# Spatial, Spectral, and Perceptual Nonlinear Noise Reduction for Hands-free Microphones in a Car

A Thesis submitted to the faculty
of
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements for
The Degree of Master of Science
in
Electrical and Computer Engineering
by

Jeffery Faneuff
July 2002

APPROVED

Dr. D. Richard Brown III, Major Advisor

Dr. Nathaniel A. Whitmal III, Committee Member

Dr. Peder C. Pedersen, Committee Member

# Abstract

Speech enhancement in an automobile is a challenging problem because interference can come from engine noise, fans, music, wind, road noise, reverberation, echo, and passengers engaging in other conversations. Hands-free microphones make the situation worse because the strength of the desired speech signal reduces with increased distance between the microphone and talker. Automobile safety is improved when the driver can use a hands-free interface to phones and other devices instead of taking his eyes off the road. The demand for high quality hands-free communication in the automobile requires the introduction of more powerful algorithms.

This thesis shows that a unique combination of five algorithms can achieve superior speech enhancement for a hands-free system when compared to beamforming or spectral subtraction alone. Several different designs were analyzed and tested before converging on the configuration that achieved the best results. Beamforming, voice activity detection, spectral subtraction, perceptual nonlinear weighting, and talker isolation via pitch tracking all work together in a complementary iterative manner to create a speech enhancement system capable of significantly enhancing real world speech signals.

The following conclusions are supported by the simulation results using data recorded in a car and are in strong agreement with theory. Adaptive beamforming, like the Generalized Side-lobe Canceller (GSC), can be effectively used if the filters only adapt during silent data frames because too much of the desired speech is cancelled otherwise. Spectral subtraction removes stationary noise while perceptual weighting prevents the introduction of offensive audible noise artifacts. Talker isolation via pitch tracking can perform better when used after beamforming and spectral subtraction because of the higher accuracy obtained after initial noise removal. Iterating the algorithm once increases the accuracy of the Voice Activity Detection (VAD), which improves the overall performance of the algorithm. Placing the microphone(s) on the ceiling above the head and slightly forward of the desired talker appears to be the best location in an automobile based on the experiments performed in this thesis. Objective speech quality measures show that the algorithm removes a majority of the stationary noise in a hands-free environment of an automobile with relatively minimal speech distortion.

# Acknowledgements

First I'd like to thank my family for their love, support, and patience. My wife Wendy has graciously listened to me talk on and on about speech enhancement. Wendy's EE background enabled her to give me insightful feedback. The list of thanks for my wife is far too long to be written here. Honey, thanks for the last 5 years of support. I love you forever. My children Nicholas, Angela, and Benjamin are my inspiration and joy. They always had a hug, kiss, and smile to help keep their Dad going.

Dr. Brown was a great thesis advisor and has given me plenty of sound advice. Our conversations helped me keep on a profitable course when there were many different roads to take. I always came out of our meetings with renewed enthusiasm and a new angle to pursue. I appreciate the comments and suggestions from Dr. Whitmal and Dr. Pedersen who were part of my thesis review committee. I would also like to thank Gary Elko, currently at Avaya, for his practical guidance and suggestions based on his years of experience in speech enhancement. Thanks to Bose for their financial support and use of equipment. Last but not least I'd like to give all the praise and glory to God from whom all wisdom comes.

**TABLE OF CONTENTS**

## LIST OF FIGURES

## LIST OF TABLES

**TABLE OF NOTATION**

| | |
|---|---|
| $c$ | Speed of sound = 342 m/s |
| $d$ | Distance between microphones |
| $dB$ | Decibels |
| $f$ | Frequency in samples / second (Hz) |
| $k$ | Frame index |
| $m_j(n)$ | Signal received at microphone $j$ |
| $M_j(\omega)$ | Spectrum of signal received at mic. |
| $\lvert M(w,k)\rvert$ | Magnitude of spectrum |
| $M(w,k)^*$ | Complex conjugate of $M(w,k)$ |
| $n$ | Discrete time index |
| $N(w,k)$ | Noise estimate for frame $k$ |
| $s(n)$ | Noise free speech |
| $\hat{s}(n)$ | Enhanced speech estimate |
| $\hat{S}(w,k)$ | Spectrum of speech estimate |
| $T(w,k)$ | Mask threshold for frame $k$ |
| $V(k)$ | VAD decision for frame $k$ |
| $\vec{v}(\theta)$ | Steering vector |
| $\vec{x}(n)$ | Vector of signals |
| $z(n)$ | Interfering noise signal |
| | |
| $\theta$ | Phase |
| $\mu_N$ | Mean of spectral noise estimate |
| $\sigma_N^2$ | Variance of spectral noise estimate |
| $\sigma_N$ | Std deviation of spectral noise estimate |
| $\tau$ | Delay between microphones |
| $\lambda$ | Wavelength |
| $\omega$ | Frequency in radians |
| | |
| AI | Articulation Index |
| AWGN | Additive White Guassian Noise |
| BF | Beamforming |
| GSC | Generalized Side-lobe Canceller |
| IS | Itakura – Saito distance measure |
| SNR | Signal-to-Noise-Ration |
| SSNR | Segmental SNR |
| SSPN | Spatial Spectral Perceptual Nonlinear |
| SS | Spectral Subtraction |
| VAD | Voice Activity Detection |

# Chapter 1

# Introduction

This thesis proposes a unique combination of algorithms aimed at suppressing noise in a hands-free phone of an automobile. The demand for hands-free phones in the noisy automobile environment requires more powerful noise suppression algorithms than those used currently in cell phones and conference phones. The availability of inexpensive processing power makes eventual implementation of more sophisticated noise suppression possible in an automobile. The following sections in the introduction describe the hands-free phone challenges in the automobile, selected historical research on speech enhancement, and motivation for the algorithm proposed in this thesis.

## 1.1 Hands-free in the car

The use of hands-free phones in the automobile is motivated by consumer demand, safety, and legal mandate as underscored by the following quote.

> "Beginning December 1, 2001 New York's six million cell phone users may no longer make quick calls home, check stock quotes, or reschedule tee times using a handheld cell phone while driving. New York isn't alone in this prohibition; 38 other states have pending laws limiting handheld cell phone use in automobiles. Plus, England, Italy, Israel, Japan, and 20 other countries have already outlawed arm anchored cellular communication.

> USA Today has estimated that cell phone use will grow from 105 billion minutes in 1998 to 554 billion minutes in 2004. The Cellular Telecommunications and Internet Association estimate of U.S. wireless subscribers will undoubtedly grow greater than its present 117 million as more cars come with wireless equipment. An amazing 70 percent of all cell phone calls in North America originate from automobiles."[1]

The current hand-held phones pose a hazard in the car and the change to hands-free phones in automobiles is already well underway, but customers will demand the same clarity they currently enjoy with hand-held phones. One solution is for people to wear headsets while talking in the car. Problems with headsets include their inconvenience and the likelihood that people will put them on and take them off while driving. Microphones mounted in the automobile are easier to use and introduce less distraction than headsets. A challenge presented with microphones installed in the car is that they are further away from the talker's mouth, which decreases the desired signal strength relative to the surrounding noise. There are many noise sources in the automobile that only exacerbate the problem including passing cars, rain, windshield wipers, engine noise, fans, music, horns, wind, road noise, reverberation, echo, and other talkers; these can all make it difficult to hear the desired speech. Things are further complicated by the auto manufacturers' desire to keep their costs down, thus limiting the amount of microphones and locations where they can be installed. The speech acquired by the microphones mounted in an automobile requires post processing to improve quality and intelligibility to the level expected by consumers who are accustomed to using hand-held cellular phones.

## 1.2 Speech enhancement research

There has been an abundance of research in the area of speech enhancement over the past 40+ years, which has been applied to noise suppression, echo cancellation, talker isolation, and enhancement for perception or recognition. Some of the successful methods used to meet the above challenges fit roughly into the categories listed below.

- **Adaptive filtering**
  Wiener filtering and adaptive filtering assumes a desired response is available and minimizes the difference in a mean-square sense between the output of the filter and the desired output. [2] Wiener filtering is optimal for stationary signals and white noise. Adaptive filters are necessary for the non-stationary signals and are commonly used for adaptive noise cancellation (ANC) [3] and echo cancellation (EC).

- **Spectral subtraction**
  Spectral Subtraction [4] uses an estimate of the noise and short-time spectral analysis to subtract the spectral components of the noise from the received signal, thus improving the signal-to-noise-ratio (SNR).

- **Beamforming**
  Beamforming [5] uses multiple microphones to keep a constant gain in a given direction while suppressing sounds from other directions. Beamformers can also steer deep nulls to block interfering signals at known locations. [6]

- **Blind signal separation**
  Blind signal separation (BSS) uses statistical measures with very little a priori information to separate a signal into its various components. This is useful to remove other talkers, noise, or interference in order to better hear the desired speech. BSS is also practical for implementation because of the relatively few assumptions required.

- **De-correlation**
  De-correlation attempts to estimate the system's transfer functions in order to separate the signals into separate channels or components. [7]

- **De-convolution**
  De-convolution attempts to separate signals from their convolution mixture and accounts for multi-path effects from reverberation. De-convolution requires a very good approximation of the channel effects, which ideally are accurately measured using known signals. [8]

- **Parametric modeling**
  Parametric modeling of the speech production system is a powerful way to characterize and enhance the speech signal. "This technique (Linear Predictive Analysis) has been the basis for so many practical and theoretical results that it is difficult to conceive of modern speech technology without it." [9]

- **Perceptual masking**.
  Human auditory perception causes some noise to be masked by the desired speech. Noise suppression algorithms can use this information to only attenuate the audible noise. [10]

There is still much room for improvement in speech enhancement despite the successful progress made using the algorithms mentioned above. The following quotes call for continued research, especially in the automotive environment.

> "The problem of enhancing speech degraded by noise remains largely open, even though many significant techniques have been introduced over the past decades."[11]
>
> "The majority of speech enhancement algorithms actually reduce intelligibility and those that do not generally degrade the quality. This balance between quality and intelligibility suggests that considerable work remains to be done in speech enhancement."[12]
>
> "The primary barrier to the proliferation and user acceptance of voice based command and communications technologies in vehicles has been noise. The consequences of noise are poor voice signal quality in far field microphones and low speech recognition accuracy for in-vehicle speech command applications. The current commercial remedies, such as noise cancellation filters and noise canceling microphones have been inadequate to deal with the multitude of real world situations, at best providing limited improvement, and at times making matters worse."[13]

This thesis focuses on combinations of algorithms as a step towards improving speech enhancement beyond the current limitations. Listed below is some of the research that has, in a similar fashion, investigated robust solutions by employing a combination of algorithms, which supports the thinking behind the proposed approach in this thesis.

- Using Wiener Filtering, Spectral Subtraction, and Beamforming simultaneously in a real car environment has produced noise reduction of almost 8 dB and significant increases in speech recognition rates.[14]
- An adaptive microphone array and spectral subtraction has been used to produce 20 dB of echo cancellation and 15 dB of noise suppression; this also had the advantage of self-calibration.[15]
- The dual excitation speech model and spectral subtraction combination is another effective combination because of the distinction made between voiced and unvoiced portions of the signal.[16]
- Yet another good mix is decomposition of the signal into eigen-spaces in the context of the Bark domain to take advantage of the masking properties of the human auditory system.[17]

The key to successfully combining algorithms is to leverage the strengths of each approach in a way that still allows them to work together toward the end goal of enhancing the speech. For instance, beamforming provides a gain based on direction, spectral subtraction is a good complementary process because it handles the difficult low frequency ranges where beamforming fails, and perceptual weighting can be used to mask artifacts introduced by spectral subtraction. The following section outlines the proposed combination of algorithms in this thesis.

## 1.3 A multi-dimensional approach

The real-world application of hands-free phones in automobiles occurs in a very noisy environment where the current one-dimensional algorithms do not offer improvement to the received speech signal. Spatial, spectral, and nonlinear perceptual (SSPN) properties can theoretically be used in a complementary fashion to suppress noise more effectively than using any one dimension alone. This thesis analyzes the interaction of beamforming, spectral subtraction, nonlinear perceptual weighting, talker isolation via pitch tracking, and voice activity detection. Careful consideration must be made when designing the combination of these algorithms in order to achieve maximum noise suppression and minimal speech distortion.

- Beamforming takes advantage of a known source location to coherently combine the desired speech from multiple microphones while canceling noise from other directions.
- Noise estimation and spectral subtraction work together to remove stationary noise in the frequency domain.
- Nonlinear weighting of critical frequency bands according to the human auditory perceptual masking characteristics minimizes the negative effects from spectral subtraction.
- Talker isolation can be accomplished using pitch and amplitude tracking, which helps suppress unwanted speech and other periodic noise sources. Talker isolation avoids falsely classifying data frames as voiced when interfering talkers are speaking while the desired talker is silent.

- Voice activity detection is critical to characterize the noise and avoid distorting or attenuating the desired speech.

The algorithms and considerations above were analyzed by running MATLAB© simulations using data measured with a uniform linear array of 4 microphones clipped on the driver's side visor of a 2001 Honda Odyssey mini-van. From these recordings taken in the automobile, road noise, engine noise, interfering talkers, and fan noise were combined with female speech at SNRs of 0, 5, and 10 dB and then processed by the SSPN algorithm. The objective speech quality measures used to analyze the results of the SSPN algorithm were Signal-to-Noise Ratio (SNR), Segmental Signal-to-Noise Ratio (SSNR), Articulation Index (AI), and Itakura-Saito distance (IS).[18]

The MATLAB© simulation results, as measured by the objective speech quality measures, prove that the SSPN algorithm attains better noise suppression and speech quality performance than either spectral subtraction or beamforming alone. Several variations of the SSPN algorithm were attempted before converging on the current design, which produced the best noise suppression while maintaining high speech quality.

The simulation results from several of these variations in the algorithm design suggest that the algorithm is very sensitive to how well the Voice Activity Detector (VAD) performs. The accuracy of the VAD was crucial to the Generalized Side-lobe Canceller and noise estimation for spectral subtraction, so they would not attenuate the desired speech. Spectral subtraction depends on the VAD to update its noise estimate. The VAD

accuracy and overall algorithm performance was improved by limited iteration with spectral subtraction, even without the talker separation in the loop.

The quality of the speech before processing the signal was improved when the location of the microphone relative to the desired talker and relative to the noise sources was taken into account because of its effect on desired signal strength versus the strength of the unwanted noise.

Details of the algorithm, underlying theory, results, and conclusions are presented in the remainder of the thesis and are organized as follows. Chapter 2 provides introductory background material for the read unfamiliar with speech modeling and auditory masking. Chapter 3 describes the proposed algorithm design, data flow, and control flow and Chapter 4 gives the details on the theory and algorithms used in noise suppression that serve as the foundation of the work presented here. Chapter 5 reports on simulation results using real-world data and compares the new algorithm with beamforming and spectral subtraction. Chapter 6 presents conclusions and suggestions for future research. Appendix A describes the psychoacoustics involved in determining human auditory perceptual masking and appendix B provides details on experiments related to microphone location and desired signal strength.

# Chapter 2

# Background material

This chapter contains background information for the reader unfamiliar with far field microphones, modeling speech, auditory masking and speech enhancement in general. A basic understanding of signal processing concepts is assumed.

## 2.1 Source to microphone distance

Microphones that are part of a hand-held or headset communication system are typically on the order of 4 centimeters (cm) from the talker's mouth. Hands-free microphones are anywhere from 20 cm to several meters from the talker's mouth, which degrades the quality of the received speech compared to handset microphones. This thesis is focuses on the application of a hands-free phone in a car using microphones approximately 24 cm from the desired talker's mouth.

How the microphone receives the signal will be different based on whether it is considered in the near-field or far-field situation. Near-field signal arrive at the microphone with spherical spreading while far field signals can be assumed to arrive as a planar wave-front, which is shown in Figure 2.1.

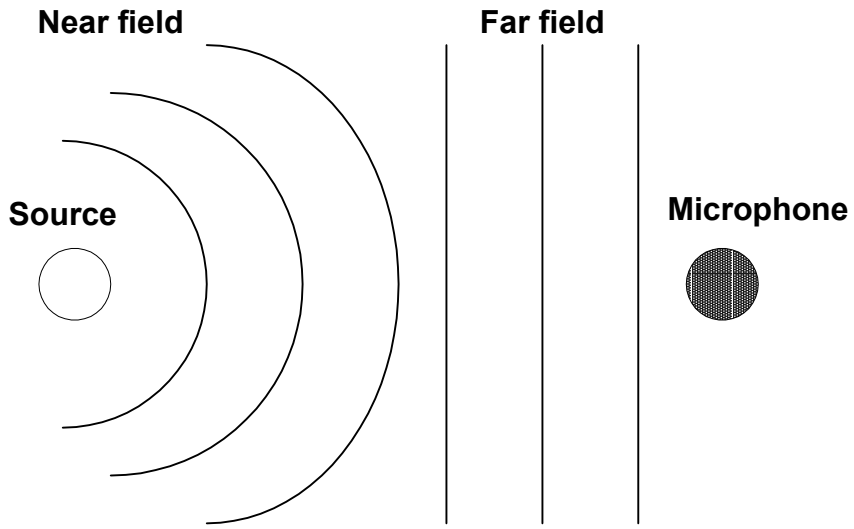


**Figure 2.1: Near field and far field wave propagation**

The near/far field transition is based upon signal wavelength, shape of the source, aperture of the receiving microphone elements, and source to microphone distance. A source is said to be in the near field if equation (2.1) is true, [19]

$$r < \frac{2L^2}{\lambda} \tag{2.1}$$

where $r$ is the radial distance from the microphone, $L$ is the aperture of the microphone array, and wavelength $\lambda$ is defined as the speed of sound, *c=342 m/s*, over frequency, *f*.

$$\lambda = \frac{c}{f} \tag{2.2}$$

If the length of the receiving microphone array is equal to a wavelength then the near-field assumption is valid for radial distances less than 2 wavelengths and the far-field assumption can be made for distances greater than 2 wavelengths. Figure 2.2 shows the

relationship of frequency to *log10* of the wavelength to give perspective of the near/far field requirements. A signal with a frequency of 340 Hz will have a wavelength of roughly 1 meter, which requires a distance of 2 meters to assume a far-field. A 1kHz signal will have wavelength of 0.34 meters, which requires a distance of 0.64 meters to assume a far-field.

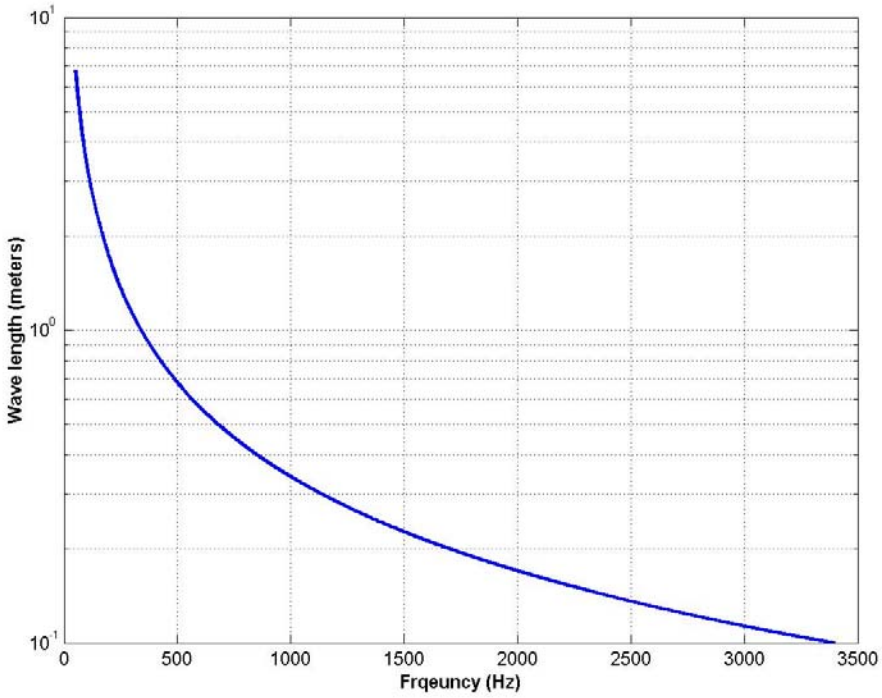


**Figure 2.2: Log10 of wavelength vs. frequency**

Inserting equation (2.2) into (2.1) yields the near field requirement of

$$r < \frac{2L^2 f}{c} \tag{2.3}$$

or

$$f > \frac{rc}{2L^2} \tag{2.4}$$

Sound Pressure Level (SPL), shown in equation (2.5), is measured on a logarithmic scale of decibels because the human ear is sensitive to a wide range of pressure variations. The threshold of audibility at 1 kHz is about $p_0 = 2x10^{-5} N/meter^2$ (0 dB SPL) and the upper limit, called the threshold of pain, is approximately 20 *N/meter*$^2$ (120 dB SPL), which represents a range of $10^6$.

$$SPL = 20\log(p/p_0)dB \tag{2.5}$$

SPL varies proportional to the inverse square of its distance from the microphone, $1/r^2$, for the far field case, assuming an omni-directional source in a free field (no reflections). This decrease in signal strength of the source is 6 dB SPL in the far field decrease for each doubling of distance as shown in Table 2.1. SPL measurements vary widely in the near field based on the microphone position and signal wavelength. Thus the inverse square law does not hold when a near field condition exists.

| Distance to talker in (cm) | Equivalent SPL in a **far** field (dB) |
|---|---|
| 4 | 40 |
| 8 | 34 |
| 16 | 28 |
| 32 | 22 |

**Table 2.1: SPL vs. distance from omni-directional source in a free field**

Noise sources, on the other hand, become stronger relative to the desired talker when using hands-free microphones instead of using hand-held or headset microphones. Noise suppression algorithms must work much harder at attenuating the noise as the distance of the talker from the microphone increases because of the weaker desired signal and stronger influence of the noise sources. Hands-free microphones impose adverse conditions for quality speech reception and are the main motivation for the advanced noise suppression algorithm proposed in this thesis. Strong noise sources also present the challenge of masking the desired speech.

## 2.2 Auditory masking

Auditory masking occurs when the listener cannot hear a particular source because it is hidden by a louder interfering sound source. Conversely, the desired source can be loud enough to hide (mask) the interference from the listener. The SSPN algorithm presented in this thesis takes advantage of the situations where the desired source masks the noise, so it does not need to suppress the noise thus reducing the risk of distorting the desired speech.

The sounds perceived by humans are affected by the direction, timing, amplitude, and frequency of the signals arriving at the ear. Loudness of sound is usually expressed as Sound Pressure Level (SPL) in decibels (dB) where a whisper is about 30 dB, normal conversation is about 60 dB, and a subway train is about 90 dB. Loudness varies depending on frequency as demonstrated by Figure 2.3 where the dashed line represents

the average threshold of audibility, which is the level a sound can be heard with no other interfering noise present. The normal frequency range of human hearing is from 20 to 20,000 Hz.

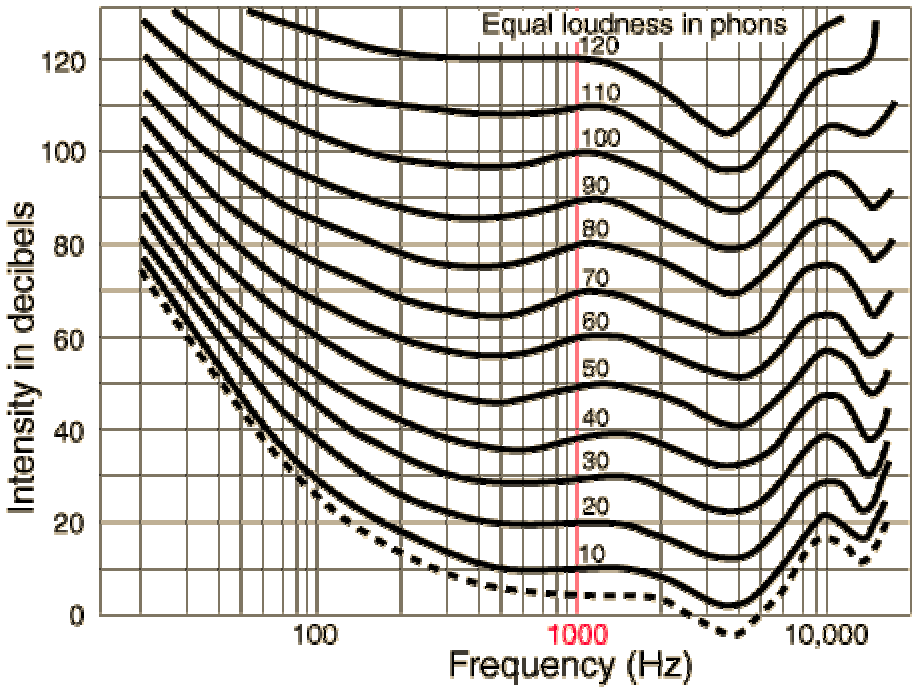


**Figure 2.3: Equal loudness curves (dashed line represents threshold of hearing)** [107]

To determine this threshold of audibility, an experiment must be performed. A typical masking experiment might proceed as follows. A short, about 400 ms, pulse of a 1,000 Hz sine wave acts as the target, or the sound the listener is trying to hear. Another sound, the masker, is a band of noise centered on the frequency of the target (the masker could also be another pure tone). The intensity of the masker is increased until the target cannot be heard and this point is then recorded as the masked threshold.[20] Another way of proceeding is to slowly widen the bandwidth of the noise without adding energy to the original band. The increased bandwidth gradually causes more masking until a certain

point is reached, at which no more masking occurs, and this bandwidth is called the critical band.[21] The masker can keep extending until it is full-bandwidth white noise and it will have no more effect than at the critical band. Critical bands grow larger as they ascend the frequency spectrum and there are many more bands in the lower frequency range, because they are smaller.[22] About 30 critical bands cover the 10 octaves of human frequency perception, this yields 30 disjoint bands.[23] [24]

The two most popular perceptually based nonlinear frequency scales are the Mel scale and the Bark scale. The Mel scale is generally used with Cepstral coefficients (Mel-Cepstrum) on a logarithmic scale and commonly found in speech recognition applications. It is based on experiments done by Stevens and Volkman in the 1940s, which is a mapping from linear frequency to the nonlinear Mels frequency scale. The Bark scale is based on critical band analysis, which maps linear frequency to the critical bands of the human auditory system as show in Table 2.2. Details of the frequency bands used in the human auditory system are described in appendix A.

| Bark of lower frequency | Lower / Upper frequency in Hz | Bark of Center | Center frequency In Hz | Bandwidth in Hz |
|---|---|---|---|---|
| 0 | 0 | 0.5 | 50 | 100 |
| 1 | 100 | 1.5 | 150 | 100 |
| 2 | 200 | 2.5 | 250 | 100 |
| 3 | 300 | 3.5 | 350 | 100 |
| 4 | 400 | 4.5 | 450 | 110 |
| 5 | 510 | 5.5 | 570 | 120 |
| 6 | 630 | 6.5 | 700 | 140 |
| 7 | 770 | 7.5 | 840 | 150 |
| 8 | 920 | 8.5 | 1000 | 160 |
| 9 | 1080 | 9.5 | 1170 | 190 |
| 10 | 1270 | 10.5 | 1370 | 210 |
| 11 | 1480 | 11.5 | 1600 | 240 |
| 12 | 1720 | 12.5 | 1850 | 280 |
| 13 | 2000 | 13.5 | 2150 | 320 |
| 14 | 2320 | 14.5 | 2500 | 380 |
| 15 | 2700 | 15.5 | 2900 | 450 |
| 16 | 3150 | 16.5 | 3400 | 550 |
| 17 | 3700 | 17.5 | 4000 | 700 |
| 18 | 4400 | 18.5 | 4800 | 900 |
| 19 | 5300 | 19.5 | 5800 | 1100 |
| 20 | 6400 | 20.5 | 7000 | 1300 |
| 21 | 7700 | 21.5 | 8500 | 1800 |
| 22 | 9500 | 22.5 | 10500 | 2500 |
| 23 | 12000 | 23.5 | 13500 | 3500 |
| 24 | 15500 | | | |

**Table 2.2:Critical Bands of the Human Auditory System [24]**

Modeling the speech signal is one approach taken to extract the speech and attenuate the noise to overcome noise masking.

