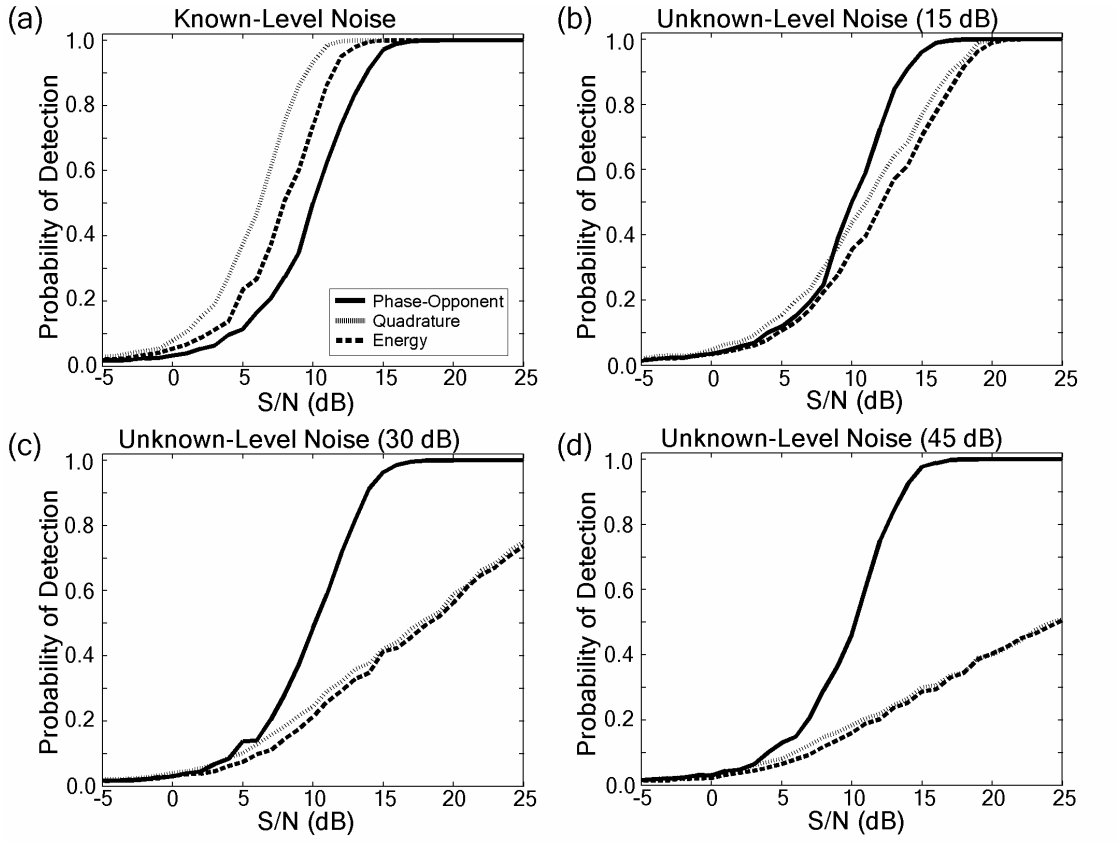


**Figure 2-6: Effect of Known and Unknown Noise Spectrum Level on Detection**

The curves of (a) were generated with a fixed, known noise spectrum level of 50 dB re 1 V. Panels (b-d) show the performance when the noise spectrum level was unknown and drawn from a rectangular distribution centered around 50 dB, with an increasing width (15 dB, 30 dB, and 45 dB for (b), (c), and (d), respectively). In all panels, the performance improved with increasing SNR. Comparing (a) to (b-d), it is obvious that the PO detector's performance was unaffected by the unknown noise amplitude conditions.

performance of the simulated quadrature detector matches the analytical solution for a quadrature detector.

Introducing an unknown-level noise, in which case the spectrum level randomly varied from trial-to-trial, resulted in the performances shown in Fig. 2-6(b-d) for the three detectors. The criterion that was chosen to achieve a $P_{FA}$ equal to 0.01 for each detector was based on each detector's output distributions. This criterion was allowed to change for each distribution of noise levels (i.e. 15, 30, and 45 dB ranges of noise level). Again, all detectors' performances improved with increasing SNR, as expected. However, the performances of both the quadrature detector and energy detector with spectral subtraction were affected by the wider distributions of noise level, while the PO detector was unaffected. As the width of the noise-level distribution increased (with a correlated increase in the signal-amplitude distribution due to the constant-SNR condition), the performances of the quadrature detector and energy detector with spectral subtraction degraded. While the SNR for which the probability of detection began to increase remained constant for all three detectors and all four conditions (~0 dB), the slopes of the $P_D$ vs. SNR functions for the quadrature detector and energy detector with spectral subtraction decreased as the width of the noise-level distribution increased (Fig. 2-6b,c,d). For a known-level noise , the slope was 11.3 %/dB; the slope decreased to 6.8 %/dB for a 15 dB range of levels, 3.4 %/dB for a 30 dB range, and 2.2 %/dB for a 45 dB range.

Comparing the panels of Fig. 2-6, the performance of the PO detector was unaffected by the unknown-level conditions. The detector's performance remained constant, with the detector showing improvement at 5 dB and reaching nearly perfect detection at

approximately 15 dB SNR. This robustness across different noise levels was the result of the PO detector's reliance on temporal information; changing the overall level of the noise while keeping the SNR constant had no effect on this temporal information.

### 2.3.3 Noise-Level Dependence of Criterion

Variations in noise level of the input had a profound effect on both the quadrature detector and energy detector with spectral subtraction, whereas the PO detector was unaffected (Fig. 2-6a-d). The effect can be examined by plotting the criterion value required for a constant $P_{FA}$ of 0.01 as a function of noise spectrum level (Fig. 2-7). The trend for both the quadrature detector and energy detector with spectral subtraction was for the criterion value to increase in proportion to the noise spectrum level, while the PO detector's criterion value remained constant across noise levels. As expected, the criterion value for the energy detector with spectral subtraction grew in direct proportion to the energy in the noise (*i.e.*, had a slope of 2 on a log-log axis), whereas the criterion value for the quadrature detector grew in direct proportion to the magnitude of the noise (*i.e.*, had a slope of 1 on a log-log axis).

### 2.3.4 Amplitude-Modulated Gaussian Noise

Fig. 2-8 shows results for an additive noise that was sinusoidally amplitude-modulated. The SNR was held constant at 10 dB, which represented a point on each detector's $P_D$ vs. SNR curve at which the performance was changing as a function of SNR. When the Gaussian noise was modulated, the performance of the PO detector and energy detector with spectral subtraction changed from that shown in Fig. 2-6(a). The quadrature detector's performance remained constant across modulation frequency.

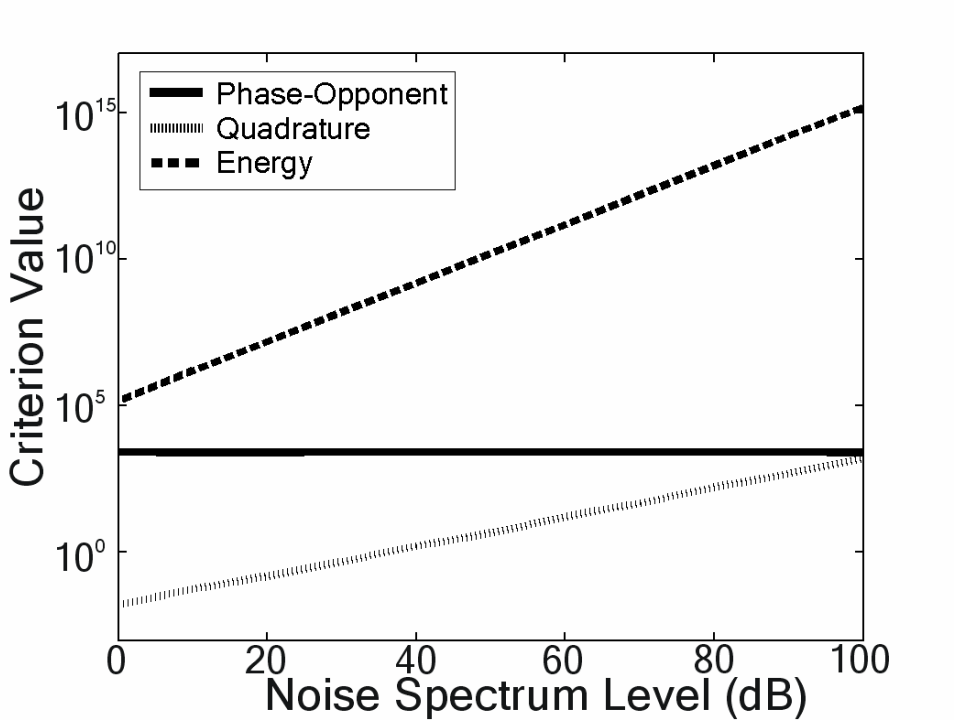


**Figure 2-7: Effect of Noise Spectrum Level on Detector Criterion**

The criterion levels to achieve a $P_{FA}$ of 0.01 are shown for all three detectors as a function of the noise spectrum level. The SNR was kept constant at 10 dB, a point that was within the transition range of all detectors. The PO detector, because of its reliance on temporal information, maintained a constant criterion across spectrum level. The quadrature and energy detectors' criteria, as expected, increased in proportion to the noise spectrum level.

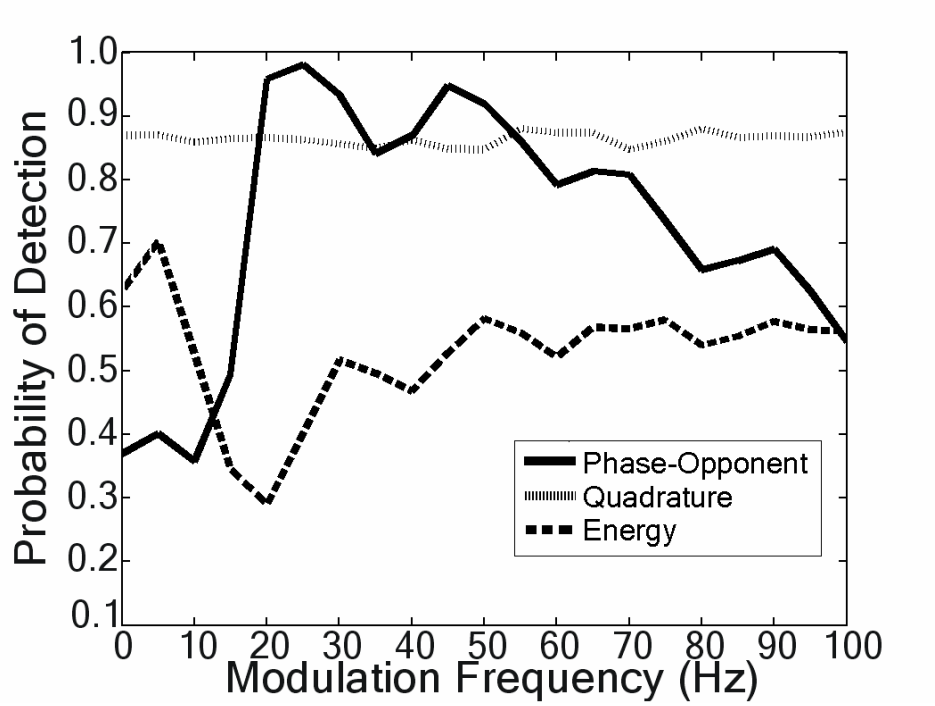


**Figure 2-8: Conditions of Amplitude-Modulated Gaussian Noise**

The probability of detection is shown as a function of the modulation frequency for all detectors. The modulated noise was a 50 dB re 1 V Gaussian noise, and the SNR was kept constant at 10 dB. The PO detector demonstrated an improvement with modulation. This improvement decreased with increasing modulation frequency, as the sluggishness in the detector began to affect the detector's ability to take advantage of the temporary increases in SNR caused by the modulation. The energy detector's performance was degraded by the modulation, because the long noise estimation time did not allow it to take advantage of the SNR increases during dips in noise amplitude. The quadrature detector was unaffected by the modulation. In all cases, the $P_{FA}$ was kept constant at 0.01.

The PO detector's performance improved with low modulation frequencies, reaching almost perfect detection at a modulation frequency of 25 Hz. As the modulation frequency increased, the performance of the PO detector gradually degraded. The PO detector was able to take advantage of the "dips" in the completely modulated noise to improve its performance. As the "dips" became more rapid (i.e., as modulation frequency increased), the sluggishness of the detector resulted in a return toward baseline performance.

Unlike the PO detector, the performance of the energy detector with spectral subtraction dropped when the noise was amplitude-modulated. Like the PO detector, however, the performance appeared to return to baseline with increasing modulation frequency. Because it estimated the energy over a long period of time (50 ms), the energy detector with spectral subtraction was unable to take advantage of the short "dips" in the noise level caused by modulation.

## 2.4 DISCUSSION

### 2.4.1 Performance for Known- and Unknown-Noise Level

When compared to both the quadrature detector and an energy detector with spectral subtraction, the PO detector had better performance for the unknown-noise level condition and for low-frequency amplitude-modulated noise. These are the conditions in which the PO detector was designed to operate, as it uses the temporal information encoded within the incoming signal which is less affected by overall level than is the energy information. The PO detector was able to perform within 8 to 10 dB of the optimal detector for the known-level noise condition, even though the primary design

was for unknown level, and the model upon which it was based is non-optimal for these conditions. While the PO detector implemented here was optimized for detection by varying the bandwidth of its two gammatone filters, it is possible that the PO detector's performance for the known-level noise condition could be further improved by optimizing the shape of the two band-pass filters.

The quadrature detector performed poorly under conditions of unknown noise level because it did not contain mechanisms for the estimation or removal of noise other than the narrowband filters in the detector. An unknown noise level resulted in wider distributions of detector responses for both the noise-only and signal-plus-noise distributions that were used to calculate the ROC curves; these wider distributions appeared in the $P_D$ vs. SNR plots as an increase in the range of SNRs over which the performance of the quadrature detector changed.

Unlike the quadrature detector, the energy detector with spectral subtraction included a simple mechanism for the estimation and removal of the noise. However, the performance of the energy detector with spectral subtraction still degraded when the noise level was unknown or fluctuating. While the detector's decision-variable distribution for noise-only trials remained fairly constant because of the subtraction of the baseline noise energy, the standard deviation of the signal-plus-noise distribution increased. This increase was because the SNR was kept constant and the signal level was varied in proportion to the noise level [to match the psychophysical roving-level condition (Kidd et al., 1989)]. Increasing the width of the noise level distributions resulted in a larger range of signal amplitudes, resulting in a wider distribution of detector responses.

The PO detector was unaffected by noise level. This result was expected, because the PO model upon which the detector was based is similarly unaffected by unknown-level noises. The temporal information used by the PO detector was encoded in the zero-crossings of the outputs of its two filters. The locations of these zero-crossings were determined by the SNR of the input, not by the overall levels of the signal or noise. Thus, scaling the two in proportion resulted in no change in the zero-crossings.

### 2.4.2 Performance with Amplitude Modulation

The PO detector's performance improved with amplitude modulation because the detector was able to take advantage of the temporary increases in SNR during modulation. During the "dips" in the envelope, the zero-crossings were dominated by the signal, allowing the PO detector response to saturate at -1. The decrease back to baseline with increasing modulation frequency was due to the sluggishness of the PO detector's low-pass filter. As the dips became smaller, the smoothing provided by the low-pass filter resulted in an inability to take advantage of the SNR increase. Increasing the cut-off frequency of the low-pass filter would result in a slower return to baseline, but would also allow the output to fluctuate more. An increase in the amount of fluctuations would result in a decrease in the probability of detection because detection is based on an estimate of the detector output's expected value.

The performance of the energy detector with spectral subtraction was degraded by amplitude modulation because it estimated the noise energy over a 50-ms window, which was too long to take advantage of quick increases in the SNR. Decreasing the amount of time over which the detector estimated the noise would result in an improvement in performance for low-frequency modulated noise at the cost of degraded performance in

unknown-level noise. The smaller estimation time would result in a less accurate noise estimate, which would change the distribution of the output for the energy detector with spectral subtraction in the unknown-level noise condition.

### 2.4.3 PO Detector Performance

The PO detector's criterion value was unaffected by the noise spectrum level. This allowed the specification of a criterion that provided a constant false-alarm rate for Gaussian noise. This is an important advantage, as the PO detector can act as a constant-false-alarm detector without extensive design effort. The appropriate criterion value can be determined by calculating the correlation of the outputs of the two PO filters in response to a given noise or by simulation.

The PO detector, however, has some limitations in its detection ability as compared to an energy detector with spectral subtraction; the signal to be detected must have a clearly defined temporal structure for the PO detector to utilize in detection. An energy detector requires only that the energy of the signal be discernable.

The robustness of the PO detector to variations in input noise spectrum level and to amplitude modulation suggests its use in situations where these conditions occur. We are currently working on adapting the PO detector for the detection of low-frequency, narrowband components of speech. The level-invariance of the PO detector is ideal for speech, as the overall level can range over the entire dynamic range of the human auditory system. In addition, the PO detector is able to take advantage of the "dips" that occur in the natural fluctuations of noise that often accompany speech.

The performance of the PO detector suggests that mimicking the mechanisms of biological sensory systems can provide detectors that outperform classical detectors in

several ways. Sensory systems are generally well suited for detection of the many signals that are encountered in real life and have been fine-tuned by evolutionary forces into powerful detectors. Digital implementations of these detectors, and other biological processes, can lead to useful algorithms that provide many of the benefits of the biological system. While these implementations are often simplified versions of the biological system, they can provide the basic functionality of the system while at the same time overcoming some of the limitations faced in biological systems.

# Chapter 3

# Determination of the Potential Benefit of Time-Frequency Gain Manipulation

## 3.1 INTRODUCTION

One of the greatest problems facing the listener with hearing loss is understanding speech in the presence of background noise. While current hearing-aid technology has done much to improve the audibility and comfort in noisy backgrounds for listeners with a hearing loss, little has been done to increase the intelligibility of speech in a noisy environment. The current study examined the potential benefits for intelligibility of a general noise-reduction (NR) strategy known as time-frequency gain reduction. Studies using this strategy have shown mixed results in intelligibility, with some showing benefits (Stein and Dempesy-Hart, 1984; Rankovic et al, 1992), and others showing none (Klein, 1989; Fabry and Van Tasell, 1990). The results of these studies, however, were based on different methods of adjusting the time-frequency gain profile. The study performed here utilized an ideal binary mask to determine the maximum benefits that could be gained by varying the time-frequency gain profile. The ideal binary mask is the time-frequency profile based only on the speech in quiet; it represents the output of an ideal detector of speech components. The use of the ideal mask allows for the evaluation of the time-frequency gain manipulation strategy without the influence of the actual detectors that may have affected previous studies.

Quantitatively, the performance of listeners in understanding speech in noise can be measured using a reception threshold for speech (RTS). The RTS is a measure of the signal-to-noise ratio (SNR) that is required to achieve a preset level of intelligibility, generally 50 or 100 percent (Moore, 2003). The RTSs of listeners with hearing loss are increased relative to normal listeners. For speech-spectrum shaped noise, the increase is 2-5 dB (Plomp, 1994), while the RTS increases 7-15 dB when the noise is amplitude modulated (Takahashi and Bacon, 1992; Eisenberg et al., 1995) or is a competing speaker (Carhart and Tillman, 1970).

A large amount of research has gone into the development of algorithms to perform noise-reduction, with the goal of restoring the RTSs of listeners with hearing loss to that of normal-hearing listeners. This goal is two-fold: to restore lost intelligibility and to improve the quality of noisy speech (Schum, 2003). The general consensus as to the fulfillment of these goals is that single-microphone NR systems perform the second goal of improving quality, without an increase in intelligibility (Levitt, 2001; Chabries and Bray, 2002; Schum, 2003). The only NR strategy that meets the first goal is that of directional microphones (Levitt, 2001; Schum, 2003, Chabries and Bray, 2002). While directional microphones can create a large increase in intelligibility when examined in a laboratory setting, the presence of reverberation in real-world listening environments limits the RTS improvement to a few decibels (Hawkins and Yacullo, 1984; Ricketts, 2000; Ricketts and Hornsby, 2003). In addition, the directional microphone requires that the noise and speech be spatially separated to achieve an increase in RTS.

Single-microphone NR systems generally attempt to increase intelligibility by increasing the SNR of the speech that is presented to the listener. Several general

strategies exist for doing this, such as Wiener filtering (Wiener, 1949), spectral subtraction (Boll, 1979), adaptive filtering (Graupe et al, 1987), speech synthesis (McAulay and Quatieri, 1986; Kates, 1994) and time-frequency gain manipulation (Ono et al, 1983; Dempsy, 1987; Klein, 1989; Fabry and Van Tassell, 1990; van Dijkhuizen et al, 1987, 1989, 1990, 1991; Rankovic et al, 1992). Wiener filtering involves estimating the characteristics of the signal and the noise, and creating a filter that optimizes the SNR of the output based on these characteristics. Under certain conditions, this filtering is similar to other single-microphone NR strategies. In spectral subtraction, the spectrum of the noise is estimated and subtracted from the noisy signal, leaving only the spectrum of the speech (Boll, 1979). Adaptive filtering is similar to Wiener filtering, but involves utilizing a time-varying filter that is varied based on the difference of the output and a noise-estimate. In speech synthesis, the signal is replaced by speech that has been synthesized based on the speech that is detected in the original signal, often by replacing the noisy frequency bands with sinewaves matched in amplitude and frequency (McAulay and Quatieri, 1986; Kates, 1994). All of these strategies, however, rely on having an accurate model of the noise and/or speech. These models are difficult to create or measure, as the type of noise in practical settings can vary greatly depending on the listening environment, and speech is a complex signal that varies with each individual speaker. While the use of all of the above strategies has been shown to improve the SNR of the noisy signal, there is no increase in intelligibility because of the distortions that are added by the processing (Boll, 1979; McAulay and Quatieri, 1986; Kates, 1994; Levitt, 2001, Schum, 2003).

Time-frequency gain manipulations can increase the SNR and intelligibility in some, but not all, cases (van Dijkhuizen et al., 1989; van Dijkhuizen et al., 1990, Rankovic et al., 1992. In this strategy, the gains of each frequency band are time-varying; when the SNR within the band is high, the gain is high, when the SNR is low, the gain for that frequency band is reduced. This strategy has been shown to be effective when the noise is limited to one frequency band (Rankovic et al, 1992), but the results conflict when the noise is wideband (Ono et al, 1983; Klein, 1989; Fabry and Van Tasell, 1990). However, the implementations of the strategies have been different, making comparison of the general strategy difficult. It is here that the ideal binary mask becomes useful, as it is a time-frequency mask that is based on the clean speech, thus allowing evaluation of the general strategy without the influence of the actual detectors or SNR measures that may have limited previous studies.

The binary mask effectively separates the signal of interest from interfering signals by reducing the gains for frequency bands when only the interfering signal is present. As its name suggests, in a binary mask the gain of each frequency band is allowed to change between two states (an "on" and "off" state). The ideal binary mask is based on the uncorrupted copy of the signal of interest; it represents an unrealistic condition, as the uncorrupted signal is generally not available. However, the ideal binary mask allows evaluation of the general strategy, and represents the upper limit of noise reduction that the strategy can produce. In addition, the time-frequency gain patterns produced by any system that attempts to detect the signal of interest can be compared to the ideal binary mask to allow a quantitative comparison. Binary masks have been shown to be effective in increasing the performance of automatic speech recognition in noise (Cooke et al.,

2001; Srinivasan et al, 2004), as well as for separating acoustical sources (Roman et al, 2003; Yilmaz and Rickard, 2004).