## 2.3 Modeling speech

Modeling the speech production system enables the speech enhancement algorithm to take advantage of certain source signal characteristics. It is important to know when the talker is speaking (voice activity detection) and knowledge of the source signal can simplify this task. Knowledge of speech production is also needed when developing algorithms that identify a particular talker from interfering talkers, called talker isolation.

This thesis uses algorithms that depend on modeling the human speech production system such as Voice Activity Detection (VAD) and pitch detection.

Speech is produced by a cooperation of lungs, glottis, vocal cords, mouth, and nose cavity and Figure 2.4 shows a cross section of the human speech organ. For the production of voiced sounds, the lungs press air through the epiglottis, the vocal cords vibrate (open and close), which interrupt the air stream and produce a quasi-periodic pressure wave. The rate at which the vocal cords vibrate determines the pitch of your voice where women and young children tend to have high pitch (fast vibration) while adult males tend to have low pitch (slow vibration). The shape of the vocal tract changes relatively slowly (on the scale of 10 msec to 100 msec) and vowel sounds such as a/e/i/o/u represent voiced speech.

The pitch impulses stimulate the air in the mouth and for certain sounds (nasals) also the nasal cavity and when these cavities resonate, they radiate a sound wave, which is the speech signal. Both cavities act as resonators with characteristic resonance frequencies, called *formant frequencies* and since the mouth cavity can be greatly changed, it is able to pronounce very many different sounds. In the case of unvoiced sounds, the excitation of the vocal tract is more noise-like and for certain *fricatives and plosive (or unvoiced)* sound, the vocal cords do not vibrate but remain constantly opened. Examples of unvoiced sounds are /f/, /th/, /h/, /p/, /t/, or /k/.

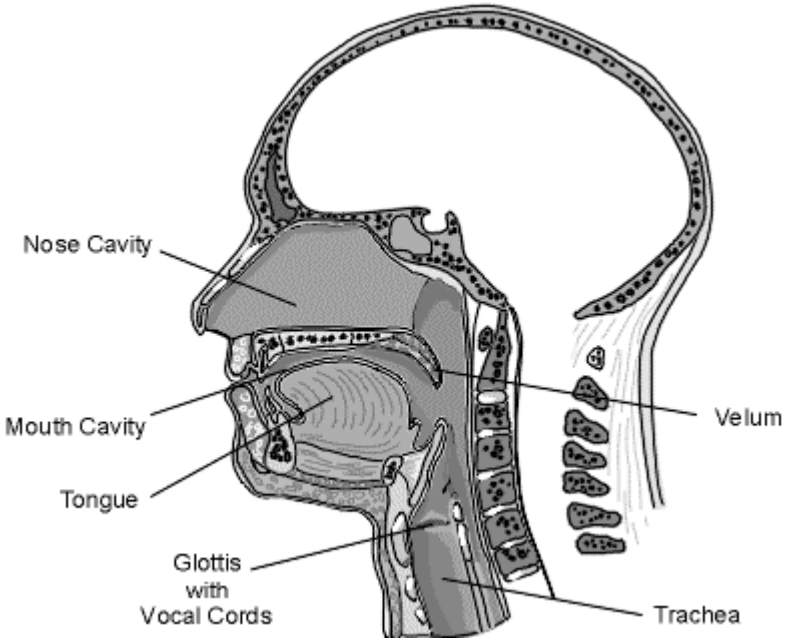


**Figure 2.4: Human Speech Production**

The speech production system can be represented with an all-pole filter and the Linear Prediction algorithm identifies the parameters associated with the all-pole system. Voiced sounds are generated with periodic pulses and unvoiced sounds are generated by white noise. Figure 2.5 shows the high-level system.

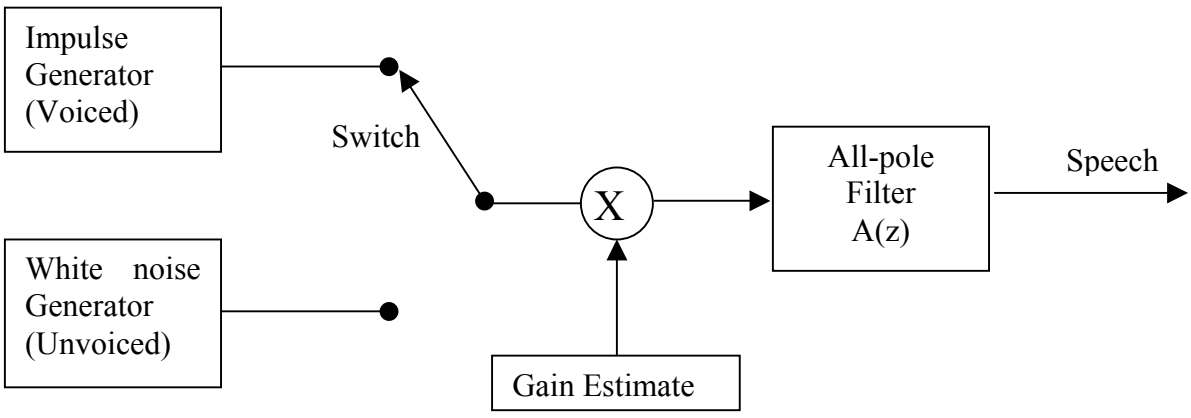


**Figure 2.5: Human Speech Production Model**

# Chapter 3

# Using spatial, spectral, and perceptual information

The application and design of the SSPN algorithm was done in the context of a hands-free phone application in an automobile where the goal of the signal processing was to attenuate the noise sources while preserving the clarity and intelligibility of the speech source. Section 3.1 presents the general problem of speech enhancement and noise suppression, Section 3.2 describes the signals and environment used to exercise the proposed algorithm and Section 3.3 describes the algorithm design, control flow, and data flow.

## 3.1 Multiple sources and sensors

Many sources of noise interference are possible in a mobile environment. These interferences are combined with a single desired speech source assuming only one person is involved in the conversation. A high-level approach to obtain a noise-free speech signal is to separate the desired speech from the additive noise with the knowledge that the speech and noise are independent. Multiple sensors can help achieve signal

separation by introducing a spatial dimension to the parameters used to identify the speech versus the noise sources. The signal-processing framework of this multiple source, multiple sensor, and associated acoustic coupling is described in Figure 3.1, where the multiple input multiple out system (MIMO) has *I+1* inputs and *J* outputs.

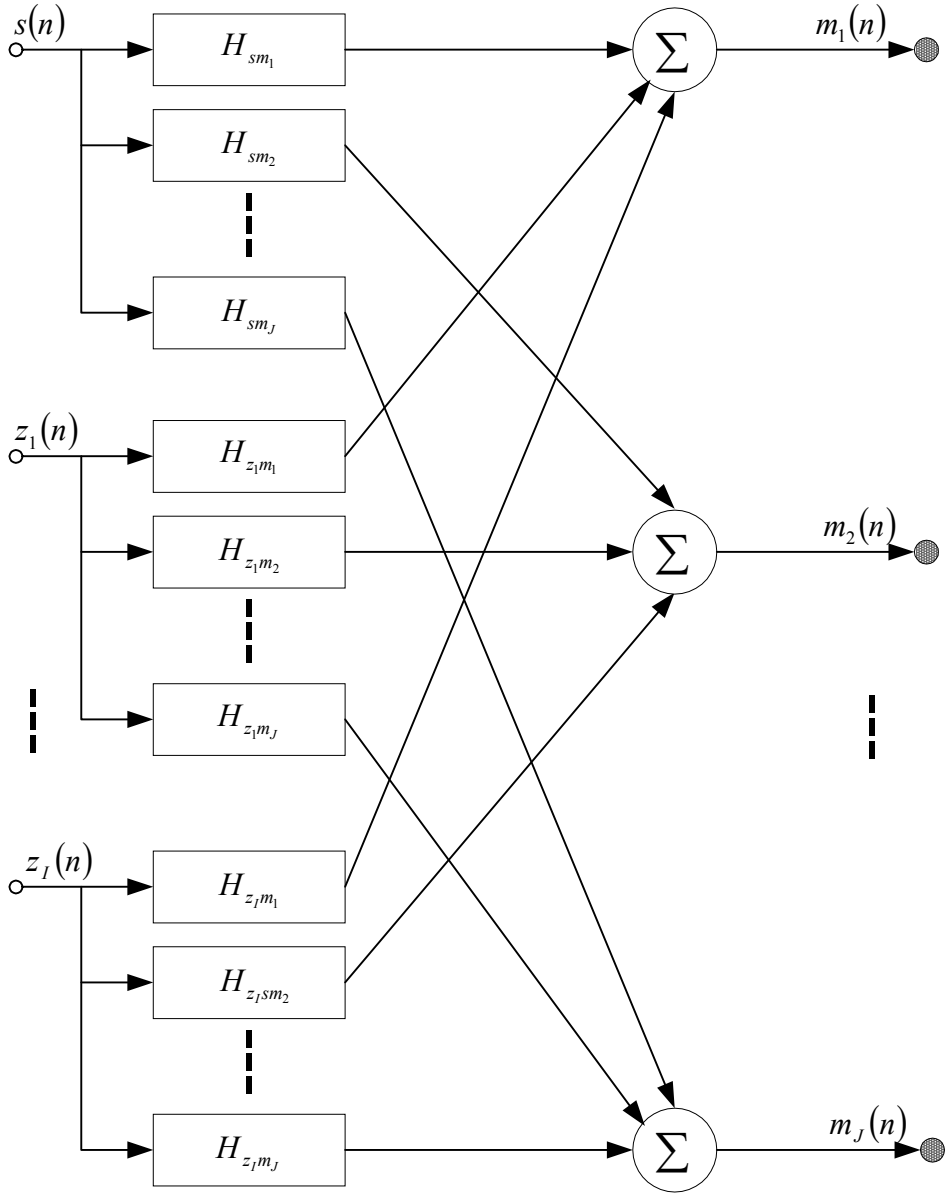


**Figure 3.1: Multiple source and sensor framework**

In Figure 3.1, the desired speech is denoted by $s(n)$, the $i^{th}$ noise source is represented by $z_i(n)$, and the signal at the $j^{th}$ microphone is represented by $m_j(n)$ and equation (3.1), where $n$ is the sample index. $H_{sm_j}$ and $H_{z_i m_j}$ are the source to sensor acoustic coupling functions for the speech and the $i^{th}$ noise source to each $j^{th}$ microphone respectively.

$$m_j(n) = s(n) * H_{sm_j} + \sum_{i=1}^{I} z_i(n) * H_{z_i m_j} \quad \textbf{(3.1)}$$

The goal of the speech enhancement algorithm is to suppress the sum of the noise contributions in equation (3.1) observed by the $J$ microphones while minimizing any distortion to $s(n)$.

## 3.2 Automobile environment and system setup

An automobile's acoustic environment was used to exercise the algorithm proposed in this thesis. The system consisted of four microphones, spaced 5cm apart, and placed directly in front of and slightly above the driver attached to the visor in the car. Some of the specific signal sources in the automobile environment are the driver's speech, engine noise, road noise, wind noise, passing cars, fan noise, and interfering talkers, which are all shown in Figure 3.2. The signal received at the four microphones in Figure 3.2 can be described by equation (3.1) where $j = 1..4$ and $I = 8$. The goal of the proposed algorithm is to remove the noise from the signal received by the four microphones while minimizing the distortion to the desired speech where the noise is considered to be independent and uncorrelated with the desired speech signal.

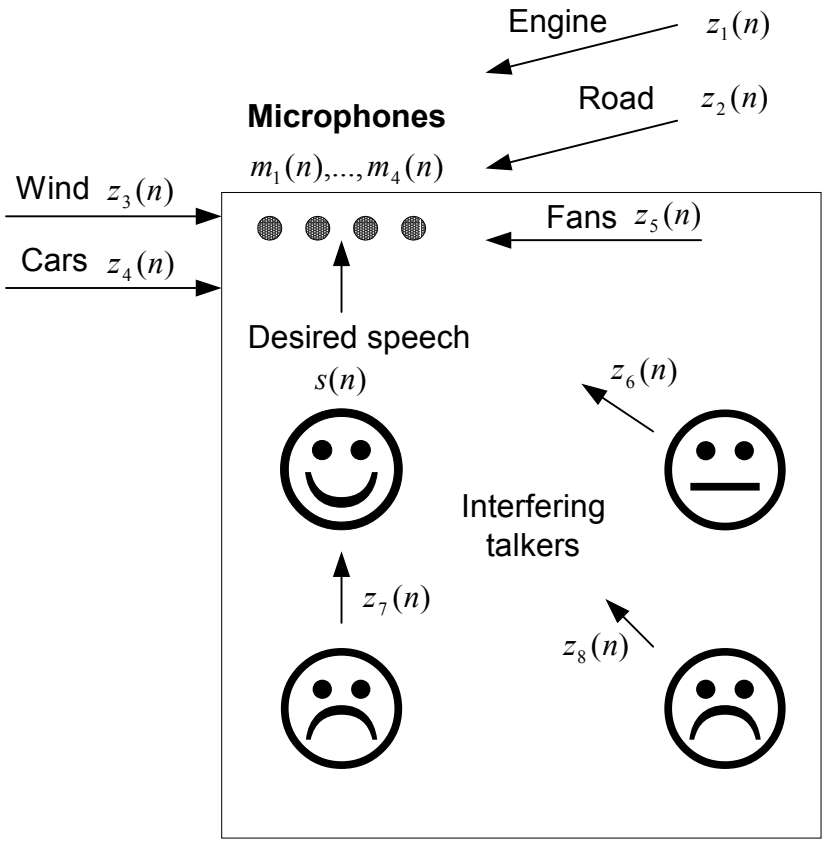


**Figure 3.2: Source signals**

The speech, $s(n)$, is assumed to be short-term stationary, which allows for the use of short-term spectral analysis to enhance the signal. The four microphones in the car provide the advantage of using signal direction as a means of discriminating between the desired signal and interference. Larger numbers of microphones were not used because the increased cost to the overall system is less attractive for selling a noise suppression hands-free solution to the automotive market. The SNR gain over a single microphone scales linearly with the number of microphones, but there is a point at which adding more microphones will have diminishing returns. Typical configurations that have been implemented in other hands-free research in the automobile are 1, 2, 4, 6, and 8 microphones.

The location of the microphones in the system used in this thesis was chosen based on prior research and the fundamental principal that microphones should be as close to the desired signal as possible and far away from the interference.[25] The placement of microphones relative to the desired signal and interference is one of the most critical aspects of the speech acquisition and enhancement system because the source signal strength reduces approximately as the inverse square of the distance from the microphone. Positioning the microphones on the ceiling above the driver was chosen because it is closer to the talker's head and further away from the engine and road noise and has been shown to yield the best results for both SNR and speech recognition rates.[26] Recordings were made of the signal strength at pairs of microphone locations and the results are reported in appendix B, which confirms that the best choice for microphone location is on the ceiling above the driver's head.

The acoustical environment described above requires the use of multiple signal and auditory properties to effectively enhance the speech, which is the strength of this work. The enhanced hands-free speech can then be used for improved speech recognition or a higher quality phone conversation. The following section (3.3) describes the proposed algorithm, detailed signal and auditory properties used by the algorithm, and why specific design choices were made.

## 3.3 Spatial, spectral, and perceptual nonlinear processing

The spatial, spectral, perceptual nonlinear (SSPN) processing algorithm uses a unique combination of several techniques with iteration and feedback to maximize speech quality. SSPN noise suppression is potentially better than what the individual techniques can achieve alone because the specific order of the signal processing and feedback used. The SSPN algorithm depends on several other functional components, which Chapter 4 describes in detail. The specific algorithms used for this thesis were chosen based on their generally good performance, widespread use, and simplicity of implementation. This chapter puts the component algorithms in context of the larger design and explains the reasoning behind their being chosen with a section dedicated to each major component as indicated in the list below.

3.3.3 GSC beamforming

3.3.4 VAD and noise estimation

3.3.5 Spectral subtraction

3.3.6 Perceptual nonlinear frequency weighting

3.3.7 Talker isolation and pitch tracking

The resulting system could be further improved if optimal techniques for each component were investigated as part of the main algorithm, but such optimizations will not be explored here in order to limit the scope of this already broad-based investigation.

The high level design of the algorithm was based on getting the most benefit from each component and increased accuracy of the VAD because the VAD plays a central role in

minimizing the distortion of the desired speech. Section 3.3.1 describes the high level design of the SSPN algorithm, while the rest of sections in Chapter 3 discuss the role for each component and design decisions for the algorithm.

### 3.3.1 SSPN algorithm

The SSPN algorithm takes advantage of *spatial*, *temporal*, *spectral*, and *auditory perceptual* properties in order to suppress the interfering noise while minimizing distortion to the desired speech. *Spatial* diversity is used to attenuate signals not originating from the desired talker. *Temporal* properties of speech production and signal energy is used to determine when there is no speech present, so the noise estimate can be updated thereby enabling *spectral* estimates of the noise to be subtracted out from the spectrum of received signal. Analysis of human *auditory perceptual* masking properties within critical frequency bands provides a perceptual model for attenuating the noise. A perceptual nonlinear weighting of the spectral gain function reduces the musical noise artifacts typically introduced by spectral subtraction. *Temporal* properties can also be used for pitch and amplitude tracking to isolate the desired speech and increase the accuracy of the VAD. The high-level steps of the SSPN algorithm are listed below and further described in the following paragraphs.

1. Beamforming
2. VAD
3. Spectral subtraction
4. Auditory perceptual mask threshold calculation
5. Nonlinearly weight spectral subtraction
6. Talker isolation
7. Iterate to improve the VAD accuracy

**Beamforming** reduces the noise in the received signal independent of voice activity detection because it is relying on the spatial diversity of the sources, which makes it a good candidate to be first in the signal processing chain because the resulting output of the beamformer can increase the accuracy of the initial voice activity detection when compared to placing the VAD first. The VAD relies on changes in signal energy to detect frames with voice, so reducing the noise energy with beamforming will allow the energy of the speech to be more prominent and easily detected. Another reason for choosing to place the beamformer first in the design is that beamforming reduces the amount of attenuation required in spectral subtraction, thus reducing possible distortion.

The results in sections 5.5 and 5.8 show that beamforming significantly reduces the noise levels with less distortion when compared to simple spectral subtraction. Thus, by placing beamforming before spectral subtraction the lower energy high frequency consonants, such as /f/t/s/h/p, are less likely to be over attenuated by the spectral subtraction part of the algorithm, this is beneficial because it has been noted that consonants are important to speech intelligibility.[27] Placing the beamformer first also reduces the computational requirements because all subsequent processing is performed on a single channel of data. The beamformer would not benefit by preceding it with other components of the algorithm because it is primarily working in the spatial domain, which is not effected by the other components. The Generalized Side-lobe Canceller (GSC) was chosen as the specific form of beamforming for the SSPN algorithm because of its superior noise suppression when compared to simple delay and sum beamforming and the details of exactly how it was used are given in section 3.3.3.

**Voice Activity Detection (VAD)**, described in section 3.3.4 and 4.2, is performed on the output signal from the beamformer. Noise estimation is done when the VAD determines that no voice is present in a frame, which is then used for half wave rectified spectral subtraction on the output from the beamformer to reduce the residual noise left in the processed signal. The beamformer used in this thesis does not perform well in the lower frequency ranges, as discussed in section 4.1, because of the fixed spacing of the microphone elements, so spectral subtraction is employed to make up for this shortcoming.

A clean speech estimate is required for the calculation of the **perceptual mask threshold**, so it must follow an initial half wave rectified **spectral subtraction** processing step. A nonlinear weighting function based on the calculated masked threshold estimate is applied to the spectral subtraction process to minimize the introduction of artifacts and distortion into the signal. This **weighted spectral subtraction** is applied to the output of the beamformer in order to obtain an improvement in noise reduction and perceptual quality over the initial half wave rectified spectral subtraction.

**Pitch based talker isolation** can be performed on the noise-reduced output of the nonlinear spectral subtraction. Better pitch estimates are possible after the first two stages of noise removal have been performed. If the pitch estimate and talker separation were done first on the noisy signal, the results could be worse than if noise removal were

not done at all. Experiments were done placing pitch detection before spectral subtraction with no benefit resulting in the algorithms noise suppression or speech quality.

**Iterating the SSPN algorithm** once improved the VAD accuracy because the VAD could makes its decisions based on a noise-reduced signal, which is supported by results in section 5.4. Iterating the algorithm more than once did not prove beneficial because too many frames were classified as speech causing the noise estimate to suffer as is further explained in section 5.4. Feeding back the output of the first pass noise reduction to the VAD can change the current VAD decision and enable a second more accurate noise reduction processing of the current frame. The GSC beamformer can then reprocess the current frame and adapt its noise cancellation filters if the VAD marks the frame as not having any speech. The new VAD decision will also give the opportunity to update the noise estimate to more accurately reflect the actual noise occurring in the current time frames, which will in turn improve the spectral subtraction results. If the first VAD decision detects a voiced frame and the second VAD decision also detects voice for the same frame, then the current frame does not need a second pass of processing because the GSC filter coefficients will not change and the noise estimate will not be updated. Figure 3.3 shows the control flow for the SSPN algorithm after the initialization is done.

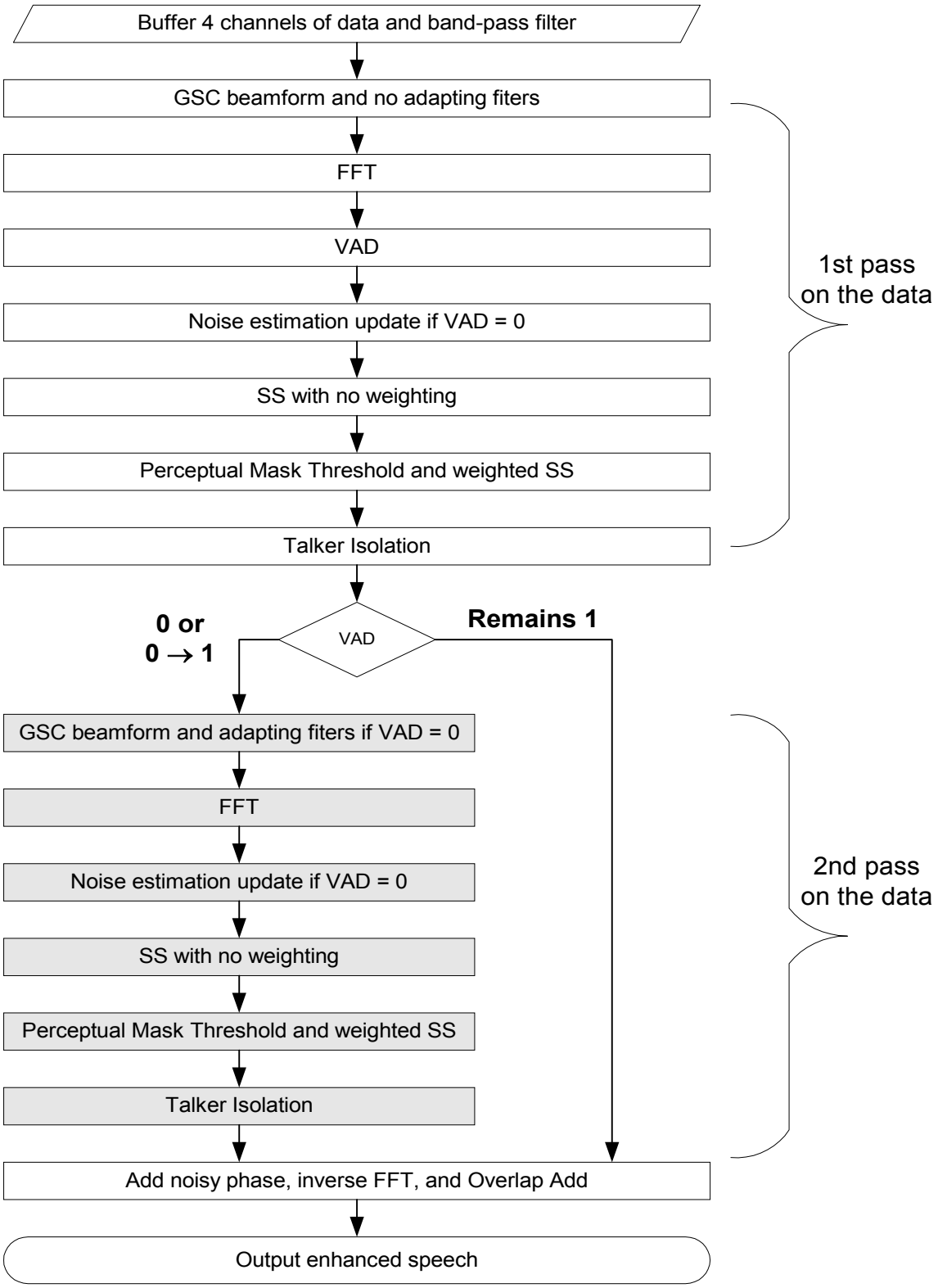


**Figure 3.3: SSPN algorithm flow chart**

The first 10 frames of data are assumed to be just noise in order for the GSC filters to adapt and the noise estimate to initialize, the initialization time is calculated in equation (3.2).

$$t_{init} = \frac{10\,frames * 128 samples / frame}{8000 samples / \sec ond} = 0.16 \sec onds \qquad \textbf{(3.2)}$$

The first and second half of the flow chart of Figure 3.3 look similar because the current frame is processed a second time based on the updated parameters. The number of iterations for VAD improvement was limited to 2 because experiments, documented in section 5.4, showed that further iteration degrades the noise suppression performance due to lack of frames classified as noise only. The first GCS pass does not adapt its filters because it has no way of knowing if the current frame contains speech. The second VAD decision tells the GSC filters if they can adapt and if the noise estimate for spectral subtraction should change. This iteration of the algorithm will cause more of the signal to be classified as speech thus reducing the risk of falsely classifying speech as noise, which would result in attenuation of the desired speech. The VAD's sensitivity to its' fixed energy multipliers, used to calculate the threshold for a speech decision, is reduced and allows the algorithm to perform more consistently over a wider range of input SNR.

Figure 3.4 shows the data flow for the first iteration of the algorithm that receives the new frame of data from all four microphones and does an initial noise removal. All four channels are band-pass filtered to limit the frequencies to that of normal human speech ranging from 50 Hz to 4 kHz. The talker's location is known because we are focusing on the driver of the car, who is broadside to the microphone array. The estimate of the

speech spectrum is fed back to the VAD, which is where the second iteration begins. Control data is indicated by dashed lines in Figure 3.4 and are the VAD decisions. The frame number is represented by the parameter $k$ and the parameter $w$ denotes the frequency index.