Application of the ideal binary mask results in several modifications of the stimulus properties that may lead to an improvement in intelligibility, depending on the parameters of the binary mask and the method of applying the time-varying gains. If the frequency resolution of the binary mask is finer than that of the impaired auditory system, the SNR of each auditory band would be improved by the reduction of noise within the frequency bands of the binary mask. For example, if the binary mask has two frequency bands per auditory band, the application of the ideal binary mask could improve the SNR by a factor of 2. This SNR increase should improve intelligibility, provided that the distortions introduced by the binary mask are not detrimental. If however, there is a match between the binary mask and auditory bandwidths, the instantaneous SNR of the auditory band would not be changed, as the binary mask does not change the frequency band when the signal is present. Articulation index (AI) theory (French and Steinberg, 1947; ANSI, 1969; Pavlovic, 1988; Mueller and Killion, 1990) would suggest that the SNR of each frequency band determines the intelligibility of the sentence; if there is no change in the bands' SNRs, there should be no increase in intelligibility.

The reduction of gains in frequency bands in which there are no speech components would also lead to a decrease in the spread of masking in adjacent bands. Listeners with hearing loss have been shown to have an increased upward spread of masking (Trees & Turner, 1986; Klein et al., 1990). By reducing the amount of masking caused by adjacent frequency bands, the internal neural representation of the signal might show an effective increase in SNR.

The application of the binary mask may also enhance certain speech cues, such as the onset and offsets of speech components. At adverse SNRs, the ideal binary mask acts to shape the noise that is present within each band with the envelope of the speech components within that band. Provided there are greater than four frequency bands covering the range of speech, these envelope cues have been shown to be all that is necessary for comprehension of speech for normal-hearing listeners (Shannon et al., 1995).

The use of the ideal binary mask on noisy speech was hypothesized to lead to an improvement in intelligibility, as indirectly measurd using the Hearing-in-Noise Test (HINT) (Nilsson et al., 1994), compared to unprocessed stimuli. The HINT provides an estimeate of the RTS, which is the SNR needed to for the listener to correctly identify 50% of the words in a given sentence. This hypothesis was shown to be valid, leading to a second experiment involving degraded versions of the ideal binary mask.

The second experiment examined how the HINT RTSs changed as a function of the degradation of the ideal binary mask. While the results of Experiment I showed that time-frequency gain manipulation improved the RTSs, Experiment II was designed to determine how well the detector must perform to realize the improvements shown in Experiment I. These degraded versions of the ideal mask represent the performance of a real-world detector; the real-world detectors are modeled as being able to detect only a fixed percentage of the total energy of the speech stimulus (as opposed to the ideal detector, which detected all of the energy). With knowledge of how the detectors should perform, it is then a matter of choosing an appropriate detector that can produce a

detection pattern similar to the binary masks found in Experiment II that lead to an improvement in RTS.

After performing these two experiments, a third experiment was performed that further analyzed the parameters of the binary mask necessary for improvement in RTS. The first two experiments had a fixed frequency- and time-resolution of the binary mask; in Experiment III, the resolution of the binary mask was altered to examine the effect of the frequency-resolution and temporal smearing on the results of Experiment I. The hypothesis for Experiment III was that a reduction in the frequency-resolution or the temporal smearing of the binary mask would result in a reduction of the benefits of the ideal binary mask shown in Experiment I.

## 3.2 METHODS

### 3.2.1 Stimuli

The stimuli consisted of 250 sentences and 250 noise samples that were obtained from the HINT (Nilsson et al., 1994). The sentences and noise were combined at SNRs ranging from -10 dB to +7 dB and passed through one of the experimental noise-reduction algorithms. Noise-reduction was obtained by manipulating each band separately, and then combining the outputs of all bands to obtain the final output. The gains were based on the ideal binary mask, which resulted in the gains being switched between two values based on whether or not a speech component was present; there was no smoothing of the gains or filtering to remove the sudden onsets and offsets of the changing gains.

The algorithm made use of an analysis-synthesis Gammatone filterbank (Hohmann, 2002) that separated the stimuli into frequency bands that were equally spaced on an

equivalent rectangular bandwidth (ERB) scale. This scale is based on the frequency-dependant bandwidths of auditory filters for human listeners (Glasberg and Moore, 1990). Filters were spaced at ½-ERB intervals over the range of 70-7000 Hz; the filters had a bandwidth of 1 ERB. The filterbank and the noise-reduction algorithm were implemented using MATLAB (Mathworks, Natick, MA).

The noise-reduction algorithm consisted of a "perfect" detector that changed the gains of the individual frequency bands based on the presence of components of the sentence. The perfect detector was based on the original, non-noisy samples of the sentences. A spectrogram was computed, with the CFs of the frequency bands matched to those of the analysis-synthesis bank. An example of this is shown in Fig. 3-1A for the sentence "Her shoes were very dirty". A single frequency band's output is shown in Fig. 3-1C, for a CF of 414 Hz (Fig. 3-1A, arrow). The energy of each frequency band was computed by filtering the input with a $4^{th}$-order gammatone filter that had ½-ERB bandwidth and that was matched in center frequency to the corresponding band of the analysis-synthesis filterbank). The ½-ERB bandwidth was used to create a spectrogram with finer frequency resolution. The output of the gammatone filterbank was then squared and low-passed filtered with a $4^{th}$-order Butterworth filter with a cut-off frequency of 300 Hz. Speech was considered to be present if the energy in the band exceeded a threshold (shown in Fig. 3-1C by the dashed line). This threshold varied across sentences, and was chosen such that 99% of the total energy contained in the speech was above the threshold. The 99% criterion was used only for Experiment I; by varying the threshold, the binary mask was degraded in Experiment II. When the energy within a band was greater than the threshold, the gain was set to unity; otherwise the gain

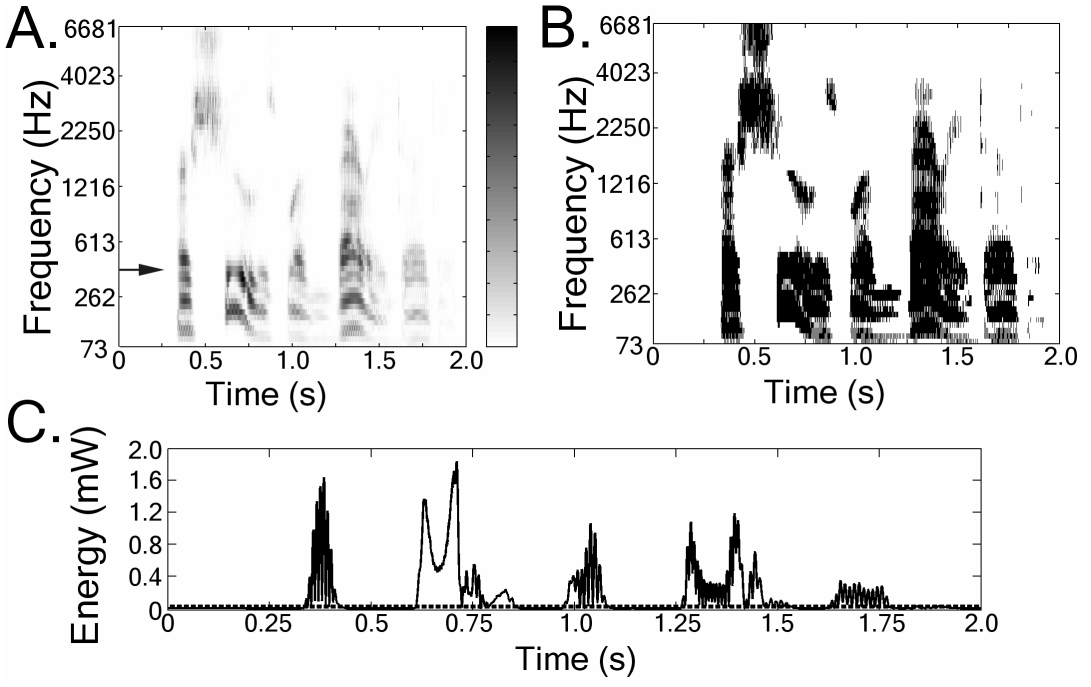


**Figure 3-1 Ideal Binary Mask Generation**

To generate the ideal binary mask, a spectrogram (A) of the clean speech ( "Her shoes were very dirty") was produced by filtering the speech with a filterbank with center frequencies matched to the analysis filterbank of the NR algorithm, but with ½-ERB bandwidth to produce less overlap. The filter output was then squared and filtered with a 300Hz low-pass filter. An example for the frequency band with a center frequency of 414Hz (the arrow in A) is shown in C. A global threshold for the sentence was then applied to each frequency band (the dotted line in C), with the gain for that frequency band set to unity when the energy within the band exceeded the threshold and 0.2 when below the threshold. For the ideal binary mask, the threshold was individually set for each sentence such that 99% of the total energy in the signal was above the threshold used. The entire ensemble of gains (the ideal binary mask) can be visualized in a manner similar to the spectrogram (B). The dark areas represent when the gain is unity.

was set to 0.2. While setting the gain to zero during noise-only time periods would result in a larger overall SNR improvement, limiting the attenuation to less than 20 dB results in reduced amounts of musical noise in time-frequency gain manipulation (Berouti et al., 1979).

The ideal binary mask is shown for the sentence of Fig. 3-1 in panel B. The dark regions represent periods of time when the gain was set to unity; for all other regions, the gain was set to 0.2. Comparing the ideal binary mask to the spectrogram, one can see that the ideal binary mask simply identified periods of time and frequency when energy was present. The output of the analysis filterbank was multiplied by the binary mask; an example is shown in Fig. 3-2. The top panel is the output of the analysis filterbank for the sentence of Fig. 3-1 with speech-spectrum noise added at a SNR of 0 dB. The bottom panel had the ideal binary mask (Fig. 3-1B) applied to it.

To reconstruct the final signal, each frequency band was delayed and scaled such that the peaks of each band's impulse response had a maximum at 4ms (Hohmann, 2002). All of the frequency bands were then added together to obtain single waveform.

#### 3.2.1.1 Experiment I

Stimuli were processed for four conditions of noise-reduction. In the first, the stimuli were simply passed through the analysis-synthesis bank without any manipulation of the frequency bands' gains. This condition served as a control, as the analysis-synthesis bank adds some minor distortions to the signals and band limits the signal (Hohmann, 2002). The three remaining conditions involved applying the ideal binary mask to varying frequency regions of the stimuli. In the second condition, the algorithm was only applied to lower frequencies (from 70-1500 Hz), and the remaining frequency bands were

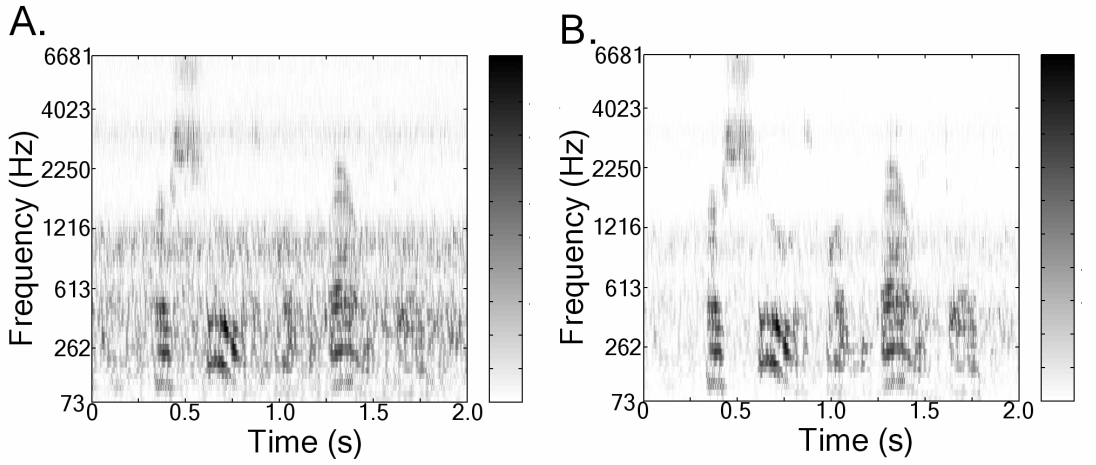


**Figure 3-2 Application of the Ideal Binary Mask**

A spectrogram for the sentence of Figure 1 ("Her shoes were very dirty") is shown (A) when speech-spectrum noise has been added at a SNR of 0 dB. The ideal binary mask derived from this sentence (Fig. 3-1 B) was applied to the sentence, with the resulting spectrogram shown in (B). Application of the ideal binary mask results in the noise between speech components being attenuated; because the binary mask is applied on a sample-by-sample basis (with 50 us sampling time), the reduction occurs both between words as well as within the words themselves.

passed without modification. Similarly, in the third condition, the algorithm was only applied to higher frequencies (1.5-7 kHz), and the lower frequency bands were passed without modification. In the last condition, the noise-reduction algorithm was applied to all frequency bands (70 Hz-7 kHz).

For all four conditions, the frequency resolution of the binary mask and analysis/synthesis filterbank was set at 2 filters per ERB. The binary mask was applied on a sample-by-sample basis, as determined by the energy contained in the signal. The time-constant of the low-pass filter used to determine the energy within each band was constant at 0.53 ms.

#### 3.2.1.2 Experiment II

In the first experiment, the ideal binary mask was used with a threshold for each sentence set such that 99% of the total speech energy exceeded it. In the second experiment, the threshold was systematically varied such that the binary mask was based on 75-95% of the sentence energy in steps of 5%, as well as a 99% condition comparable to Experiment 1. All other aspects of processing were the same as in Experiment I. An illustration of the effect of changing the threshold on the binary mask is shown in Fig. 3-3 which shows the ideal binary mask, as well as binary masks based on 85% and 75% of the speech energy. The thresholds used are shown in Fig. 3-3B, where the energy for a single frequency band (414 Hz CF; see arrow in Fig. 3-3A) is plotted as function of time. The threshold was determined for each sentence individually; this was done to ensure that the sentences used in each track were processed using binary masks based on an identical percentage of speech energy exceeding the threshold.

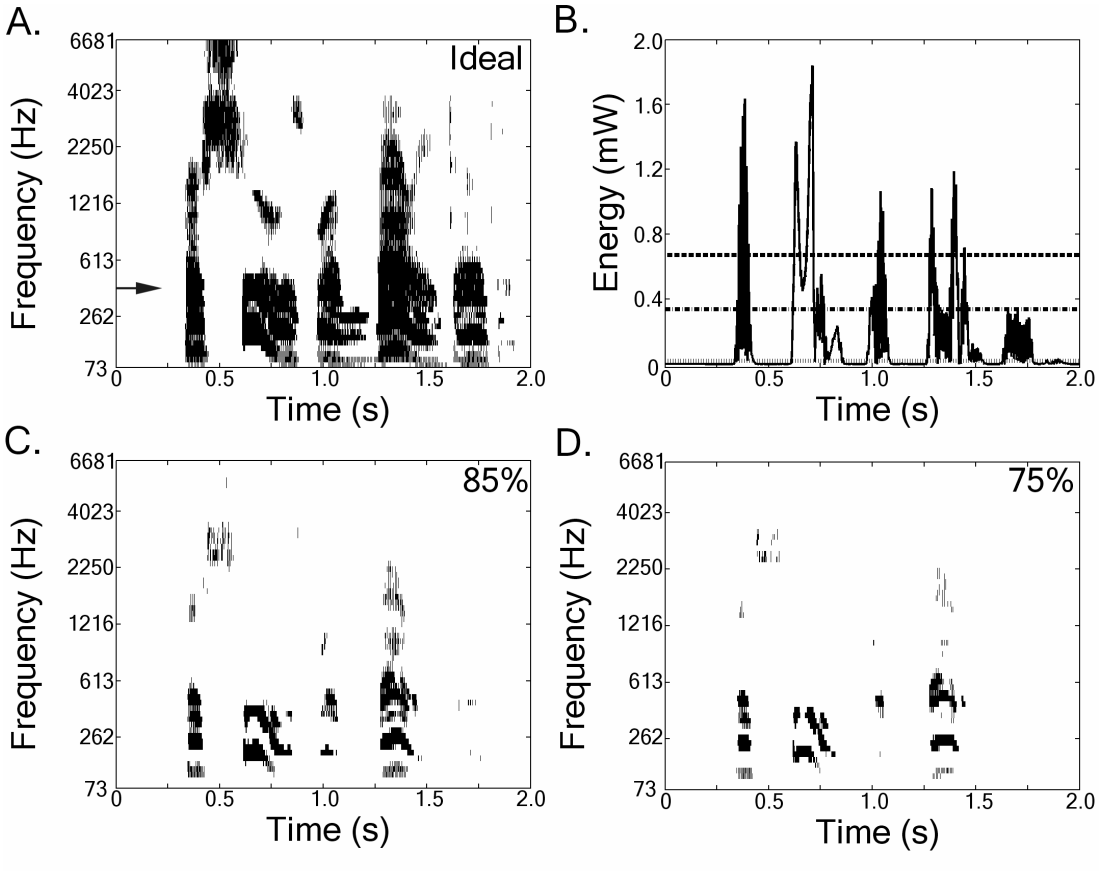


**Figure 3-3 Degradation of the Ideal Binary Mask**
The ideal binary mask for the examples sentence is shown in (A). The ideal binary mask was determined as in Experiment I, with the threshold shown by the dotted line in (B) for the frequency band with CF of 414 Hz (the arrow in A). By varying the threshold used to determine the binary mask, the ideal mask was degraded to represent more realistic detectors. When the threshold was adjusted such that 85% of the total speech energy was above it (the dot-dash line in B), the binary mask shown in (C) was produced. This binary mask was more selective than the ideal mask, as can be seen by the narrower regions of unity gain. When the threshold was raised further (dashed line in B., such that 75% of the total speech energy is above the threshold), the binary mask shown in (D) was produced. With this higher threshold, many of the speech components that are visible in the ideal mask are missing; these missing regions would be periods when a realistic detector would miss the speech components in noise. The thresholds used to determine the non-ideal binary masks were specific to each sentence; however, the amount of total speech energy above the thresholds was kept constant across sentences.

The binary mask was then applied to the stimuli as in Experiment I. Based on the results of the first experiment, and because of the limited number of stimuli available and large number of different percentages used, the processing in Experiment II was performed only on the low-frequency bands (70-1500 Hz). The frequency- and time-resolution of the binary mask was the same as in Experiment I.

#### 3.2.1.3 Experiment III

The methods of Experiment III were similar to those of the low-frequency condition of Experiment I, with the difference being the frequency-resolution of the binary mask and the application of a temporal smearing to the binary mask. In addition to the control condition, stimuli were produced for three separate conditions: one with a reduced frequency resolution (1 filter per ERB), and two with temporally smeared binary masks. The temporal smearing of the binary mask was accomplished by varying the way in which the gains were applied to the analysis/synthesis bank. While in Experiment I the gains were applied on a sample-by-sample basis, for Experiment III the gain was kept at unity for a set amount of time whenever the energy-threshold was exceeded for that band. This had the effect of removing some fast variations in the temporal patterns of the gain changes; while the gains could rapidly change to unity, they returned to the 0.2 state slowly. The temporal smearing was accomplished by convolving each frequency band of the ideal binary mask with a rectangular pulse. The two pulse durations used were 15ms and 100ms.

### 3.2.2 Listeners

Listeners for Experiment I consisted of 8 subjects, 3 normal-hearing listeners and 5 listeners with sensorineural hearing loss. Normal-hearing listeners had thresholds lower than 15 dB HLfrom 250-6000 Hz. Listeners for Experiment II consisted of 10 subjects, 3 normal-hearing listeners and 7 listeners with hearing loss. Normal-hearing listeners had thresholds as in Experiment I, and the listeners with hearing loss were also classified as having moderate sensorineural hearing losses. Listeners for Experiment III consisted of 2 subjects, both listeners with hearing loss. Normal-hearing listeners were aged 21 to 27 (mean: 23.8); listeners with hearing-loss were aged 68 to 88 (mean: 78.5).

Individual audiograms for the listeners with hearing loss for all experiments are shown in Fig. 3-4. All listeners with hearing loss had bilateral, symmetrical (less than 10 dB difference between ears) sensorineural hearing losses. All but two of the listeners with hearing loss were experienced hearing-aid wearers. All listening was performed unaided for the listeners with hearing-loss, with no spectral shaping performed. All subjects were paid for their participation.

### 3.2.3 Procedure

Listeners were seated in a double-walled, sound-attenuating booth (IAC). Stimuli were presented in the free-field by a speaker located 1 m in front of the listener. Stimuli were presented through a TDT System II 16-bit D/A system and digital attenuator (TDT PA4), and amplified by a Crown D-75A power amplifier. The level of the stimuli was adjusted for each subject to ensure that the stimuli were audible, yet remained comfortable for the listener. Presentation levels varied from 77 dB (A) to 87 dB (A).

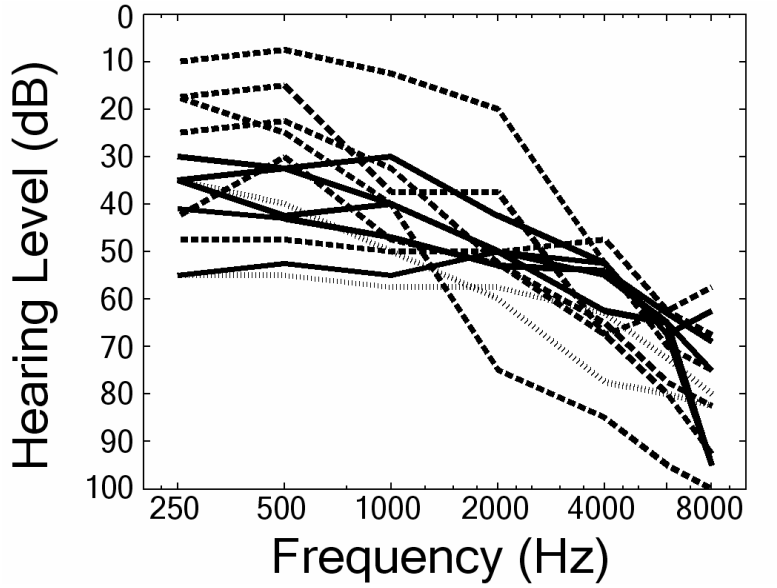


**Figure 3-4 Listeners with Hearing Loss' Audiograms**

Listeners audiograms are shown, with listeners used in Experiment I, II and III having solid, dashed, and dotted lines respectively.

The HINT procedure (Nilsson et al., 1994) was used to estimate the RTS of speech in speech-spectrum shaped noise. Briefly, the subjects were presented with a sentence in speech-spectrum shaped noise, and asked to repeat back the sentence to the experimenter. The SNR of the sentences was varied in a track to determine the RTS, with each track consisting of a total of 20 sentences. The initial sentence was repeated until the subject was able to correctly repeat the entire sentence; subsequent sentences were presented once, with the SNR varying depending on the subject's response. SNR was varied by 4 dB for the first 4 sentences, and by 2 dB for the remainder of the track. During each track, the SNR had a lower bound of -10 dB. At the lower SNRs, application of the binary mask resulted in the noise present in each frequency band having an envelope that was equal to the speech within that band; preliminary testing showed that normal-hearing listeners could understand the speech when only these envelope cues were present (i.e., even when the speech was not actually present), thus a lower bound on the SNR was necessary. Listeners are able to understand speech when noise is separated into frequency bands and amplitude-modulated with the speech's envelopes, even when there is no speech actually present (Shannon et al., 1995). Each subject was presented each HINT track only once to minimize any learning effects on the RTS. The final RTS was determined by taking the average of the SNRs of the final 16 sentences, as well as the SNR at which the 21$^{st}$ sentence would have been presented (Nilsson et al., 1994).