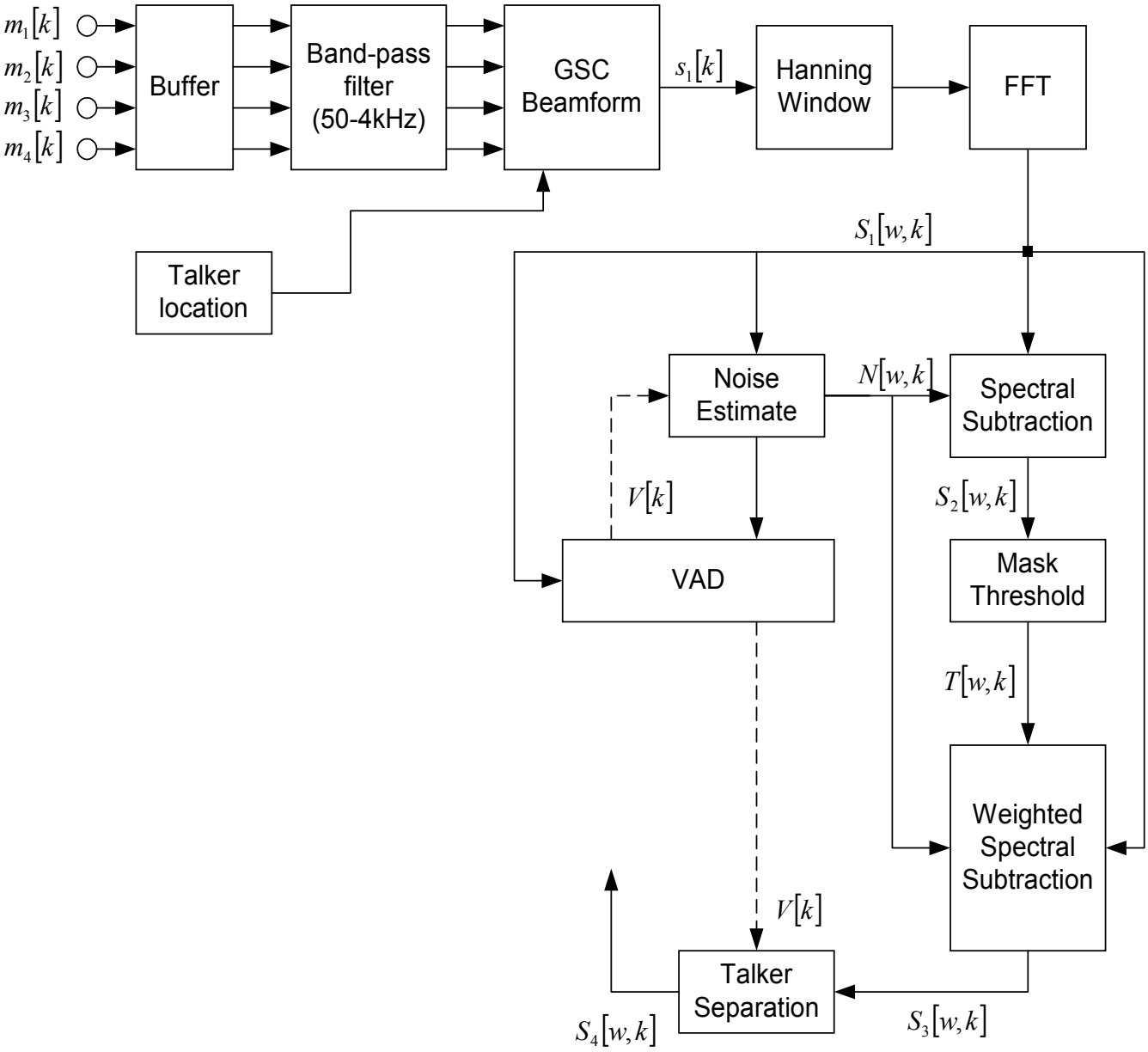


**Figure 3.4: SSPN first iteration on current frame**

Figure 3.5 shows the data flow for the second iteration of the algorithm. The iteration begins by making a new VAD decision based on the initial speech estimate. The stored buffers are used to reprocess the current frame of data in the GSC where the filters will adapt if the VAD does not detect speech. The output of the GSC updates the noise estimate for spectral subtraction if the VAD has not detected speech. Spectral subtraction and talker isolation is performed as in the first iteration. The speech estimate has the phase information added to it from the phase at the output of the beamformer. An inverse FFT and overlap-add processing then converts the signal to time domain output.

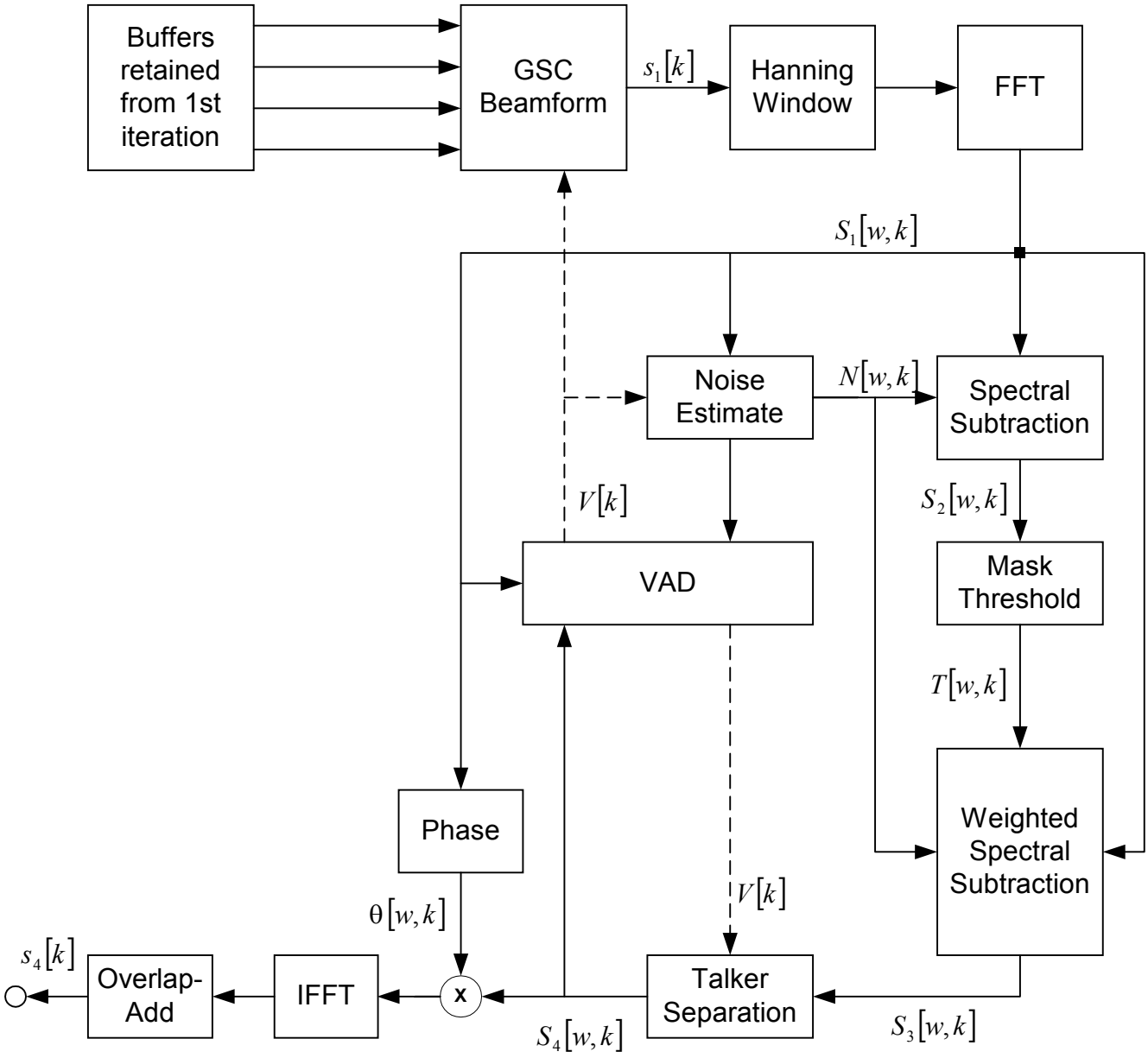


**Figure 3.5: SSPN second iteration on current frame**

Figure 3.6 shows the data flow with all the processing blocks represented, where the central role of the VAD is very apparent in this view of the algorithm.

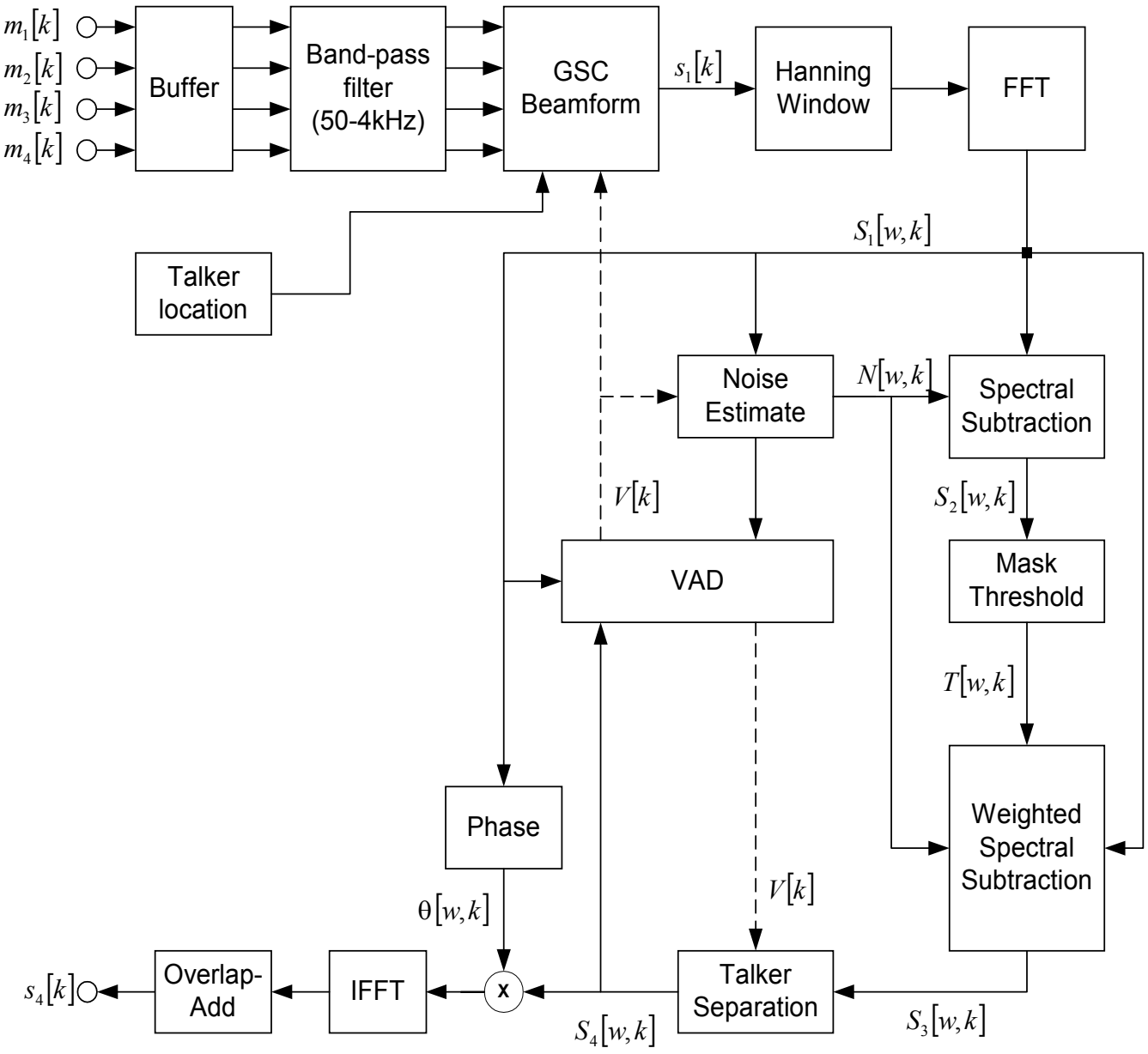


**Figure 3.6: SSPN algorithm with all signal paths**

### 3.3.2 Initial signal processing

Data is received from all four microphones, which are separated by a relatively small known distance *d*. All four channels are divided into 50% overlapped buffers, which is similar to the buffering used by Boll.[4] All input signals are then band-pass filtered to limit the frequency range to 50 to 4 kHz for subsequent processing. Frequencies below 300Hz are dominated primarily by noise from the engine and road [28] and a lower cutoff frequency of 350Hz is an example of preprocessing that has been used in car environment.[29] The lower cutoff frequency could be adapted based on operating conditions, but will remained fixed at 50 Hz for the work done in this thesis.

The GSC beamformer is then applied to attenuate the noise, which results in an enhanced single output channel of data. A Hanning window is applied to the output of the GSC because it has lower side-lobes in the frequency domain and thus less spectral leakage than a rectangular window. The 50% overlap facilitates the use of the overlap add (OLA) method to synthesize the signal, after spectral processing is complete, in a way that maintains the temporal characteristics of the input.[30] The window size is chosen to be approximately twice as large as the expected pitch frequency for accurate frequency resolution. The window buffer length, *L*, is 256 points (32ms) and input frame lengths are 128 points (16ms) when the sampling rate is 8khz. Short windows must be used in order to take advantage of the short-term stationary properties of the speech and noise because beyond approximately 48ms (3 x 16ms frames) the processing degrades the signal quality.

The data window is then transformed with an FFT for the full length *L* of the window. Zero padding can be used to increase spectral resolution and reduce the effects of temporal aliasing caused by circular convolution and could help in the talker isolation processing. However, the additional frequency resolution may not help with removing the noise as shown by Boll [4] and lower resolution methods, such as Bartlett's spectrum estimation, can actually help because of the reduced variance in the frame-to-frame estimates.[31] There is clearly a tradeoff between spectral resolution and variance. Magnitude averaging of the noise estimate spectrum over 2 or 3 frames has been effectively used in SS to reduce variance.

### 3.3.3 GSC beamforming

Generalized Side-lobe Canceling (GSC), described in section 4.1.3, is performed to strengthen the signal coming from the desired talker and attenuate noise from other directions. The beamformer has the advantage of reducing noise while introducing very little distortion to the desired speed and will improve down stream VAD decision and talker isolation tasks.

The first 10 frames of data are assumed to be silence in order to initialize the filters in the GSC. After the initialization phase, the GSC will process each frame twice because the GSC does depend on VAD information in order not to adapt its filters when speech is present, so it will avoid attenuating the speech. The first pass will not adapt the filters, but will simply use the filter coefficients resulting from past noise only frames. The filter will adapt during the second pass on a frame of data if the VAD indicates that no speech

is present. Once the frame has been processed a second time, the next frame of data will enter the GSC.

### 3.3.4 VAD and noise estimation

#### 3.3.4.1 Noise estimate during silent frames

The noise spectrum is estimated during silent frames where the VAD decides between noise only silence and speech; this is especially effective when the noise is slowly varying. Silent frame are considered as containing no speech, but only background noise. There are many pauses in natural speech, which allows the noise estimate to be updated quite frequently. An energy detection based VAD is used similar to the one described in section 4.2.2. The SSPN algorithm in this thesis will estimate the noise during silent frames because it is relatively effective and simple to implement.

#### 3.3.4.2 Continuous noise estimatation experiment

More sophisticated methods of noise estimation have been extensively studied and would in fact improve the overall system performance, but are beyond the scope of this thesis. Waiting for silent frames is not effective when the noise is varying rapidly, so estimating the noise during speech activity is required. Some examples of continuous noise estimation are adaptive [32] and MMSE (minimizing a conditional mean square error)[33], and two channel techniques that work to obtain a separate noise channel.[34]

An unsuccessful attempt was made to use beamforming to create a separate noise and speech channel that was to be used for continuous noise estimation. Spectral subtraction was to use the noise channel as a continuous noise estimate weighted by the estimate of how much noise leaked into the speech channel. The noise estimate is continuously modified and weighted differently depending on whether the frame is classified by a VAD as containing speech or not. The speech estimate from the signal separation during frames with no speech acts as an indication of how much noise leaks into the speech channel. This information is used to scale the noise estimate during speech frames prior to spectral subtraction. An approximate procedure is outlined below.

A silent frame contains only background noise. The variable $\beta$ is the percentage of the noise that leaks into the speech channel $L_1(w)$.

$$L_1(w) = \beta \cdot N_1(w) \quad \textbf{(3.3)}$$

$$I_1(w) = (1-\beta) \cdot N_1(w) \quad \textbf{(3.4)}$$

The speech frame contains speech + noise. The variable $(1-\alpha)$ is the percentage of the speech that leaks into the noise channel $I_2(w)$.

$$L_2(w) = \alpha \cdot S(w) + \beta \cdot N_2(w) \quad \textbf{(3.5)}$$

$$I_2(w) = (1-\alpha) \cdot S(w) + (1-\beta) \cdot N_2(w) \quad \textbf{(3.6)}$$

Equations (3.7) and (3.8) describe an algorithm for removing the residual stationary noise. In reality the noise subtracted would be averaged over a few frames, but this shows the general idea.

$$L(w) = L_2(w) - L_1(w) = \alpha \cdot S(w) + \beta \cdot N_2(w) - \beta \cdot N_1(w) \quad \textbf{(3.7)}$$

$$L(w) \approx \alpha \cdot S(w) \quad \textbf{(3.8)}$$

The above approach did not improve the results and actually made things worse in some cases. The reasons for the poor performance were the poor correlation of the noise in the two channels, leakage of the desired speech into the noise channel, and an inaccurate estimate of the noise leaking into the speech channel. In the end, simple noise estimation during silent frames using a single channel gave dramatically better results than the attempt at continuous noise estimation.

### 3.3.5 Spectral subtraction

An introduction to spectral subtraction concepts is presented in section 4.3. The first spectral subtraction operation in the SSPN algorithm of Figure 3.6 used fixed weighting of the noise estimate and half-wave rectification. This is equivalent to setting $a(w) = 1$ and $b(w) = 0$ in Figure 3.7. This step is necessary because a clean speech estimate is needed to calculate the masking threshold. Half-wave rectified spectral subtraction produces very noticeable musical noise artifacts when the input SNR approaches 0 dB, but its output, $S_2(w,k)$, is only used to calculate the masked threshold and not as part of the final speech estimate. The input to the weighted spectral subtraction is the original speech estimate, $S_1(w,k)$, before the initial spectral subtraction is applied.

Figure 3.7 shows the details of the weighted spectral subtraction block used in the SSPN algorithm of Figure 3.6 where $S_1(w,k)$ is the initial speech channel input to the algorithm

and $N(w,k)$ is the noise estimate to be subtracted. The auditory perceptual masking threshold, $T(w,k)$, adjusts the over-subtraction factor $a(w)$ and noise floor factor $b(w)$.

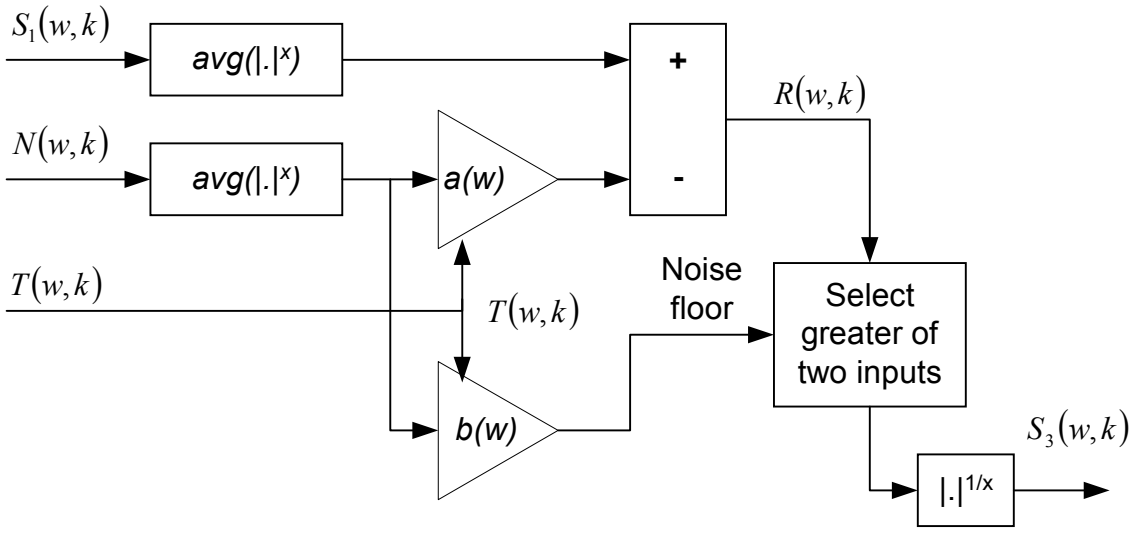


**Figure 3.7: Generalized Spectral Subtraction**

The magnitudes are taken and raised to a power *x* for the signal $S_1(w,k)$ and noise $N(w,k)$. The received signal $S_1(w,k)$and noise $N(w,k)$ are then averaged over three frames to reduce the frame-to-frame variability that causes musical noise. The averaging is skipped when the received VAD detects speech. The high-energy speech will tend to mask the musical noise in these frames. This approach avoids smearing of the spectral peaks in the speech.[35]

The subtraction multiplier $a(w)$ and noise floor factor $b(w)$ are determined by interpolating between the minimum and maximum values using the threshold &

weighting function $T(w,k)$. The linear interpolation functions are represented by $F_a$ and $F_b$ in equations (3.9) and (3.10) respectively.

$$a(w,T(w)) = F_a(a_{\min}, a_{\max}, T(w)), \qquad a_{\min} = 1,\ a_{\max} = 6 \tag{3.9}$$

$$b(w,T(w)) = F_b(b_{\min}, b_{\max}, T(w)), \qquad b_{\min} = 0,\ b_{\max} = 0.02 \tag{3.10}$$

Both low noise and little distortion are desired, but the more the noise is suppressed the bigger the speech distortion gets. The range of values for the multipliers will balance a tradeoff between residual noise allowed through and speech degradation. The values in equations (3.9) and (3.10) for the ranges of a(w) and b(w) were used initially based on results reported by Virag.[91] However, setting $a_{max}$ = 3 was found to produce better results for the data sets processed in this thesis. The exponent was set to $x = 2$ as successfully done in related research. Spectral subtraction takes place in equation (3.11) and the noise floor is calculated in equation (3.12).

$$R(w) = (|S_1(w)|^x - a(w,T(w))|N(w)|^x) \tag{3.11}$$

$$\text{Noise floor} = b(w,T(w))|N(w)|^x \tag{3.12}$$

The noise estimate is multiplied by a function in the experimentally determined range of 0 to 0.002 resulting in a noise floor that keeps the noise level low, but also to leaves enough noise in the signal to prevent large discontinuous jumps in the frame-to-frame audible noise. If $R(w)$ is greater than the noise floor, then it is chosen as the output. The noise floor is selected as the output if it is greater than $R(w)$. This ensures that there are no negative spectral components and that the noise remains fairly constant when speech

is not present. The last step is to take the xth root of the signal to yield the output $S_3(w,k)$.

### 3.3.6 Perceptual nonlinear frequency weighting

Perceptual nonlinear frequency weighting requires the perceptual masking threshold to be calculated as described in section 4.4. A smoothing function is applied to the resulting masked threshold function to remove any discontinuities that could introduce artifacts during the spectral subtraction. The estimate of the masking thresholds need to be closely matched to the desired speech, therefore, an initial estimate of the clean speech is necessary. The spectral subtraction with half-wave rectification is a sufficient estimate of the clean speech.

The implementation requires some numerical checking that is not mentioned in the general description of the mask threshold algorithm. It is possible for the spread linear threshold to be zero, so the algorithm must guard against calculating a *log(0)* that would produce negative infinite values. Because energy in a band can be zero and the renormalization process divides by the band energy, the algorithm must also guard against a divide by zero.

### 3.3.7 Talker isolation and pitch tracking

A talker isolation algorithm was not implemented for this thesis because it would have expanded the scope of this work beyond the time available. Instead, some experiments

were performed with pitch detection before and after the SSPN algorithm to better understand how to incorporate such a feature into the design. Continued research on the effects of talker isolation on the SSPN results should be carried out in future work, the pitch tracking and talker separation algorithm described in section 4.5 is an excellent place to start. Beamforming and spectral subtraction are inadequate for removing interfering speech as is shown in the results of sections 5.5 and 5.6 respectively. Attenuating interfering talkers is an important part of the solution for the automobile especially in multi-passenger vehicles in which children may be riding, such as mini-vans.

The results of pitch detection experiments with SSPN, reported in section 5.10, suggest that an advanced talker isolation algorithm based on pitch tracking would perform better after the initial noise removal is done. If talker isolation is done before noise removal, then inaccuracies in the pitch estimate could severely attenuate the desired speech and not block the interfering speech well enough. Trying to incorporate the pitch tracking to modify spectral subtraction did not improve noise suppression or speech quality as noted below.

For the single talker case, it was thought that the desired speech could be further improved by pitch tracking used to modify the spectrum. In fact, the use of pitch tracking to adjust the gain in spectral subtraction does not improve the signal, but makes it worse due to inaccuracies in the pitch estimate. The perceptual masked threshold weighting already adjusts the SS gain to avoid attenuating the strong periodic portions of the

speech. The perceptual masked threshold weighting also does a very good job at eliminating the musical-noise artifacts typically introduced by spectral subtraction.

A single talker pitch detection algorithm based on the autocorrelation method was implemented for the experiments. This was done to compare pitch detection on the original clean speech, speech + noise, and enhanced signal. The pitch detector was also used to experiment with a modified spectral subtraction as reported in section 5.10. The autocorrelation method with center clipping is described in Figure 3.8.

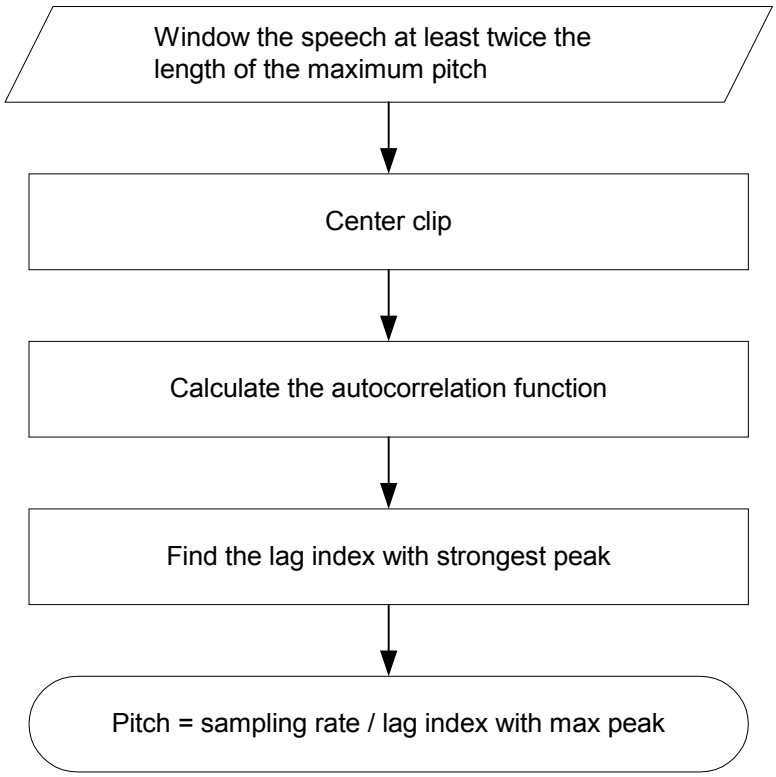


**Figure 3.8: Autocorrelation method for pitch detection**

The pitch detection is disabled when the VAD indicates a frame is unvoiced because there is no pitch to detect. The first formant frequency, which is often near or below the pitch frequency, can interfere with detection. Spectral flattening via center-clipping has been shown to remove the formant structure and enable more accurate pitch detection.[36] Center clipping is described in equation (3.13) where $c$ is the center clip threshold. The clipping threshold calculated as a percentage of the maximum value in the speech frame where 0 % represents no clipping and 100% would remove the whole signal.

$$s_c(n) = \begin{cases} s(n) + c & , s(n) \le -c \\ 0 & , -c \le s(n) \le +c \\ s(n) - c & , s(n) \ge +c \end{cases} \quad \textbf{( 3.13)}$$

The autocorrelation function is defined by equation (3.14) where $m = 0,1,...M$ is the lag numbers and $N$ is the frame size.

$$R(m) = \sum_{n=0}^{N-1-m} s_c(n) s_c(n+m) \quad \textbf{( 3.14)}$$

Normal human pitch ranges from 50 to 900 Hz, so only points that would produce a result in that range were considered. For the case of an 8kHz sample rate, the point between 9 and 160 are the limits of 888Hz and 50 Hz respectively. Again, the above pitch detection algorithm was used to experiment with placement of the talker isolation component of SSPN and was not actually implemented as part of the SSPN simulation used to obtain the results in Chapter 5. The output of the perceptually weighted spectral subtraction was fed back to the VAD for the first iteration and sent to the output via the phase multiplication, inverse FFT, and overlap add synthesis.

### 3.3.8 Adding phase, inverse FFT, and overlap add

The noisy phase from the spectrum of the beamformer output is used as the phase of the speech estimate after the weighted spectral subtraction. It has been shown by Lim[64] and others that the human auditory system is less sensitive to phase distortions, so using the noisy phase as an approximation of the clean speech phase does not significantly degrade the speech quality.[37] An inverse Fourier transform and overlap add is then done to reconstruct the speech estimate in the time domain.

## 3.4 Real-time implementation comments

The computational load on a system is measured in the number of instruction that processor aggregate needs to execute per unit time and typically described in Mega-Instructions Per Second (MIPS). Memory requirements are not easily determined because it depends on how much buffer re-use can be accomplished and other code optimizations. The SSPN algorithm will consume a large amount of a processing power because it is a combination of several fairly complex methods. An exact characterization of an SSPN real-time implementation is beyond the scope of this thesis, but a very rough approximation can easily be done. The resource requirements will vary based on sampling rate, frame size, processor architecture, and programming optimizations, so the following descriptions should be taken in that context. Each of the major components of the SSPN algorithm, listed below, can have their real-time requirements described based on prior work.

1. Generalized Side-lobe Canceller (GSC) beamforming
2. 256 point FFT and inverse FFT
3. Energy based voice activity detector
4. Spectral Subtraction
5. Auditory masking threshold calculation

**GSC beamforming** consumes are large amount of processing resources by itself, with one implementation requiring 2 Motorola 56001 DSP chips, which provide from 26 to 40 MIPS.[48] If longer adaptive filters or more channels are used the processing requirements become even higher.

A **256-point FFT** is almost negligible if an optimal implementation is used on a DSP processor designed for such operations. One example of a 256 16-bit complex FFT requires only 0.27 MIPS on an Analog Devices DSP. The **256-point inverse FFT** has similar complexity compared to the FFT and consumes roughly the same amount of cycles.

An **energy based Voice Activity Detector (VAD)** does not need many resources and is used in many resource-limited applications such as wireless phones. VAD is commonly used as part of a voice compression algorithm or echo cancellation algorithm, where the VAD consumes on the order of less than a single MIP.

**Spectral subtraction (SS)** is also an inexpensive operation, which is why it is such a popular noise suppression algorithm. A brief survey of SS implementation shows that SS requires anywhere from 4 to 10 MIPS.

**Calculation of the auditory masking threshold** add complexity to spectral subtraction type algorithms and not been widely adopted for this reason.[69] However, MIPS consumption for the auditory masking threshold calculation have not been widely published. One of the descriptions mentioned the use of a single ATT&T DSP32C processor capable of 20 MIPS.

**Summing up the worst-case scenario** of the above MIPS approximations totals 72 MIPS, so an iteration of the SSPN algorithms should consume processing resources in a range near 72 MIPS. If the SSPN algorithm is iterated twice to improve the VAD, then the MIPS consumption will approach about 100. These are very crude approximations that provide only a general sense of the SSPN processing needs. A more thorough investigation is required to accurately characterize the real-time resource requirements for the SSPN algorithm.

# Chapter 4

# Algorithms used for noise suppression

This chapter describes the detailed theory behind the methods used in the SSPN algorithm and some other methods to be considered and compared.

## 4.1 Beamforming

Beamforming and multiple microphones allow spatial diversity to be used in the SSPN algorithm and the only part of the algorithm to use this dimension. Beamforming is the term used for steering an array of sensors to have unity gain in the direction of the desired source while attenuating signals originating from other directions. Source localization can be important to improve the effect of beamforming, but for the purposes of the simulations in this thesis an approximate location of the desired source is used and no attempt is made to optimize results based on a better estimate of the desired taker's location. Section 4.1.1 explains the implications of approximating source location and how more accurate estimates could be obtained, especially on non-stationary sources.

Delay and sum beamforming is an important part of the Generalized Side-lobe Canceller (GSC) and is described in section 4.1.2. Section 4.1.3 talks about the GSC, which is used

as part of the SSPN algorithm and is evaluated independently for comparison with spectral subtraction and SSPN.

### 4.1.1 Source localization

Source localization is very important when using spatial information to improve the quality of the desired signal acquired by hands-free microphones because determining the area where you expect the desired signal to be allows you to amplify the signal from that direction while attenuating signals from other directions. The receiver can also steer deep nulls to block interferences if their locations are known. The approach to source localization is very dependent on the application scenario. A conference room can have multiple desired talkers anywhere in the room and at varying distances from the microphone depending on the room size with the additional challenge of tracking moving sources if the talker is walking around. A more constrained environment like the automobile reduces the possible locations of the desired speech because they must be seated in the car.

This thesis is concerned with hands-free speech of the driver in a car, so the talker's location will vary only by seat position, possibly by the person leaning to one side, and the person's height. Interferences from other talkers seated in the car can also be fairly well located. This *a priori* information about source locations in the car allows constraints to be put on the area searched and thus avoids steering to the wrong source and simplifies computations. Sensitivity to location errors will depend on how aggressive the beamformer is trying to attenuate signals coming from undesired

directions. It has been shown that, for a source in close proximity to the microphones, the array aiming location must be accurate to within a few centimeters to prevent high frequency roll off in the received signal. [38]

There are many source localization methods that vary in complexity, robustness to environment, and performance. Three general categories for the methods are listed below.

- **Steered response of a beamformer**
  Maximum Likelihood (ML) steered beam response location estimators steers an array to various locations and searches for a peak in output power.

- **High-resolution spectral estimation concepts**
  Harmonic Enhanced 2-D MUSIC is a high-resolution method that has been effectively applied to hands-free speech in a car. [39]

- **Time-Difference-Of-Arrival (TDOA)**
  TDOA is one of the most popular because of its low complexity and relatively good performance under general conditions. The Cross-power Spectrum Phase Analysis (CSPA) is a method used to calculate the TDOA.

TDOA is described in detail here because it is a good candidate for use with a microphone array in an automobile. The description of a sound wave impinging upon an array of microphones is shown in Figure 4.1

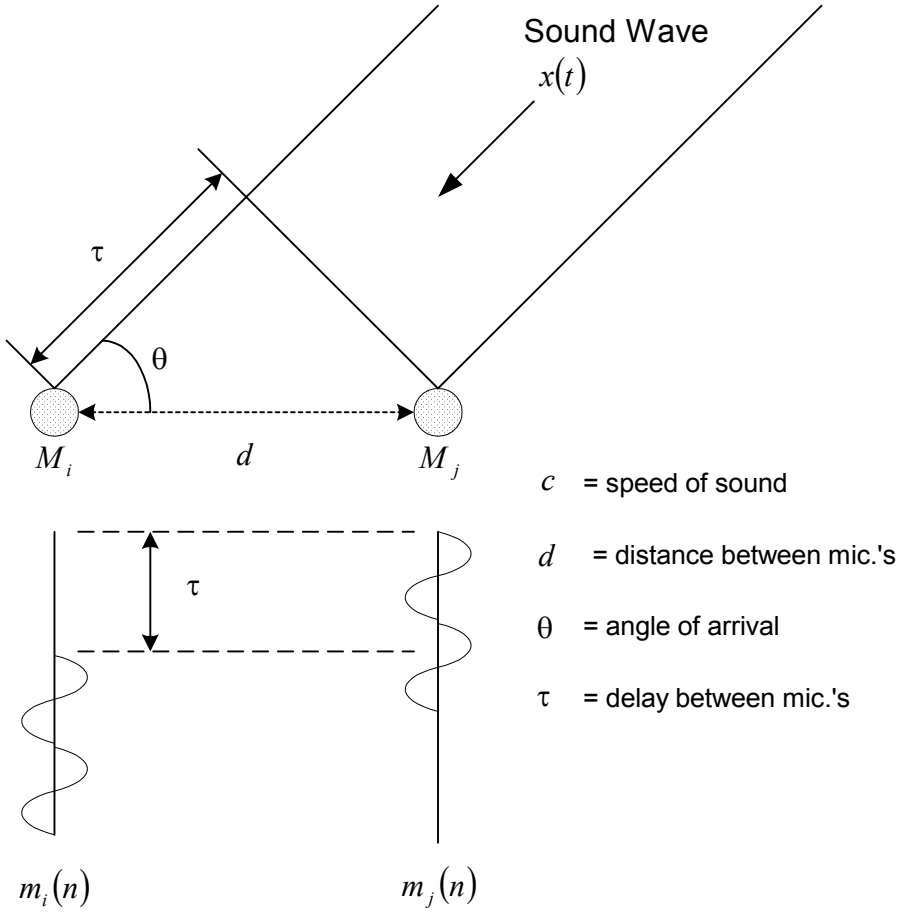


**Figure 4.1: Time Difference of Arrival**

**The *first* assumption** is that the source is in the far field with respect to the microphones, so the received signals can be treated as plane waves. When the source is close to the array (in the near field), the time it takes the signal to propagate to each sensor results in a curved wave-front; thus, the time-delays depend on the exact source location. On the other hand, as the source moves further away (into the far field), the wave-front becomes planar, so the time-delays depend only on the source direction without any range information. The near-far situation, shown in Figure 2.1, depends on the source distance as well as the array aperture (spatial dimension), spacing among the sensors, and the wavelength of interest. When the sensor spacing becomes small, a curved wave-front can

be well-approximated by a planar wave-front. As a result, generally only angle estimation (DOA) is possible in the far-field for a single array, while source localization is possible only in the near-field.[40]

Given a single source of sound that produces a time varying signal $x(t)$ each microphone will receive the following signals

$$m_i(t) = \alpha_i x(t - \Delta_i) + z_i(t) \quad \textbf{(4.1)}$$

$$m_j(t) = \alpha_j x(t - \Delta_j) + z_j(t) \quad \textbf{(4.2)}$$

where $\Delta_i$ and $\Delta_j$ are the time delays it takes for sound to propagate from the source to microphone and $z_i(t)$ and $z_j(t)$ are the noise signals present. The Time Difference of Arrival (TDOA) is $\tau = \Delta_i - \Delta_j$.

Finding the TDOA via Cross-power Spectrum Phase Analysis (CSPA) is divided into the following steps, shown in Figure 4.2. [41]

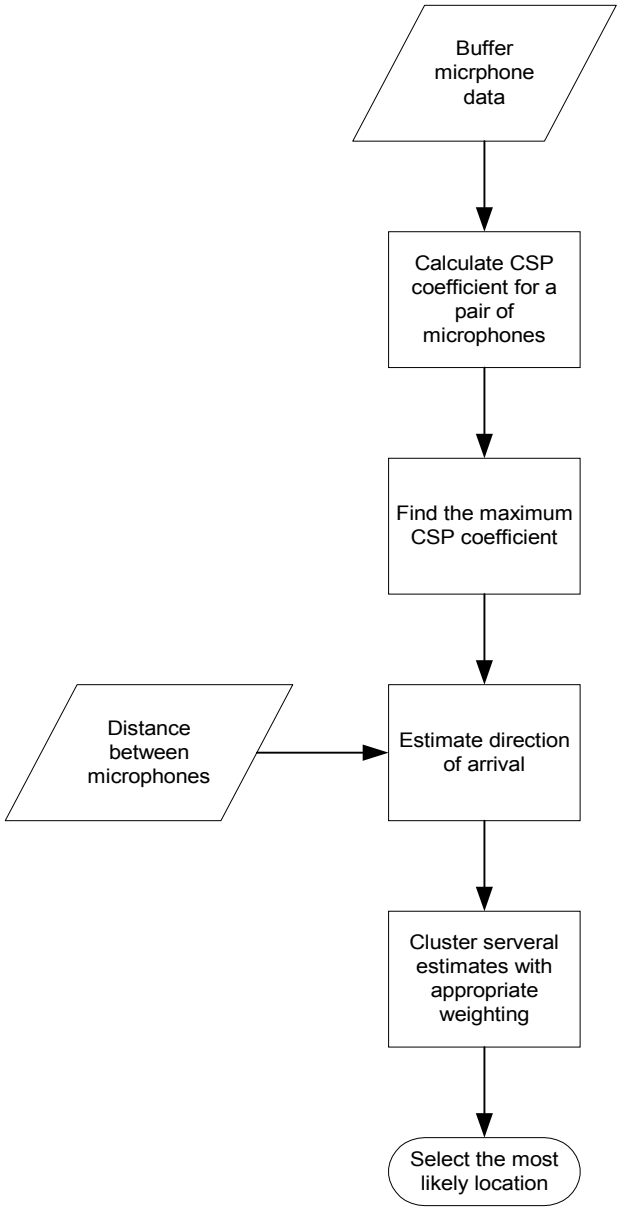


**Figure 4.2: TDOA CSP steps**

The Fourier transform of the captured signals is:

$$m_i(t) \leftrightarrow M_i(\omega) = \alpha_i X(\omega) e^{-jw\Delta_i} + Z_i(\omega) \quad \textbf{(4.3)}$$

$$m_j(t) \leftrightarrow M_j(\omega) = \alpha_j X(\omega) e^{-jw\Delta_j} + Z_j(\omega) \quad \textbf{(4.4)}$$

**The *second* assumption** is that the average energy of the source signals is greater than that of the interfering noise and is shown in equation (4.5) where the over-bar represents the average operation.

$$\alpha_i{}^2 \overline{|X(\omega)|^2} >> \overline{|Z_i(\omega)^2|} \quad \textbf{(4.5)}$$

The cross correlation of $m_i(t)$ and $m_j(t)$ is calculated in (4.6) and maximized when $\psi$ is equal to the TDOA.

$$R_{ij}(\tau) = \int_{-\infty}^{\infty} m_i(t) m_j(t-\psi) dt \quad \textbf{(4.6)}$$

Converting the cross correlation to the frequency domain and expanding yields equation (4.7) where $M_j^*(\omega)$ is the complex conjugate of $M_j(\omega)$.

$$R_{ij}(\tau) \leftrightarrow S_{ij}(\omega) = M_i(\omega) M_j^*(\omega) \quad \textbf{(4.7)}$$

$$S_{ij}(\omega) = \left(\alpha_i X_i(\omega) e^{-j\omega\Delta_i} + Z_i(\omega)\right)\left(\alpha_j X_j(\omega) e^{-j\omega\Delta_j} + Z_j(\omega)\right)^* \quad \textbf{(4.8)}$$

Because of the second assumption, the cross terms of equation (4.8) can be considered negligible, hence (4.8) reduces to equation (4.9).

$$S_{ij}(\omega) \approx \alpha_i \alpha_j |X_i(\omega)|^2 e^{-j\omega(\Delta_i - \Delta_j)} \quad \textbf{(4.9)}$$

The delay, $\tau_{ij}$, can be found by evaluating the equation (4.10) where $F^{-1}$ is the inverse Fourier transform.

$$\tau_{ij} = \max_{\tau} R_{ij}(\tau) = \max_{\tau} F^{-1}\{S_{ij}(\omega)\} \quad \textbf{(4.10)}$$

Equation (4.11) is the cross-power spectrum phase (CPSP) function.

$$CPSP_{ij}(w) \cong \frac{M_i(\omega)M_j^*(\omega)}{|M_i(\omega)||M_j(\omega)|} \quad \textbf{(4.11)}$$

Taking the inverse Fourier transform and converting to the discrete domain gives:

$$cpsp_{ij}(t) = IDFT\left\{\frac{DFT\{m_i(t)\}DFT\{m_j(t)\}^*}{|DFT\{m_i(t)\}||DFT\{m_j(t)\}|}\right\} \quad \textbf{(4.12)}$$

The *cpsp* function, in typical conditions, is delta-like and has a peak that $\tau = \tau_{ij}$.

The angle of arrival can then be found using the calculation in equation (4.13) where $c$ is the speed of sound at approximately 340 meters/second, $\tau$ is the estimated delay, $d$ is the distance between microphones, and $F_s$ is the sampling rate.

$$\theta = \cos^{-1}\left(\frac{c \cdot \tau / F_s}{d}\right) \quad \textbf{(4.13)}$$

Finally, several locations are clustered and analyzed to determine the most likely direction of arrival. Various clustering and optimization methods can be applied to choose the best DOA. Some frequency specific considerations should be applied rather than assuming uniform processing across all bands. Noise power can be severe below 200 kHz, so in the absence of noise reduced signal it makes sense to discard information

at those frequencies. It can also be expected that frequencies with greater magnitude come from the source because the desired source dominates the average energy per frame and would suggest weighting portions of $S_{ij}(w)$ with greater magnitude more heavily in order to increase accuracy.[42]

### 4.1.2 Delay and sum beamforming

Multiple microphones can have their data processed to form a receptive beam that enhances signals from certain directions and suppresses signals from other directions. The spacing between the microphones and different delays of the signals to each microphone are used to spatially discriminate between signals. A beamformer can be considered a spatial filter. Figure 4.3 shows the processing for a delay and sum beamformer.

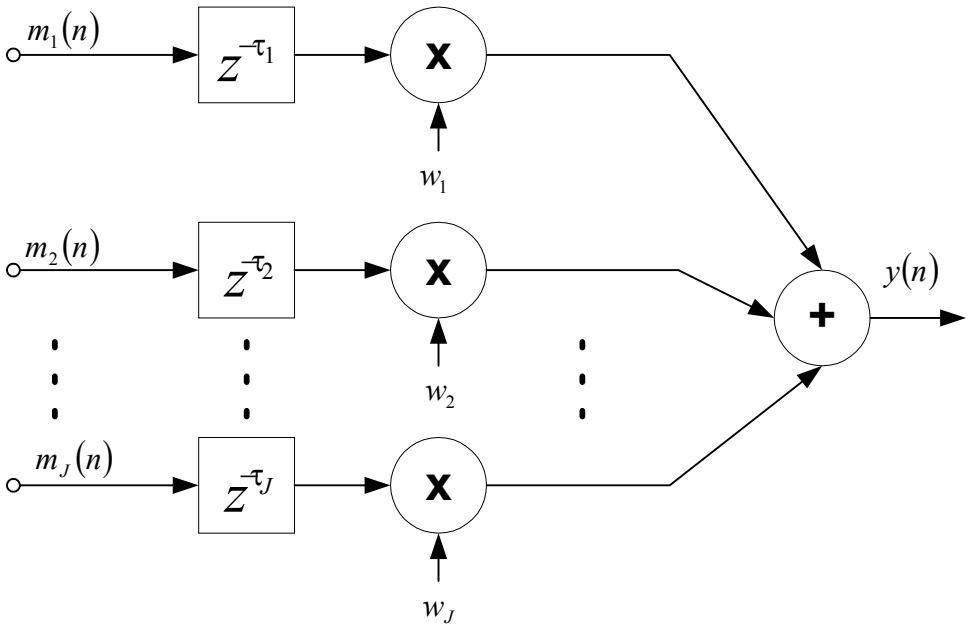


**Figure 4.3: Conventional Delay and Sum Beamformer**

Figure 4.1 shows the physical layout of the microphones and signals. The delay between microphones is described by equation (4.14) where $c = 342 m/s$ is the speed of sound, $d_i$ is the distance between the microphones, and $\theta_i$ is the angle of arrival. If the microphones are in a straight line and have equal spacing between them, then it is referred to as a Uniform Linear Array (ULA). A ULA simplifies the equations by allowing them to use constant distance and delays respectively.

$$\tau_j = \frac{d_j \cos\theta_j}{c} \quad \textbf{( 4.14)}$$

The array of received signals at each microphone is

$$\vec{m}(n) = \left[m_1(n), m_2(n), m_3(n), ..., m_J(n)\right]^T \quad \textbf{(4.15)}$$

The source signal received by the *j*th microphone with a delay $\tau_j$ is

$$m_j(n) = h_j(n,\theta_s) * s(n-\tau_j) \quad \textbf{(4.16)}$$

where $h_j(n,\theta_s)$ is the impulse response of the *j*th microphone.

$$\vec{m}(n) = \sqrt{J}\vec{v}(\theta_s)s(n) \quad \textbf{(4.17)}$$

The array response vector $\vec{v}(\theta_s)$ is derived directly from equations (4.16) and (4.17). Equation (4.18) is often used in equations that analyze the response of the delay and sum beamformer.

$$\vec{v}(\theta_s) = \frac{1}{\sqrt{J}}\left[e^{-j2\pi F_c \tau_1(\theta_s)}, e^{-j2\pi F_c \tau_2(\theta_s)}, ..., e^{-j2\pi F_c \tau_J(\theta_s)}\right]^T \quad \textbf{(4.18)}$$

Substituting equation (4.14) into equation (4.18) and recognizing that $F_c = c/\lambda$ yields

$$\vec{v}(\theta_s) = \frac{1}{\sqrt{J}}\left[e^{-j2\pi d_1 \cos(\theta_s)/\lambda}, e^{-j2\pi d_2 \cos(\theta_s)/\lambda}, \ldots, e^{-j2\pi d_J \cos(\theta_s)/\lambda}\right]^T \quad \textbf{(4.19)}$$

Equation (4.20) is the output of the delay and sum beamformer where $m_j$ is the signal received at each microphone, $J$ is the number of microphones, and $w_j$ is the weighting coefficient for a microphone $j$. The weights can be used to shape the received signal to further attenuate signals coming from undesired locations.

$$y(n) = \frac{1}{J}\sum_{j=0}^{J-1} w_j m_j(n - \tau_j) \quad \textbf{(4.20)}$$

The "far field assumption" is made here, which assumes the incident signals are plane waves as described in section 2.1 and section 4.1.1. One can assume a planar wave if the source to sensor distance is at least twice the aperture of the sensor array. The aperture of the microphone array is the distance between the first and last microphone. A Uniform Linear Array (ULA) with 4 microphones and uniform spacing of 5 cm would have an aperture of 15 cm, for example.

A source is said to be in the near field if equation (4.21) is true

$$r < \frac{2L^2}{\lambda} \quad \textbf{(4.21)}$$

where $r$ is the radial distance from the microphone, $L$ is the aperture of the array, and wavelength $\lambda$ is defined as the speed of sound over frequency.

$$\lambda = \frac{c}{f} \quad \textbf{(4.22)}$$

Inserting equation (4.22) into (4.21) yields the near field requirement of

$$r < \frac{2L^2 f}{c} \tag{4.23}$$

or

$$f > \frac{rc}{2L^2} \tag{4.24}$$

Given the following example:

$$r = 0.24m$$
$$L = 0.15m$$
$$c = 342m/s$$

Equation (4.24) then tells us that a near field assumption is valid for frequencies above 1824 Hz. Conversely, a far field assumption is valid for frequencies below 1824 Hz. The range of frequencies that can be considered in the far field condition becomes higher as the source's distance $r$ from microphone increases. The maximum frequency in the far field decreases as the array aperture $L$ increases, which represents a tradeoff because a larger array aperture provides better spatial resolution.[43]

A microphone array increases in its ability to distinguish between closely spaced sources as its aperture increases because of a narrower main beam and narrower side-lobes for larger aperture, as can be seen in Figure 4.4. The degrees of freedom also increase with number of microphones, where 5 microphones produce a single main lobe plus four nulls and 10 microphones produce one main lobe and 9 nulls. The beam response of equation (4.25) is computed by applying the beamformer weights to a set of array response vectors from all possible angles, $-90^\circ \le \phi < 90^\circ$.

$$W(\phi) = \vec{w}^H \vec{v}(\phi) \tag{4.25}$$

Figure 4.4 shows the magnitude of the beam response $|W(\phi)|^2$ for $w_j = 1/\sqrt{J}$ uniformly weighted beamformers where *J=5* and *J=10* respectively.

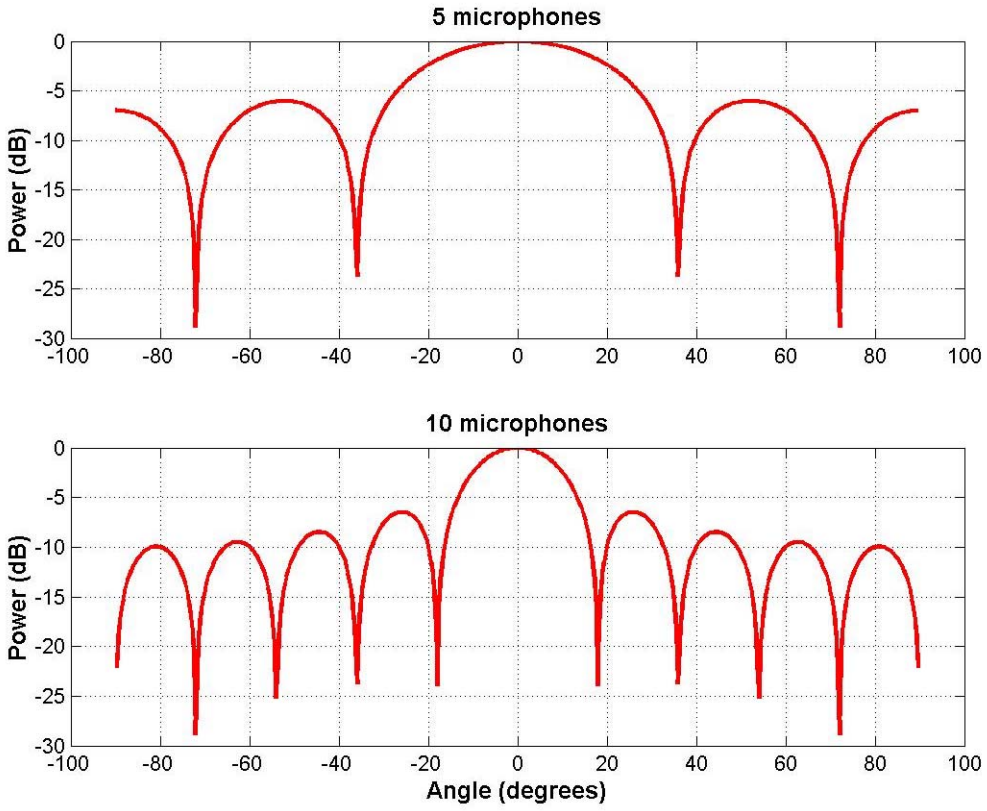


**Figure 4.4: Beam pattern for different apertures**

The spacing between elements has a direct effect on the frequency resolution of the array. In order to avoid aliasing in the spatial frequency domain the spacing between elements should be less than half the wavelength of interest, which is shown in equation (4.26). Figure 4.5 shows that multiple main lobes with unity gain are created when the distance between microphones increase beyond half of a given wavelength. Multiple main lobes make certain angles of arrival indistinguishable from others, which is the definition of spatial aliasing.

$$d \leq \frac{\lambda}{2} \tag{4.26}$$

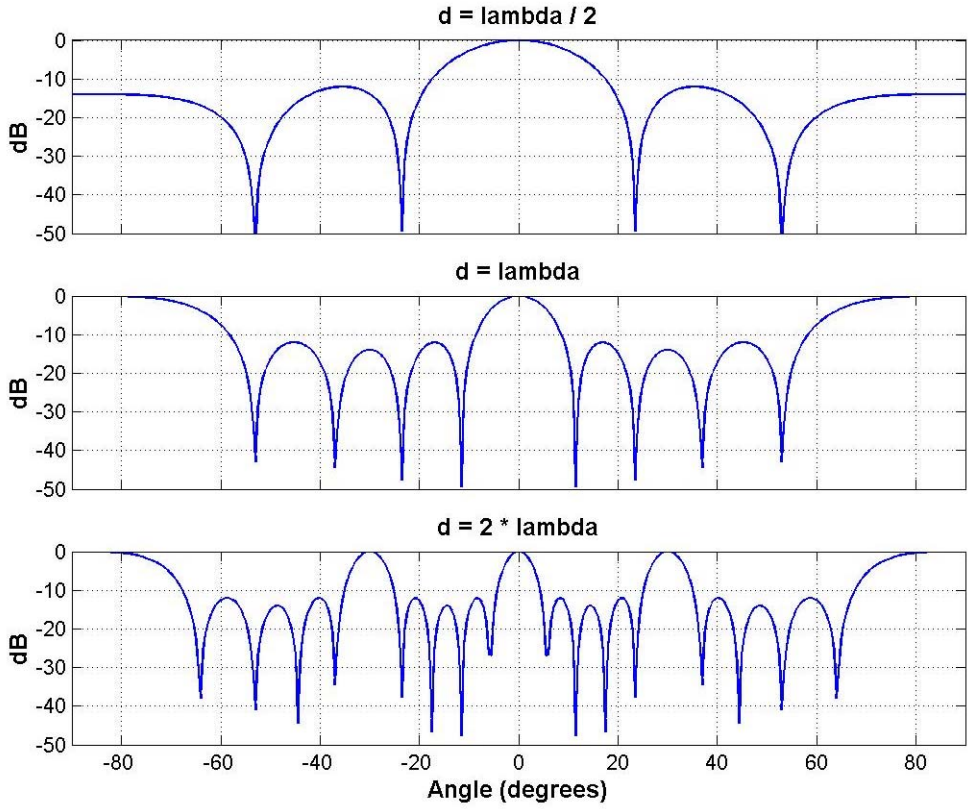


**Figure 4.5: Spatial aliasing beamformer**

To obtain un-aliased spatial frequency resolution at 100 Hz the spacing between microphones would have to be less than or equal to 1.7 meters. However, for a frequency of 3400 Hz the inter-microphone spacing would have to be less than or equal to 0.05 meters, which would also be acceptable for frequencies lower than 3400 Hz.