Subjects were given 2 tracks of unprocessed HINT sentences to familiarize themselves with the testing procedure to ensure that the overall level was comfortable, and to ensure that subjects performed within normal bounds of the HINT. After the initial familiarization, subjects were presented with the processed stimuli. For each

condition, the RTS was the average of 2 HINT tracks, each of which consisted of 20 sentences. The order of presentation for the experimental tracks was randomized for each subject. Subjects were allowed to take short breaks between tracks as necessary.

## 3.3 RESULTS

### 3.3.1 Experiment I

Figure 3-5 shows the RTSs measured for individual listeners with hearing loss for all four conditions involving the ideal binary mask, along with a group average. Individual subjects' PTAs are indicated next to the subject identifier. Individual normal-listener's RTS are shown in Fig. 3-6. For the unprocessed condition, the difference between the average listener with hearing loss and normal-hearing listener was 3.6 dB, matching previous work for long-term speech-spectrum shaped noise (Plomp, 1994). Examining all subjects, the effect of the processing was a decrease in the RTS, indicating an improvement in the listener's ability to understand speech in noise. All of the listeners with hearing loss showed the greatest reduction in RTS for the combined condition, followed closely by the low-frequency condition. The amount of reduction for the high-frequency condition was smaller than for the low-frequency or combined conditions and had a greater variability between subjects. Simple linear regression showed no significant relation between the RTS and PTAs of the listeners with hearing loss ($R^2$ ranged from 0.25 to0.56 with p values from 0.15 to 0.40)

Normal-hearing listeners also showed a decrease in RTS for all conditions of processing (Fig. 3-6). For the combined condition, all normal-hearing listeners were operating near the minimum SNR that was used (-10 dB). The arrows of Fig. 3-6

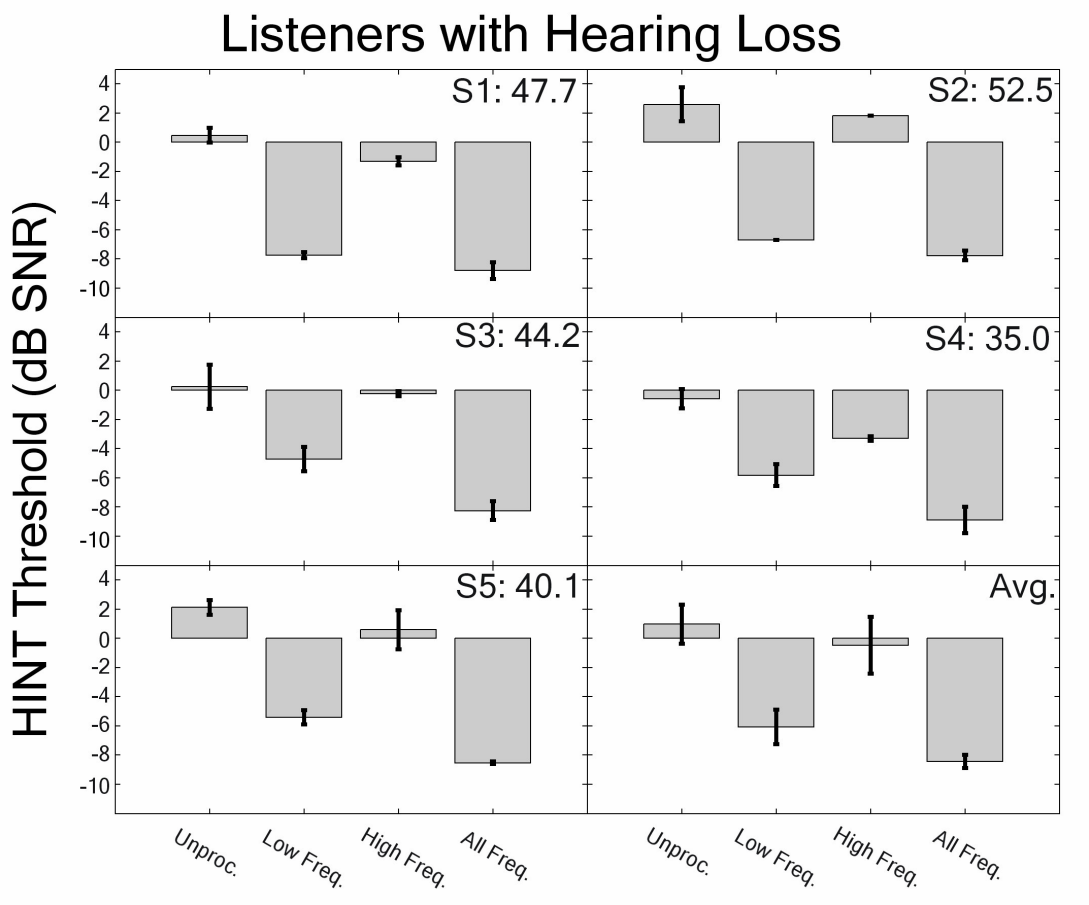


**Figure 3-5 HINT Thresholds for Listeners with Hearing Loss**

Individual HINT thresholds are shown as a function of the processing condition, as well as an average RTS. Listeners' pure tone average (PTA) for 500, 1000 and 2000 Hz tones are also shown. For the unprocessed condition, RTSs ranged from -0.6 dB to 2.6 dB. For all conditions where the ideal binary mask was applied to the stimuli, the RTS improved for all subjects compared to the unprocessed condition. The amount of decrease was greatest for the all-frequency condition and smallest for the high-frequency condition. Subjects' unprocessed RTSs were elevated compared to those obtained for normal listeners (Fig. 3-6).

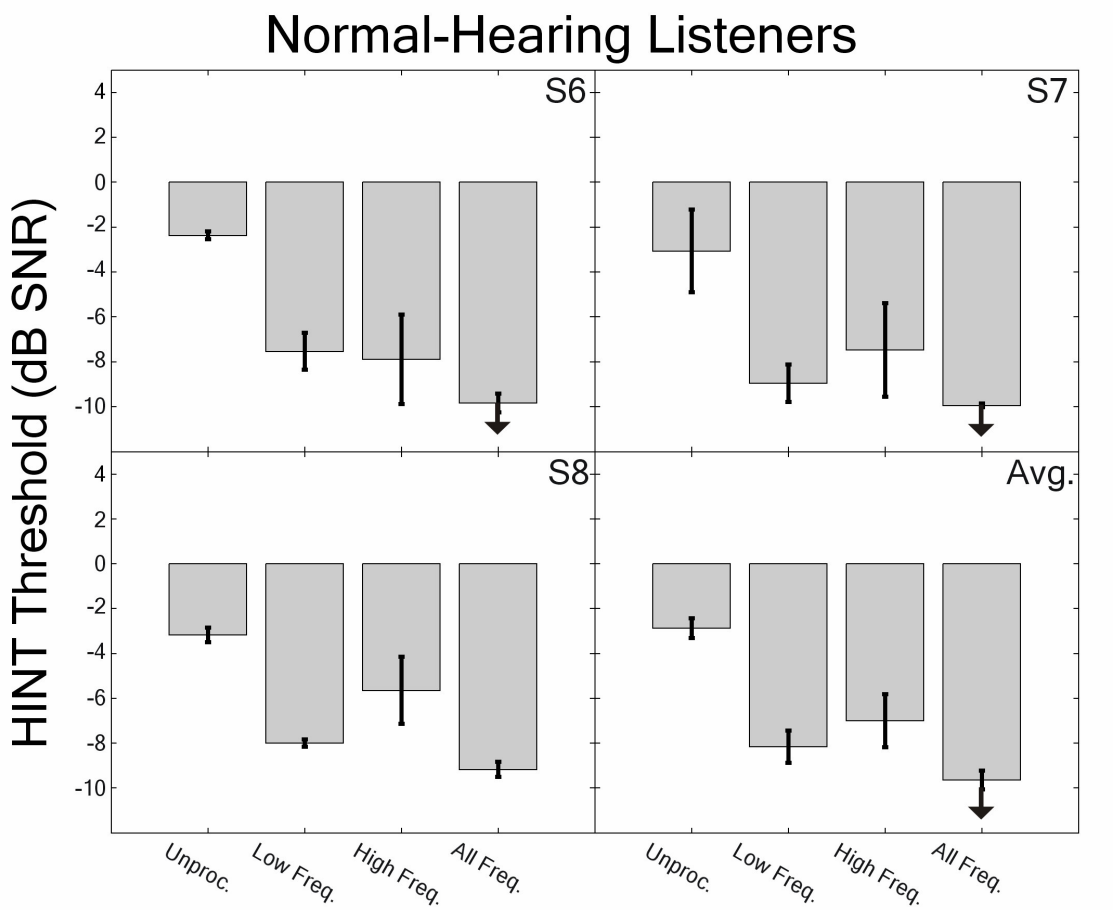


**Figure 3-6 HINT Thresholds for Normal Listeners**

The HINT thresholds are shown for all conditions for the normal subjects, as well as a group average. All subject's unprocessed RTSs fell within the norms for the HINT (Nilson et al, 1994). Normal listeners had a large improvement for all conditions tested; for the all-frequency condition, both listeners performed at or near the limit of the SNRs used (-10 dB). The arrows indicate that the subjects hit the floor of the processed SNRs, where only the temporal envelope cues were available (see text).

indicate that the results are based on tracks in which the listener reached the limit of SNRs used in the processing and their scores would have likely been lower if the SNR had not been limited to -10 dB. Unlike the listeners with hearing loss, the normal-hearing listeners showed an improvement in RTS for the high-frequency condition.

Comparing the two subject groups, the listeners with hearing loss derived more benefit from the low-frequency and combined-frequency conditions than did the normal-hearing listeners, with the latter showing a larger improvement in the high-frequency condition. The large improvement for the listeners with hearing loss was due to their higher unprocessed RTSs; the improvement of normal-hearing listeners was also limited by minimum SNR used (-10 dB). However, the listeners with hearing loss were not affected by the floor imposed by the processing.

### 3.3.2 Experiment II

The results for Experiment II are summarized in Figs. 3-7 to 3-9. The RTSs for listeners with hearing loss are plotted as a function of the amount of energy above the threshold of the degraded binary mask. Individual subjects' PTAs are shown in the legend. As the binary mask approached the ideal mask, all subjects with hearing loss showed a decrease in RTS. The RTSs obtained for the ideal mask (the rightmost point on the curves) matched the results found in Experiment I, which were obtained with a different set of listeners. A weak, but not significant trend for the RTS to increase with increasing PTA was seen.

The change in RTS for the listeners with hearing loss is shown in Fig. 3-8. Here, the results of Fig. 3-7 are replotted by subtracting out each subject's unprocessed score. While the absolute values of the scores for individual subjects varied (Fig. 3-7), when the

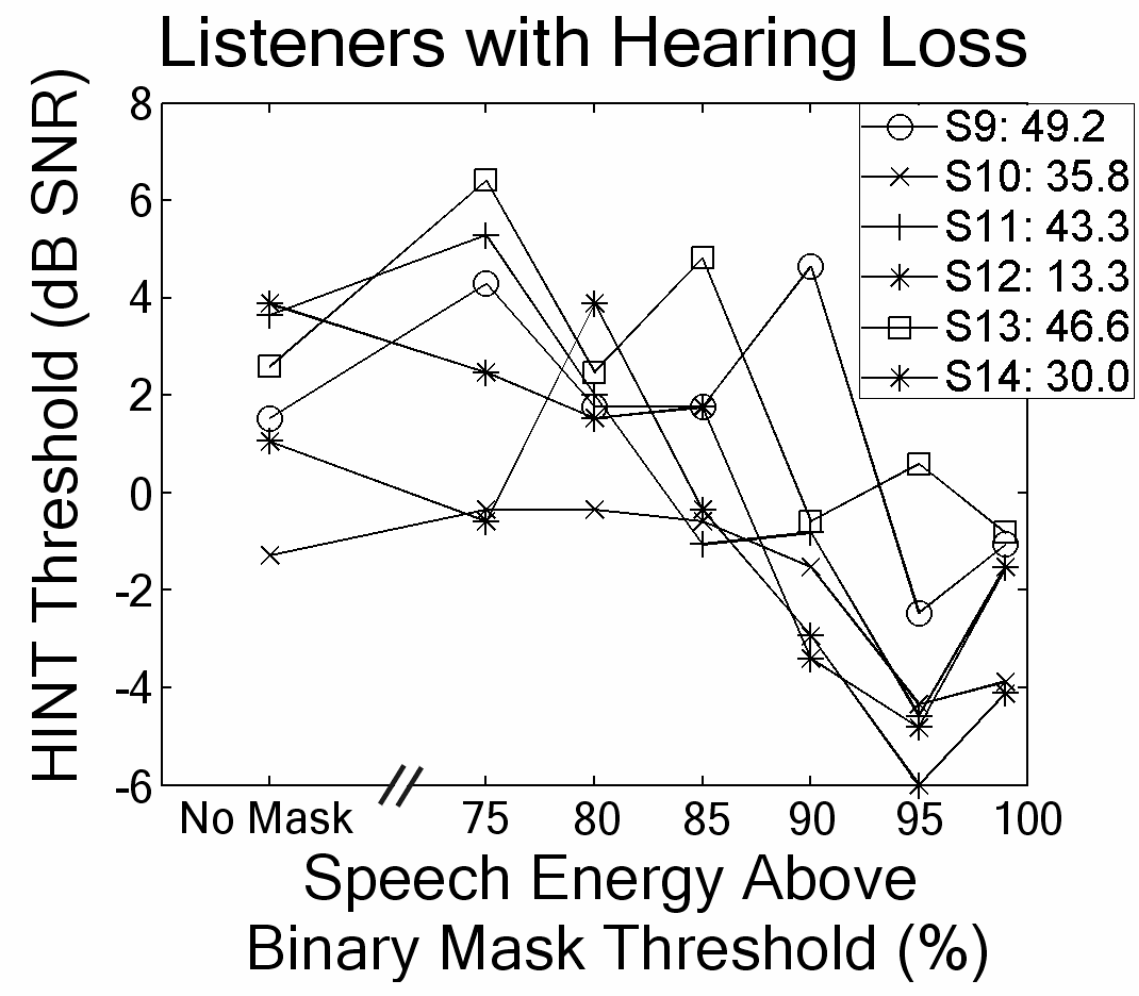


**Figure 3-7 Effect of Degrading the Ideal Binary Mask on Listeners with Hearing Loss**

The ideal binary mask was degraded by increasing the threshold applied to the spectrogram of the clean speech when generating the binary mask (as shown in Fig. 3-3). This was done for each sentence, such that for each HINT track, the amount of speech energy above the threshold was constant. Shown are subject's HINT thresholds as a function of the amount of degradation, expressed as the percent speech energy above the threshold used. All subjects showed a decreasing trend in RTS as the degraded mask approached the ideal binary mask. Subjects' performance for the ideal mask (the rightmost point of each line) matched the results found in Experiment I.

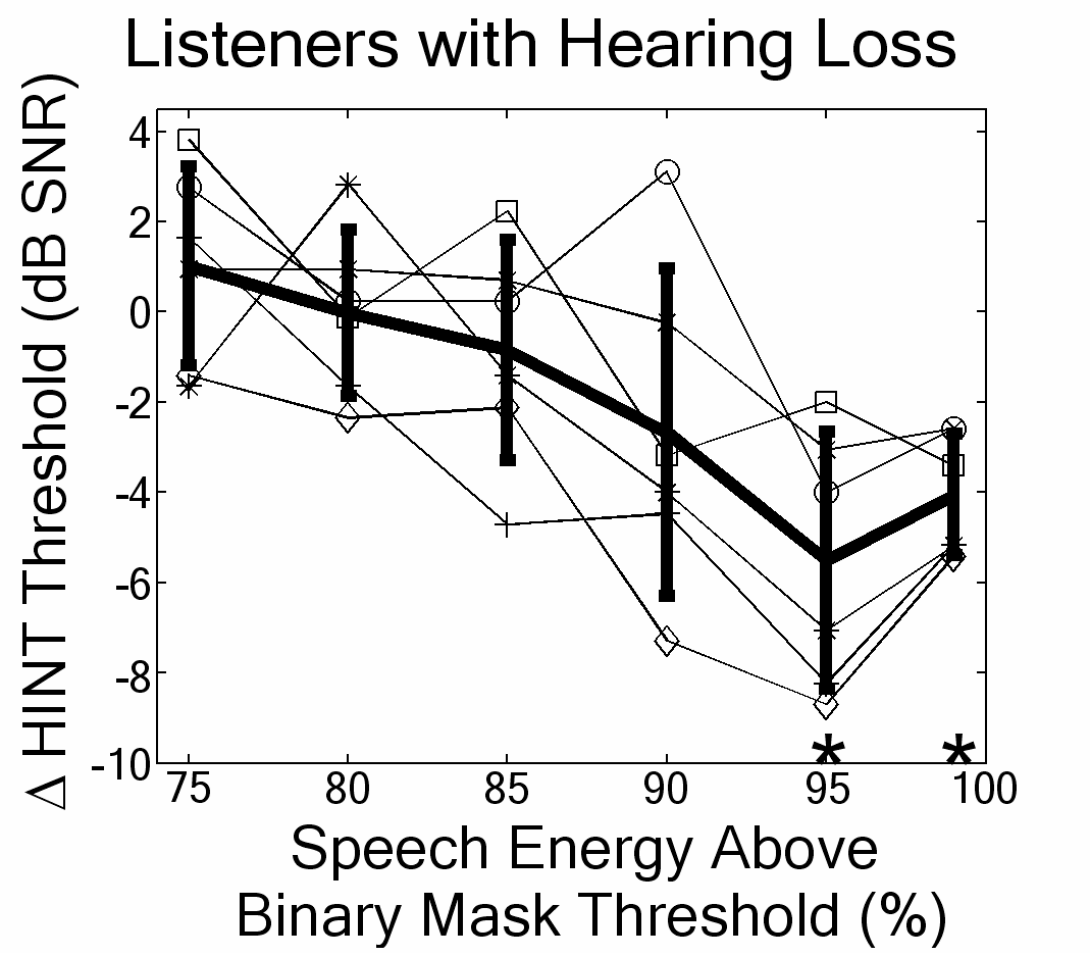


**Figure 3-8 Change in HINT Score with Ideal Binary Mask Degradation**

Shown are the changes in HINT threshold for listeners with hearing loss as a function of the degree of processing. The solid line represents the mean (± 1 s.d.) change for all listeners with hearing loss. Listeners with hearing loss showed an improvement when the binary mask was based on greater than 90% of the speech energy; below this level, there were either no changes from the unprocessed condition, or a slight increase in RTS. A paired-t test was used to determine statistically significant differences from the unprocessed condition ($p < 0.05$, denoted by the asterisks).

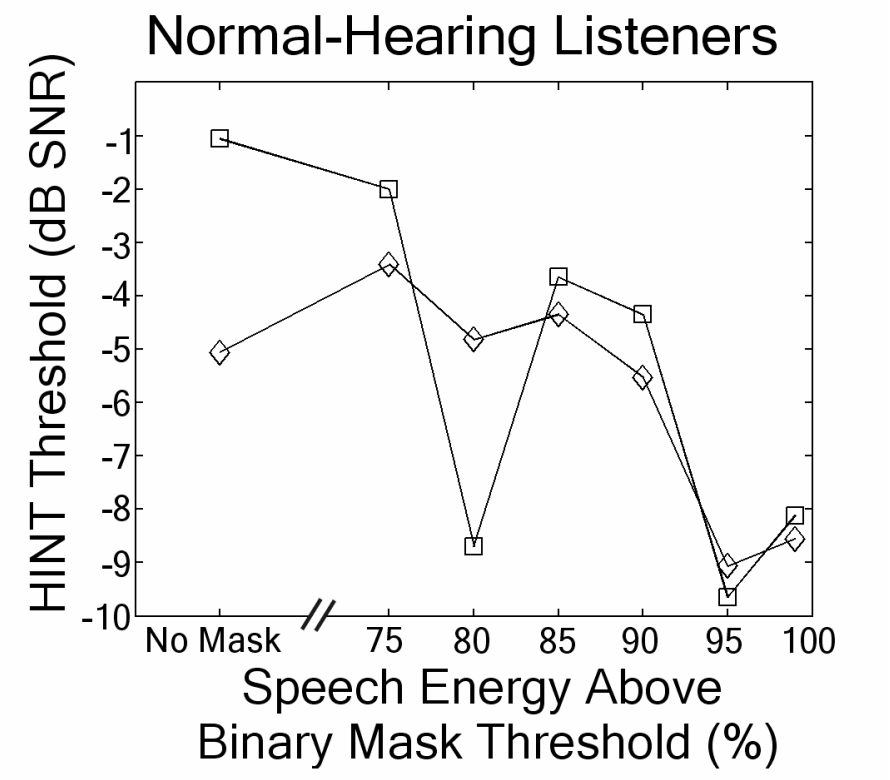


**Figure 3-9 Effect of Degrading the Ideal Binary Mask on Normal-Hearing Listeners**

Similar to the listeners with hearing loss, normal-hearing listeners' HINT thresholds decrease as the binary mask approaches the ideal binary mask. Unlike listeners with hearing loss, at the higher percentages of energy above the binary mask threshold, normal-hearing listeners reached the minimum SNR used in the study (-10 dB).

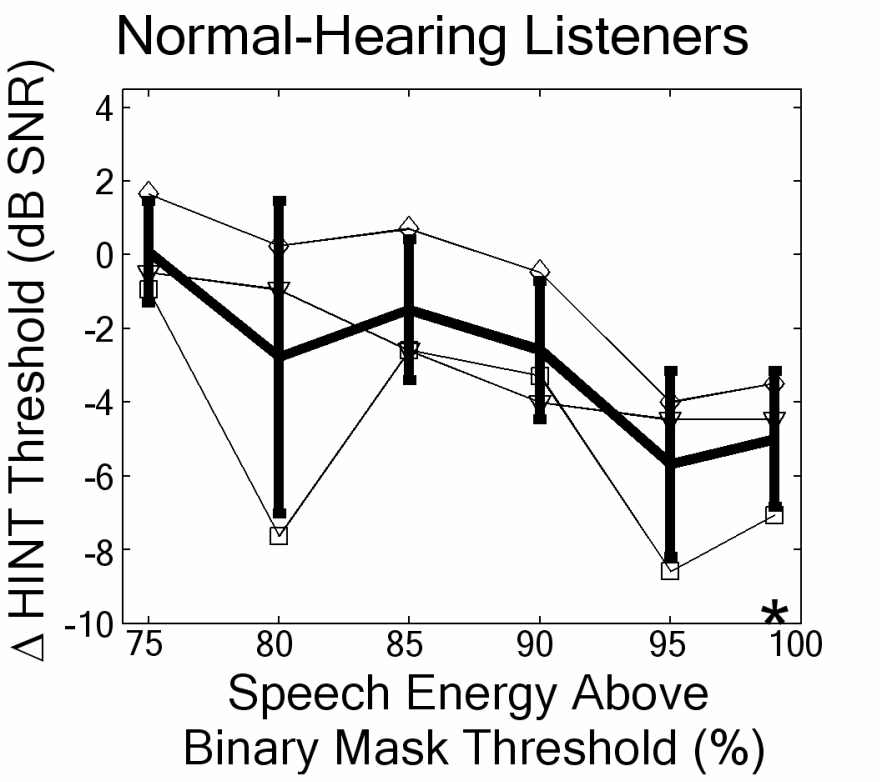


**Figure 3-10 Change in HINT Threshold with Ideal Binary Mask Degradation**

The change in HINT threshold is shown for listeners with normal hearing as a function of the amount of speech energy used in the derivation of the binary mask. The solid line represents the mean (± 1 s.d.) change for all normal-hearing subjects. A paired-t test indicated that the only condition statistically different from the unprocessed condition was 99% ($p < 0.05$, denoted by the asterisk).