Ideally, the beamformer would have the same response over all frequencies of interest, which is referred to as Constant Directivity Beamforming (CBD). Nested sub-arrays of

microphones and forming multiple beams are the two major techniques used to achieve a constant response over a broad range of frequencies. The problem that CBD is trying to overcome is the fixed inter-element spacing of microphones, which causes the main beam to be wider (with less resolution) at lower frequencies and spatial aliasing to occur at higher frequencies. Spatial selectivity start to be lost at below about 800 Hz for a delay & sum beamformer using a Uniform Linear Array (ULA) with 5 microphones and uniform spacing of 5 cm.[44]

The beamforming gain is defined as the ratio of the SNR for one microphone to the SNR for the array, where

$$G_{bf} \cong \frac{SNR_{array}}{SNR_{one}} = J \frac{\left|\vec{w}^{H}\vec{v}(\theta_s)\right|^2}{\left|\vec{w}\right|^2} \quad \textbf{(4.27)}$$

($G_{bf}$) is strictly a function of the angle of arrival ($\theta_s$) of the desired signal, the beamforming weight vector ($w$), and the number of microphones ($J$). Equation (4.27) shows that $G_{bf}$ scales linearly with the number of microphones $J$ [45] and equation (4.19) is the array response vector, $v(\theta_s)$, for the array in look direction $(\theta_s)$ where $T$ represents the vector transpose operation.

$d$ = distance between microphones
$c$ = speed of sound which is about 342 m/s
$\lambda$ = wavelength of the signal

### 4.1.3 Generalized Side-lobe Canceller (GSC)

The Generalized Side-lobe Canceller is used in the SSPN algorithm and is evaluated independently for comparison purposes. The GSC has higher noise suppression than the delay & sum beamformer because it can adaptively steer nulls toward locations of interfering sources in addition to the steering of the main beam in the direction of the desired source. This higher noise suppression and adaptation to the noise sources was the motivation for choosing the GSC instead of the delay & sum beamformer for the SSPN algorithm. Theoretical advantage of the GSC was confirmed by the simulations results using the signals measured in the car.

The beamforming problem can also be formulated at a constrained minimization [46] where the goal is to minimize the array output subject to the constraint of unity gain in the desired direction. The output of the array can be described by equation (4.28) where $\vec{m}$ is the vector of inputs received at the array microphones.

$$y(n) = \vec{w}^H \vec{m}(n) \tag{4.28}$$

Equation (4.29) defines the output covariance matrix, which is often used to compute the weights of an optimal filter. The symbol $H$ represents the complex conjugate (Hermitian) transpose operation.

$$R = E\left[\vec{m}(n)\vec{m}^H(n)\right] \tag{4.29}$$

The goal is to minimize the contribution of the noise, which is done by adapting the filters to minimize the output power of the array while maintaining unity gain in the look direction. The power to minimize is shown in equation (4.30) subject to the constraint of

unity gain in the look direction in equation (4.31) where is $v(\theta)$ defined in equation (4.19).

$$P = E\left[\left|y(n)\right|^2\right] = w^H E\left[\vec{m}(n)\vec{m}^H(n)\right]\vec{w} = \vec{w}^H R\vec{w} \tag{4.30}$$

$$\vec{w}^H \vec{v}(\theta) = 1 \tag{4.31}$$

The solution to this problem is the Linearly Constrained Minimum Variance Filter (LCMVF) using the weights of equation (4.32).

$$\vec{w} = \frac{R^{-1}\vec{v}(\theta)}{\vec{v}^H(\theta)R^{-1}\vec{v}(\theta)} \tag{4.32}$$

Griffiths & Jim[47] reformulated the problem as a natural separation of the constraint and the minimization described as follows. The *Griffiths & Jim* beamformer is also known as the Generalized Side-lobe Canceller (GSC). In the GSC the constraint is inserted in the direct signal path and an unconstrained Least Mean-Square (LMS) algorithm is used to minimize the output energy in $y_{out}$. The architecture of a 4-microphone GSC is shown in Figure 4.6 where the four microphones outputs are time delay steered to produce 4 signals, which ideally have the desired signal in phase with each other. These four signals are then sent to the blocking matrix whose purpose is to block out the desired signal from the lower path of the GSC. The blocking matrix produces 3 signals that are fed into adaptive FIR filters.

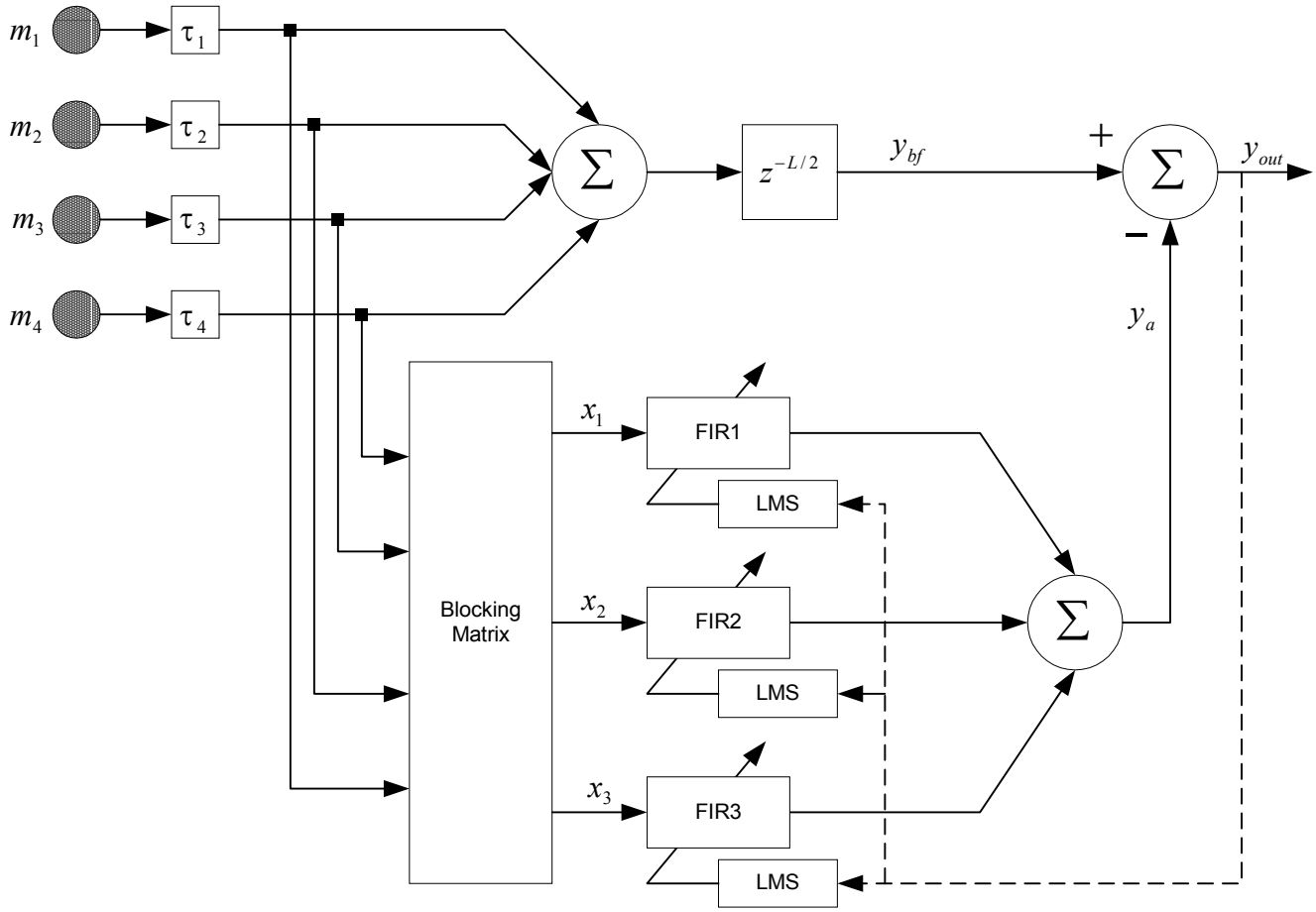


**Figure 4.6: Generalized Side-lobe Canceller**

Equation (4.33) is an example of a blocking matrix that simply takes the difference of the input channels and assumes the inputs are perfectly in phase where $\omega_0 = 0$.

$$B = \begin{bmatrix} e^{-j\omega_0} & -1 & 0 & 0 \\ 0 & e^{-j2\omega_0} & -1 & 0 \\ 0 & 0 & e^{-j3\omega_0} & -1 \end{bmatrix} \quad \textbf{(4.33)}$$

The matrix $B$ represents a bank of band-rejection filters, each of which is tuned to an angular frequency $\omega_0$. The function of the matrix $B$ is to cancel interference that leaks

through the side-lobes of the band-pass filter in the direct path. A detailed development of the blocking matrix of the GSC can be found in the adaptive filter theory textbook by Simon Haykin.[2]

The bottom branch of Figure 4.6 contains 3 filtered versions of the noise to be subtracted from the top branch. The top branch of the GSC is the beamformed signal delayed by *L/2* to be in phase with the noise signals after the FIR filters, where *L* is the number of taps in the filters. The filters in the bottom branch must adapt to approximate the noise in the top branch. The Normalized Least-Mean-Square (NLMS) algorithm is used to update the coefficients of the FIR filters for the implementation of the GSC used in this thesis.

Ideally the lower path would not contain any of the desired signal and the filters would adapt to cancel the noise source. Unfortunately, reverberation and multi-path effects cause only a portion of the desired signal to arrive at the array from the steering direction. Multi-path combined with an energy minimization criteria results in signal cancellation because the free filter coefficients are partially adapted to minimize power from the desired signal.[48]

The NLMS algorithm is a stochastic gradient algorithm used for linear adaptive filtering similar to the LMS algorithm, where the tap weight vector, $\vec{w}_n$, represents an estimate whose expected value approaches the optimal Wiener solution.[49] Minimizing the output of the GSC in equation (4.34) is done by adapting the coefficients using the NLMS algorithm.

$$y_{out}(n) = y_{bf}(n) - y_a(n) \tag{4.34}$$

Given the filter tap-input vector $x(n)$ and desired response $y_a(n)$, determine the tap-weight vector $w(n+1)$ so as to minimize the squared Euclidean norm of the change

$$\delta w(n+1) = w(n+1) - w(n) \tag{4.35}$$

in the tap weight vector $w_{k+1}$ with respect to its old value $w_k$, subject to the constraint

$$w(n)^H x(n) = y_a(n) \tag{4.36}$$

The squared norm of the change $\delta w(n+1)$ in the tap weight vector $w(n+1)$ may be expressed as

$$\|\delta w(n+1)\|^2 = \delta w^H(n+1)\delta w(n+1) \tag{4.37}$$

$$\|\delta w(n+1)\|^2 = [w(n+1) - w(n)]^H [w(n+1) - w(n)] \tag{4.38}$$

$$\|\delta w(n+1)\|^2 = \sum_{j=0}^{Order-1} |w_j(n+1) - w_j(n)|^2 \tag{4.39}$$

The remaining derivation of the NLMS algorithm goes on to define a complex cost function based on the squared norm $\|\delta w(n+1)\|^2$ and uses *Lagrange multipliers* to find the solution to the minimization of equation (4.39). [50]

The vectors $\vec{w}_{1,n}$, $\vec{w}_{2,n}$, and $\vec{w}_{3,n}$ contain the filter coefficients of FIR1, FIR2, and FIR3 respectively at time *n*, they can be combined to create an overall filter coefficient vector

$\vec{w}_n$ in equation (4.40). Similarly an overall input vector can be created as shown in equation (4.41).

$$\vec{w}_n = \begin{bmatrix} \vec{w}_{1,n} \\ \vec{w}_{2,n} \\ \vec{w}_{3,n} \end{bmatrix} \tag{4.40}$$

$$\vec{x}_n = \begin{bmatrix} \vec{x}_{1,n} \\ \vec{x}_{2,n} \\ \vec{x}_{3,n} \end{bmatrix} \tag{4.41}$$

NLMS update of the coefficients at time *(n+1)*, where $\mu$ is the step-size is:

$$\vec{w}_{n+1} = \vec{w}_n + \frac{\mu}{\|\vec{x}_n\|^2} \vec{x}_n y_{out,n} \tag{4.42}$$

The current output at time *k* is

$$y_{out,n} = y_{bf,n} - y_{a,n} \tag{4.43}$$

where

$$y_{a,n} = \vec{w}_n^H \vec{x}_n \tag{4.44}$$

The delay *(L/2)*, corresponding to half the filter's propagation time, is introduced to the $y_{bf}(k)$ to ensure that the middle of each of the adaptive filters at time k corresponds to $y_{bf}(k)$. This means that both samples were generated from the same input samples before their paths split.

## 4.2 Voiced Activity Detection (VAD)

The purpose of Voice Activity Detection (VAD) is to determine whether a frame of the captured signal represents voiced, unvoiced, or silent data. Voice activity detection ideally is aware of the human speech production system, so it can differentiate between silence, unvoiced, and voiced sounds. Voiced sounds are periodic in nature and tend to contain more energy than unvoiced sounds, while unvoiced sounds are more noise-like and have more energy than silence. Silence has the least amount of energy and is a representation of the background noise of the environment. The VAD plays a central role in the SSPN algorithm in that its accuracy dramatically affects the noise suppression level and amount of speech distortion that occurs.

Applications of VAD include speech recognition, voice compression, noise estimation/cancellation, and echo cancellation. Speech recognition is concerned with finding out exactly when a word or utterance begins and ends; these are called the speech endpoints. The speech recognizer requires accurate endpoints in order to achieve good performance when pattern matching. Voice compression uses VAD to reduce the required bit rate needed to accurately transmit a voice stream. The percentage of time that a VAD detects the presence of speech is called the voice activity factor, VAF, which can range from 44% down to 36%.[51] During typical conversation, talk spurts comprise only 31.5% of each party's speech and the remaining 68.5% is silence.[52] The wireless phone voice compression standards such as GSM (Global System for Mobile Communications), EVRC (Enhanced Variable Rate Coder), and ITU (International Telecommunications Union) G.729 are examples of algorithms that use VAD.[53], [54], [55]

Single channel noise estimators require a VAD to know when to update the noise reference and a VAD is used in many spectral subtraction algorithms, which works well as long as the noise is slowly varying. This single channel application of the VAD is exactly what is used in the SSPN algorithm where the VAF range of 36 to 44% for typical speech patterns enables the accumulation of an accurate stationary noise reference.

The choice of a VAD algorithm for the SSPN algorithm required a trade off of delay, sensitivity, accuracy, and computational cost. Some of the measures used in VAD algorithms are the Itakura LPC distance, energy levels, spectral energy distribution, timing, pitch, zero crossing rates, cepstral features, adaptive noise modeling, and periodicity. Noise frames have also been detected by measuring the spectral difference over a number of time periods.[56] Many of the algorithms also assume that the first 4 to 10 frames are silence in order to initialize the noise estimate, where 10 frames or 160 ms was used in the VAD for the SSPN algorithm. Most VADs operate well in the 5 to 10 dB SNR range with a few advanced algorithms reaching to between -5 and 0 dB [57] and the VAD used in this thesis worked reasonably well for the purposes of a noise estimate down to 0 dB input SNR. An algorithm by Rabiner and Sambur was experimented with for this thesis, which is noted for good performance for speech endpoint analysis and low computational cost uses a combination of zero crossings and energy level detection.[58] However, it became too conservative in classifying noise at low SNR and did not update the noise reference frequently enough, this decreased the amount of noise suppression the SSPN algorithm could achieve.

### 4.2.1 Energy level detection

Energy level detection in the frame is one of the simplest and earliest measures used for VAD, and was chosen as the VAD method used for the SSPN algorithm in this thesis. Other research has extended the energy calculation to dual and multiple spectral sub-bands within each frame.[55] [59] The initial noise spectrum, mean, and variance are calculated assuming the first 10 frames are noise only. Thresholds are calculated for speech and noise decisions and all statistics are gradually updated when a noise frame is detected. The update factors $\alpha$ and $\beta$ can be tuned and have been set to 0.95 in previous experiments.[59] The steps for the VAD algorithm used for all calculations in this thesis are outlined in Figure 4.7, where the same VAD was used in order to make fair comparisons between spectral subtraction and SSPN. More advanced VAD algorithms documented in the literature are mentioned, in the paragraphs following the detailed description of energy detection, as possibilities to investigate in future research to improve the SSPN algorithm.

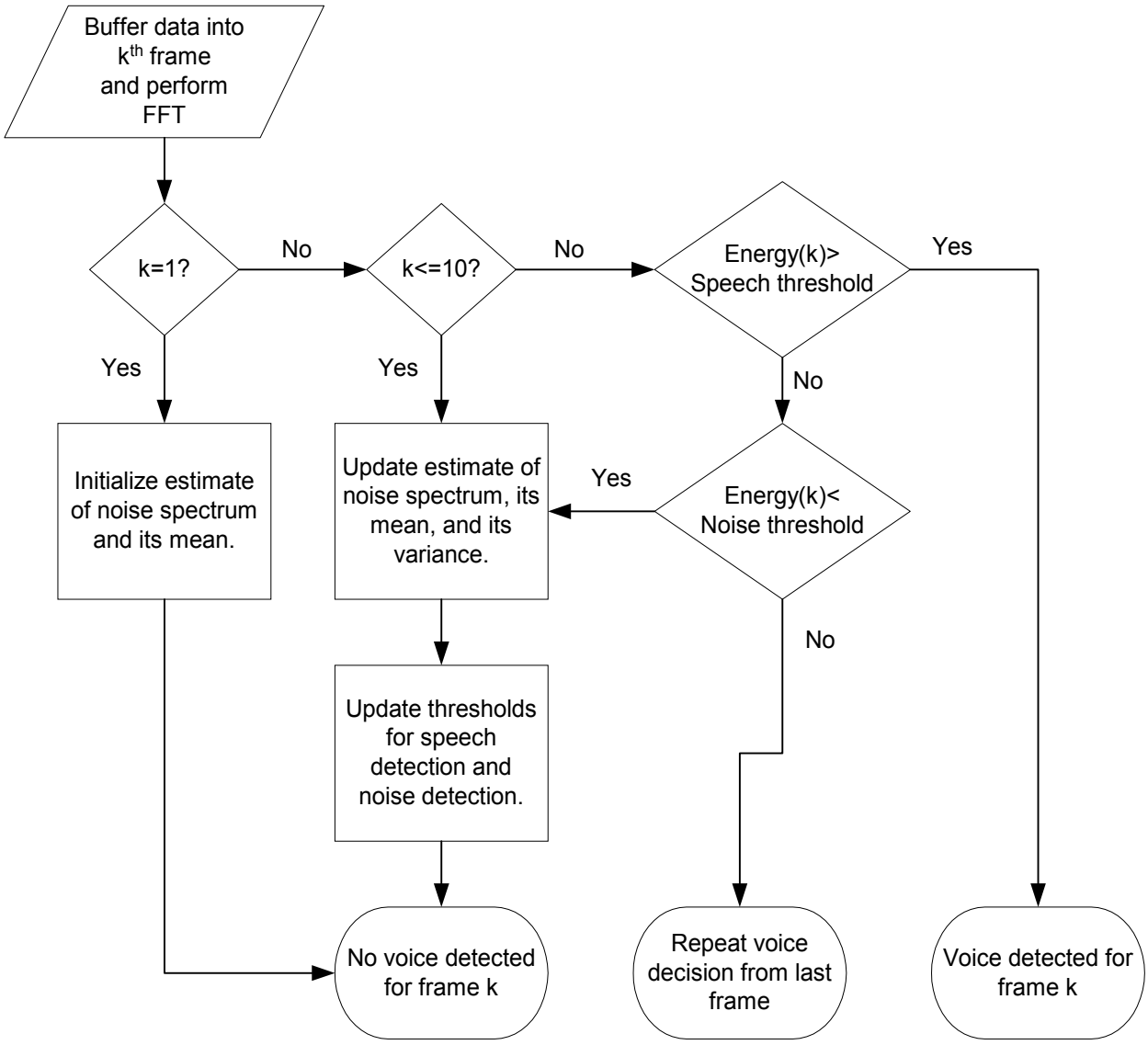


**Figure 4.7: VAD using energy detection**

The equations used in the VAD of Figure 4.7 are described next, where these equations were used directly as part of the SSPN VAD.

1. **Buffer data** into the $k^{th}$ frame, *x(n, k),* and transform to the frequency domain.

$$X(w,k) = FFT(x(n,k)) \tag{4.45}$$

2. **Initialize** the noise spectrum and noise mean for $k$ =1.

$$N(w) = X(w,k) \tag{4.46}$$

$$\mu_N = \frac{1}{L}\sum_{w=0}^{L-1} N(w) \tag{4.47}$$

3. **If VAD = 0, then update** the noise spectrum, mean, and standard deviation for frame. Frames 2 through 10 are assumed to be noise in order to get a good initial average of the stationary noise in the environment.

$$N(w) = \alpha N(w) + (1-\alpha)X(w,k) \tag{4.48}$$

$$\mu_N(k) = \frac{1}{L}\sum_{w=0}^{L-1} N(w) \tag{4.49}$$

$$\mu_N = \beta\mu_N + (1-\beta)\mu_{N(k)} \tag{4.50}$$

$$\sigma_N = \left(\beta\sigma^2{}_N + (1-\beta)\mu_{N(k)}{}^2\right)^{1/2} \tag{4.51}$$

The mean of he noise estimate is $\mu_N$, the standard deviation of the noise estimate is $\sigma_N$, and the noise estimate variance is represented by $\sigma^2{}_N$.

4. **Update thresholds** if a frame does not contain speech, using the mean and variance of the noise estimate where threshold settings are adjusted using the multipliers $\alpha_S$ and $\alpha_N$, which can be adapted and set experimentally. Optimally adapting these VAD thresholds has been the subject of recent research [57], but

was not attempted in this thesis because sensitivity to the thresholds was reduced by the iteration of the algorithm as mentioned in section 5.4.

$$Thresh_S = \mu_N + \alpha_S \sigma_N \quad \textbf{(4.52)}$$

$$Thresh_N = \mu_N + \alpha_N \sigma_N \quad \textbf{(4.53)}$$

5. **VAD decisions** can be made with a speech threshold determination where if the signal energy exceeds twice the standard deviation above the mean of the noise, then the frame is classified as speech. If the signal energy falls within some fraction of the noise standard deviation, then it is classified as noise and modifies the reference accordingly. If neither speech nor noise is chosen, then the last frame's decision is repeated for the current frame.

   if(Energy(k) > $Thresh_S$), VAD(k) = 1

   if(Energy(k) < $Thresh_N$), VAD(k) = 0

   else VAD(k) = VAD(k-1)

The plot in Figure 4.8 shows the sentence, "Nonlinear speech processing", and was sampled at 8kHz and the frame size was 128 samples or 16ms. The solid line superimposed on the plot shows the portions of the signal that are classified as speech when it is one and noise when it is zero. From this plot you can visually conclude that the algorithm performs quite well on average, but has some trouble at transitions.

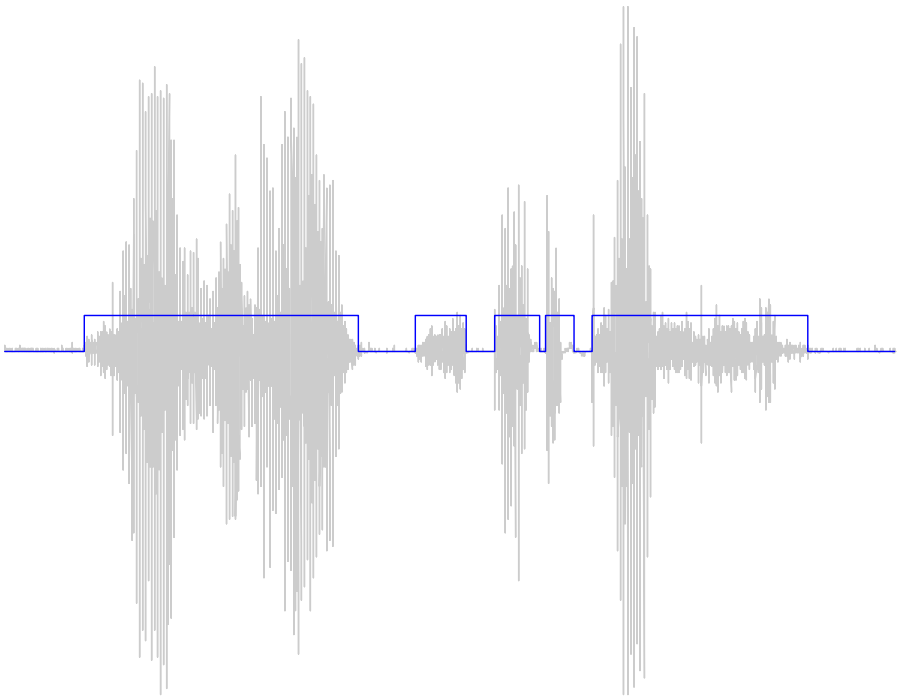

**Figure 4.8: Voice Activity Detection**

### 4.2.2 Other VAD algorithms

Other VAD algorithms have been developed to improve accuracy of detecting the beginning and ending of speech segments. The additional measures of voice activity can be considered for possible improvements to the SSPN algorithm.

Zero crossing rates can be calculated for each frame and compared with a threshold. The zero crossing rate of noise is assumed to be considerably larger than that of speech. This assumption works well at high SNR values, but has problems at low SNR and in the

presence of periodic noise (interfering talkers).[60] Zero crossing was added to the criteria used for the VAD in the SSPN algorithm, but it had the affect of not classifying enough frames as non-speech, so the noise estimate was not updated frequently enough, especially at lower input SNR values near 0 dB. The zero crossing rate was taken out of the VAD for the final SSPN algorithm used to report the results in Chapter 5.