*change* in HINT score was examined, a clear trend appeared, with the change in RTS decreasing as the percentage of energy increased. The heavy solid line shows the mean change for all subjects with hearing loss, with error bars denoting ±1 s.d.. While individual subjects showed improvement in RTS at each percentage, a paired-t test showed no significant difference ($p < 0.05$) until the percentage was equal to or greater than 95% (indicated by the asterisks in Fig. 3-8). If the outlier from subject 10 was removed (the open circle of Fig. 3-8), the difference at 90% becomes significant ($p = 0.02$).

Similar to the listeners with hearing loss, normal-listeners' RTSs improved as the binary mask became closer to the ideal binary mask (Fig. 3-9). Normal-listeners had a lower RTS for the entire range of percentages, as expected. The scores are replotted as a change in RTS, along with an average change, in Fig. 3-10. A paired-t test between the processed and unprocessed conditions showed significant differences ($p<.05$) for the normal-hearing listeners only for the ideal binary-mask condition, although only a small number of normal-hearing listeners were tested.

### 3.3.3 Experiment III

The result of reducing the frequency-resolution and temporally smearing the ideal binary mask is shown in Fig. 3-11. When the frequency resolution was decreased to 1 filter per ERB, both subjects showed an improvement in RTS over the control condition. The magnitude of this improvement matched that found in Experiment I.

The RTSs obtained for both subjects for the temporally smeared binary masks are also shown in Fig. 3-11. The two conditions represent gains that are forced to slowly return back to their attenuating state, thus removing many of the fast variations in the

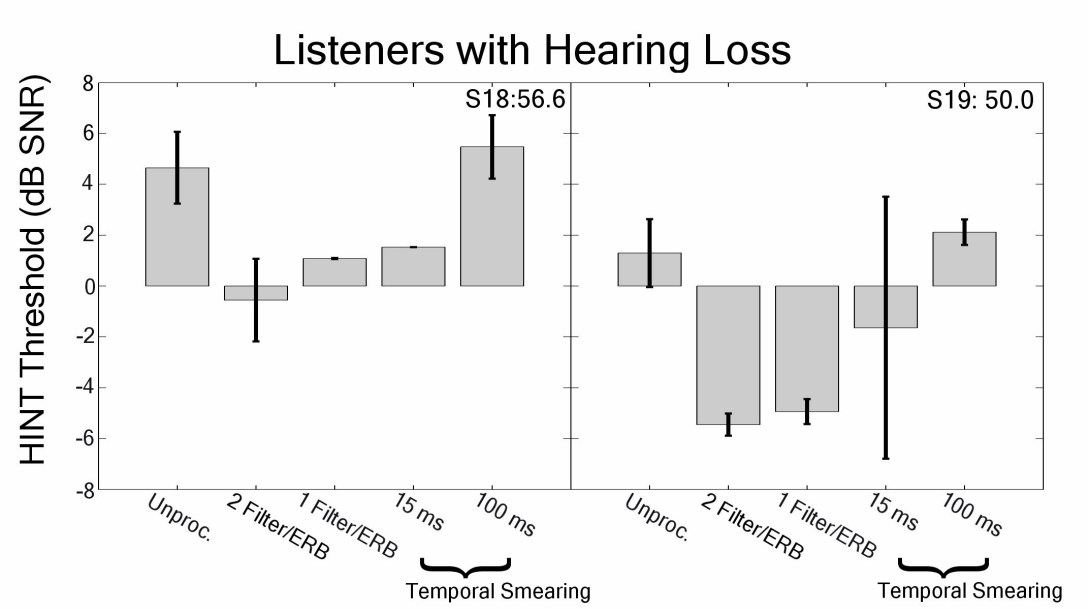


**Figure 3-11 HINT Thresholds for Changing Frequency-Resolution and Temporal Smearing**

The HINT thresholds for two additional subjects with hearing loss are shown after changing the frequency-resolution of the binary mask as well as temporally smearing the binary mask. Both subjects showed an improvement for the conditions of Experiment I (2 filters/ERB). When the frequency-resolution was decreased to 1 filter/ERB, both subjects still showed an improvement in RTS. When the binary mask was temporally smeared with a 15 ms rectangular window, the subjects showed different results; one subject showed an improvement similar to Experiment I, with the other showed a very variable response. When the binary mask was smeared with a 100 ms rectangular window, both subjects RTSs returned to their control values.

temporal properties of the gain changes. Here, the two subjects showed a similar pattern for the 100-ms condition, with little or no change from the control condition. For the 15-ms condition, the first subject showed an improvement similar to that obtained in the one-filter-per-ERB condition, which used a temporal-resolution matching Experiment I. The change seen in the second subject was difficult to determine, as the subject's variability was high; for the first track, the SNR improved to a level similar to the one-filter-per-ERB condition, whereas the second track resulted in a RTS higher than that of the control.

## 3.4 DISCUSSION

The results of Experiment I demonstrated that the strategy of time-frequency gain manipulation could improve the RTS of speech in speech-spectrum noise, provided that the gains were manipulated based on the ideal detector. This result is in contrast to some past studies that have shown no improvement using this technique (Klein, 1989; Fabry and Van Tasell, 1990). There are three possible differences in past studies that may be the cause of this contrast (Ono et al, 1983; Dempsy, 1987; Klein, 1989; Fabry and Van Tassell, 1990; van Dijkhuizen et al, 1987, 1989, 1990, 1991; Rankovic et al, 1992): reliance on non-ideal detectors, usage of fewer frequency bands for their gain manipulations, and sluggish manipulation of the gains. The combination of these differences may have led to the large improvements shown in the current study; replicating the current study with the frequency bands used in previous studies (or in current hearing aids) would enable a more direct comparison.

One of the interesting results of Experiment I was the difference for the ideal binary mask between normal-hearing listeners and those with hearing loss. While the normal-hearing listeners reached the limits of processing (-10 dB), listeners with hearing loss did not, even when the ideal binary mask was applied to the full frequency range of speech. At the lower levels of SNR, below about -6 dB, the output was dominated by the noise, which had been given a speech-like envelope by the manipulation of the gains. The inability of the listeners with hearing loss to use this information suggests an impairment in their ability to process envelope cues compared to normal-hearing listeners. While studies have shown that listeners with hearing loss generally perform similarly to normal-hearing listeners in AM detection, it is possible that the wider tuning of the impaired auditory system could have a larger effect on the wide-band envelope cues that are contained in speech.

Additionally, listeners with hearing loss did not show an improvement when the ideal binary mask was applied to the high-frequency (> 1500 Hz) region, unlike the normal-hearing listeners. The addition of higher-frequency components should have a large effect on the subjects' scores; that it did not suggests that the information contained in these regions was unavailable to the listeners with hearing loss. This lack of improvement for these listeners may have resulted from two factors: that the processed speech was inaudible due to the listener's hearing loss, or because of the increased amount of masking resulted from the hearing loss. That all subjects show additional improvement when the ideal binary mask covers the entire frequency range suggests that the masking effect of the low-frequencies in the high-frequency-only condition is the explanation for the lack of improvement under this condition.

By degrading the ideal binary mask in Experiment II, it was found that over 90% of the energy must be detected in order to see an improvement in RTS. To detect this high of a percentage, however, would require a very good detector that may not be realizable in the computationally-limited environment of a hearing-aid. However, overall energy detected might not be the best cue. Additional experiments should be done to examine the effect of detecting various other cues in speech, such as onsets and offsets, the transitions of formants, as well as envelope cues. It is possible that a detector would not have to detect the high percentage of energy shown in Experiment II if it were to detect some of the salient cues mentioned above.

The results of Experiment III, while based on a small number of subjects, help to further define the binary mask needed for improvement in RTS. The frequency resolution results suggest that a lower frequency resolution can still provide an improvement in RTS; because the SNR within each band at a frequency-resolution of 1 filter per ERB is not improved when speech is present within that band, the results suggest that improvement in the RTS is the result of other factors, such as a reduction in masking of adjacent bands

The temporal smearing results of Experiment III suggest that any time-frequency gain manipulation must be relatively fast (15 ms or less) for there to be an improvement in the RTS. This result may explain why many of the current time-frequency gain manipulation algorithms used in current hearing aids provide very little in the way of intelligibility increase, as these algorithms are typically slow-acting (~20ms to a few seconds).

The results of the 3 experiments have practical application in the development of time-frequency gain algorithms for use in hearing applications. The results of

Experiment I demonstrate that this sort of algorithm will work; these results also show that benefits to RTS can be achieved by processing only the low-frequency components (< 1.5 kHz) of speech. This allows the use of simpler detectors, because the low-frequency, narrowband components of speech are generally easier to detect than the noise-like, high frequency components. The results of Experiment II set the guidelines for the performance of the detectors to be used in a real-world algorithm. To show an improvement in RTS, the real-world detector would have to detect at least 90-95% of the energy of the speech. Experiment III defines the necessary frequency-resolution of the overall system, as well as the temporal speed of the gain changes.

Practically speaking, time-frequency gain manipulations often result in poorer sound quality of the output. This degradation in quality was seen in the current study, with normal-hearing listeners often describing the stimuli as "machine-sounding" or artificially generated speech. This "machine-like" quality is the result of the rapid transitions of the gains, as well as the low SNRs of the stimuli. For the low SNRS (-10 to -6 dB) presented to the normal-hearing listeners, the output was dominated by the noise, which was roughly shaped by the gain manipulations to mimic speech; the resulting outputs were similar in quality to cochlear-implant demonstrations, in which noise is modulated with a speech envelope (Shannon et al., 1995). Interestingly, the listeners with hearing loss were generally less sensitive to the quality of the stimuli; informal comments after listening did not often involve quality, unlike normal-hearing listeners' comments. This may have been because the stimuli were at a level where the actual speech still dominated the output, or because of the impairments associated with hearing loss. However, the degradation of quality should not be an issue with a real-world

implementation, as the detectors used in such an algorithm would probably not reach the performance levels (i.e., ideal detection at -10 dB SNR) that the current study used. At the SNRs that real-world detectors would probably achieve, the signal that remains in the output would dominate over the speech-envelope imposed on the noise. Informal listening suggests that at these SNRs, the quality of the processed speech is at an acceptable level.

The use of binary masks allow for the evaluation of important cues in understanding speech in noise, and provides a platform for the evaluation of NR algorithms. The ideal binary mask provides for an absolute limit of benefit that can be achieved by manipulating the gains of frequency bands. By systematically manipulating the binary mask, one can determine the necessary cues or regions of the stimulus that are needed for intelligibility. Once these cues are determined, quantitative comparisons can be made to the detection patterns that are produced by experimental detectors, allowing for the design and testing of such detectors without having to perform intelligibility testing with them. By simply comparing the detection patterns of the experimental detector to the binary masks that produce increases in intelligibility, one can make a prediction of how that detector would perform.

# Chapter 4

## The Development and Testing of a Phase-Opponent Noise-Reduction Algorithm

### 4.1 INTRODUCTION

Listeners often have difficulty understanding speech in a noisy environment. This difficulty is exaggerated when the listener has some form of hearing loss. An increase of 2-5 dB (Plomp, 1994) in signal-to-noise ratio (SNR) is required for listeners with hearing loss to achieve the same level of intelligibility as normal-hearing listeners. This increase is partly due to the increased thresholds of the listener with hearing-loss, but even with sufficient amplification, the necessary SNR increase is reduced by only 1-2 dB (Peters et al., 1998; Bentler and Duve, 2000). To help overcome this, modern hearing aids have begun to implement various noise-reduction (NR) algorithms to provide an increase in SNR in addition to amplification.

NR algorithms have been developed for many applications other than speech in noise; however, speech in noise represents a particularly difficult class of signals, as in most cases the spectral information of both the speech and the noise overlap. In addition, the acoustical parameters of speech fluctuate depending on the physical characteristics of the speaker, as well as cultural factors. The noises encountered in real life are also non-stationary. These factors rule out many traditional NR algorithms, such as Wiener filtering (Wiener, 1949).

NR algorithms that are suitable for hearing-aid use fall into two categories: directional microphone approaches or single-channel algorithms. Current thinking holds

that only directional microphones can improve intelligibility (Levitt, 2001; Chabries and Bray, 2002; Schum, 2003), while both categories can improve the subjective quality or ease-of-listening (Schum, 2003). To this end, the majority of research involving NR and hearing aids has been directed at improving directional microphones.

Directional microphones, in a laboratory setting, can provide substantial increases in signal-to-noise ratio (SNR) and the resulting increases in intelligibility for listeners with hearing loss (Mueller and Johnson, 1979; Soede et al., 1993; Valente et al., 1995). However, when taken outside of the laboratory setting, these increases diminish in size (although still present) due to real-world considerations such as reverberation (Hawkins and Yacullo, 1984; Ricketts, 2000; Ricketts and Hornsby, 2003). In addition, for the directional microphone to provide any benefit at all, the speech source and noise must be spatially separated (Levitt, 2001).

Single-channel NR algorithms, though more difficult to implement, can be used when the speech and noise can not be spatially separated or when the physical size of the hearing aid prevents the use of directional microphones (i.e., in a completely in-the-canal hearing aid). While single-channel NR algorithms can significantly improve the SNR of a noisy signal, when tested on listeners with hearing-loss, there are no improvements in intelligibility. Most of the single-channel algorithms currently developed are intended for use in automatic-speech-recognition (ASR) systems, which also suffer from degraded performance in noise. While these systems can improve SNR, a key difference between the ASR systems and listeners with hearing-loss is the SNRs that are needed for understanding. While a listener with hearing-loss requires a SNR from 0 to 5 dB for understanding when the noise is speech spectrum shaped (Plomp, 1994), ASR systems

generally require SNRs in excess of 15 dB to perform adequately (Gong, 1995; Lippmann, 1997). The difference in target SNRs for the two systems (hearing-aid and ASR) makes it difficult to compare across NR algorithms that are developed for ASR, as they generally do not perform well at the SNRs at which the hearing-aid must perform.

Most single-channel NR algorithms are based on spectral subtraction (Boll, 1979). In spectral subtraction, the incoming signal is divided into segments and a short-term Fourier transform used to convert the segment to the frequency domain. An estimate of the noise spectrum (generally obtained from a pause in the speech) is then subtracted from the segment's spectrum. The remaining spectrum is then combined with the phase information from the noisy signal and an inverse FFT applied to obtain a noise-reduced signal. The resulting signal, while having less noise than the original, does have some residual noise left because of differences between the estimate of the noise-spectrum and the instantaneous spectrum of the noise. This residual noise is often described as "musical noise" because of its sound quality. This musical noise can often be more distracting than the original noise, which is a possible reason that this NR algorithm results in no improvement in intelligibility (Cappa, 1994).

Improvements to the basic spectral subtraction algorithm outlined above attempt to improve the subtraction by using models of speech (Ephraim & Malah, 1984). Here, a statistical model of the speech is used to calculate the spectrum of the speech in each segment of the signal. The estimate of the speech spectrum is then compared to the segment, and those frequency bands that are dominated by the noise are attenuated. The segment is then subject to an inverse fast-Fourier transform (IFFT). By using a

minimum-squared-error estimator for the estimation of the SNR of each frequency band, this NR algorithm can show an improvement in the amount of noise removed.

Most current NR algorithms are derivatives of spectral subtraction; the improvements made by these algorithms generally involve reducing the musical noise. Current research involves a more psychophysical approach, with the consideration of perceptual masking of the musical noise (Tsoukalas et al., 1997; Arehart et al, 2003).

Tsoukalas et al. (1997) have demonstrated a single-channel NR algorithm that can result in an improvement in the intelligibility for normal-hearing listeners. Arehart et al. (2003) extended this work to show that a similar algorithm can also show small improvements (2-8%) for listeners with hearing-loss. These algorithms work by using auditory masked thresholds (AMT) to determine the frequency bands that are suppressed by a spectral-subtraction-type algorithm. Those frequency bands that are not already masked by the speech are suppressed.

The current study also used a variant on spectral subtraction. The phase-opponent (PO) NR algorithm performs spectral subtraction by splitting the incoming signal into frequency bands and attenuating bands believed to contain only noise. Unlike traditional spectral subtraction, PONR does not use a direct estimate of the SNR to determine the gain of each frequency band; instead, a physiologically inspired detector is used to determine the presence or absence of speech. The use of the PO detector is similar to the AMT-NR algorithms, in that the decision to attenuate any given frequency band is determined by a variable that is more closely related to the human auditory system than a measurement of the SNR of that band. By using a detector that matches the physiology,

the hypothesis is that the output of the PONR algorithm will result in an improvement in intelligibility and quality of speech in noise for listeners with hearing-loss.

# 4.2 METHODS

## 4.2.1 Phase-Opponent Noise-Reduction (PONR) Algorithm

The PONR algorithm was based on time-frequency gain manipulation, which is a form of spectral subtraction. The algorithm divided the incoming signal into separate frequency bands, and varied the gain of each frequency band depending on the presence of speech components within that band. The presence of speech components was determined by PO detectors. The PO detector is an ideal match for a speech detector, as it is a narrowband detector that uses the temporal information to detect the speech; this allows the detector to be used under a wide range of noise levels and conditions without having to determine a spectral estimate of the noise. A flow chart of the PONR algorithm is shown in Fig. 4-1. The algorithm contained three main parts: an analysis filterbank, a bank of phase-opponent (PO) detectors, and a synthesis stage.

### 4.2.1.1 Analysis Stage

The analysis stage of the PONR algorithm was based on that of Hohmann (2002). This stage separated the incoming waveform into separate frequency bands that could then be individually manipulated to achieve an improvement in the SNR of the overall signal.

The analysis filterbank consisted of 59, $4^{th}$-order gammatone filters that covered a frequency range of 100-7000Hz. The filters bandwidths changed according to the equivalent rectangular bandwidth (ERB) of the human ear, a logarithmic scale that results

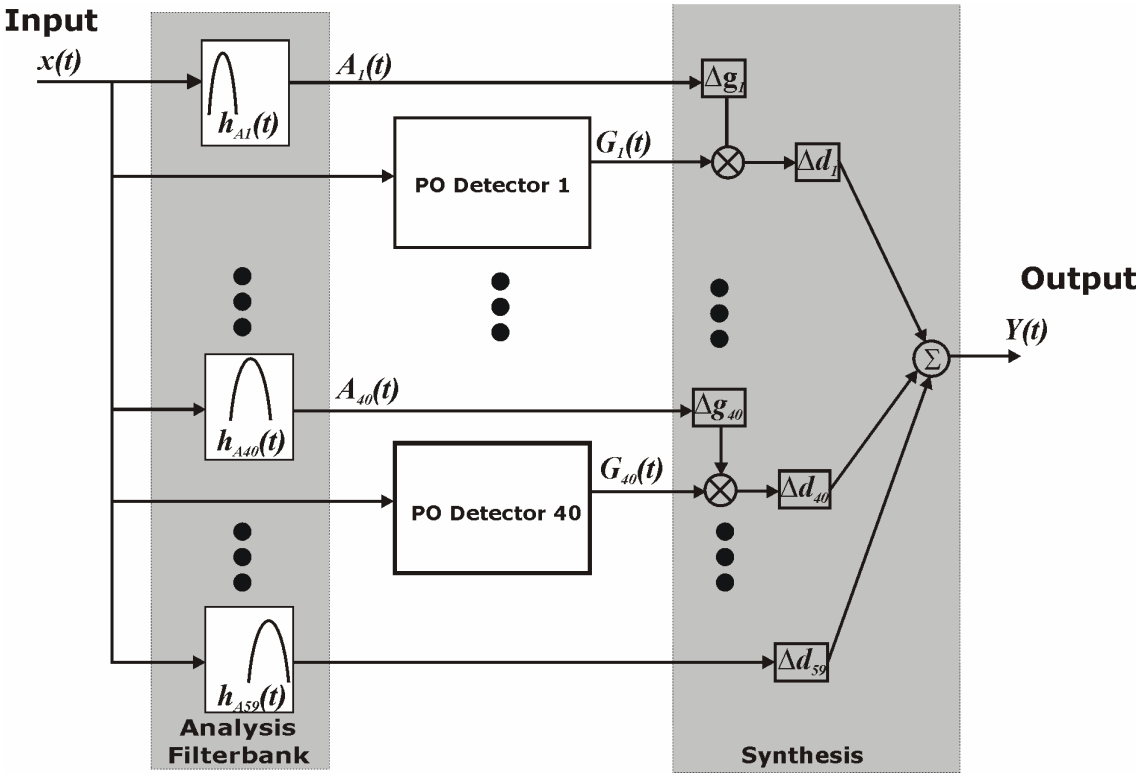


**Figure 4-1 Flow Diagram of the Phase-Opponent Noise Reduction Algorithm**

The PONR algorithm consisted of three stages: the analysis filterbank, a bank of PO detectors, and a synthesis stage. The analysis stage consisted of a bank of 59 gammatone filters that separated the incoming signal into separate frequency bands. These bands covered the frequency range of 100Hz-7000Hz in ½ ERB increments. Bands below 2.5kHz (#1-40) had their own PO detector that attempted to detect any speech components that fell within that band. The PO detectors controlled the gains of the individual bands; bands without a PO detector (#41-59) had a static gain of unity. The synthesis stage applied the gains determined by the PO detectors, and resynchronized each of the band to account for the individual bands' group delay. The bands were then scaled and summed together to produce the final, noise-reduced output.

in filters with lower center frequencies having narrower bandwidths than filters with higher center frequencies. The analysis filters had a bandwidth of one ERB, but were spaced such that adjacent filters were ½ ERB apart. The auditory filters of a listener are at least one ERB wide; by using an analysis filterbank that can change the SNR within a given ERB, it is possible to achieve an improvement in the instantaneous SNR of the listener's auditory filters by manipulating the gains of the two analysis filters that cover each auditory filter.

The output of the $k^{th}$ filter of the analysis filterbank can be expressed as:

$$A_k(t) = x(t) * h_{Ak}(t)$$

$$h_{Ak}(t) = t^3 e^{-\frac{t}{\tau_k}} \cos(2\pi f_k t) u(t)$$

$$\tau_k = \frac{1}{2\pi(1.019 * B_k)}$$

where $h_{Ak}(t)$ is the impulse response of the $k^{th}$, $4^{th}$-order gammatone filter with a center frequency $f_k$ and bandwidth of $B_k$ and u(t) is the unit step function.

The single-ERB bandwidths of the gammatone filters were chosen because they more closely resemble the human auditory system. As stated above, by spacing these filters at ½ ERB intervals, it was possible to improve the SNR within the listener's single-ERB wide auditory filter. The ½-ERB spacing also allowed for efficient use of a smaller number of filters than would the use of equal-bandwidth filters. This smaller number of filters resulted in faster processing times, and was more feasible for the long-term goal of the use of the PONR algorithm in a digital hearing aid.

### 4.2.1.2 General PO Detector

The presence or absence of speech components in each of the frequency bands was determined by a PO detector for that frequency band. An individual PO detector is shown in Fig. 4-2. The PO detectors used were based on the same principles of the PO detectors of Chapter 1, but were implemented in slightly different way. The PO detector of Fig. 4-2 achieved its phase-opponency through the use of an allpass filter (Deshmukh, personal communication) that was applied to the output of a single gammatone filter. The allpass filter for the $k^{th}$ band was a second order filter of the form:

$$p_k = R_k e^{j(2\pi f_k)}$$

$$H_{ap_k}(z) = \frac{(z^{-1} - p_k^{*})(z^{-1} - p_k)}{(1 - p_k z^{-1})(1 - p_k^{*} z^{-1})}$$

The allpass filter, as its name implies, has unity gain across frequency; it is the phase response of the filter that changes with frequency. The phase response of the allpass filter is shown in Fig. 4-3. The phase goes from 0 to -360°, with the slope of the transition controlled by the location of its pole ($p_k$); moving the pole closer to the unit circle (ie, changing $R_k$) results in a steep phase transition. The angle of the pole determines the approximate location of the -180° point of the phase transition; by placing the pole at the appropriate angle, the -180° point can be made to match with the center frequency of the single gammatone filter, which had a bandwidth equal to 3 times the ERB of its center frequency (chosen based on optimizations detailed in the results section). Thus the resulting output of the allpass filter will be 180° out-of-phase with its input in the vicinity of the pole.

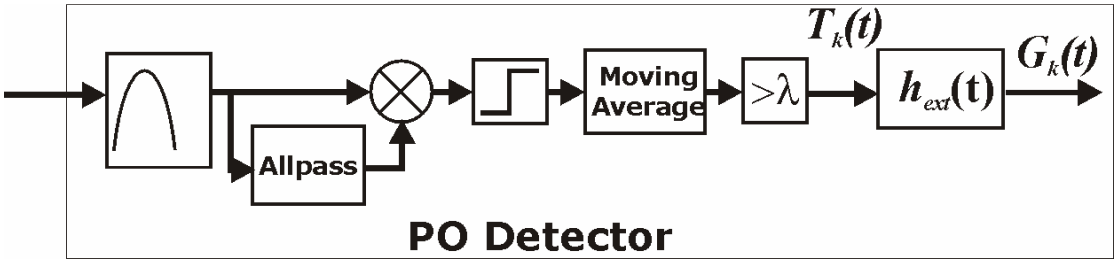


**Figure 4-2 Flow Diagram of a Phase-Opponent (PO) Detector**

The phase-opponent detector was based on the phase-opponent model of Carney et al. (2001). It consisted of a 4$^{th}$-order gammatone filter with a center frequency matched to the frequency band it was associated with, with a bandwidth equal to three times the ERB at the center frequency. The output of the gammatone filter was subject to an allpass filter that provided the 180° phase-shift needed to produce phase-opponency. The output of the gammatone and allpass filters was then multiplied and subject to a signum function. The signum function removed any magnitude information present in the signals, resulting in the reliance on only temporal information. The output of the signum function was smoothed using a moving average filter, and then compared to a threshold to determine if a speech component was present. The output of the thresholder was then convolved with a rectangular pulse $h_{ext}(t)$, to prolong the high gain state that detection resulted in for 10 ms afterwards

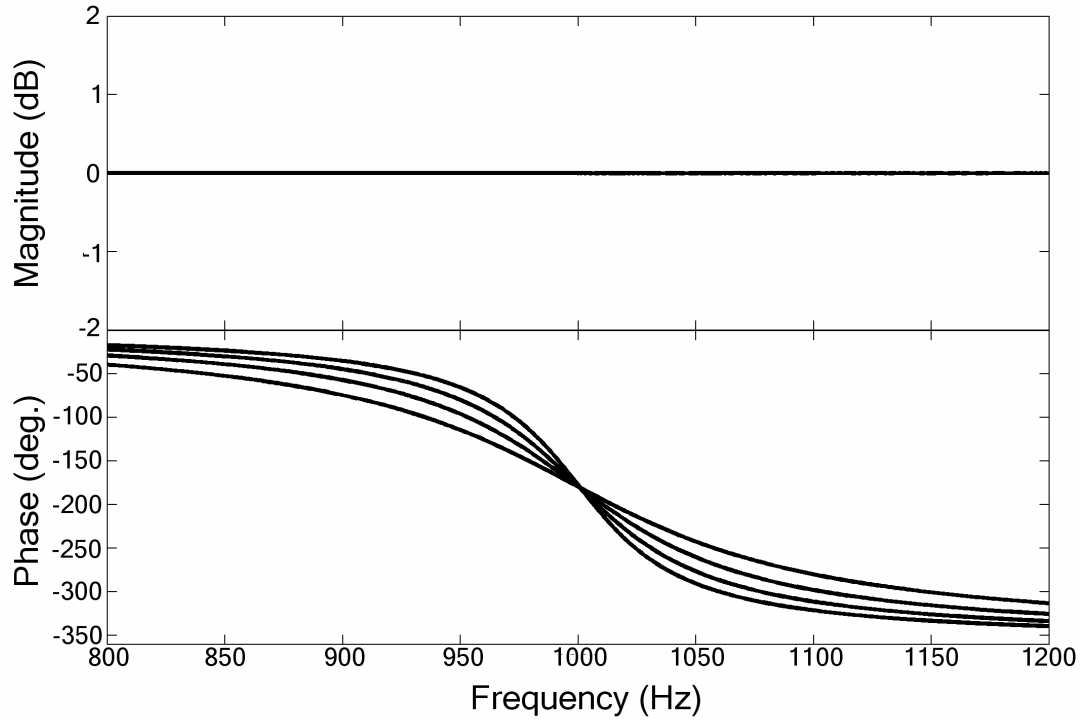


**Figure 4-3 Transfer Function of a PO-Detector Allpass Filter**

The magnitude and phase responses for the allpass filter of the PO detector tuned to 1 kHz are shown for the four conditions used in the experiment. While the magnitude was flat, the phase response of the filter varied with frequency, going from 0 to 360°. The phase response of the filter was 180° at 1 kHz, providing the phase-opponency necessary for detection. As the magnitude of the allpass filter's pole was changed, the slope of the phase transition changed; as the pole was moved closer to the unit circle, the transition became steeper. The four slopes shown are: 180, 150, 120 and 90°/ERB.

Multiplying the output of the 4th-order gammatone filter with the allpass filter output resulted in the same phase-opponency as the PO detector of Chapter 1 that used two separate gammatone filters to accomplish the phase difference. The advantage of using the allpass filter was that it allowed for a more direct control of a key parameter, the phase-difference between the inputs to the multiplier, independent of the detectors' bandwidth (which was controlled by the single gammatone filter).

With the two-filter PO detector of Chapter 1, this phase-difference could not be changed independently of the detector's overall bandwidth. This was because the phase-difference was achieved by the relative spacing and bandwidth of the two gammatone filters. To achieve a phase-difference over a narrower frequency range, the two gammatone filters had to be spaced closer together, with narrow bandwidths. In this case, the detector had a narrower bandwidth, as the two filters were narrow. With the allpass filter, to achieve a phase-difference over a narrower frequency range, only the magnitude of the pole was changed, leaving the bandwidth of the single gammatone filter unchanged.

The output of the gammatone filterbank and the allpass filter were multiplied together and subject to a signum function. Recall that for a PO detector, when the two inputs to the multiplier (the output of the gammatone and allpass filters) are out-of-phase the output is negative. Therefore, it is only how the sign of the multiplier output changes with time that is important, and not the magnitude of the multiplier output. The signum function, which simply returns the sign of the input, results in only temporal information passing through later stages of the detector. The output of the signum function was then subjected to a moving average filter to smooth it.

The relationship between the PO detector's gammatone filter bandwidth and the allpass filter's phase transition determines the detector's performance. The correlator output, before the signum function, was derived based on the simple system shown in Fig. 4-4A. The expected value of the output can be expressed as (Van Trees, 1971):

$$E[y(t)] = E[x_1(t)x_2(t)]$$
$$= R_{x_1x_2}(\tau)$$

where $R_{x1x2}(\tau)$ is the cross-correlation function between the gammatone filter and the allpass filter.

The cross-correlation function between the gammatone filter and allpass filter can be further simplified because the PO detector only relies on the zero-lag of the cross-correlation function, (i.e. tau = 0):

$$R_{x_1x_2}(\tau) = \int_0^\infty h_2(\eta)R_{x_1}(\tau-\eta)d\eta$$
$$= \int_0^\infty h_2(\eta)R_{x_1}(-\eta)d\eta$$

where $h_2(\eta)$ is the transfer function of the allpass filter and $R_{x1}(-\eta)$ is the autocorrelation of the output of the gammatone filter, which can be determined by the following equation:

$$R_{x_1}(\varphi) = \int_0^\infty\int_0^\infty h_1(\eta)h_1(\alpha)R_x(\varphi+\alpha-\eta)d\eta d\alpha$$

where $R_x$ is the autocorrelation of the input noise. The above equations were simulated in MATLAB to examine the expected value of the correlator response as a function of both

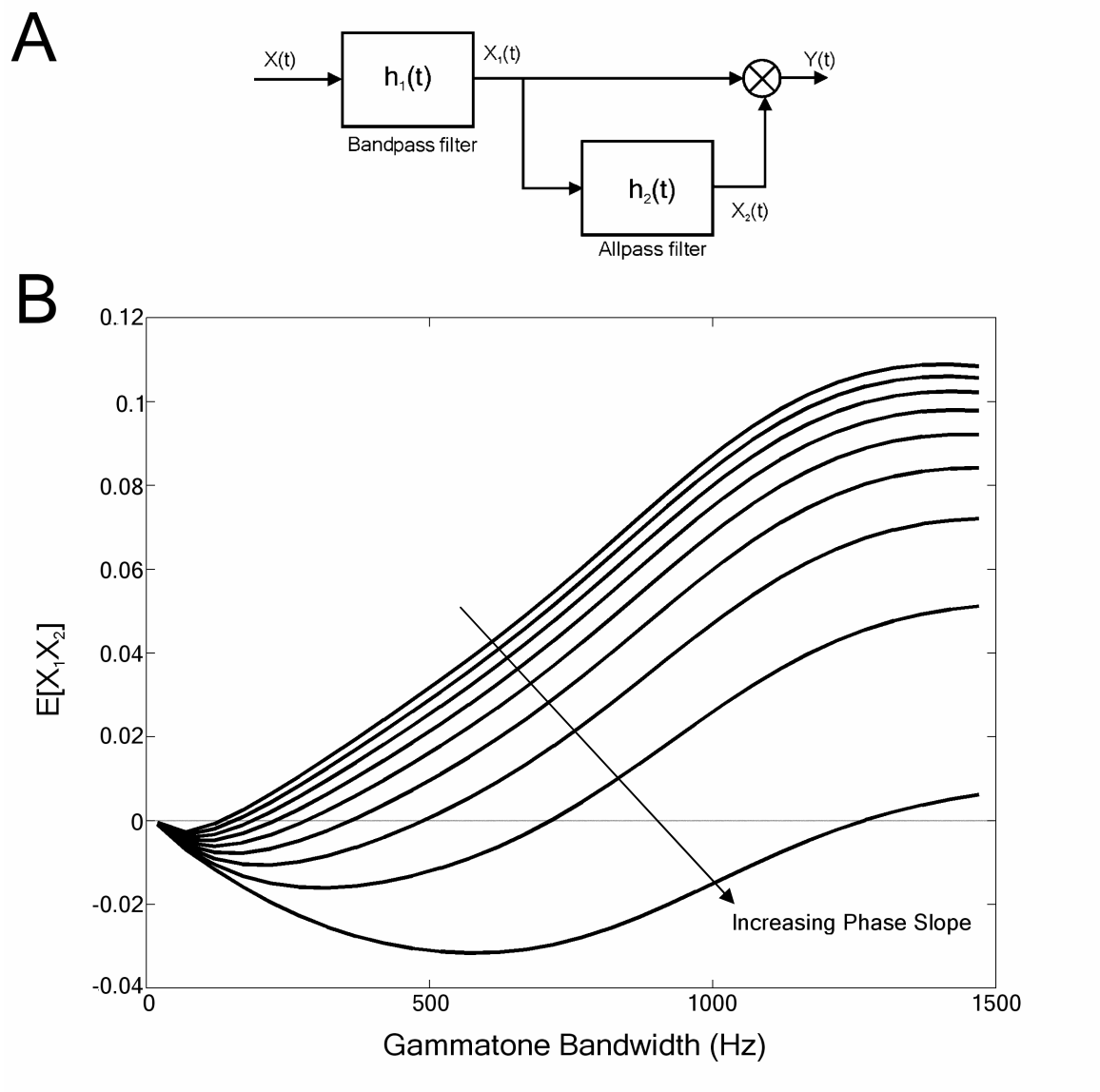


**Figure 4-4 Analytical Results for PO Detector's Correlator**

Shown in the top panel is a simplified version of the PO detector, before the signum function. Using this arrangement, an analytical expression was found for the expected output of the correlator, which is shown in (B) for allpass filter slopes from 20-180°/ERB. As the gammatone filter was made broader, the frequency regions outside the allpass phase transition dominated, leading to a positive expected value. At narrower bandwidths, the allpass filter output dominated, leading to a negative expected output. Because the addition of a narrowband signal resulted in a decrease in correlation, the gammatone filter required a sufficiently large bandwidth such that the output without such a narrowband signal was positive.

allpass filter slope and gammatone filter bandwidth. The results are shown in Fig. 4-4B, where the expected value of the correlator response is plotted as a function of the gammatone bandwidth. When the bandwidth of the filter was large compared to the frequency range of the allpass filter, the expected value of the output of the correlator was positive; the frequency regions outside of the allpass filter's phase transitions dominated. As the bandwidth was made smaller, the expected value of the output became negative, and the allpass filter's phase transition dominated. Because the output of the PO detector was subject to a signum function, the location of the zero-crossing of the expected value was important; if the bandwidth of the gammatone filter was too small, the output of the correlator was always negative. The presence of a narrowband signal in the phase-transition of the allpass filter resulted in the output becoming negative; therefore, the bandwidth of the gammatone filter had to be sufficiently large to guarantee that the output was generally positive when no signal was present.

Taken together, the multiplier, signum function, and moving average filter combined to perform a running correlation between the output of the gammatone and allpass filters. When a signal was present, the output of the correlation was pulled towards -1. The output of the correlation was compared to a threshold $\lambda$ to determine the presence of a speech component in the frequency region covered by the PO detector.

Each PO detector controlled the gain of the frequency band to which it was tuned. When a signal was present (i.e., the output of the correlator was below the threshold for the band), the gain for the band was set to unity for the period of 10 ms. This extension of the gain was done to improve the performance of the detector; the PO detector reliably

detected the onsets of speech components, but often failed to detect the middle parts of a speech component. By extending the gain, the detector filled in the missed signal.

The gain of the each frequency band was set to 0.2 whenever that band's PO detector determined there were no speech components in the band. The reduction in gain attenuated that frequency band, which was presumed to contain only noise. The gain was not set to zero because this leads to an increase in the amount of "musical" noise (Berouti, 1979). By limiting the amount of attenuation, it has been shown that the amount of musical noise can be reduced (Berouti, 1979). Mathematically, the gain ($G_k(t)$)of the $k^{th}$ band was derived based on the correlator output $C_k$ by the following equation:

$$G_k(t) = 0.8(T_k(t) * h_{ext}(t)) + 0.2,$$

$$\text{where } T_k(T) = \begin{cases} 0, C_k(t) > \lambda \\ 1, C_k(t) \le \lambda \end{cases}$$

$$\text{and } h_{ext}(t) = u(t)u(t - 0.01)$$

$$u(t) = \begin{cases} 1, t \ge 0 \\ 0, t < 0 \end{cases}$$

The matrix $G_k$(t) was referred to as the binary mask, because at each point in time, the value can had one of two values: 1.0 or 0.2. Examples of four such masks are shown in Fig. 4-5. Areas in black represent time-frequency positions that had unity gain, all other areas were multiplied by 0.2. The binary mask was similar to a mask used in semiconductor manufacture; multiplication by the binary mask results in the passing of

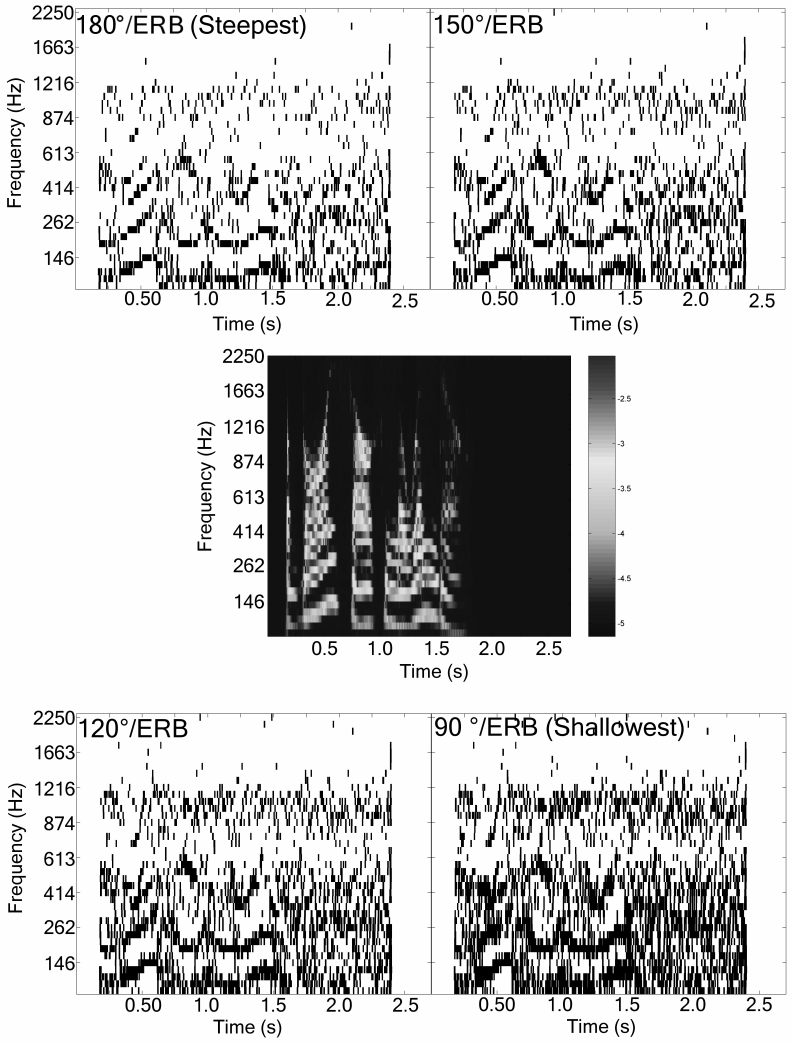


**Figure 4-5 Examples of Binary Masks**

A spectrogram of the sentence "A boy fell from the window" is shown in the middle of the figure. Around the spectrogram are the binary masks produced by the PONR algorithm for the 4 allpass filter slopes used (180, 150, 120, and 90°/ERB) when long-term speech spectrum noise was added at a 0 dB SNR. The detection of the speech components can be clearly seen for all four conditions for times around 0.5 s and 1-1.5 s, where the binary mask closely resembles the spectrogram. The presence of false alarms is illustrated in the time period of 1.75-2.4 s. As the allpass filter slope was made shallower, the detection performance increased, as seen by the larger areas that correspond to the spectrogram. Note the widths of the harmonics, going from only one frequency band in the 180°/ERB condition to 2-3 bands in the 120 and 90°/ERB condition. The number and rate of false alarms also increased with allpass filter slope, as can be seen in the increasing dense pattern of pixels in the time period after 1.75 s.

energy where speech was detected, and attenuating all other regions. The effect of the PONR algorithm is shown in Fig. 4-6.

### 4.2.1.2a PO Detectors vs. Frequency

The PO detectors were matched to the lower frequency bands of the analysis filterbank, with each frequency band having 1 PO detector. The PO detectors only covered those frequency bands below 2.5 kHz (bands 1-40). The higher frequency bands were not subject to detection because above 2.5 kHz, speech components become more noise-like in quality. The PO detectors were unable to detect these components because the detectors required that the signal have a more defined temporal structure.

The gammatone filter of the PO detector had a center frequency equal to that of its associated frequency band. The bandwidth of the PO detector's gammatone filter was equal to three times the ERB of its center frequency. This bandwidth was very wide compared to that of the analysis filter for that band; this wide bandwidth was necessary for the proper functioning of the detector, as mentioned above.

The absolute slope of the allpass filter was changed as a function of frequency, to account for the non-uniform bandwidths of the analysis filterbank. However, the slopes were equal when expressed in terms of degrees/ERB. Four slopes were used in the processing: 180, 150, 120 and 90 °/ERB. The allpass filter always had its -180° point matched to the center frequency of the band.

The duration of the smoothing filter was also changed as a function of frequency, with the duration set to the maximum of either one period of the center frequency or 3 ms.

The threshold for the presence of a speech component within individual PO detector's frequency band was kept constant across PO detectors; the threshold ($\lambda$) was set at -0.55.

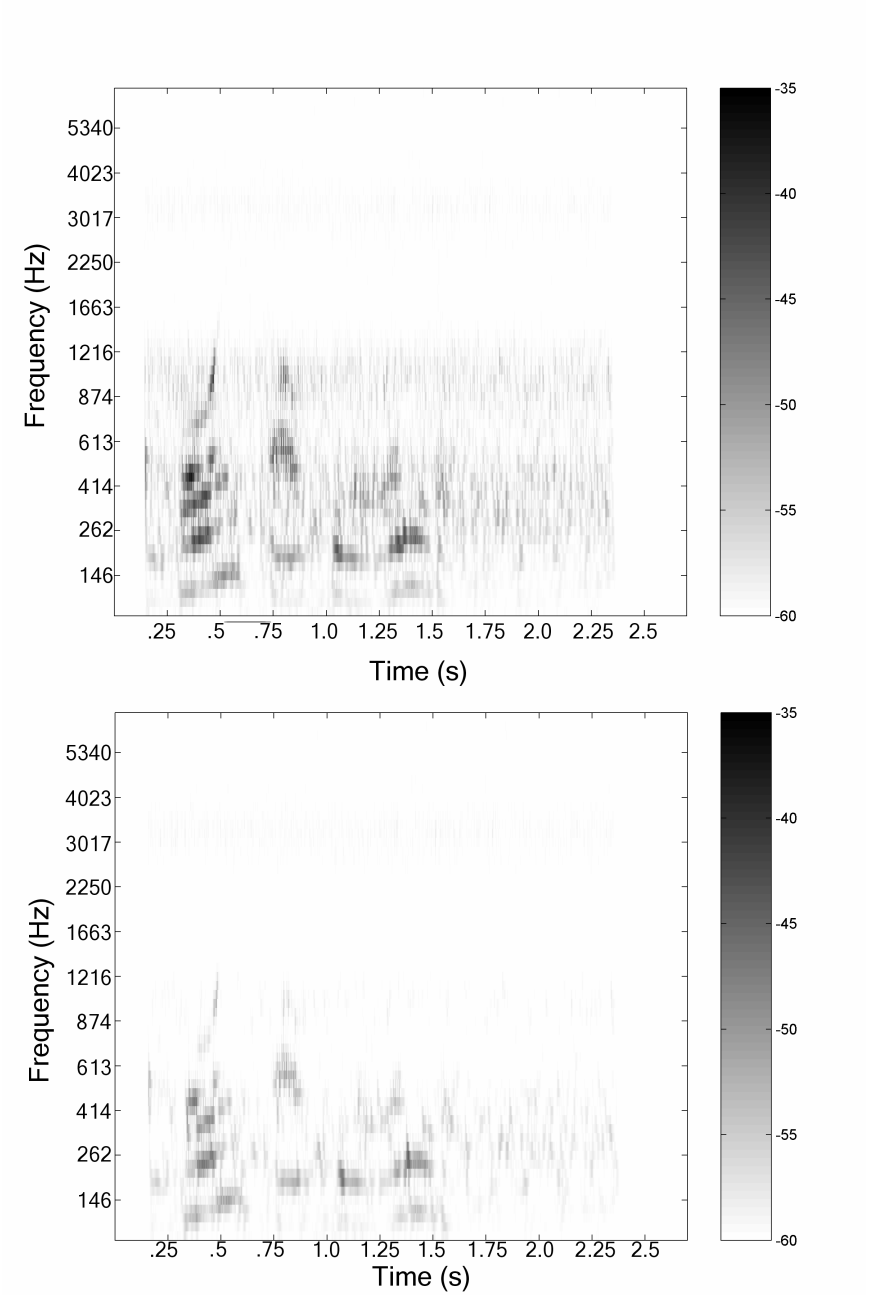


**Figure 4-6 Application of the PONR Algorithm**

The top panel is a spectrogram for "the boy fell from the window" at a SNR of 0 dB. The bottom panel is a spectrogram of the output from the PONR algorithm with an allpass slope of 180°/ERB. While the noise was attenuated, the speech was as well.