Periodicity is a major indicator that speech is present in the current frame. However, care must be taken to understand the possible noise present in the environment because interfering talkers, tones, or other periodic noise would cause false indications of speech. Pitch tracking and other measures can minimize the problem of periodic noise. Tucker designed a VAD based on periodicity that operates successfully even at 0 dB SNR and has moderate performance as low as -5 dB.[61] The detector uses a least-squares periodicity estimator, LSPE, on the input signal and triggers when a significant amount of periodicity is found. Irwin investigated the LPSE optimum method for measuring periodicity and discusses the required preprocessing.[62] Periodicity was not used as a VAD indicator for the SSPN algorithm because it does not detect unvoiced non-periodic speech utterances and cannot accurately locate the boundaries of the speech. Periodicity should be used in combination with other VAD methods or for speech applications that can afford a margin of error to account for missed unvoiced speech.

Hangover time is needed to prevent certain low-energy unvoiced speech sounds from being confused with background noise. If the noise level is high enough speech sounds such as /f/, /th/, or /h/ can be confused with the noise and there are also extremely short

pauses during active speech plosives such as /p/, /t/, and /k/, where a detector could prematurely declare the start of a silent frame. Hangover is the amount of time that the VAD will delay its decision to declare silence where more delay will cause more frames to be counted as speech and allow less updates of the noise estimate. Less delay will allow the noise reference to be updated more frequently, but may result in a higher speech threshold and mask the low energy speech mentioned above. This tradeoff must be tuned according to the specific application and environment.[63] The use of hangover time was experimented with for the VAD in the SSPN algorithm with mixed results. Hangover helped reduce distortion of the speech if the speech detection thresholds were set high in an attempt to aggressively attenuate the noise because the hangover would prevent some speech from being falsely classified as noise. However, if the speech detection threshold was set low enough, then the speech was already being detected and the only effect of the hangover was to reduce the frequency of the noise update thus suppressing less noise. The hangover was not included in the SSPN algorithm for the results reported in Chapter 5 because the algorithm took the conservative approach by classifying more frames as speech by setting the speech detection threshold fairly low. Another reason the hangover was not needed is the iteration of the SSPN algorithm tended to expand the number of frames classified as speech as reported in section 5.4.

## 4.3 Spectral subtraction

### 4.3.1 General spectral subtraction

Generalized spectral subtraction (GSS) was used for the SSPN algorithm to allow the parameters of the noise subtraction to be modified by the auditory perceptual masking threshold function. GSS also has the ability to fall back to simple spectral subtraction with half wave rectification, which is used in the SSPN algorithm prior to calculating the masking threshold. Spectral subtraction uses the short-term spectral magnitude of the noisy speech and an estimate or reference of the noise signal. Most single channel spectral subtraction methods use a voice activity detector (VAD) to determine when there is silence in order to get an accurate noise estimate and the noise is assumed to be short-term stationary so that noise from silent frames can be used to remove noise from speech frames. In order to estimate the clean speech frame a phase estimate is also required, but Wang and Lim have determined that it is sufficient to use the noisy phase spectrum as an estimate of the clean speech phase spectrum.[64] Figure 4.9 shows the signal flow for spectral subtraction where $m(k)$ is a frame of unprocessed noisy data, $k$ is the frame index, $\omega$ is the frequency index, $M(\omega,k)$ is the spectrum of the frame, $N(\omega,k)$ is the spectrum of the noise estimate, $\hat{S}(\omega,k)$ is the spectrum of the speech estimate, and $\hat{s}(k)$ is the speech estimate frame in the time domain.

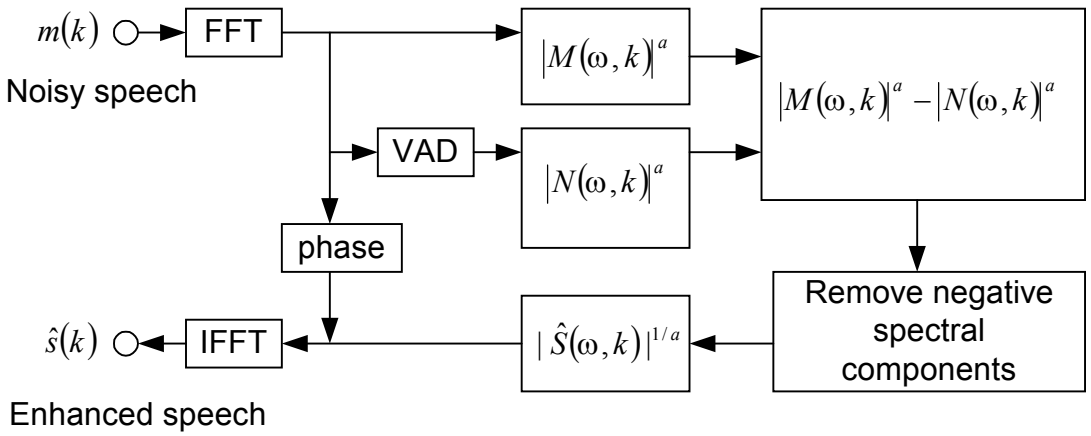


**Figure 4.9: Single Channel Spectral Subtraction**

The most frequently referenced paper on spectral subtraction was written by Boll in 1979 and uses four steps in its spectral subtraction algorithm.[4] First, magnitude averaging across frames is done to reduce spectral errors, which helps avoid the creation of musical tones. Longer averages will further decrease the error, but if too many frames are averaged then the short-term stationary assumption on the speech is no longer valid. The averaging of more than three half overlapped windows with total time duration of 38.4ms will reduce intelligibility. The second step is half-wave rectification to remove the noise, but this also degrades the speech signal and introduces musical noise. The third step is for residual noise reduction, which done by is replacing the current frame value with a minimum value from adjacent frames. This approach is based on the theory that noise will vary more than the speech from frame to frame. Taking the minimum is only used when the magnitude of the speech estimate is less than the maximum noise residual calculated during non-speech activity. If the value is higher than the maximum noise level, then it is considered speech and left alone. Taking the minimum will retain the speech signal below the noise maximum because it varies slowly, but taking the

minimum of varying noise will suppress the noise value. The fourth step is to use additional signal attenuation during non-speech activity. It was found that leaving some noise present in the non-speech frames created a more perceptibly pleasing transition between non-speech and speech frames, with one example of attenuating noise frames by –30dB found to be a reasonable amount.

A major challenge in spectral subtraction is to obtain low variance amplitude spectrum estimates. Variance in the spectral estimate can be reduced using Bartlett's method at the cost of frequency resolution.[31] Since the spectral subtraction technique is frame-based and uses FFT's, frame effects must be considered. An FFT corresponds to a critically sampled filter bank. A circular convolution (which comes from the FFT and IFFT operations) results in discontinuities between frames, but can be avoided using correct lengths of the filter and data frame.[31]

After subtraction the spectral magnitude is not guaranteed to be positive where the possibilities to remove the negative components are by half-wave rectification (setting the negative portions to zero), full wave rectification (absolute value), or weighted difference coefficients. Half-wave rectification is commonly used but introduces "musical" tone artifacts in the processed signal. Full wave rectification avoids the creation of musical tones, but is less effective at reducing noise. Much of the spectral subtraction research has focused on how to remove or reduce the creation of musical tones while maximizing the suppression of noise.[65] [66] The SSPN algorithm prevents the

negative spectral components from accruing by weighting the spectral gain function according the masking threshold and a lower limit of zero.

The noise residual will typically appear as randomly spaced narrow bands of magnitude spikes and have a magnitude between zero and a maximum value measured during non-speech activity. It will sound like the sum of a tone with random fundamental frequencies when it is transformed back to the time domain and during speech activity these tones will be heard where the speech does not mask them. A simple way to reduce the musical tones is to over-subtract the noise estimate from the signal, but this will also eliminate low energy speech information.[67] Low energy unvoiced speech is particularly important to hearing impaired listeners, so it is best to minimize this type of signal degradation.[27],[68] Other methods explored with varying success are critical band analysis[69], sub-frame randomization[70], iterative spectral subtraction[71], post-processing spectral classification[67], non-linear spectrum estimation[72], and minimum mean square error estimation.[73] Critical band analysis with an auditory perceptually weighted spectral subtraction gain function was used in the SSPN algorithm, which was very effective in eliminating the introduction of musical noise artifacts.

### 4.3.2 Noise estimation

The SSPN algorithm uses an estimate of the stationary noise based on the frame classification provided by the VAD. Other noise estimation techniques are mentioned below as a possibility for future research into improving the noise suppression performance of the SSPN algorithm. Noise estimation is an important part of spectral

subtraction for removing the unwanted interference from the signal. The noise sources can be broadband (white) background noise, interfering talkers, or narrow band signals. Stationary noise can be estimated over longer time frames to obtain better accuracy and noise that is non-stationary requires a quickly converging adaptive algorithm or estimation during the current frame. Naturally, the accuracy of the estimated noise spectrum will determine to a great deal how much residual noise is left after processing the signal through algorithms like spectral subtraction. The measure of the noise estimate is also a key to the performance of the voice-activity detector because the speech detection threshold is often based on the noise statistics.

#### 4.3.2.1 Noise reference channel

A noise reference channel was not used in the SSPN algorithm, but represents an interesting possibility for enhancing the algorithm to cancel non-stationary interferences like passing cars. Multiple microphone systems enable the signal processing to perform signal separation to obtain a noise reference channel free of the desired signal where this separation provides a continuous estimate of the noise, so non-stationary noise can be tracked and removed. A typical use of the noise reference signal is for an adaptive noise cancellation system, such as those using the Least Mean-Square (LMS) algorithm as shown in Figure 4.10.[2], [74], [75]

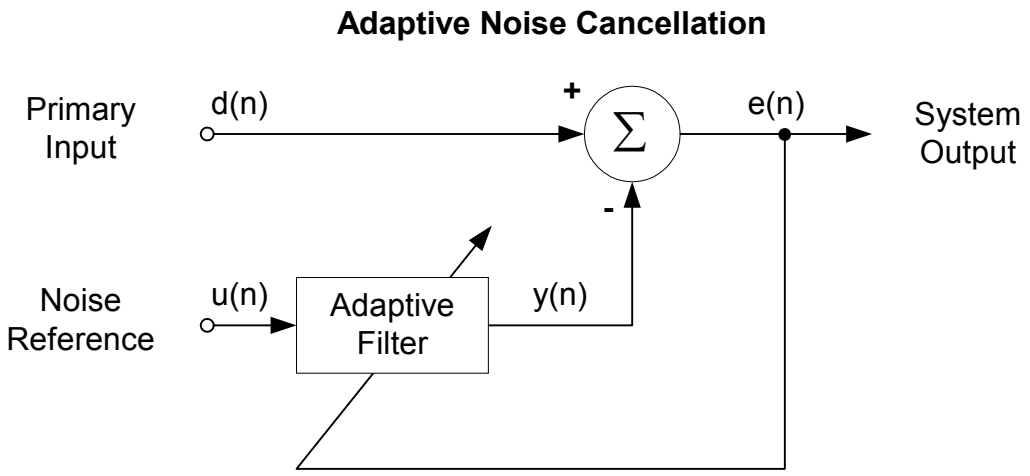


**Figure 4.10: Adaptive Noise Cancellation**

Signal separation can be done statistically, spectrally, spatially, and by estimating the source to sensor transfer functions to de-correlate and the general problem is to determine the coupling between the signals based on a given criteria and then undo it to achieve the signal separation. De-correlation of the received signal to perform channel separation has received some attention in recent literature and yet another area to explore for improved noise suppression algorithms.[76], [77], [78]

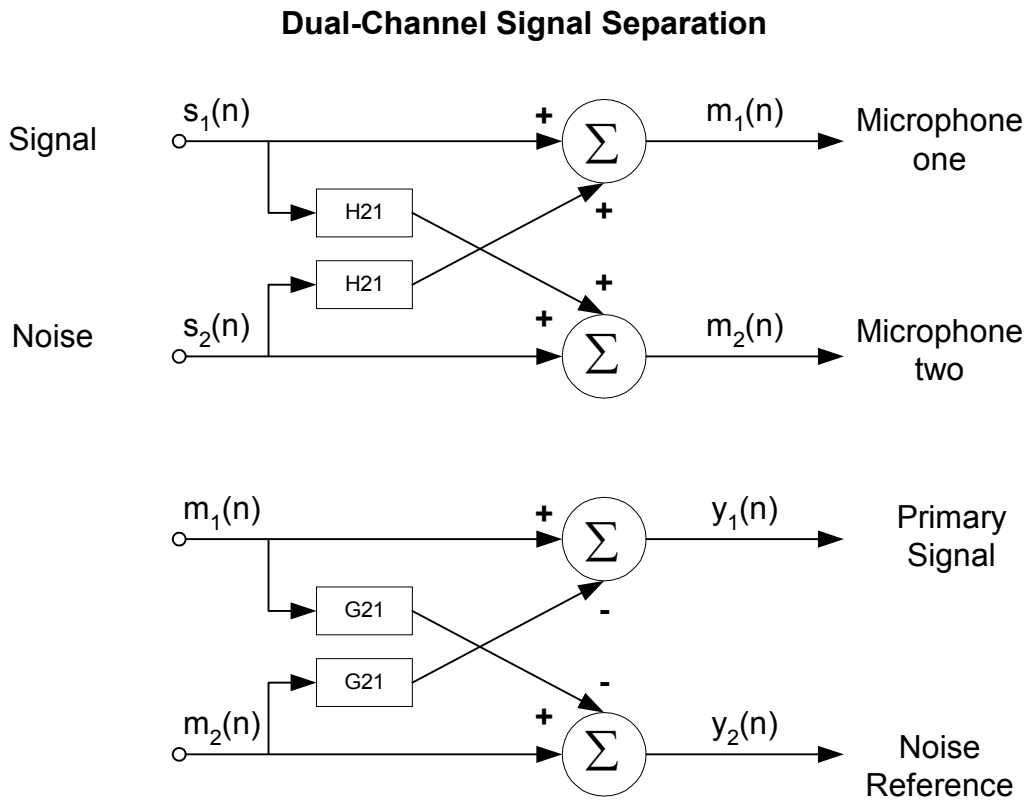


**Figure 4.11: Dual Channel Signal Separation**

Figure 4.11 describes one framework used for separating noise and the desired signal where the observed signals at each microphone are shown in equations (4.54) and (4.55). The coupling between channels is represented by $H_{12}(s_2(n))$ and $H_{21}(s_1(n))$.

$$m_1(n) = s_1(n) + H_{12}(s_2(n)) \quad \textbf{(4.54)}$$

$$m_2(n) = s_2(n) + H_{21}(s_1(n)) \quad \textbf{(4.55)}$$

The goal is to subtract out the coupling effects between channels by using an estimate of the relationships between sources and sensors.

$$y_1(n) = m_1(n) - G_{21}(m_2(n)) \quad \textbf{(4.56)}$$

$$y_2(n) = m_2(n) - G_{12}(m_1(n)) \quad \textbf{(4.57)}$$

Assuming the source signals are statistically independent has been used as the criteria for signal separation. The decoupling filters, $G_{12}$ and $G_{21}$, are adjusted so that the reconstructed signals $y_1(n)$ and $y_2(n)$ are statistically independent, which should approximately represent the original source signals $s_1(n)$ and $s_2(n)$, then higher order statistics are used to measure the degree of independence achieved.

Single microphone implementations simply place the microphone as close to the desired source as possible and far away from the noise source. Multi-microphone systems offer a lot of flexibility for placement and potential for improving the noise estimation accuracy. The microphone designated for the noise reference should be far away from the desired source for good channel separation, but not so far that the noise at one microphone is not correlated with the noise at the microphone for the desired signal. If the noise and speech channel are not well separated, then a portion of desired signal will be cancelled. If the noise is not correlated well between the microphones, then noise statistics will not be well matched and the signal will either be degraded or more noise will remain. These two contradicting goals must be balanced and are very application dependent.

The SSPN algorithm can be used independently in multiple seat locations in the car, but could possibly be improved by taking advantage of the expanded set of microphones and additional spatial diversity. It has been shown that there is some correlation of noise between the area facing the driver's seat and other locations in the car such as the front passenger seat.[79] This situation allows the placement of microphones in front of both the driver and front seat passenger. When the driver is speaking the passenger

microphone can be used as the noise reference and when the passenger is speaking the driver side microphone can be used as the noise reference. Simply placing the microphones in an optimal location is not enough because the correlation of the noise between the microphones will not always be high because if one window is down in the car and the other is not for example, then the noise will vary a lot based on location. The correlation of the noise between the microphones is also frequency dependent, where experiments have shown that low frequency components of the noise are highly correlated, and that the correlation decreases gradually as the frequency is increased until it vanishes for frequencies higher that about 2 kHz.[79]

#### 4.3.2.2 Noise estimation during silence

The main advantages of noise power estimation during non-speech frames in conjunction with a VAD are the low complexity, low computational costs, and implementation with a single microphone. The disadvantages are that non-stationary noise cannot be tracked and the VAD performance degrades significantly at lower SNR. The long-term average spectral mean of the noise is calculated to estimate the stationary noise and the new noise estimation data is introduced slowly to the current estimate to avoid sharp frame-to-frame changes that would worsen the musical noise artifacts present in some spectral subtraction algorithms.[80]

A sub-band implementation could eventually be used for the SSPN algorithm to improve the noise estimation and computational efficiency. Noise estimation during non-speech has also been used for sub-band processing algorithms where low frequency bands have

high coherence, so they can use a strategy that adapts during silent frames. In some of the higher frequency bands the speech information has higher coherence than the noise and can be continually adapted.[81] Another example that the SSPN algorithm could consider for noise estimation is a microphone array that makes the speech vs. non-speech decision for each microphone, so only the microphones with significant noise content update the overall estimate of the noise.[82]

#### 4.3.2.3 Continuous noise estimation

Continuous noise estimation could expand the SSPN algorithm's ability to suppress severe and non-stationary noise sources. More sophisticated techniques are required when the signal to noise ratio, SNR, is low or the noise is highly non-stationary.[83] If the noise estimate is too low, then there will be excess residual noise passed through the system, but if the noise estimate is too high, then speech sounds will be degraded and perhaps be worse than if the noise was still present. Current research is aimed at incorporating soft-decision schemes, which are also capable of updating the noise Power Spectral Density (PSD) during speech activity. Some of the successful methods will be described below, which could be used with the SSPN algorithm proposed in this thesis.

1. *Multi-channel signal separation*
2. *Minimum statistics*
3. *Energy clustering*
4. *Weighted average of past sub-band spectral magnitudes*

*Multi-channel signal separation* to obtain a noise reference channel is a common technique to get an accurate noise estimate during speech. This method is described in section 4.3.2.1. Another effective channel separation technique is to use a two-channel

beamformer where a blocking matrix filters out the desired speech leaving the rest of the signal as the noise estimate. This use of a beamformer is more effective than the traditional application because it works well even when the number of sources exceed the number of microphones.[84] Microphone placement is also of great concern when using arrays of microphones to perform one of the many beamforming algorithms. The array needs to be able to locate and track signal and noise sources. The array should be placed close to the desired source as in the single microphone case. However, the spacing between microphones has an effect on the frequency resolution of the array and thus its ability to enhance the signal.[85]

*Minimum statistics* noise estimation is based on the observation that even during speech activity a short-term power spectral density estimate of the noisy signal frequency decays to values that are representative of the noise power level. There is a fundamental assumption that during speech pause or within brief periods in between words and syllables the speech energy is close to zero. Thus, the noise floor can be estimated by tracking the minimum power within a finite window large enough to bridge high power speech segments. The low energy envelope of the signal is tracked within frequency bands. Some of the challenges to this technique are:

- Calculating an optimal smoothing of the estimate
- Accounting for a bias towards lower values
- Delays in tracking during periods of increasing power

Listening tests show that this approach outperforms a VAD plus soft-decision updating during speech activity. The minimum statistics algorithms also preserved weak voiced

sounds such as the consonants /*m*/ and /*n*/ and had dramatic improvements when the input signal was music.[86], [87]

*Energy clustering* is based on the analysis of histograms of energy values within different frequency bands. Each band can be assumed to have two modes. A low energy mode related to speech pauses and a high energy mode related to the speech. The energy distribution in a band is analyzed using this two-mode approach. Either a two-centroid clustering algorithm or a fit to Gaussian probability density function is used to detect the mode for each band. This analysis can be done during speech and can also determine the SNR within a given band. One of the challenges of this method is that the two modes tend to merge when low SNR conditions are present. This blending of the modes causes an underestimate of the noise level that must be compensated.[88]

*Weighted average of past sub-band spectral magnitudes* and *analysis of their histograms* are two more methods for noise estimation without a VAD. The *first* method calculates the weighted sum of past spectral magnitudes $X_i$ in each sub-band *i*. The weighting is done by a first order recursive system, $\hat{N}_i(k) = (1-\alpha)\cdot X_i(k) + \alpha \cdot \hat{N}_i(k-1)$, where $X_i(k)$ denotes the spectral magnitude at time *k* in sub-band *i* and $\hat{N}_i(k)$ is an estimation of the noise magnitude. Higher values occur at the onset of speech, so a threshold $\beta \cdot \hat{N}_i(k-1)$ is introduced where $\beta$ takes on values between 1.5 and 2.5. When the spectral component $X_i(k)$ exceeds the threshold this is considered as an approximate detection of speech and the recursive algorithm is stopped. The *second* approach looks at

the histograms of the spectral values in each sub-band. Values below the same threshold of the first method are classified as noise. Past values of noise segments totaling 400 ms are evaluated to determine the distribution in 40 frequency bins. The noise level is estimated as the maximum of the distribution in each sub-band. The estimated noise values are smoothed to remove any spikes. The second method provides a more accurate noise estimate than the first, but requires more computation.[83]

## 4.4 Perceptual nonlinear frequency weighting

The SSPN algorithm uses perceptual nonlinear weighting of the gain function used for spectral subtraction, which enables it to aggressively attenuate the noise while avoiding the introduction of annoying artifacts to the speech signal. SNR, signal to noise ratio, is the most broadly used criteria for reducing noise in a received speech signal and has been very successful, but it is limited because inaccuracy of the noise estimate can cause either excess residual noise or distortion of the signal. Taking advantage of the human auditory system's characteristics can help mitigate the effects of residual noise and render the speech to be more perceptually pleasing to the ear because the distortion of the signal is minimized by not processing noise that is effectively inaudible. The human auditory system performs some form of frequency signal analysis and reconstruction when listening to a signal present in noise, so enhancement algorithms can follow a similar process. The short-time spectral amplitude, STSA, enhancement methods can take advantage of how people perceive the frequencies instead of just working with SNR. There has been considerable work done in the area of perceptual masking during the past decade and some examples can be found in these references.[11], [89], [90], [91], [92], [93], [94], [95],

[96], [97] A description of the psychoacoustics behind the perceptual masking threshold is described in Appendix A.

This section describes how the SSPN algorithm calculates the masking threshold for purposes of weighting the spectral subtraction gain function. Perceptual speech enhancement techniques have the problem that there is no clean speech reference or accurate spectral noise estimate in order to determine exact auditory masking thresholds. If the clean-speech masking threshold is too high then more noise will be left in the signal, but if the clean-speech masking threshold is calculated too low, then information about the desired signal will be lost. Spectral subtraction is commonly used to obtain an estimate of the clean speech from which the masking thresholds are calculated and what is used in the SSPN algorithm.

A similar approach has been proposed which calculates the masking threshold of the noise and noisy speech, compares the two, and then performs a subtraction to obtain an estimate of the clean speech threshold. The problem introduced by the distortion of spectral subtraction can be avoided by performing the threshold calculation *before* the subtraction is done.[89] This approach was not attempted in this thesis due the added complexity that would be introduced to the algorithm.

The steps required to calculate the masking threshold are taken directly from the paper by Johnson [98] and are shown in Figure 4.12 with the data flow of the mask threshold algorithm shown in Figure 4.13. The diagrams and equations are an accurate

representation of the algorithm used in the SSPN processing to obtain the results in Chapter 5.

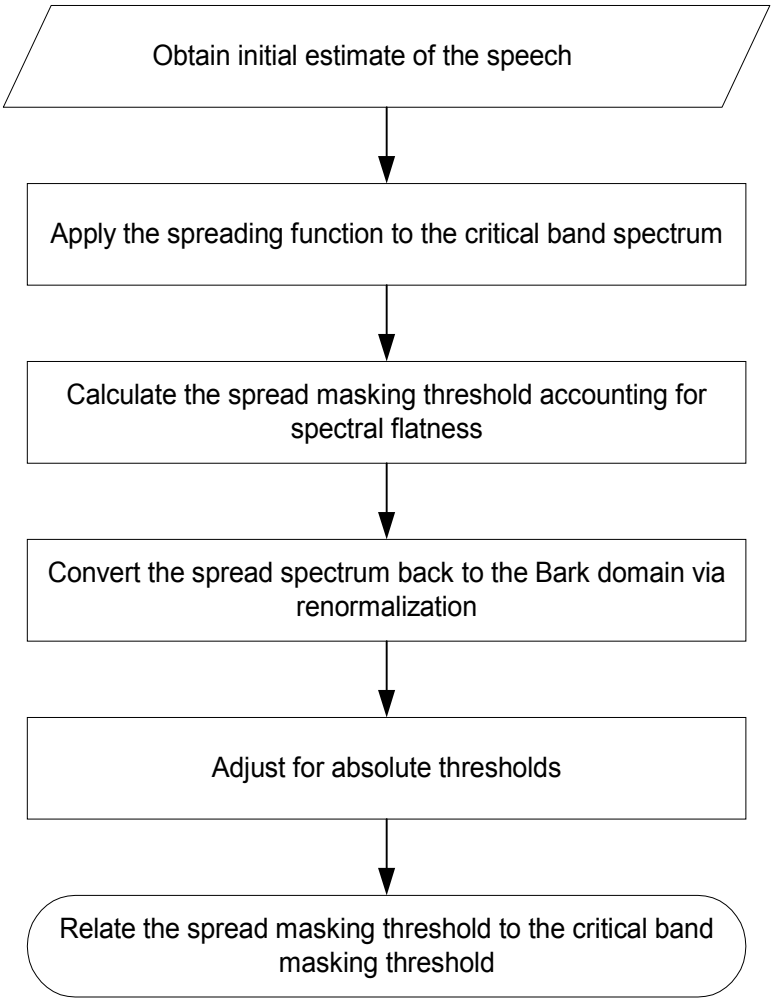


**Figure 4.12: Steps for mask threshold calculation**

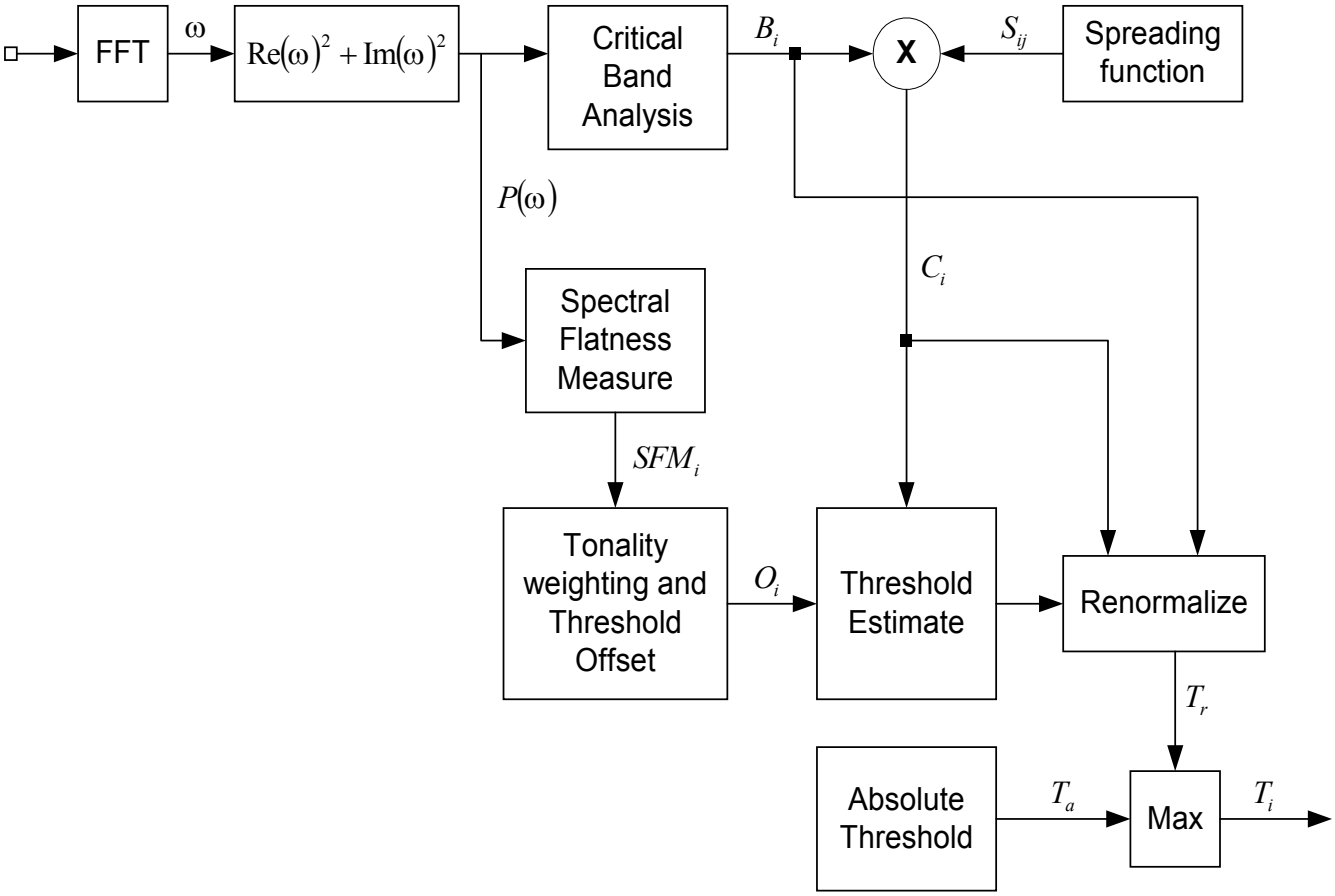


**Figure 4.13: Masking Threshold Calculation**

Critical Band Analysis partitions the power spectrum into critical bands according to Table 2.2 presented in the section 2.2. The power spectrum is calculated from the frequency data as in equation (4.58).

$$P(\omega) = \mathrm{Re}^2(\omega) + \mathrm{Im}^2(\omega) \qquad \textbf{(4.58)}$$

The energy in each critical band is summed in equation (4.59) where $Bi$ is the energy for critical band $i$, $bli$ is the lower frequency for the band, and $bhi$ is the upper frequency for the band.

$$Bi = \sum_{w=bli}^{bhi} P(\omega) \tag{4.59}$$

The number of critical bands used will depend on the bandwidth of the signal in question. Humans can only perceive frequencies between 20 Hz and 20kHz, so that places a bound on the range of frequencies to consider. There would be 22 critical bands for an 8kHz signal that is sampled at the Nyquist rate of 16kHz.

The spreading function is used to estimate the effects of masking across critical bands. The spreading function is calculated as $|(j-i)| <= 25$, where $i$ is the Bark frequency of the masked signal and $j$ is the Bark frequency of the masking signal. The term Bark is used to indicate the frequencies of one critical band as defined in Table 2.2. The spreading function is put into matrix form, $S_{ij}$, and convolved with critical band energies $B_i$. The spread critical band spectrum, $C_i$, is given in equation (4.60) where $*$ is the convolution operator.

$$C_i = S_{ij} * B_i \tag{4.60}$$

There are different masking thresholds based on spectral flatness of the signals.

1. Tone masking noise is estimated as (14.5 + $i$) dB below $C_i$,

   where $i$ is the Bark frequency.

2. Noise masking a tone is estimated as 5.5 dB below $C_i$ uniformly across the critical band.

Spectral Flatness Measure, SFM, is defined in equation (4.63) as the ratio of the geometric mean, $G_m$, of the power spectrum to the arithmetic mean, $A_m$, of the power spectrum. Arithmetic mean is given in equation (4.61) and geometric mean is given in equation (4.62).

$$A_m = \frac{P(\omega_1) + P(\omega_2) + P(\omega_3) + ... + P(\omega_n)}{n} \quad \textbf{(4.61)}$$

$$G_m = \sqrt[n]{P(\omega_1) \cdot P(\omega_2) \cdot P(\omega_3) \cdot ... \cdot P(\omega_n)} \quad \textbf{(4.62)}$$

$$SFM_{dB} = 10 * \log_{10} \frac{G_m}{A_m} \quad \textbf{(4.63)}$$

The coefficient of tonality in equation (4.64), $\alpha$, is calculated where an SFM = $SFM_{dbmax}$ = -60 dB indicates the signal is very tone-like and an SFM = 0 indicates the signal is more noise-like. For example an SFM = -30dB would result in $\alpha = 0.5$.

$$\alpha = \min\left( \frac{SFM_{dB}}{SFM_{dB\max}}, 1 \right) \quad \textbf{(4.64)}$$

The offset in equation (4.65), $O_i$, for the masking energy in each band, is determined by using the tonality to weight the masking thresholds for tones and noise.

$$O_i dB = \alpha * (14.5 + i) + (1 - \alpha) * 5.5 \quad \textbf{(4.65)}$$

The spread threshold estimate is then calculated using equation (4.66).

$$T_i = 10^{\log_{10}(C_i) - \left(\frac{O_i}{10}\right)} \tag{4.66}$$

The spreading convolution must now be undone and the threshold converted back to the Bark domain. De-convolution is unstable due to the shape of the spreading function and would introduce undesired artifacts into the signal, so renormalization is used instead to remove the increased energy added to each band by the spreading function. Renormalization multiplies each $T_i$ by the inverse of the energy gain, assuming a uniform energy of 1 in each band.

Critical band noise thresholds that are lower than the absolute threshold of hearing are changed to equal the mean of the absolute threshold of hearing for that band, so it does not make sense to calculate a mask threshold for something that cannot be heard anyway. The absolute threshold of hearing has been measured with several experiments and is given as an estimated curve plotted versus frequency in Figure 4.14.[24], [99]

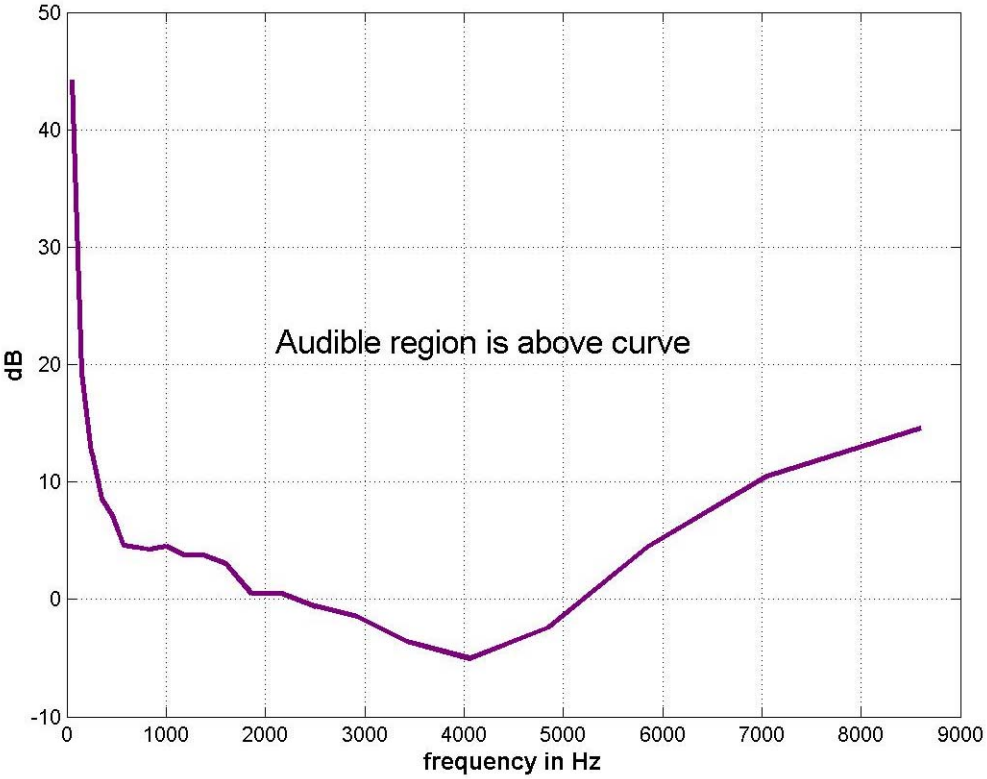


**Figure 4.14: Absolute threshold of hearing in a free field [24], [100]**

## 4.5 Talker isolation via pitch tracking

The SSPN algorithm proposes the use of talker isolation after initial noise removal is done by the GSC beamforming and weighted spectral subtraction, which, in theory, should allow the talker isolation to perform better than if were operating on the original noisy signal. Experiments using pitch detection, described in section 3.3.7 and reported in section 5.10, were done with the SSPN algorithm to gain insight on how a talker isolation algorithm based on pitch tracking might perform using the noise reduced signal compared to the noisy signal. A full pitch tracking and talker isolation algorithm was not

implemented because the complexity of such an algorithm would have expanded the scope of this thesis beyond the initial goals of designing the higher-level framework. Integrating a talker isolation program into the SSPN algorithm is the next logical step of continued research on this algorithm and some specifics of a potential algorithm are discussed in the follow paragraphs.

A typical goal of signal separation is to isolate the desired speech from other speech sources and noise. Separating the sources is crucial since the desired and undesired speech have similar spectra and comparable amplitudes. The challenge of extracting a single talker from a signal with multiple talkers has been referred to as the "cocktail party effect" where the name comes from the amazing ability of the human auditory system to focus on an individual talking in a crowded noisy room by using binaural cues to spatially focus on the desired speech.

An excellent example using a combination of techniques for advanced pitch tracking and talker isolation is the work by Luo and Denbigh.[101] They use frequency and amplitude continuity to track the desired talker. Binaural spatial cues are used to discriminate pitch frequencies that are too close to resolve spectrally. Room reverberation can severely degrade the performance of speech enhancement algorithms and reported results often neglect this important measure. The multi-path effects of an enclosure can introduce false peaks and can nullify or split genuine spectral peaks of the speech. These multi-path effects make it much more difficult to perform pitch tracking when multiple signals have similar frequency content. Multi-path reverberation also adversely effects signal

separation based purely on directional information because of the large variations across the frequency bands of the inter-aural time difference. Their results show an average 40% increase in intelligibility for low SNR speech of –6dB and –12dB. The algorithm by Luo and Denbigh is presented in Figure 4.15.

The algorithm described in Figure 4.15 does not separate the signal above 3kHz based on the assumption there is little power in the speech at those frequencies. Also if two speech sources have simultaneous unvoiced frames, then they must be separated based only on inter-aural time differences because they lack the required periodicity for pitch tracking.

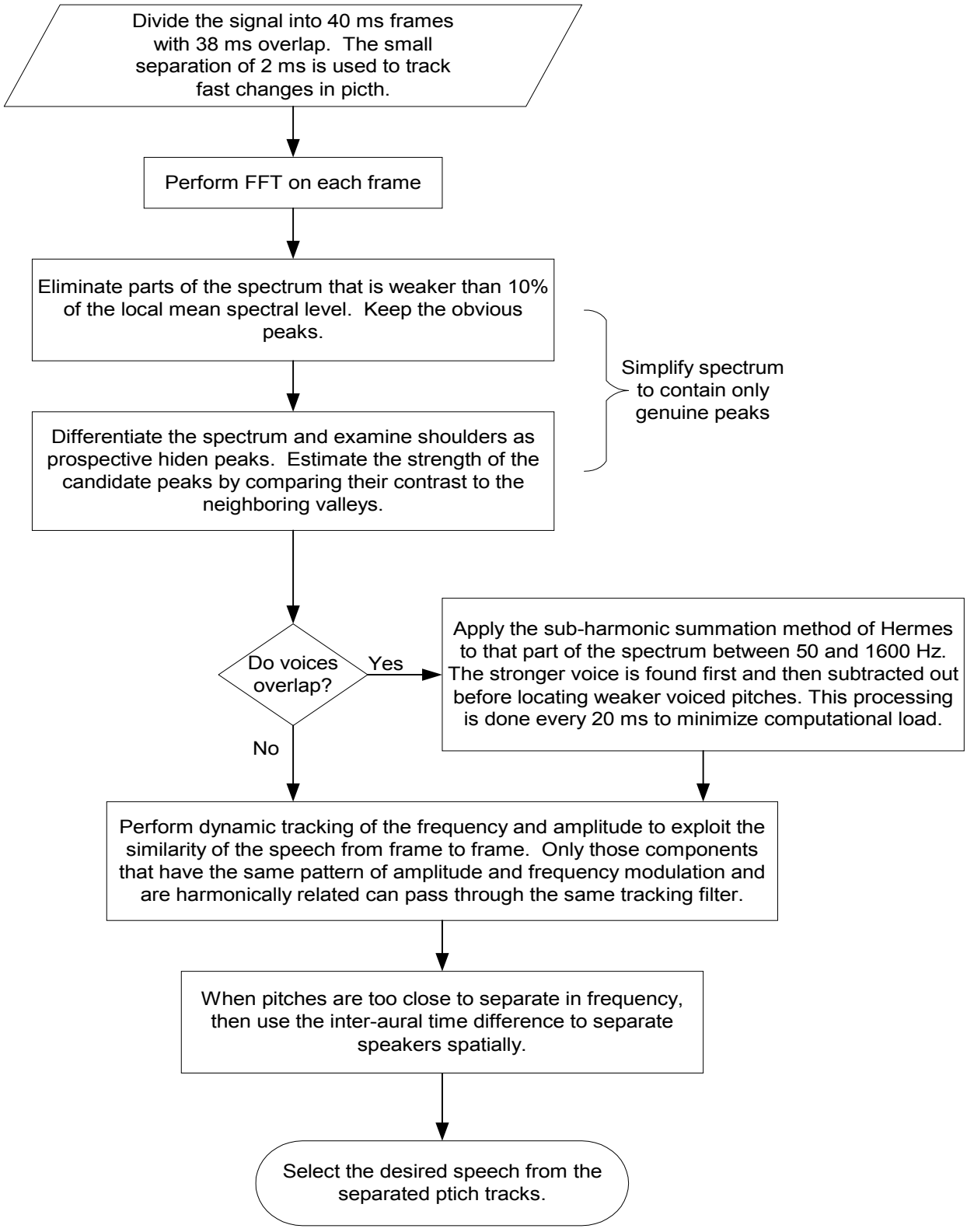


**Figure 4.15: Talker separation algorithm by Luo and Denbigh**

# Chapter 5

# Results

Evaluating the SSPN algorithm using real data collected in an automobile was an important consideration in this thesis, so the performance could be characterized beyond what could be learned from computer generated white noise. The eventual deployment of the SSPN algorithm in a hands-free application in the automobile will have a higher chance of success because research is based on data from that environment. Ideally, extensive subjective listening tests would have been performed on the results, but that is a time consuming and costly process, so only informal listening was done to verify objective quality measures. The objective speech quality measures provided a consistent method for comparing results and evaluating different initial SNR conditions.

Section 5.1 provides the details of the real-world data that was used in order to test the algorithms' performance for the intended application of a hands-free phone in an automobile. MATLAB© simulations and the spectrum of the input signals are explained in section 5.2. The objective speech quality measures that were used are described in section 5.3. Results for the proposed SSPN algorithm, beamforming and spectral subtraction, are reported in sections 5.4, 5.5, and 5.6 respectively and compared in

section 5.8. Because the VAD plays such an important role, it's accuracy was closely analyzed in section 5.9. Pitch detection results are reported in section 5.10. Statistical analysis of SSNR results using all the data in Table 5.2 is presented in section 5.11. The results reported in this chapter support the expected theoretical strengths and weaknesses of the algorithms as described in Chapter 3 and Chapter 4. Some ideas were proven not to work and insights were gained with respect to the quality of the speech and performance of the VAD.

## 5.1 Measurements

The measurements were made in a 2001 Honda Odyssey minivan where Figure 5.1 shows the microphone setup in the van. A uniform linear array of 4 Larson-Davis BNK omni-directional microphones was mounted between the visor and ceiling slightly above and in front of the driver. The center of the array was about 38cm from the talker's mouth and the spacing between the microphones was 5cm to allow for the beamforming to have good spatial resolution up to 3420Hz and a total aperture of 15cm. A position close to where the microphones might be permanently installed in an automobile is protruding from beneath the visor toward the windshield, so this position was used for the recordings.

**Figure 5.1: Microphone setup in van**

Calibration between the microphones is important for the subsequent beamforming to be effective, so this was done to ensure they had equal gains. The microphones were attached to Larson-Davis 2200C pre-amplifiers with Larson-Davis 5-pin EXC010 microphone cables and gains on the pre-amps were set to 10dB at all times. The pre-amps then fed their outputs through a BNC to ¼ inch jack connector cable to the line inputs of a 4-track TASCAM PortaStudio-424-III analog taper recorder where the trim level on the line inputs to the recorder were set to the top position for the clean speech recordings and lowered two notches for the noise recordings in order to avoid clipping.

14.4 volts dc was supplied from the van to a 1750 Watt PortaWattz DC to AC inverter by StatPower Technology Corp. that was plugged into the power socket of the car to supply power for the recording equipment and a 6 amp fuse was in the socket plug. Figure 5.2 describes the layout of analog recording components, which were borrowed from Bose Corporation.

Later, the 4-track analog tape recordings were then played simultaneously into a Midiman Delta-10-10 sound card in a PC that digitized the signal at a 16kHz sampling rate with 16-bits of resolution. All recording clips were digitized to be of the same time duration to make subsequent digital mixing easier and the components used to digitize the recordings are shown in Figure 5.3

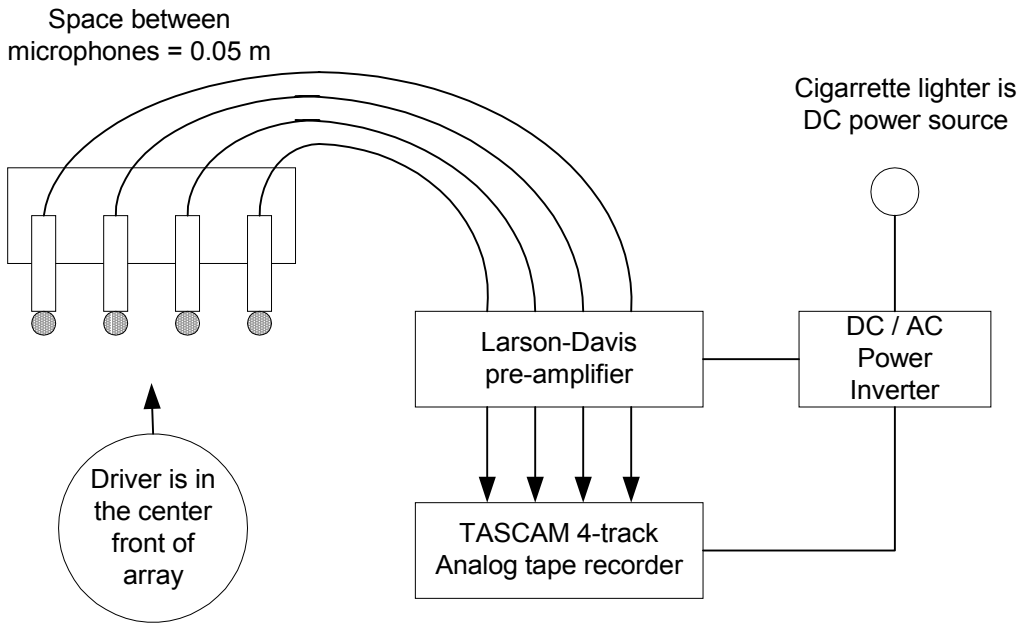


**Figure 5.2: Analog recording setup**

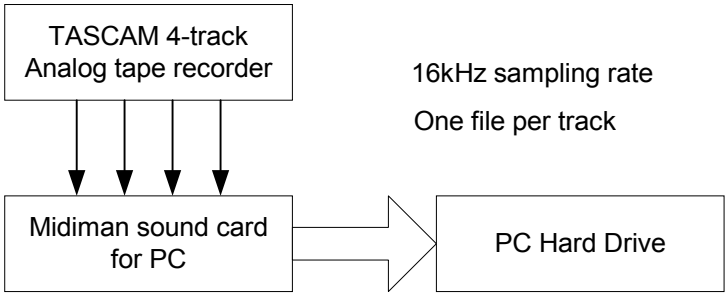


**Figure 5.3: Digitizing analog data**

Single close talking headset microphone recordings of the clean speech were made, so it would be possible to compare the results of microphone array processing to alternative solutions. Ideally, noise recordings with the single microphone would have been done also, and these may be done at a later date. Headsets achieves good speech quality simply because it is close to the source, but present an inconvenience to the user. The headset microphone was an Optimus 33-3012 manufactured in the Philippines and was connected to an Optimus 170 Mhz FM wireless transmitter where an Optimus wireless receiver had its volume set to 5/10 and was connected through a mono RCA to ¼ inch cable to the TASCAM recorders line 1 input. A single microphone was 1½ inches from the talker's mouth.

Adult male and female voices were recorded, using the array of microphones and the single headset microphone, in a quiet parked car with the windows up and down. The different speech recordings are summarized in Table 5.1. On a quiet clear evening, the recordings were made using the phrase, "We are getting off exit 12 near Bose, so we will be there in ten minutes. Turn up the radio. This is my favorite song."

| Speech measurements | Adult | Window positions | Single headset microphone | Array of microphones | Tape Locations |
|---|---|---|---|---|---|
| FUA | Female | Up | | X | 20-45 |
| FDA | Female | Down | | X | 45-66 |
| MDA | Male | Down | | X | 66-94 |
| MUA | Male | Up | | X | 94-121 |
| FUS | Female | Up | X | | 121-150 |
| FDS | Female | Down | X | | 150-180 |
| MDS | Male | Down | X | | 180-206 |
| MUS | Male | Up | X | | 206-235 |

**Table 5.1: Clean speech measurements**

The noise measurements were made separately without the driver speaking and are summarized in Table 5.2. The noise was recorded on a day when a steady light rain fell, so there is noticeable splashing in the street and windshield wiper sounds evident in some of the recordings. The rain can also be heard hitting the windshield and roof, which was significant because the microphones were located on the ceiling near the windshield.

| Noise ID | Location | Speed | Window positions | Radio | Fan | Other talkers | Wipers | Tape Position |
|---|---|---|---|---|---|---|---|---|
| Q0UIW | Quiet road | 0 | Up | Off | Off | Yes | On | 235-285 |
| Q0DIW | Quiet road | 0 | Down | Off | Off | Yes | On | 285-335 |
| H55UIW | Highway | 55 | Up | Off | Off | Yes | On | 335-445 |
| H55UW | Highway | 55 | Up | Off | Off | No | On | 445-495 |
| H55DW | Highway | 55 | Down | Off | Off | No | On | 495-545 |
| H55UFW | Highway | 55 | Up | Off | High | No | On | 545-595 |
| H55U | Highway | 55 | Up | Off | Off | No | Off | 595-645 |
| H55D | Highway | 55 | Down | Off | Off | No | Off | 645-695 |
| H55UF | Highway | 55 | Up | Off | High | No | Off | 695-745 |
| H55UI | Highway | 55 | Up | Off | Off | Yes | Off | 745-795 |
| EOUI | Engine off | 0 | Up | Off | Off | Yes | Off | 795-895 |
| EODI | Engine off | 0 | Down | Off | Off | Yes | Off | 895-955 |
| Q0UR | Quiet road | 0 | Up | On | Off | No | Off | 955-1037 |
| Q0UF | Quiet road | 0 | Up | Off | On | No | Off | 1037-1137 |
| Q0DR | Quiet road | 0 | Down | On | Off | No | Off | 1137-1188 |
| Q30U | Quiet road | 30 | Up | Off | Off | No | Off | 1188-1238 |
| Q30D | Quiet road | 30 | Down | Off | Off | No | Off | 1238-1288 |
| D0U | Downtown | 0 | Up | Off | Off | No | Off | 1288-1355 |
| D0D | Downtown | 0 | Down | Off | Off | No | Off | 1355-1410 |
| D20UW | Downtown | 20 | Up | Off | Off | No | On | 1410-1464 |
| D20DW | Downtown | 20 | Down | Off | Off | No | On | 1464-1562 |

**Table 5.2: Car noise measurements**

The clean speech was then added to the different types of noise that were measured to form the input signals to the enhancement algorithms. Adding the noise and speech after

measuring provided a consistent "clean" speech reference to compare against the improvements made to the different noise corrupted signals. Known SNR levels can be used as inputs to the simulations because of the separate noise and speech recordings. Only a limited amount of simulation data is presented in the following sections because it is sufficient to demonstrate the performance of the algorithms and results did not differ considerably using the other data.

## 5.2 Simulation

The simulations were done in MATLAB© using real data recorded in the car sampled at 16 kHz and down-sampled to 8kHz to reduce the simulation time. The phrase, "turn up the radio", said by a woman was used in all the tests; it had a duration of 2.5 seconds or 20,000 samples. The different noise types were also from recordings in the car except for the additive white Gaussian noise (AWGN), which was computer generated. The signals and noise were mixed on the computer and the SNR of the signals was adjusted to the known values of 0, 5, and 10 dB before running the enhancement algorithms to simplify analysis of the results.

Figure 5.4 shows the Power Spectral Density of the signals where a FFT was used on the entire 2.5 seconds length of the signal sampled at 8 kHz. The frequencies below 1 kHz contain most the energy for the signals involved and most of the speech energy is in the lower frequencies with peaks around the pitch of the desired talker just below 200 Hz. About 90% of the road noise is contained below 120 Hz. The fan noise has significantly more energy around 200 Hz, which caused more problems than the road noise when

mixed with the speech. The interfering talker noise has strong harmonics at 300 and 600 Hz, which corresponds well to the expected pitch of the children talking. The Additive White Gaussian Noise (AWGN) energy is fairly well distributed across the spectrum by definition and causes more problems at low SNR than colored noise because it will be more likely to mask the speech in a given spectral band. The insights gained from Figure 5.4 agree with the results calculated after processing the signal combined with different noise types.