This threshold was chosen based on optimizations of the overall system, as explained in the results section.

#### 4.2.1.3 Synthesis Stage

To resynthesize the signal, the gains were applied to the individual frequency bands and the bands combined. To apply the gains, the outputs of the analysis filterbank were synchronized with the gain profile for each band by delaying the output of each analysis filter. This delay of the analysis filters was necessary because of the delay introduced by the PO detector. Because the bandwidth of the PO detector varied with frequency, the delay introduced by the PO detector also varied with frequency. The delay for the $k^{th}$ band was computed based on the group delay of the PO detector's gammatone filter, which was approximated by the equation:

$$g_k = 4\tau_k = \frac{4}{2\pi(1.019 * B_k)}$$

where $B_k$ is the bandwidth of the $k^{th}$ PO detector's gammatone filter bandwidth. The four is from the order of the gammatone filter, and the approximation of $\tau$ from the gammatones bandwidth from Patterson (1987)

After delaying the analysis filterbank outputs, the outputs were multiplied by the gains determined by the PO detectors. The masked waveforms were then further delayed and scaled, such that the impulse responses of the analysis filterbanks' filters aligned with a delay equal to the largest group delay of the PO detector plus 4 ms. The 4 ms is necessary to account for the delays introduced by the analysis filterbank; 4 ms was chosen based on the results of Hohmann (2002) who demonstrated that this delay was

sufficient for quality reconstruction. The waveforms were also scaled to achieve a near uniform gain across frequency when the bands were summed (Hohmann, 2002). The final output, given the gain from the $k^{th}$ PO detector ($G_k(t)$) and the output of the $k^{th}$ analysis filter ($A_k(t)$), was:

$$y(t) = \sum_{k=1}^{40} G_k(t) A_k(t - d_k) S_k + \sum_{k=41}^{59} A_k(t - d_k) S_k$$

$$d_k = g_1 - g_k + d_{Ak}$$

with the values of $S_k$ and $d_{Ak}$, the band scaling and delay factors respectively, derived from Hohmann (2002). The first summation was over the lower-frequencies (<2.5kHz) and the second summation resulted in the addition of the original higher frequency components with no change, other than a delay to resynchronize them with the lower frequency bands.

### 4.2.2 Stimuli

The stimuli consisted of the 250 sentences and associated noises of the Hearing-in-Noise-Test (HINT) (Nilsson et al., 1994). The sentences and noises were combined at SNRs ranging from -10 dB to 10 dB, and passed through the PONR algorithm, with allpass filter slopes of 180, 150, 120 and 90°/ERB. The sentences were also passed through the analysis and synthesis stages without the application of the binary mask; this was the control condition, in which the only processing that affected the speech was that introduced by the analysis and synthesis stages.

### 4.2.3 Listeners

Both normal-hearing listeners and listeners with hearing loss were used. Normal-hearing listeners consisted of 5 subjects, 1 male and 4 female, aged 20 to 27 (mean: 22.8). All normal-hearing listeners had thresholds less than 15 dB HL for frequencies between 250 and 8000 Hz.

Audiograms for listeners with hearing loss are shown in Fig. 4-7. A total of 5 listeners with hearing loss were used, 2 males and 3 females aged 68 to 78 (mean 71.5). Subjects were all classified as having mild-to-moderate sensorineural hearing losses. All subjects with hearing-loss had bilateral, symmetrical (less than 10 dB difference) losses, and were experienced hearing-aid wearers. Listeners with hearing loss performed all listening unaided, with no spectral shaping applied to the stimuli. All subjects were paid for their participation, and all experiments were approved by the Syracuse University Institutional Review Board.

### 4.2.4 Experimental Procedure

Listeners were seated in a double-walled sound-attenuating booth (Acoustic Systems). Preprocessed stimuli were presented through a TDT System II 16-bit D/A system and digital attenuator (TDT PA4), and amplified by a Crown D-75A power amplifier. Stimuli were presented through a speaker 1 m directly in front of the listener. The level of the stimuli was adjusted for each listener with hearing loss to ensure that the stimuli were audible, yet remained comfortable for the listener; presentation levels varied from 65 dB (A) to 90 dB (A). For listeners with normal hearing, a level of 65 dB (A) was used.

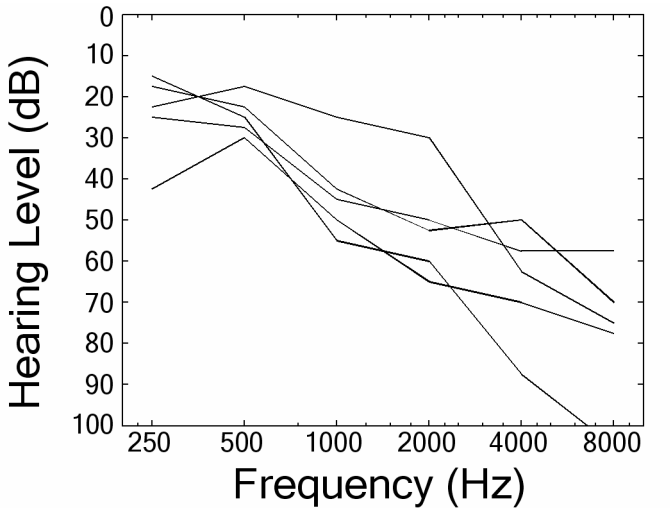


**Figure 4-7 Audiograms for Listeners with Hearing Loss**

### 4.2.4.1 HINT Procedure

The standard HINT procedure (Nilsson et al, 1994) was used to measure the reception threshold for speech (RTS), which is the SNR required for 50% intelligibility. The HINT test was begun by presenting a sentence (either processed or the control condition) at a SNR of -10 dB. The listener was asked to repeat the sentence; the SNR was raised in 4-dB increments until the listener was able to repeat the first sentence of the track with 100% accuracy. The next sentence was then played at a SNR 4 dB down from the first sentence. Again, the listener was asked to repeat the sentence; if the listener repeated the sentence correctly, the next sentence was played at a lower SNR. If the response was incorrect, the next sentence was played at a higher SNR. The remaining sentences of the 20-sentence track were not repeated, with the SNR changing by in steps of 4 dB for the first four sentences of the track, and 2 dB steps for the remaining sentences. The final RTS was the average of the SNRs of the last sixteen sentences, as well as the SNR at which the 21$^{st}$ sentence would have been presented (Nilsson et al., 1994).

RTSs were obtained from all subjects for five conditions: the control condition that had only the analysis/synthesis stages, and PONR with allpass filter slopes of 180, 150, 120, and 90°/ERB. For each condition, the RTS was obtained using two 20-sentence tracks. Listeners were given two 20-sentence tracks of unprocessed HINT to familiarize them with the task; after this familiarization, the remaining ten tracks were presented in random order to each listener.

#### 4.2.4.2 Preference Testing

After subjects were finished running all of the HINT sentences, a preference test was performed, similar to Hanson (2002). Subjects were presented with two HINT sentences, one of which was unprocessed while the other was processed with one of the PONR systems; the SNR was set at 2 dB above the RTS found using the HINT for the control condition (using only the analysis/synthesis stages with no gain changes). The order in which the two sentences were presented was random. The subject was asked to choose which sentence they would prefer to listen to in an everyday situation, and then rate the strength of preference in one of three categories: weak, moderate, or strong (Hanson, 2002). Twenty-five presentations were made for each allpass filter slope using a pool of the first 25 HINT sentences. The order in which the comparisons were made was randomized for each subject, with the randomization across all of the allpass filter slopes.

## 4.3 RESULTS

### 4.3.1 PONR Algorithm Performance

The parameters of the PONR algorithm were initially based on a visual comparison of the detection patterns produced by the detectors to the spectrograms of the clean speech. The parameters included the bandwidth of the PO detectors' gammatone filters, the slope of the allpass filter, the threshold (which was kept constant across bands), and the amount of time that the output was "extended" after a detection. Once the parameters were initially selected, a more systematic search was performed. The search involved varying all parameters except the allpass filter slope around the initial parameters and obtaining the binary mask for that parameter set. The allpass filter slope was kept constant at 120°/ERB. A constant allpass filter slope was used because once the remaining

parameters were set, the slope was then varied to change the detection performance of the bank of PO detectors, as described below.

From the binary mask and the spectrogram of the clean speech, the percentage of energy was calculated by summing the energy in the spectrogram for those regions that had unity gain in the binary mask across time and frequency and dividing by the total energy of the spectrogram. This percentage of energy detected is analogous to the probability of a correct detection, but weights the correct detections by the amount of energy contained in the sample correctly identified. Similarly, the percentage of false alarms was determined by summing the number of samples that had unity gain in the binary mask, but contained no energy in the spectrogram. This sum was then divided by the total number of samples in the spectrogram that contained no energy. The difference between the percentage of energy detected and the percentage of false alarms was then used to determine the parameters used in the PONR algorithm; those parameters leading to the largest difference (~50%) were used. An example spectrogram of the sentence "A boy fell from the window" at 0 dB SNR is shown in Fig. 4-6 before (top panel) and after processing with PONR (bottom panel).

The percentage of energy detected and the percentage of false alarms are illustrated in Fig. 4-8 as a function of the allpass filter slope. Also shown is the difference between the two, which shows a peak at around 110°/ERB. As the filter slope was made steeper (towards the right in Fig. 4-8), the percentage of energy decreased as the PO detectors become more selective. Conversely, as the allpass filter slope was made shallower (towards the left of Fig. 4-8), the percentage of false alarms increased. This trend of increasing energy detected and false alarms is illustrated in Fig. 4-5; binary masks for the

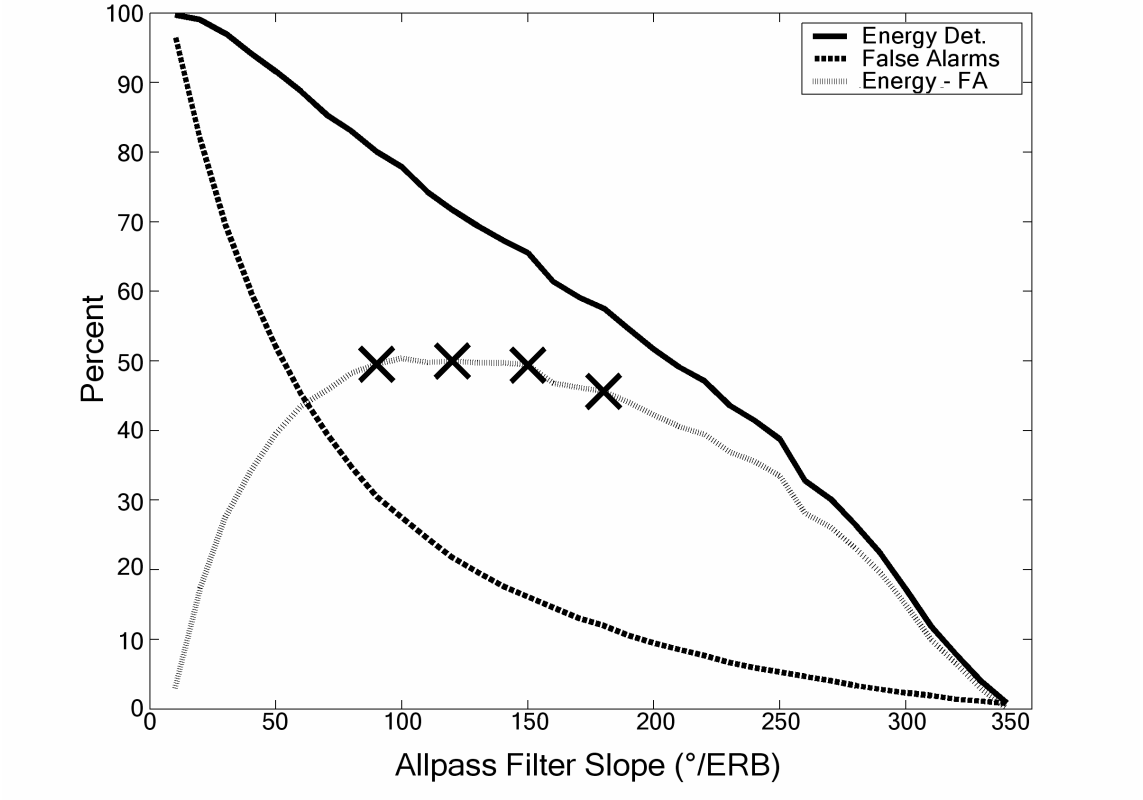


**Figure 4-8 Detection Performance of the PONR Algorithm**

The stimuli that was used to determine the detection performance shown was the sentence "A boy fell from the window" with long-term speech-spectrum noise added at a 0 dB SNR. The solid line is the percentage of energy detected as a function of the allpass filter slope; the dashed line shows the percentage of false alarms, with the difference between the two the dotted line. The difference between the energy detected and false-alarm percentage was used to optimize the parameters of the system. Both the amount of energy detected and the percentage of false alarms decreased with increased allpass filter slope. Because the percentage of false alarms decreased faster than the energy detected, a peak in the difference was obtained located around 110°/ERB. The four allpass filter slopes used during the experiment are shown with the large Xs, and represent a range of both percent energy detected and false-alarm rates.

four slopes used in the PONR are shown, along with the spectrogram of the speech-in-quiet. Examining the steepest slope (180°/ERB), one can see that the components of the speech were correctly identified around 0.50s as well as between 1.0-1.5s. However, comparing the binary mask to the spectrogram, the regions where the binary mask detected these components was rather selective in frequency, with detection occurring in a single frequency band while the spectrogram showed energy across 2-3 frequency bands. The false alarms present in the system can be seen for times greater than 1.75s, where there are regions of unity gain, even though no speech energy was present in the spectrogram. As the slope became shallower, the speech component detection became broader and better matched the spectrogram, which resulted in an increase in the percentage of energy detected. However, concomitant with this increase was an increase in the number of false alarms, seen as the growing density of black regions for times greater than 1.75s.

The application of each of the four allpass slopes resulted in an improvement of the SNR of the output when compared to the input SNR. The application of the PONR resulted in both the reduction of noise, as well as parts of the speech signal. To get an accurate measurement of the output SNR, the effect of the system on both the signal and noise must be known. To calculate the SNR improvement, the method suggested by Umapathy and Parsa (2003) was used. Briefly, the sentence was added to the noise (N+S) and run through the PONR algorithm. The sentence was then subtracted from the noise (N-S), and run through the PONR algorithm again. The two outputs (N+S and N-S) were then averaged together, which resulted in a cancellation of the signal, leaving only the noise that remained after application of the PONR. Similarly, the two outputs

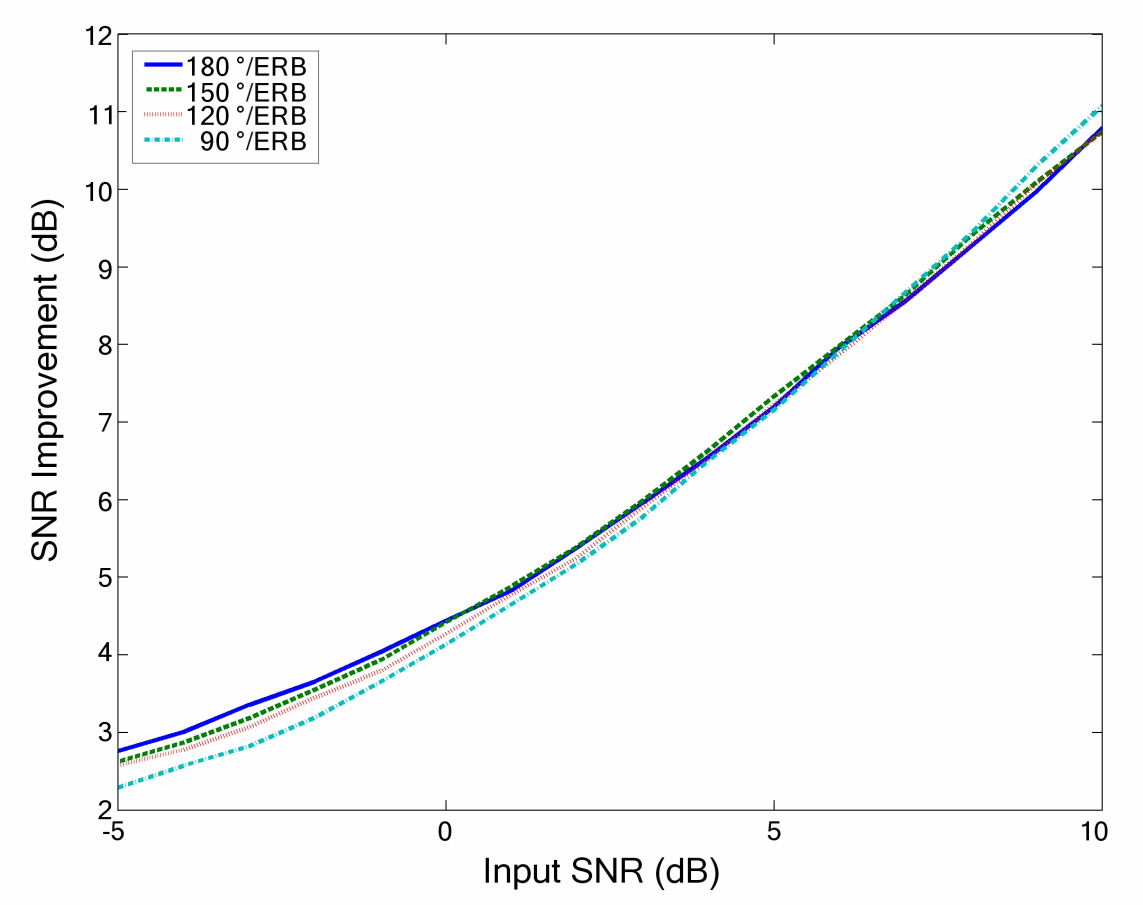


**Figure 4-9 SNR Improvement Obtained with the PONR Algorithm**

Plotted is the SNR improvement as a function of the input SNR, with the slope of the allpass filters used in the PONR algorithm as a parameter. The method used (Umapathy and Parsa, 2003) doesn't take into account distortions to the speech. As the allpass filter slope was made shallower, the SNR improvement decreased, as the shallower allpass filter resulted in more false alarms (Fig. 4-8).

were subtracted, which resulted in the cancellation of the noise, leaving only the processed sentence. The SNR of the output was measured from these two signals. The SNR improvement as a function of the input SNR of the sentence "A boy fell from the window" is shown in Fig. 4-9. At lower SNRs (around 0), all of the PONR systems produced 2-3 dB of SNR improvement. The shallower the slope of the allpass filter, the smaller the improvement of the SNR. This was because the shallower allpass filter resulted in more energy detected as well as an increase in false alarms. The net effect of the false-alarms was to reduce the amount of noise attenuated, as the gain during the false-alarms was unity.

### 4.3.2 HINT Thresholds

RTSs for the listeners with hearing loss are shown in Fig. 4-10. For all subjects, the application of the PONR system appeared to provide no improvement from the control condition, in which there where no gain changes applied to the stimulus. With the exception of the 180°/ERB condition for S4, none of the subjects' RTSs for the processed conditions were statistically different from the control condition (paired t-test, $p < 0.05$). A slight, but not significant, trend for the HINT scores in the processed condition to decrease with shallower allpass slope was evident; this trend was expected, as the shallower the allpass phase slope, the greater the percent energy detected and the greater the percentage of false alarms (Fig. 4-5). As percentage of both energy detected and false alarms increased, the binary mask became a single gain that covers the entire time and frequency axis, essentially resulting in a return to the unprocessed condition.

Shown in Fig. 4-11 are the RTSs for the listeners with normal-hearing. Compared to the listeners with hearing loss, the normal-hearing listeners had a better overall RTS.

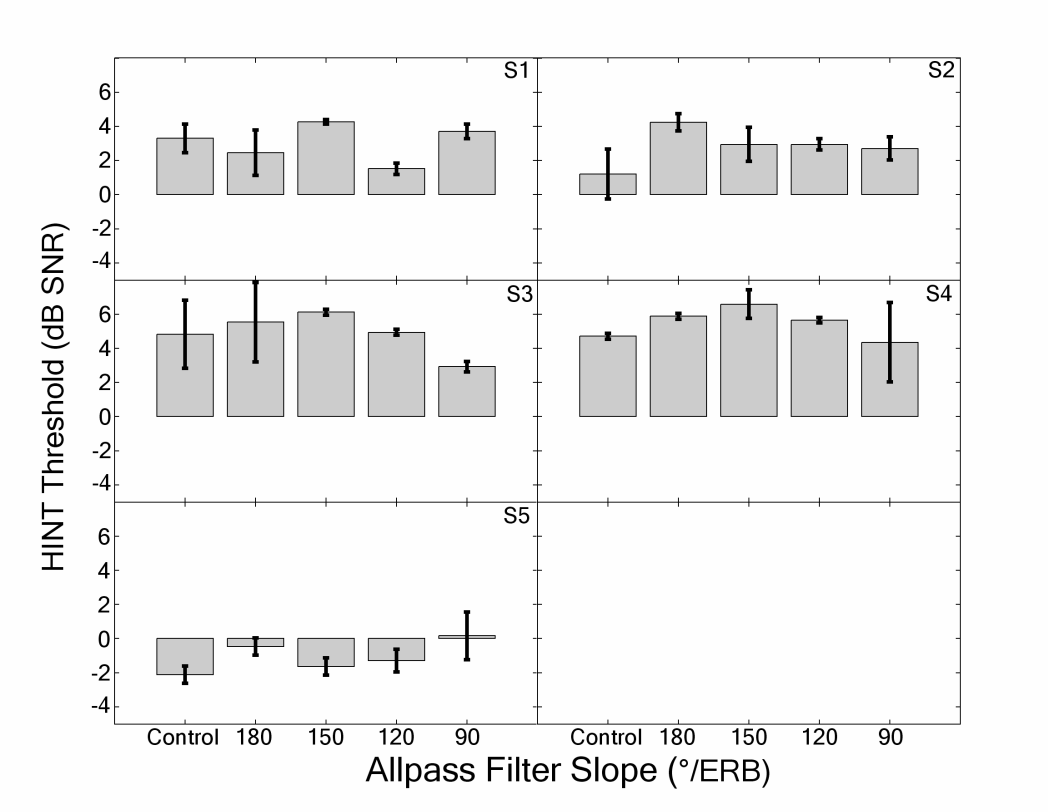


**Figure 4-10 Listener with Hearing Loss HINT Thresolds**

Plotted are the listener's HINT threshold for the control (unprocessed) and 4 processed conditions (PONR with allpass filter slopes of 180, 150, 120 and 90°/ERB). None of the listeners showed an improvement with application of the PONR algorithm; all but the 180°/ERB condition of S4 showed no significant difference from the control condition ($p < 0.05$).

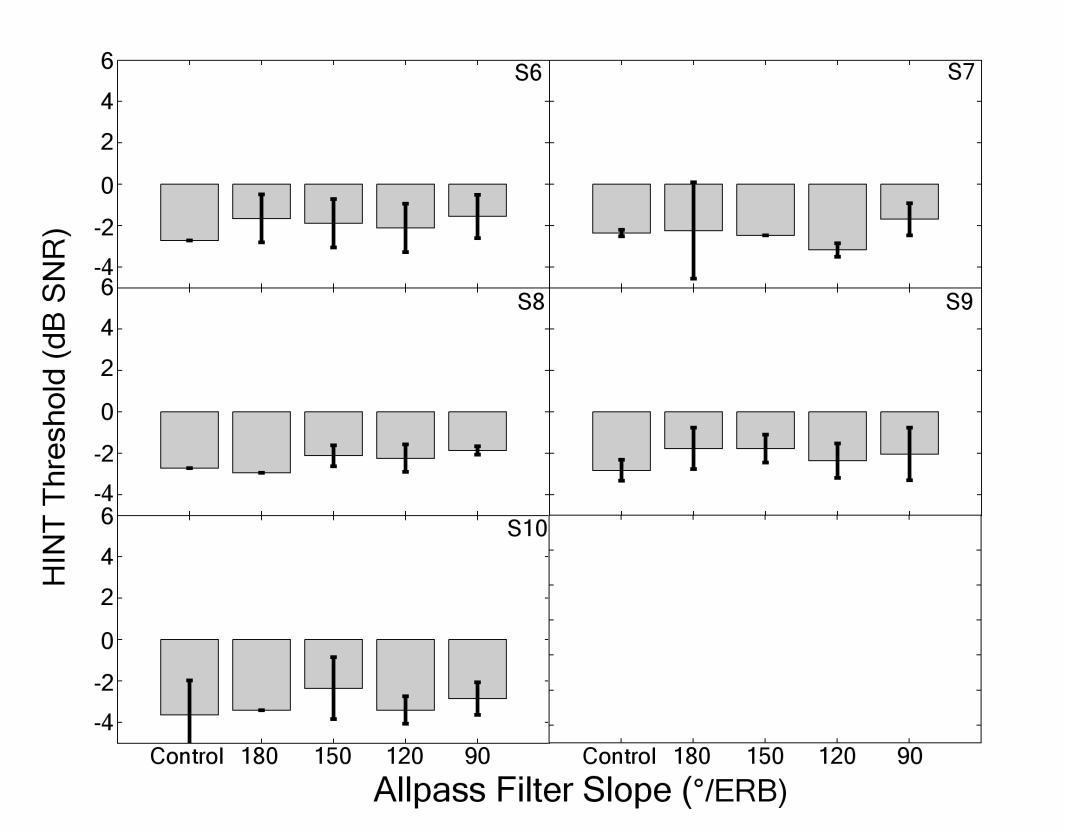


**Figure 4-11 Normal-Hearing Listener HINT Thresholds**

As in Fig. 4-10, the HINT thresholds are shown for the control (unprocessed) and 4 processed conditions (PONR with allpass filter slopes of 180, 150, 120 and 90°/ERB). The general performance of the normal-hearing listeners was better than that of the listeners with hearing-loss. However, none of the listeners showed an improvement in RTS with processing; RTSs were not significantly different ($p < 0.05$) for the processed condition than for the control (unprocessed) condition.

Similar to the listeners with hearing loss, the normal-hearing listeners showed no statistical difference in RTS between any of the processed stimuli and the unprocessed stimuli (paired t-test, $p < 0.05$). The processed conditions appeared to have a greater standard deviation than the unprocessed condition for most of the subjects, but the results are based on only 2 HINT tracks (40 sentences).

### 4.3.3 Preference Testing

Three of the five listeners with hearing loss were available to perform preference testing. All three listeners' preference scores are shown in Fig. 4-12. One of the subjects preferred the PONR-processed stimuli, while the remaining two subjects preferred the unprocessed stimuli. With two exceptions (S3, 90°/ERB and S5, 120°/ERB), all of the listeners' preferences were statistically different from 0. While both S4 and S5 showed an overall preference for the unprocessed stimuli, they occasionally chose the processed stimuli, but always with a weak preference. All listeners claimed that the overall noise was less, but that the speech was less clear. S4 and S5 both indicated that they chose based on the overall clarity of the sentence.

The preference scores of four of the normal-hearing subjects (S7-10) are shown in Fig. 4-13. Three of the listeners showed a clear preference for the unprocessed stimuli, while one showed an overall preference for the PONR processed sentences. All normal-hearing listeners' preference scores were statistically different from 0. The trend for three of the listeners (S7, S8, S9) was for the preference to decrease as the allpass slope was decreased. All listeners indicated that the amount of noise present in the stimulus contributed to their preference; the comments regarding the noise were different,

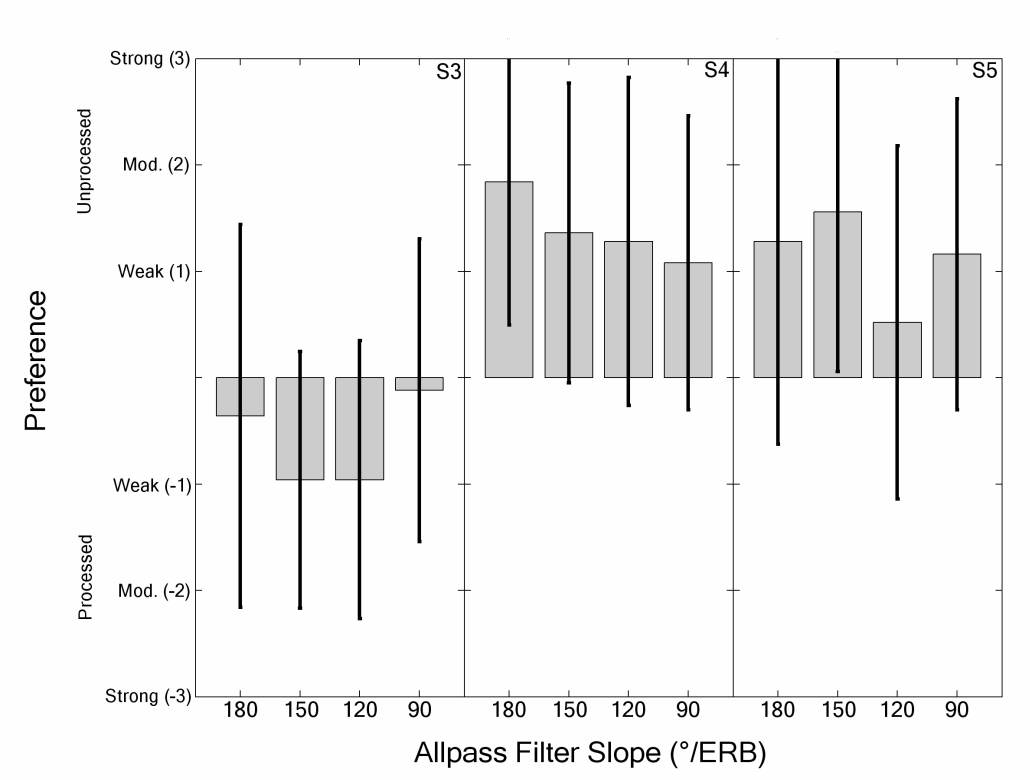


**Figure 4-12 Preference Scores for Listeners with Hearing Loss**

The preference scores for the listeners are shown for the four allpass filter slopes used in the PONR algorithm. Bars above the axis indicated a preference for the unprocessed stimuli, and bars going below the axis indicate a preference for the processed stimuli. Listener S3 preferred the PONR processed stimuli for the three steepest slopes, and was undecided for the shallowest (90°/ERB). Listeners S4 and S5 preferred the unprocessed stimuli for all conditions. S4 showed a trend for a decrease in preference with shallower allpass slopes; this trend matches the amount of false alarms present.

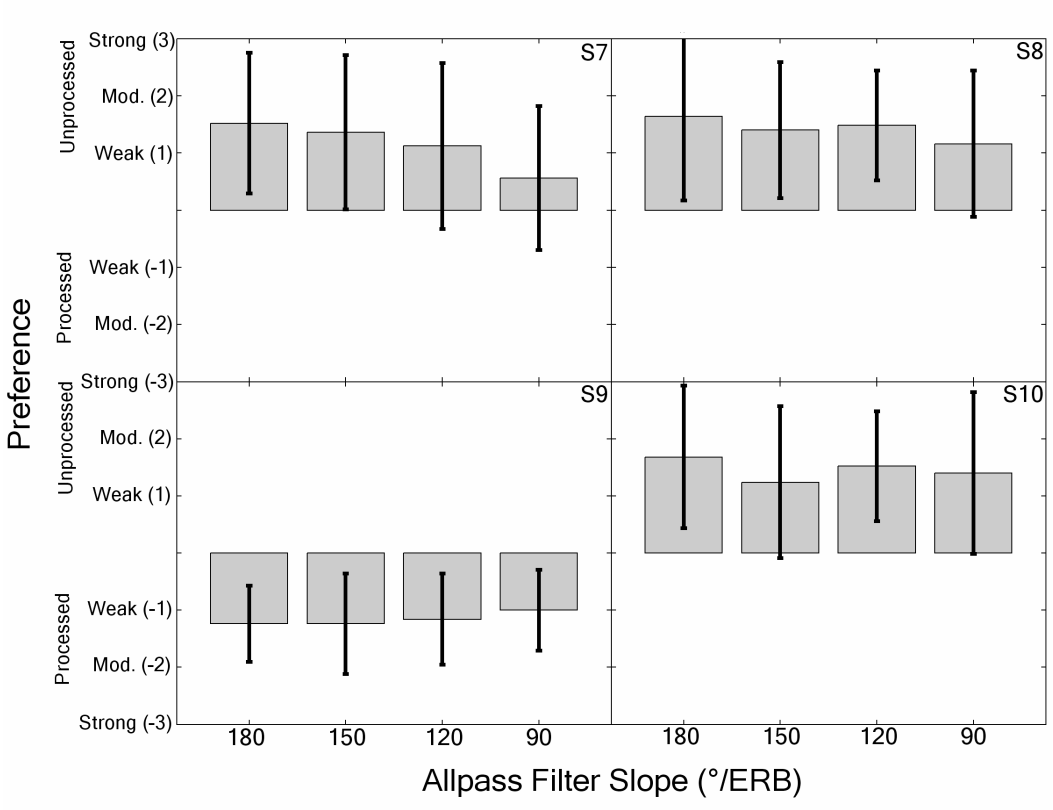


**Figure 4-13 Preference Scores for Normal-Hearing Listeners**

Only S9 showed a preference for the PONR processed stimuli, while the remaining three normal-hearing listeners showed a preference for the unprocessed stimuli. All preferences were statistically different from 0 ($p < 0.05$). Listeners S7-9 showed a trend similar to S4, with preference decreasing with shallower allpass filter slopes.

however, for individual listeners. Listeners S7 and S8 indicated that they preferred the higher-pitched noise that was present in the unprocessed stimuli over the lower-pitched noise that was the result of the false alarms of the PONR system. All three listeners who preferred the unprocessed sentences also stated that they preferred the "clearer" sentence. Listener S9 claimed to pick the sentence based on the ease of understanding the sentence as well as the quality of noise.

All listeners, whether normal-hearing or with hearing loss, showed substantial variation in their preferences. The individual preference scores for S7, a listener with normal-hearing, are shown in Fig. 4-14. The fluctuations observed across sentences may lead to further identification of parameters, as the detection patterns of those sentences for which PONR processing was preferred can be examined. The maximum correlation between the preference patterns across allpass filter slopes for all subjects was approximately 0.5, but were generally much lower. Examining the correlations of patterns across subjects, the maximum correlation was 0.47.

## 4.4 DISCUSSION

Application of the PONR algorithm resulted in an increase of the SNR of the HINT stimuli that were used; this SNR increase, however, did not result in improvement in any of the listeners' RTSs. While they were not statistically significant changes, most listeners' RTSs slightly increased with the PONR processing. This lack of improvement is similar to almost all NR algorithms that have been developed, or are currently being used.

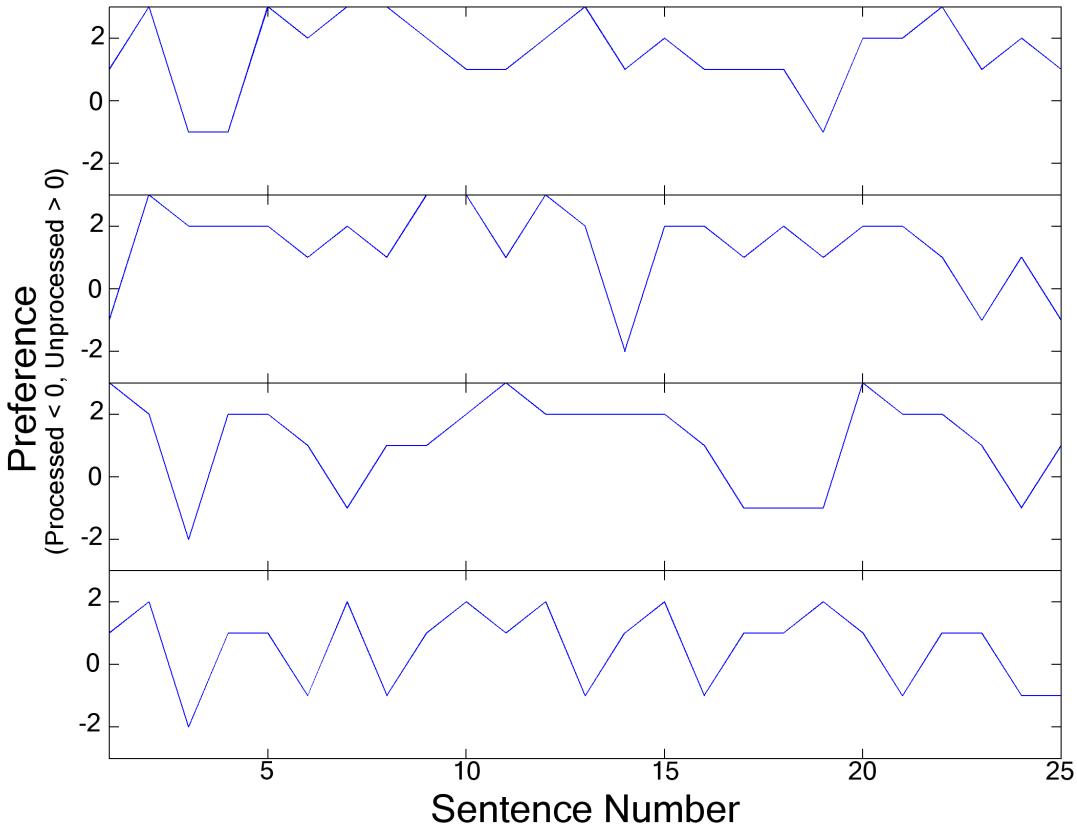


**Figure 4-14 Preference Pattern for Listener S7**

The preferences for the 25 sentences are shown, with each allpass filter slope having one panel, for an individual listener with normal-hearing (S7). Note that while the listener prefered the unprocessed stimuli overall, they did occasionally choose the processed stimuli. The substantial variations in the preference pattern, however, were not correlated across allpass filter slopes, nor were the patterns correlated across listeners (not shown).

The lack of improvement for PONR is not completely unexpected, given the results of Chapter 2. The parameters used in the current implementation of the PONR algorithm result in approximately 80% of the speech energy being detected by the PO detectors (Fig. 4-8). The results of Chapter 2 suggest that 90-95% of the energy must be detected to show an improvement in RTS; to achieve this high level of detection using the current system, the allpass slope could be adjusted, but this would result in a large percentage of false alarms (~70%, Fig. 4-8). Examining the binary mask with the allpass filter phase slope adjusted to achieve these values (95% energy detected, ~70% false alarms) reveals that the PONR algorithm essentially attenuates the frequency bands between 1.4 kHz and 2.5 kHz; these parameters were not used because the algorithm is no longer acting dynamically, but is instead acting as a static filter.

As it is based on a physiological model of detection, it was believed that the PONR system might be able to perform at a lower level than the 95% needed in Chapter 2. If the PONR system was able to detect the energy in a manner similar to humans, the lesser amount of energy detected might produce an improvement in RTS, as it would detect the features that a human would presumably be detecting and using. The lack of improvement in RTS suggests that this was not the case, suggesting that the PO detectors might be performing a different detection task than are the human listeners.

However, the use of a physiologically based detector also poses additional risks in a NR algorithm. If the NR algorithm was using the same detector as the listener, any noises that produced a false alarm in the NR algorithm would also result in the listener having a false alarm. The net result would be that the NR algorithm would reduce the

noise, but the noise remaining might be the noise that is the most effective masker of the physiological system!

As with most methods of NR that are similar to spectral subtraction, the PONR system produced residual noise that has a "musical" quality that was the result of the false alarms in the PO detectors. These false alarms are potentially a more effective masker of the listener, for the reason explained above. Examining the RTSs of the subjects, however, shows that the RTSs remain unchanged across allpass filter slopes, and each of the four conditions had a different amount of false alarms. If the false alarms were a better masker, the RTSs should be the worse for the condition with the most false alarms (the shallowest allpass slope), which was not the case. Unfortunately, the amount of energy detected changed with allpass filter slope in the opposite manner as the false alarms, confounding the issue. Future studies should be performed in which the amount of energy is held constant and the false alarms changed to further examine the net effect of the false alarms.

To achieve an improvement in RTS, the PONR algorithm must be adjusted. The simplest way to do this is to adjust the parameters of the system. The parameters in the current version were optimized based on the difference between energy detected and false alarm rate; a better optimization strategy could be developed that took into consideration other factors such as the overall detection pattern or the continuity of formant tracking.

One of the biggest improvements might be the method of producing a 180° phase shift between the inputs to the PO detector's correlator. While the allpass filter provided this shift, it was only for a specific frequency; the remaining frequencies within that band had slightly different phase shifts, depending on the slope of the allpass filter. The variation

in phase-difference with frequency led to a degradation of detection performance near the edges of the detector's frequency range. As speech is a dynamic signal, it is possible that the movement of formants could shift the formants to these regions of degraded performance. For detection to improve, the system would need a method of producing a constant 180° shift over only the frequency region in which the detector is tuned.

Additionally, further complexity can be added to the system. In the current system, each of the frequency bands' PO detector acted in isolation, basing its decision on what was occurring in that band only. A potential improvement of the system could be seen if the PO detectors were allowed to share information with one another; post-processing involving the outputs of multiple detectors could be utilized to decrease the number of false alarms that were present, while still allowing the speech to be detected. Such post-processing would have to take into account the complex patterns of speech, and examine the detection patterns of all of the PO detectors to see if they match the patterns of speech. By using post-processing, it might be possible to adjust the parameters of the PONR algorithm to allow for better detection of the speech energy (e.g., reducing the allpass slope) without resulting in more false alarms.

Unlike the NR systems used today, most of the listeners did not prefer the PONR processes stimuli over the unprocessed stimuli. While the listeners described the processed stimuli as having less noise, the main complaint was that the processing resulted in speech with reduced clarity. This reduction in clarity could potentially be improved through the detection of more of the speech energy, through the methods described above. In addition, it is possible that the use of 2 filters per ERB may have contributed to the decline in speech quality; using 1 filter per ERB may improve the

speech quality. The PONR system attempted to detect the narrowband components of speech, generally the formants. By reducing the gains in channels between formants, the noise was attenuated; however, the harmonic structure of the speech was also disrupted, as harmonics between formants were also attenuated. It has been previously shown that when speech is modeled by sine waves, the resulting stimuli are often perceived as being of lesser quality than natural speech (McAulay and Quatieri, 1986; Kates, 1994).

In addition, the main goal of the PONR algorithm was the improvement of the listeners' RTSs. To this end, most of the optimizations done to the system were done to improve the percentage of energy detected versus the false alarm rate; it is probable that this optimization resulted in parameters that resulted in some of listeners' preference for the unprocessed stimuli. Many of the parameters can be changed to accommodate the "ease-of-listening" or overall preference of the listener. These would include making the system less aggressive in the removal of noise, perhaps by using one filter per ERB, slowing the gain changes, or attenuating the noise less.

Overall, the PONR algorithm failed to improve the performance of the listeners, both in RTS or in preference. However, application of the algorithm did not degrade the listeners' performances. The results suggest that with further optimization of the PO detectors, it may be possible to improve the performance. Care must be taken to ensure that these optimizations do not result in a decreased preference, as was shown here for a majority of the listeners.

# Chapter 5

## Summary and Discussion

There exists a large amount of information in the literature about the auditory system, and how it detects various stimuli, both psychophysically and physiologically. However, a large gap exists between many of the attempts to solve the problems of listeners with hearing loss and this body of literature. Most of the attempts at solving the speech-in-noise problem have been based on various digital signal-processing methods that do not take into account detailed information gathered about the auditory system. This dissertation attempted to use some of that information, in the form of the PO model, to achieve an improvement in the performance of listeners with hearing loss. While the overall performance of the PONR algorithm was disappointing, the information that was obtained from analyzing the system leads to some interesting questions.

### 5.1 Physiologically Based Detectors

The PO detectors derived in Chapter 2 possessed many interesting qualities. The detectors performed within a few dB of the optimal detector for a tone in white noise, yet were able to exceed the performance of many classical detectors when the noise amplitude was unknown or the noise was amplitude-modulated. The PO detector was unaffected by unknown noise amplitude, or amplitude modulation. In situations where

these conditions may occur, the PO detector is ideally suited. The use of the PO detector is limited, however, to those signals that have a defined temporal structure.

The more general result of the PO detector is that the development of detectors based on physiological detection mechanisms can result in detectors that are extremely good at performing detection, generally near optimal conditions. Most sensory systems have been shaped by evolutionary forces to be near-optimal; signal processing implementations of these systems benefit from these optimizations, while at the same time can overcome some of the limitations of the biological system.

However, the PO detector was unable to detect the required amount of speech energy (90-95%) needed to show improvement. This demonstrates just how difficult the problem of speech detection in noise is; it is likely that additional intelligence is needed in the system that can take advantage of the overall patterns of detection from a large number of simpler detectors of speech characteristics

## 5.2 Limits of Time-Frequency Gain Manipulation

The results of Chapter 3 indicated that the common thinking of many in the hearing-aid community is wrong; a single-microphone NR algorithm can potentially increase the intelligibility of a listener with hearing loss. Unfortunately, the detection performance needed to achieve these gains was higher than expected. This was probably because the overall energy detected was not a good parameter for quantification of the detection performance. Many psychophysical studies have shown that energy is not a reliable cue for many tasks (Kidd et al., 1989; Richards 1992, 2002); it is unlikely that the overall energy is the optimal cue for detection of speech in noise.

Time-frequency gain manipulation resulted in improvements in listeners' RTSs even when the frequency resolution was decreased to 1 filter/ERB. This suggests that the need to increase the frequency resolution of hearing-aids might not exist for the purposes of improving intelligibility. Because of the already limited frequency-resolution of the listener with hearing-loss, it is likely that improvements in almost any metric would not improve with increasing frequency resolution.

Smearing the temporal information, however, had a large effect on the RTSs of the listeners. The results suggest that any system that attempts to improve intelligibility through the use of time-frequency gain manipulation must be able to detect and change the gains of individual frequency bands quickly. This suggests that many of the current NR algorithms in use in hearing-aids today will never show an improvement in intelligibility (regardless of optimizations), as they are all slow acting.

The next step in the process is to further examine the effects of degrading the ideal binary mask. By systematically degrading the ideal binary mask in both time and frequency, the perceptual weighting of the various features of speech can be calculated. This weighting has been done in a crude form with current indexes such as the articulation index (AI) (French and Steinberg, 1947; ANSI, 1969; Pavlovic, 1988; Mueller and Killion, 1990) or speech intelligibility index (SII) (ANSI, 1997). These current indexes attempt to weight the contributions of individual frequency bands to the overall understanding of speech. To fully understand how a listener uses the acoustical features of speech, these types of analysis must be extended into the time-domain. The use of the ideal binary mask allows a simple, yet powerful method for doing this.

Several methods of determining speech intelligibility based on neural models exist, such as the neural articulation index (NAI) (Bondy et al., 2004) or the spectro-temporal modulation index (STMI) (Elhilali, 2004). Both of these indexes use a neural model to determine the intelligibility of speech, similar to the AI or SII, but have been shown to be more robust to various forms of distortion that have an effect on listeners, but not on AI or SII. Evaluating the effects of the binary mask on these indexes may help to separate the effects of the PONR algorithm in terms of benefits from removing the noise and the costs of degrading the incoming speech signal.

All of the experiments performed in Chapter 2 were done unaided, without any spectral shaping to account for individual listeners' audiograms. This lack of amplification may have led to the result that application of the ideal binary mask to the higher frequency regions did not improve performance for listeners with hearing loss' RTSs. Repeating the experiments using spectral shaping to account for the listener's hearing loss may result in further improvements in RTS for the high frequencies.

## 5.3 Phase-Opponent Noise-Reduction

Simply put, the PONR algorithm developed in this dissertation failed to show an improvement when applied to the noisy speech stimuli. This result is consistent with previous work, as well as with Chapter 3. The PO detectors used in the PONR algorithm were unable to detect the 90-95% of the speech energy that Chapter 3 suggested was necessary for improvement in HINT scores. However, it is possible that the results from the PONR algorithm are the result of the large number of false-alarms that the PO detectors produce. Further optimizations of the PO detectors may reduce this number,

which may lead to an improvement in performance. The current PO detectors used in the PONR algorithm had many of their parameters, such as the threshold and bandwidth (ie, 3 times the ERB), fixed with center frequency. Allowing these parameters to change with frequency may improve performance.

In addition, the analysis filterbank and PO detectors were fixed. Allowing the filterbank to dynamically change and reposition to more closely align with the incoming speech's narrowband components may also improve performance. The PO detectors used were also operating in isolation; allowing the detectors to share their outputs, and using additional information about the patterns of speech might improve the detection performance of the overall system.

All of the processing that was performed for Chapters 2 and 3 was done off-line; the sentences were processed and stored on a CD for testing. The development of a real-time system that could perform these types of processing would be beneficial, as it would allow the parameters to be varied as the experiments are occurring. With the current experiments, the processing could take as many as two to three days to create a new set of stimuli with different parameters.

The PONR algorithm made use of the PO detector, a physiological model of the detection of tones-in-noise. It is possible, and highly likely, that the human auditory system uses a separate mechanism for the understanding of speech than it does for the detection of tones, which are an unnatural stimuli that are static and well-defined. The ability of human listeners to understand speech, even when the noise is a competing speaker, also suggests that the PO detector may not be the best detector; the PO detector's performance would be severely degraded under such conditions, as the temporally

defined structure of the competing speech would result in a response from the PO detectors.

The use of a better physiological detector of speech would probably aid in the development of a NR algorithm that could show improvements in speech intelligibility. The question then becomes, what models of speech detection currently exist? Unfortunately, no physiological models do. To that effect, there are still relatively few studies at the level of the auditory nerve that examine the information carried by the AN to higher centers for speech in noise (Delgutte, 1980; Voigt et al., 1982; Sachs et al, 1983; Delgutte and Kiang, 1983; Shamma, 1985; Geisler and Gamble, 1989; Silkes and Geisler, 1991; Geisler and Silkes, 1991). Those that do exist generally find no discernable method of how the speech is encoded at the lower SNRs that are perceptually achieved by human listeners. Further studies must be performed, in the hope that this information could lead to better models.

Current neural models have begun to be used for hearing-aid design. Sachs et al (2002) and Bondy et al. (2004) have demonstrated algorithms that attempt to restore the normal neural representation in models of the impaired auditory system. Dong et al. (2004) have demonstrated a speech enhancement algorithm based on speech segregation derived from psychophysical and physiological models of the auditory system. These studies have shown promising results, but the proposed algorithms haven't been fully tested on listeners with hearing loss. The improvement of neural models, both for the normal as well as the impaired auditory system, could result in improved algorithms that follow the trend of restoring normal neural representations.

# Bibliography

ANSI S3.5 (1969 R 1986). "American national standard methods for the calculation of the articulation index". New York: ANSI.

ANSI S3.5 (1997). "Methods for calculation of the speech intelligibility index." New York: ANSI.

Arehart, K.H., Hansen, J.H.L., Gallant, S., and Kalstein, L. (2003). "Evaluation of an auditory masked threshold noise suppression algorithm in normal-hearing and hearing-impaired listeners," Speech Comm. 40, 575-592.

Bentler, R.A. and Duve, M.R. (2000). "Comparison of hearing aids over the 20th century," Ear Hear. 21 (6), 625-39.

Berouti, M., Schwartz, R., and Makhoul, J. (1979). "Enhancement of speech corrupted by acoustic noise," IEEE ICASSP 4, 208-211.

Boll, S.F. (1979). "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech 27 (2), 113-120.

Bondy, J, Becker, S., Bruce, I, Trainor, L., and Haykin, S (2004). "A novel signal-processing strategy for hearing-aid design: neurocompensation," Sig Processing 84, 1239-1253.

Bondy, J., Bruce, I. C., Becker, S., and Haykin, S. (2004). "Predicting speech intelligibility from a population of neurons," in *NIPS 2003 Conference Proceedings: Advances in Neural Information Processing Systems 16*, eds. S. Thrun, L. Saul and B. Schölkopf, MIT Press, Cambridge, MA, pp. 1409–1416.

Cappa, O. (1994). "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," IEEE Trans. Speech Audio Process. 2, 345-349.

Carhart, R.C. and Tillman, T.W. (1970). "Interaction of competing speech signals with hearing losses," Arch. Ototlaryngol. 91, 273-279.

Carney, L.H., Heinz, M.G., Evilsizer, M.E., Gilkey, R.H. and Colburn, H.S. (2002). "Auditory Phase Opponency: A Temporal Model for Masked Detection at Low Frequencies," Acustica 88, 334-347.

Chabries, D. and Bray, V. (2002). "Use of DSP Techniques to Enhance the Performance of Hearing Aids in Noise," in *Noise Reduction in Speech Applications*, ed. G. Davis, Boca Raton, FA: CRC Press LLC, 379-392.

Cooke, M., Green, P., Ljubomir, J. and Vizinho, A. (2001). "Robust automatic speech recognition with missing and unreliable acoustic data," Speech Comm. 34, 267-285.

Delgutte, B. (1980). "Representation of speech-like sounds in the discharge patters of auditory-nerve fibers," J. Acoust. Soc. Am. 68 (3), 843-857.

Delgutte, B. and Kiang, N.Y. (1984). "Speech coding in the auditory nerve: V. Vowels in background noise," J. Acoust. Soc. Am. 75 (3), 908-918.

Dempsey, J.J. (1987). "Effect of automatic signal-processing amplification on speech recognition in noise for persons with sensorineural hearing loss," Ann Otol Rhinol Laryngol. 96 (3 Pt 1), 251-253.

Dillon, H. and Lovegrove, R. (1993). "Single microphone noise reduction systems for hearing aids: A review and an evalution" In Studebaker G.A. and Hochberg, I. (eds.) *Acoustical factors affecting hearing aid performance, second edition.* Boston: Allyn and Bacon.

Dong, R., Bondy, J., Bruce, I., and Haykin, S. (2004). "Single-microphone speech enhancement using speech stream segregation," in *Abstracts of the IHCON 2004 International Hearing Aid Research Conference*.

Dreschler, W.A. and Plomp, R. (1985). "Relations between psychophysical data and speech perception for hearing-impaired subjects. II," J. Acoust. Soc. Am. 78, 1261-1270.

Edwards, B.W., Hou, Z., Struck, C.J., and Dharan, P (1998). "Signal Processing Algorithms for a new, Software-based, Digital Hearing Device," Hear Jour 51 (9), 44-52.

Eisenberg, L.S., Dirks, D.D., and Bell, T.S. (1995). "Speech recognition in amplitude-modulated noise of listeners with normal and listeners with impaired hearing," J. Speech Hear. Res. 38, 222-233;

Elhilali, M, Chi, T., and Samma, S.A. (2004). "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," Speech Comm 41, 331-348.

Ephraim, Y. and Malah, D. (1984). "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. Acoust. Speech Signal Process. ASSP-32, 1109-1121.

Fabry, D.A. and Van Tassel, D.J. (1986). "Masked and filtered simulations of hearing loss: Effects on consonant recognition," J. Speech Hear. Res. 29, 170-178.

Fabry, D.A. and Van Tassel, D.J. (1990). "Evaluation of an articulation-index based model for predicting the effects of adaptive frequency response hearing aids," J. Speech Hear Res 33, 676-689.

Festen, J.M. and Plomp, R. (1983). "Relations between auditory functions in impaired hearing," J. Acoust. Soc. Am. 73, 652-662.

Fletcher, H. (1940). "Auditory Patterns," Rev. Mod. Phys, vol. 12, 47-65.

Fowler, E.P. (1936). "A method for the early detection of otosclerosis," Arch. Otolaryngol. 24, 731-741.

Geisler, C.D. and Gamble, T. (1989). "Responses of 'high-spontaneous' auditory-nerve fibers to consonant-vowel syllables in noise," J. Acoust. Soc. Am. 85 (4), 1639-1652.

Geisler, C.D. and Silkes, S.M. (1991). "Responses of 'lower-spontaneous-rate' auditory-nerve fibers to speech syllables presented in noise. II: Glottal-pulse periodicities," J. Acout. Soc. Am. 90 (6), 3140-3148.

Gong, Y. (1995). "Speech recognition in noisy environments: A survey," Speech Comm. 16, 261-291.

Glasberg B.R. and Moore B.C.J. (1986). "Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments," J. Acoust. Soc. Am. 79, 1020-1033.

Glasberg B.R. and. Moore B.C.J (1990). "Derivation of auditory filter shapes from notched-noise data" Hear. Res. 47, 103-138.

Graupe, D., Grosspietsch, J.K. and Basseas, S.P. (1987). "A single-microphone-based self-adaptive filter of noise from speech and its performance evalution," J Rehabil Res Dev. 24 (4), 119-126.

Hawkins, D.B. and Yacullo, W.S. (1984). "Signal-to-noise ratio advantage of binaural hearing aids and directional microphones under different levels of reverberation," J Speech Hear Disord. 49 (30, 278-286.

Hanson, M. (2002). "Effects of multi-channel compression time constants on subjectively perceived sound quality and speech intelligibility," Ear Hear. 23 (4), 269-380.

Hippenstiel, R.D. (2002), *Detection Theory Applications and Digital Signal Processing*. Boca Raton, FA: CRC Press LLC, Ch. 5-7.

Hohmann, V. (2002). "Frequency analysis and synthesis using a Gammatone filterbank," Acustica 88, 334-347.

Johnson, D.H. (1980). “The Relationship Between Spike Rate and Synchrony in Responses of Auditory-Nerve Fibers to Single Tones,” *J. Acoust. Soc. Am.*, vol. 68, 1115-1122.

Kalman, R.E. (1960). "A New Approach to Linear Filtering and Prediction Problems," Trans. ASME—Jour. Basic Eng. 82, 35-45.

Kates, J.M. (1994). “Speech enhancement based on a sinusoidal model,” J speech Hear Res. 37 (2), 449-464.

Kay, S.M. (1993). *Fundamentals of statistical signal processing*. Englewood Cliffs, N.J.: Prentice-Hall PTR.

Kidd., G., Mason, C.R., Brantley, M.A., and Owen, G.A. (1989). “Roving-Level Tone-in-Noise Detection,” J. Acoust. Soc. Am., vol. 86 (4), 1310-1317.

Klein, A.J., Mills, J.H. and Adkins, W.Y. (1990). “Upward spread of masking, hearing loss, and speech recognition in young and elderly listeners,” J. Acoust. Soc. Am. 87 (3), 1266-1271.

Klein, A.J. (1989). “Assessing speech recognition in noise for listeners with a signal processor hearing aid,” Ear Hear. 10 (1), 50-57.

Kuk, F., Ludvigsen, C. and Paludan-Müller, C. (2002). “Improving hearing aid performance in noise: Challenges and strategies,” Hear. Jour. 55, 34-46.

Leek, M.R. and Summers, V. (1996). “Reduced frequency selectivity and the preservation of spectral contrast in noise,” J. Acoust. Soc. Am. 100, 1796-1806.

Levitt, H., Bakke, M., Kates, J., Neuman, A., Schwander, T. and Weiss, M. (1993). “Signal processing for hearing impairment,” Scand Audiol Suppl. 38, 7-19.

Levitt, H. (2001). “Noise reduction in hearing aids: a review,” J Rehabil Res Dev. 38 (1), 111-121.

Lippmann, R.P. (1997). “Speech recognition by machines and humans,” speech communication 22, 1-15.

McAulay, R.J. and Quatieri, T.F. (1986). “Speech transformations based on a sinusoidal representation,” IEEE Tran Acoust Speech Sig Proc 34 (6), 1449-1464.

Moore, B.J.C. (1997). *An Introduction to the Psychology of Hearing*. New York, NY: Academic Press, Inc.

Moore, B.C.J. (2003). "Speech processing for the hearing0mpaired: successes, failures, and implications for the speech mechanisms," Speech Comm. 41, 81-91.

Mueller, H.G. and Johnson, R.M. (1979). "The effects of various front-to-back ratios on the performance of directional microphone hearing aids," J Am Aud Soc. 5 (1), 30-34.

Mueller, H.G. and Killion, M.C. (1990). "An easy method for calculation the articulation index," Hear. Jour. 9, 14-17.

Nilsson, M., Soli, S.D., and Sullivan, J.A. (1994). "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. 95, 1085-1099.

Noordhoek, I.M., Houtgast, T. and Feston J.M. (2001). "Relations between intelligibility of narrow-band speech and auditory functions, both in the 1-kHz region," J. Acoust. Soc. Am. 109, 1197-1212.

Ono, H., Kanzaki, J. and Mizoi, K. (1983). "Clinical results of hearing aid with noise-level-controlled selective amplification," Audiology 22 (5), 494-515.

Patterson, R.D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (1988). "Implementing a Gammatone Filter Bank," SVOS Final Report: The Auditory Filter Bank.

Pavlovic, C.V. (1988). "Articulation index predictions of speech intelligibility in hearing aid slection," ASHA 8, 63-65.

Peters, R.W., Moore, B.C.J., and Baer, T. (1998). "Speech reception thresholds in noise with and without spectral and temporal disp for hearing-impaired and normally hearing people," J. Acoust. Soc. Am. 80, 50-57.

Phonak (2000). "Claro Fine-Scale Noise Canceler," www.phonak.com/index.cfm?article_id=2083.

Plomp, R. (1994). "Noise, amplification, and compression: Considerations of three main issues in hearing aid design," Ear Hear. 15, 2-12.

Rankovic, C.M., Freyman, R.L., and Zurek, P.M. (1992). "Potential benefits of adaptive frequency-gain characteristics for speech reception in noise," J. Acoust. Soc. Am. 91 (1), 354-362.

Richards, V.M. (1992). "The detectability of a tone added to narrow bands of equal-energy noise," J. Acoust. Soc. Am. 91 (6), 3424-3435.

Richards, V.M. (2002). "Varying feedback to evaluate detection strategies: the detection of a tone added to noise," J Assoc Res Otolaryngol 3 (2), 209-221.

Ricketts, T.A. and Hornsby, B.W. (2003). "Distance and reverberation effects on directional benefit," Ear Hear. 24 (6), 472-484.

Ricketts, T. (2000). "Impact of noise source configuration on directional hearing aid benefit and performance," Ear Hear. 21 (3), 194-205.

Robertson, G.H. (1967). "Operating Characteristics for a Linear Detector of CW Signals in Narrow-Band Gaussian Noise," Bell Syst. Tech. J., vol. 46, 755-774.

Roman, N., Wang, D. and Brown, G.J. (2003). "Speech segregation based on sound localization," J. Acoust. Soc. Am. 114 (4), 2236-2252.

Sachs, M.B., Voigt, H.F. and Young, E.D. (1983). "Auditory nerve representation of vowels in background noise," J Neurophysiol. 50 (1), 27-45.

Sachs, M.B., Bruce, I.C., Miller, R.L. and Young, E.D. (2002). "Biological basis of hearing-aid design," Ann Biomed Eng 30 (2), 157-168.

Schum, D.J. (2003). "Noise-reduction circuitry in hearing aids: (2) Goals and current strategies," Hear. Jour. 56 (6), 32-41.

Shamma, S.A. (1985). "Speech processing in the auditory system. I: The representation of speech sounds in the responses of the auditory nerve," J. Acoust. Soc. Am. 78 (5), 1612-1621.

Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," Science 270, 303-304.

Silkes, S.M. and Geisler, C.D. (1991). "Responses of 'lower-spontaneous-rate' auditory-nerve fibers to speech syllables presented in noise. I: General characteristics," J. Acoust. Soc. Am. 90 (6), 3122-3139.

Soede, W., Bilsen, F.A. and Berkhout, A.J. (1993). "Assessment of a directional microphone for hearing-impaired listeners," J. Acoust. Soc. Am. 94 (2 Pt 1), 799-808.

Srinivasan, S., Roman, N. and Wang, D.L. (2004). "On binary and ratio time-frequency masks for robust speech recognition," Proc. ICSLP 2004, 2541-2544

Stein, L.K. and Dempsey-Hart, D. (1984). "Listener-assessed intelligibility of a hearing aid self-adaptive noise filter," Ear Hear. 5 (4), 199-204.

Steinberg, J.C. and Gardner, M.B. (1937). "The dependency of hearing impairment on sound intensity," J. Acoust. Soc. Am. 9, 11-23.

Takahashi, G.A. and Bacon, S.P. (1992). "Modulation detection, modulation masking, and speech understanding in noise in the elderly," J. Speech Hear. Res. 35, 1410-1421.

Trees, D.E. and Turner, C.W. (1986). "Spread of masking in normal subjects and in subjects with high-frequency hearing loss," Audiology 25 (2), 70-83.

Tsoukalas, D.E., Mourjopoulos, J.N. and Kokkinakis, G. (1997). "Speech enhancement based on audible noise suppression," IEEE Trans. speech audio Process. 5 (6), 497-513.

Umapaty, K. and Parsa, v. (2003). "Objective evaluation of noise reduction algorithms in speech applications," Paper #Z6-2, 115th *Audio Engineering Society Convention*, New York, USA.

Valente, M., Fabry, D.A. and Potts, L.G. (1995). "Recognition of speech in noise with hearing aids using dual microphones," J Am Acad Audiol. 6 (6), 440-449.

van Dijkhuizen, J.N., Anema, P.C. and Plomp, R. (1987). "The effect of varying the slope of the amplitude-frequency response on the masked speech-reception threshold of sentences," J. Acoust. Soc. Am. 81 (2), 465-469.

van Dijkhuizen, J.N., Festen, J.M. and Plomp, R. (1989). "The effect of varying the amplitude-frequency response on the masked speech-reception threshold of sentences for hearing-impaired listeners," J. Acoust. Soc. Am. 86 (2), 621-628.

van Dijkhuizen, J.N., Festen, J.M. and Plomp, R. (1990). "Speech-reception threshold in noise for hearing-impaired listeners in conditions with varying amplitude-frequency response," Acta Otolaryngol Suppl 469, 202-206.

van Dijkhuizen, J.N., Festen, J.M. and Plomp, R. (1991). "The effect of frequency-selective attenuation on the speech-reception threshold of sentences in conditions of low-frequency noise," J. Acoust. Soc. Am. 90 (2 Pt 1), 885-894.

Van Rooij, J.C.G.M., and Plomp, R. (1990). "Auditive and cognitive factors in speech perception by elderly listeners. II. Multivariate analyses," J. Acoust. Soc. Am. 88, 2611-2624.

van Schinjndel, N.H., Houtgast, T. and Festen, J.M. (2001). "Effects of degradation of intensity, time, or frequency content on speech intelligibility for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. 110, 529-542.

Van Trees, H.L. (1971). *Detection, estimation, and modulation theory*. New York: Wiley.

Virage, N. (1999). "Single channel speech enhancement based on masking properties of the human auditory system," IEEE Trans. Speech Audio Process. 7 (2), 126-137.

Voigt, H.F, Sachs, M.B. and Young, E.D. (1982). "Representation of whispered vowels in discharge patters of auditory-nerve fibers," Hear Res. 8 (1), 49-58.

Wiener, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. New York, NY:Wiley.

Whalen, A.D. (1971). *Detection of Signals in Noise*. New York, NY: Academic Press, Inc.

Yilmaz, O. and Rickard, S. (2004). "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. Sig. Process. 52 (7), 1830-1847.

# VITA

NAME OF AUTHOR: Michael Clark Anzalone

PLACE OF BIRTH: Rockville Centre, New York

DATE OF BIRTH: September 11, 1977

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:
Syracuse University, Syracuse, New York
Boston University, Boston, Massachusetts

DEGREES AWARDED:

Master of Science in Bioengineering, 2001, Syracuse University
Bachelor of Science in Biomedical Engineering, 1999, Boston University

AWARDS AND HONORS

University Fellow, Syracuse University, 2002-2005

PROFESSIONAL EXPERIENCE:

Teaching Assistant, Department of Bioengineering and Neuroscience, Syracuse University, 1999-2001
Research Assistant, Department of Bioengineering and Neuroscience, Syracuse University, 2003, 2005