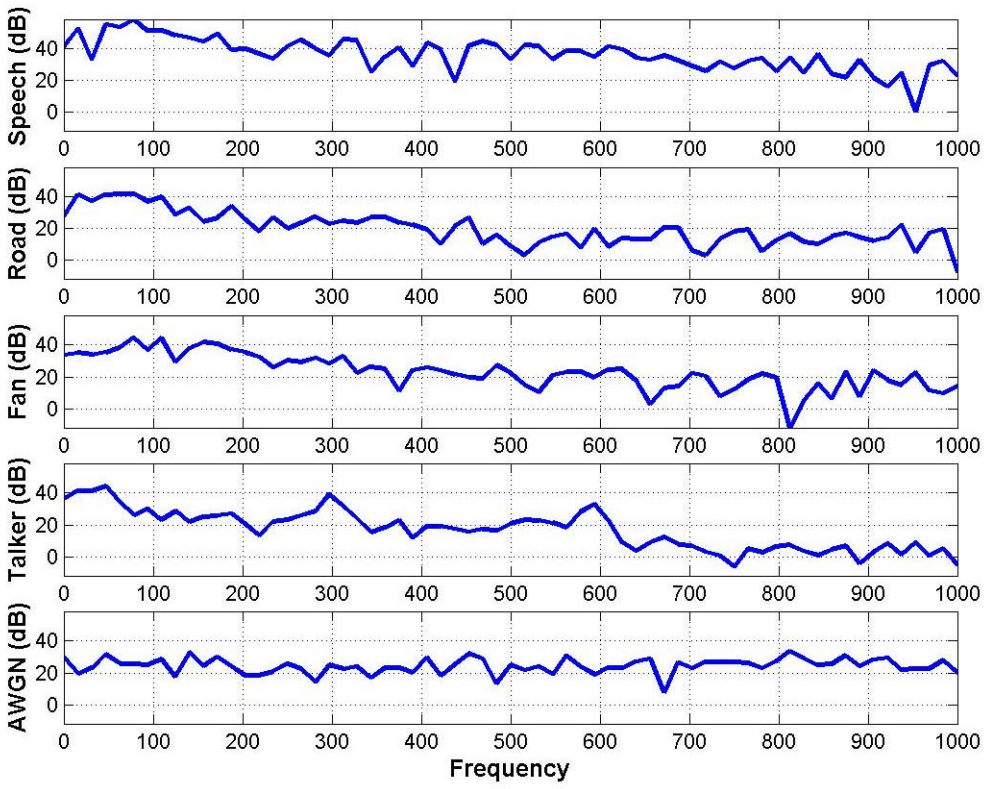


**Figure 5.4: PSD of speech and noise signals**

The speech spectrum and noise spectrum are compared more closely in Figure 5.5, Figure 5.6, Figure 5.7, and Figure 5.8.

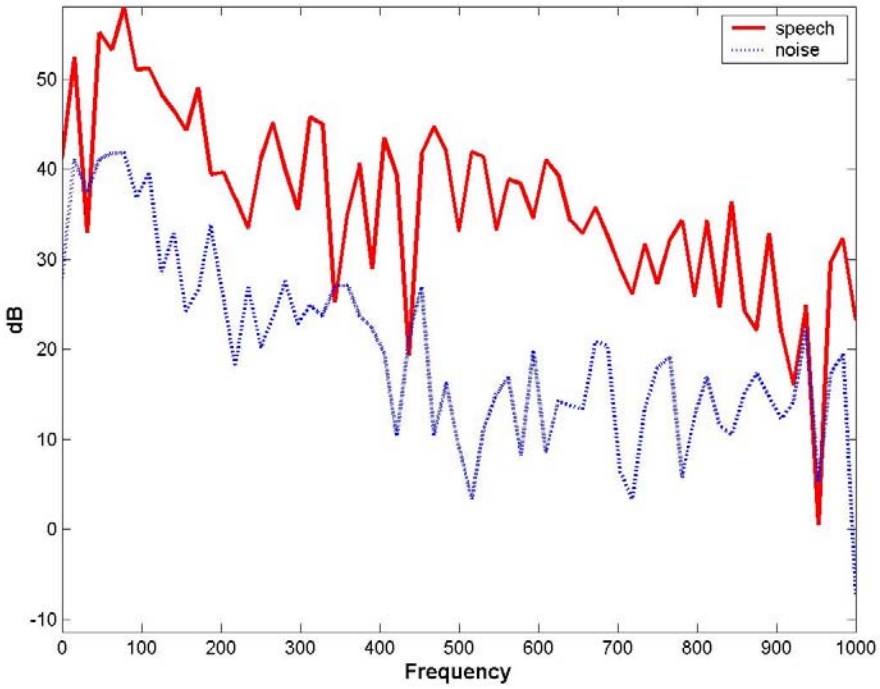


**Figure 5.5: Speech and road noise spectrums**

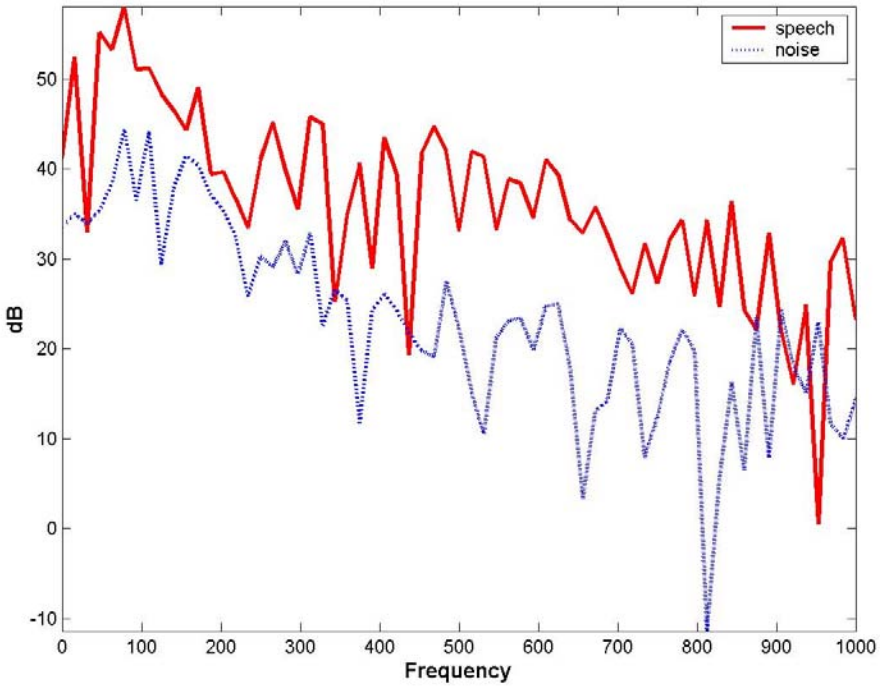


**Figure 5.6: Speech and fan noise spectrums**

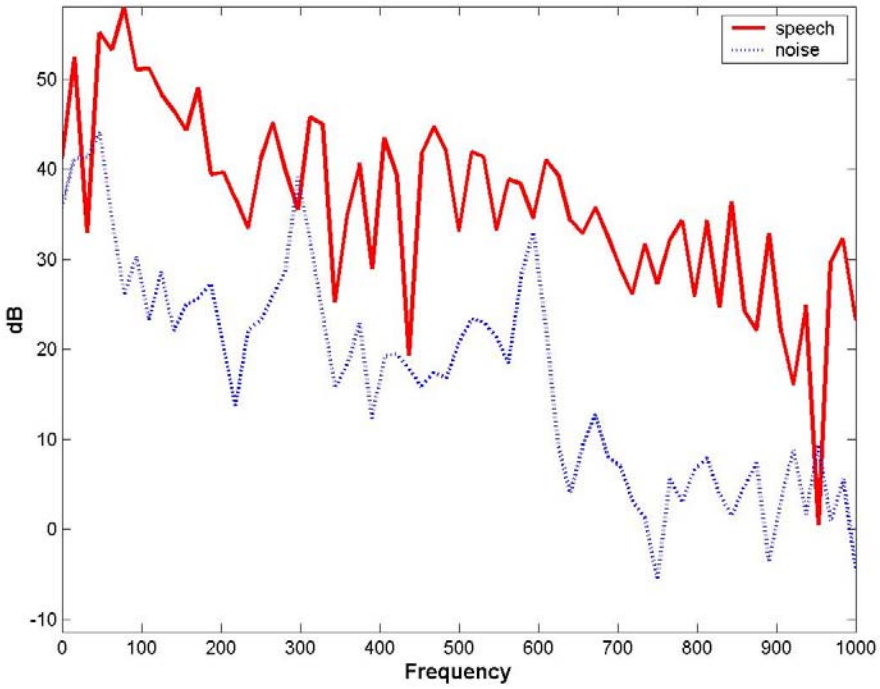


**Figure 5.7: Speech and talker noise spectrums**

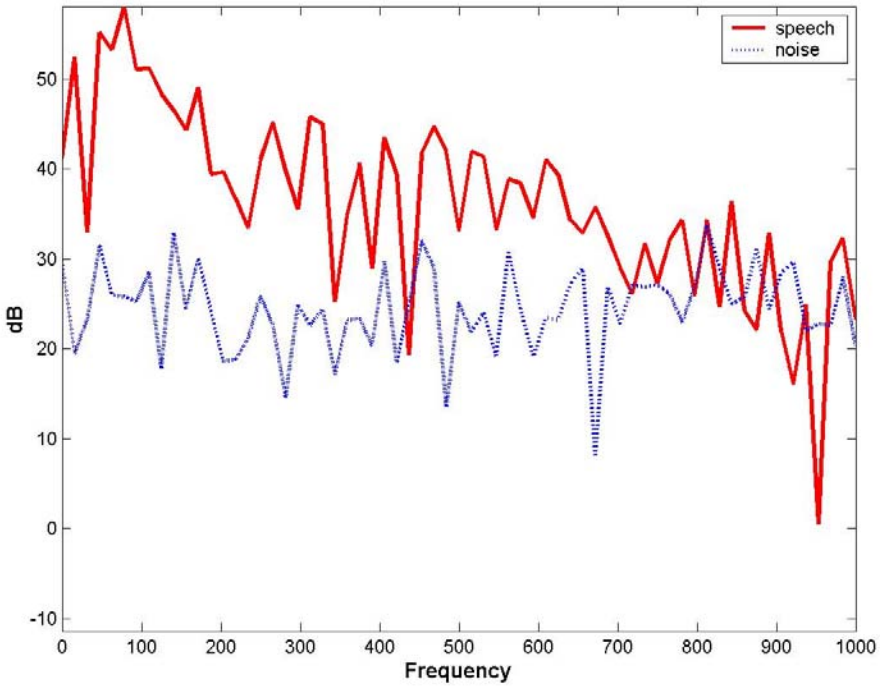


**Figure 5.8: Speech and awgn noise spectrums**

The voice activity detector (VAD) and frames-size was kept constant through all the simulation runs because they have such a large impact on the results. Using the same VAD enables fair comparison between regular spectral subtraction and the enhanced algorithm. MATLAB© M-files were also used to calculate the objective speech quality measures and create the plots to visualize the comparison of results, which allowed for the simulation flow shown in Figure 5.9 to be done entirely in MATLAB© for each noise type.

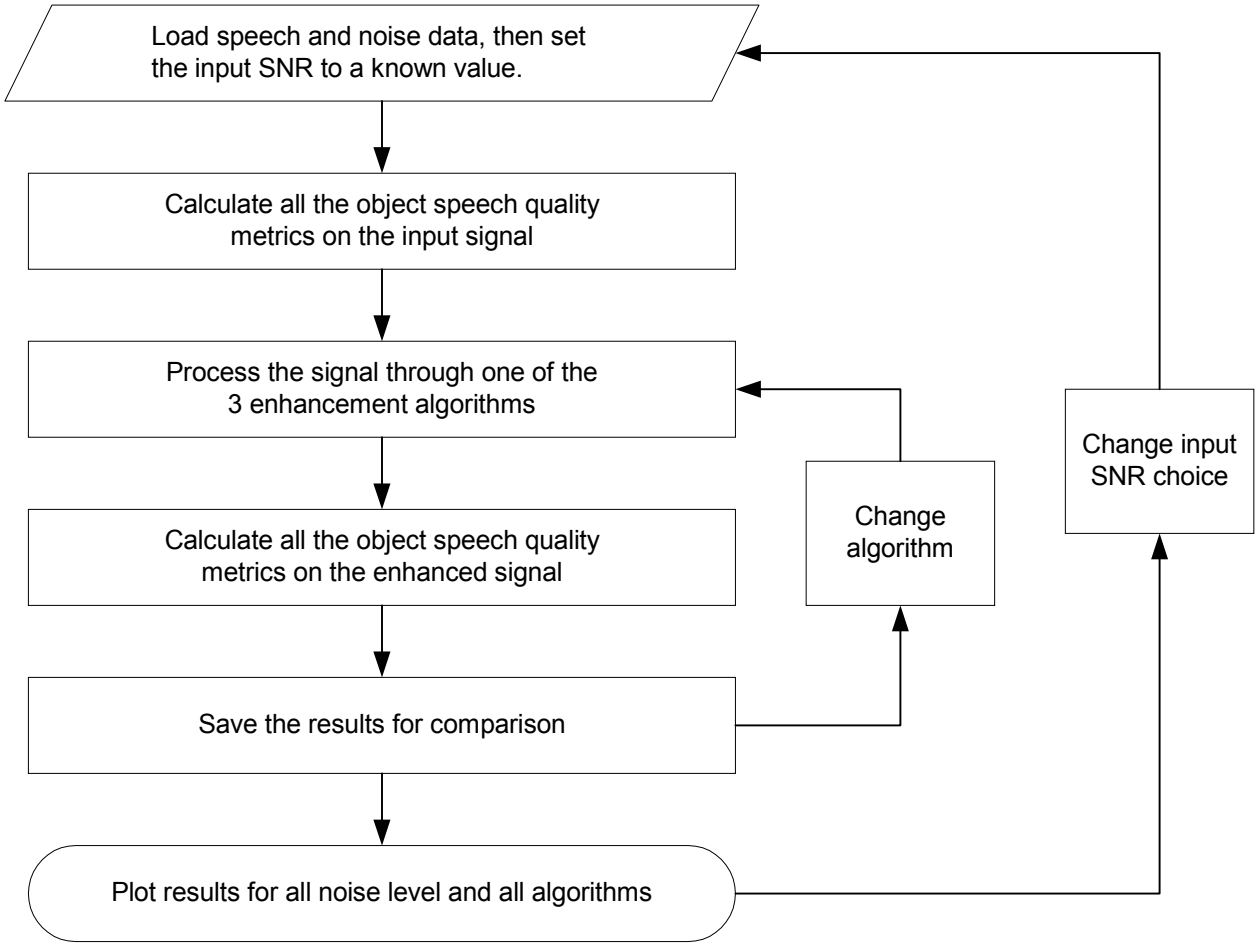


**Figure 5.9: Simulation flow chart**

## 5.3 Speech quality measures

The person who listens to the speech is ultimately the one who decides its quality, thus subjective listening tests are the best way to judge the performance of an algorithm. Commonly used subjective tests are the Mean Opinion Score (MOS), Diagnostic Acceptability Measure (DAM), and Diagnostic Rhyme Test (DRT). The challenge with subjective measures is that a large number of people tested under consistent conditions are required to get valid results. Objective measures overcome this burden by allowing a computer to analyze the speech quality.

Objective quality measures do not correlate 100% to subjective tests, but they can provide rough analysis to help understand how well an algorithm is performing. Some objective measures correlate fairly well to subjective tests and when used in combination can come even closer. The objective measures used in this thesis were chosen for their good correlation to subjective tests and general acceptance in the speech enhancement research community.

| Objective Speech Quality Measure | Correlation to Subjective Tests |
|---|---|
| Signal-to-Noise Ratio (SNR) | 24% |
| Segmental SNR (SSNR) | 77% |
| Articulation Index | 67% |
| Itakura-Saito Distance | 59% |

**Table 5.3: Objective Speech Quality Measure Correlation to Subjective Tests [102]**

The correlation measures in Table 5.3 were calculated against a database of subjective speech quality test data accumulated by Quackenbush, where the subjective quality test used was the Diagnostic Acceptability Measure (DAM). All the objective quality measures cited in Table 5.3 require the original speech for their calculations. The speech and the noise used in this thesis were recorded separately in the same environment in order to have the required clean speech reference when computing objective quality measures.

### 5.3.1 SNR and SSNR

Signal-to-Noise (SNR) ratio is the most popular measure of signal quality, but it may be necessary to correct phase errors in the signal estimate in order to achieve correct time alignment. The definition for SNR is equation (5.1) where $s(n)$ is the original clean speech and $\hat{s}(n)$ is the speech estimate.

$$SNR = 10\log_{10}\frac{E_s}{E_{Error}} = \frac{\sum_n s^2(n)}{\sum_n \left[ s(n) - \hat{s}(n) \right]^2} dB \qquad \textbf{(5.1)}$$

It has been shown that SNR, as defined in equation (4.1), is a poor estimate of subjective speech quality, so other methods are required to obtain a better characterization of the processed speech. SNR taken over short speech segments and then averaged together is called Segmental SNR (SSNR) and is more closely correlated to subjective speech

quality. Segmentation has the effect of applying equal weight to the loud and soft portions for the speech signal. The definition for $SSNR$ is equation (5.2) where $M$ is the number of frames.[18]

$$SSNR = \frac{1}{M}\sum_{j=0}^{M-1} 10\log_{10} \frac{\sum_{n=m_j-N+1}^{m_j} s^2(n)}{\sum_{n=m_j-N+1}^{m_j} \left[ s(n) - \hat{s}(n) \right]^2} dB \qquad \textbf{(5.2)}$$

If there are intervals of silence in the clean speech, even the smallest amount of error in the estimate will result in very large negative SNR values for that frame. Calculating SSNR only on frames that contain speech is a way to avoid the presence of these large negative values and a simple energy detection VAD was used in this thesis for that purpose.

### 5.3.2 Articulation Index

The Articulation Index (AI) measures only the intelligibility of the speech estimate compared to the original clean speech, which was researched as early as 1918 by Fletcher [103], proposed by French and Steinberg in 1947, and then further developed by Kryter in 1962.[104] The signal is divided into frequency bands, which are all given equal weight in the AI calculation because each band is considered to have an equal contribution to intelligibility. The Bark scale requires 18 bands for a 4 kHz bandwidth or 8kHz sampling rate. An approximation of the AI measure is defined in equation (4.3) as described by Deller, Hansen, and Proakis.[105] The SNR per critical band is calculated and averaged

together to obtain the AI result where 1 corresponds to upper limit of intelligibility and is 30 dB for the calculations performed in this thesis.

$$AI = \frac{1}{B}\sum_{j=1}^{B}\frac{\min\{SNR_j, 30\}}{30} \quad \textbf{(5.3)}$$

Equation (5.3) defines the AI approximation where $B$ is the number of critical bands and the SNR per critical band is represented by $SNR_j$. A problem with the definition of AI in equation (5.3) is that it is possible to obtain negative AI results and AI is typically reported on a scale from 0 to 1. This thesis will set any negative AI results to a lower limit of intelligibility equal to zero, which only occurs for the unprocessed AWGN case at 0 dB input SNR.

### 5.3.3 Itakura-Saito distance

The Itakura distance measure is based on the dissimilarity between all-pole modes of the reference and estimate speech signal where the all-pole model of speech using Linear Prediction (LP) is described in section 2.3. The Itakura distance is a representation of the short-term spectral differences between two frames of speech. It has been demonstrated that differences between two spectra in formant locations and formant bandwidths cause phonetic differences, which implies that a better speech spectrum envelope produces better perceptual quality.[106] The form of the measure used in this thesis is referred to as the Itakura-Saito (IS) distance, which is defined in equations (5.4) and (5.5).[18] A smaller IS measure is better, unlike the other objective quality measures, because the smaller IS metric represents less spectral distortion from the original clean speech.

$$IS = \int_{-\pi}^{\pi} \left[ e^{V(\theta)} - V(\theta) - 1 \right] \frac{d\theta}{2\pi} \quad \textbf{(5.4)}$$

$$V(\theta) = \log\left( \frac{\sigma_s}{\left| A_s\left(e^{j\theta}\right) \right|^2} \right) - \log\left( \frac{\sigma_y}{\left| A_y\left(e^{j\theta}\right) \right|^2} \right) \quad \textbf{(5.5)}$$

Further explanation of the IS distance offers some insight into how to interpret the results of comparing the noisy speech or processed speech to the clean speech reference. The IS distance is also referred to as a Log Likelihood ratio based on the $p^{th}$ order all-pole model of the speech in equation (5.6) where the speech is divided into short segments of approximately 16 ms.

$$s_r(n) = \sum_{i=1}^{p} a_r(i) s_r(n-i) + Gu(n) \quad \textbf{(5.6)}$$

| | |
|---|---|
| $s_r(n)$ | Clean reference speech |
| $a(i)$ | Coefficients of the all-pole filter |
| $G$ | Filter gain |
| $u(n)$ | Unit variance white filter excitation |

The log likelihood ratio compares the two windowed speech forms using the autocorrelation matrix and LPC coefficient vectors.

$$d(k) = \log\left( \frac{\vec{a}_y R_s \vec{a}_y^{\,T}}{\vec{a}_s R_s \vec{a}_s^{\,T}} \right) \quad \textbf{(5.7)}$$

| | |
|---|---|
| $d(k)$ | Distance for frame $k$ |
| $\vec{a}_y$ | Noisy speech LPC coefficient vector $(1, -a_y(1), -a_y(2), \ldots, -a_y(p))$ |
| $\vec{a}_s$ | Clean speech LPC coefficient vector $(1, -a_s(1), -a_s(2), \ldots, -a_s(p))$ |
| $R_s$ | Autocorrelation matrix for the clean speech |
| $x^T$ | Transpose operation |

An alternative development of the log likelihood ratio is to look at it as a filtering operation, where the inverse filters are represented in equation (5.8) and (5.9).

$$A_s(z) = 1 - \sum_{i=1}^{p} a_s(i) z^{-i} \tag{5.8}$$

$$A_y(z) = 1 - \sum_{i=1}^{p} a_y(i) z^{-i} \tag{5.9}$$

The corresponding prediction error or residual from inverse filtering the clean speech through both filters is shown in equations (5.10) and (5.11).

$$e_s(n) = s(n) - \sum_{i=1}^{p} a_s(i) s(n) \tag{5.10}$$

$$e_y(n) = s(n) - \sum_{i=1}^{p} a_y(i) s(n) \tag{5.11}$$

The log likelihood ration for a frame can be written as the ratio of the power in the residuals in equation (5.12) and re-written as equation (5.13) using Parseval's relation.

$$d(k) = \log\left( \frac{e_y(k)^2}{e_s(k)^2} \right) \tag{5.12}$$

$$d(k) = \log\left( 1 + \int_{-\pi}^{\pi} \left| \frac{A_s\left(e^{j\omega}\right) - A_y\left(e^{j\omega}\right)}{A_s\left(e^{j\omega}\right)} \right|^2 \frac{d\omega}{2\pi} \right) \tag{5.13}$$

This shows that the spectral differences between the clean reference speech and noisy or processed speech are most heavily weighted when $1/\left|A_y\left(e^{j\omega}\right)\right|$ is large, which is generally near the formant peaks of the speech.[18] It has also been noted in general that spectral estimates using linear prediction are always biased towards the pitch harmonics.[106]

The properties of the IS measure help explain the IS results reported for the algorithms in the rest of this section. The type of additive noise would have to significantly effect spectral region of the speech near the pitch harmonics or formant frequencies in order to produce large IS distances. Beamforming is working primarily in the spatial domain to achieve its noise suppression and does not dramatically affect the envelope of the speech, which explains why it results in smaller IS measures than the other enhancement algorithms. Spectral subtraction and SSPN, in contrast to beamforming, do modify the spectrum directly to achieve noise suppression. Subsequently, the IS distance measures are larger for the SS and SSPN results because they are attenuating the noise in discrete frequency bands, which also attenuate the speech harmonics to a degree. The logic above explains why the IS distance is sometimes smaller for the noisy speech than for the speech processed by the enhancement algorithms.

## 5.4 SSPN Algorithm

### 5.4.1 Simulation results

The SSPN algorithm's main advantage is the perceptually weighted nonlinear spectral subtraction and the beamformer because the talker isolation was not implemented. The beamforming results must be reported separately because the objective speech quality measures all require phase alignment of the input and output signals and this was not possible with the beamformer. The beamformer was evaluated by sending the noise and speech through separately. The results reported here are for the SSPN algorithm without

the beamformer, so the beamforming gains in quality will have to be assumed approximately additive to the gains presented here.

The SSPN algorithm was iterated once, as shown in Figure 3.4, to achieve the results shown in Table 5.4, Table 5.5, and Table 5.6. On the second iteration show in Figure 3.5, the first speech estimate is used as input to the VAD and the algorithm reprocesses the original signal. The arrows shown next to the names of the objective quality measures indicate whether higher or lower values are better. An up arrow, ↑, hints that large values are better, as in SNR. A down arrow, ↓, hints that a smaller value is better, as in IS.

| | **Original noise + speech** | | | | **After SSPN** | | | |
|---|---|---|---|---|---|---|---|---|
| Noise Type | Road & Engine | Fan | Interfering talkers | AWGN | Road & Engine | Fan | Interfering talkers | AWGN |
| SNR ↑ | **0** | **0** | **0** | **0** | 12.20 | 6.39 | 4.77 | 5.84 |
| SSNR↑ | -3.77 | -3.41 | -3.47 | -4.37 | 7.88 | 4.09 | 2.86 | 1.75 |
| AI ↑ | 0.32 | 0.20 | 0.11 | 0 | 0.42 | 0.31 | 0.16 | 0.04 |
| IS ↓ | 0.77 | 0.94 | 0.85 | 15.80 | 0.77 | 1.12 | 1.05 | 18.74 |

**Table 5.4: SSPN Results at 0 dB SNR**

| | **Original noise + speech** | | | | **After SSPN** | | | |
|---|---|---|---|---|---|---|---|---|
| Noise Type | Road & Engine | Fan | Interfering talkers | AWGN | Road & Engine | Fan | Interfering talkers | AWGN |
| SNR ↑ | **5** | **5** | **5** | **5** | 16.8 | 13.70 | 9.47 | 9.85 |
| SSNR↑ | 1.23 | 1.58 | 1.53 | 0.63 | 12.30 | 9.97 | 7.16 | 5.49 |
| AI ↑ | 0.49 | 0.37 | 0.27 | 0.01 | 0.62 | 0.47 | 0.32 | 0.17 |
| IS ↓ | 0.51 | 0.64 | 0.58 | 11.00 | 0.39 | 0.71 | 0.73 | 13.56 |

**Table 5.5: SSPN Results at 5 dB SNR**

| | Original noise + speech | | | | After SSPN | | | |
|---|---|---|---|---|---|---|---|---|
| Noise Type | Road & Engine | Fan | Interfering talkers | AWGN | Road & Engine | Fan | Interfering talkers | AWGN |
| SNR ↑ | **10** | **10** | **10** | **10** | 20.74 | 17.86 | 14.22 | 15.00 |
| SSNR↑ | 6.23 | 6.59 | 6.53 | 5.63 | 15.96 | 13.75 | 11.15 | 10.50 |
| AI ↑ | 0.65 | 0.53 | 0.44 | 0.18 | 0.76 | 0.59 | 0.47 | 0.33 |
| IS ↓ | 0.36 | 0.45 | 0.44 | 7.49 | 0.30 | 0.56 | 0.63 | 6.43 |

**Table 5.6: SSPN Results at 10 dB SNR**

The algorithm offers better improvement based on the noise type and this performance bias is consistent across most measures and starting SNR values. SSPN suppresses the road noise best, does well removing fan noise, not very good suppressing AWGN, and is worst at removing interfering talkers.

The fan noise spectrum has more energy near the same frequency as the speech energy, so it will affect the speech quality more than the road noise. The decrease in performance for AWGN can be attributed to the VAD not detecting the speech frames as accurately as for the other noise types. The interfering talker noise passes right through and can be clearly heard in the processed signal because it is classified as speech and not subtracted out, which underscores the need for talker isolation. Virtually no musical noise artifacts can be heard in the processed signals even at 0 dB SNR, this indicates that the perceptual weighting of the spectral subtraction gain function is performing as expected.

### 5.4.2 Iteration results

Voice activity detection is generally more accurate for higher SNR, thus VAD accuracy should improve when using a noise reduced signal versus the original noisy speech.

Experiments iterating the SSPN algorithm were motivated by the desire to calculate the VAD on a noised reduced signal. Voice activity detection was improved by iterating the algorithm twice, as shown in Figure 5.10, and also made the VAD less sensitive to the fixed energy thresholds for detection of speech vs. noise.

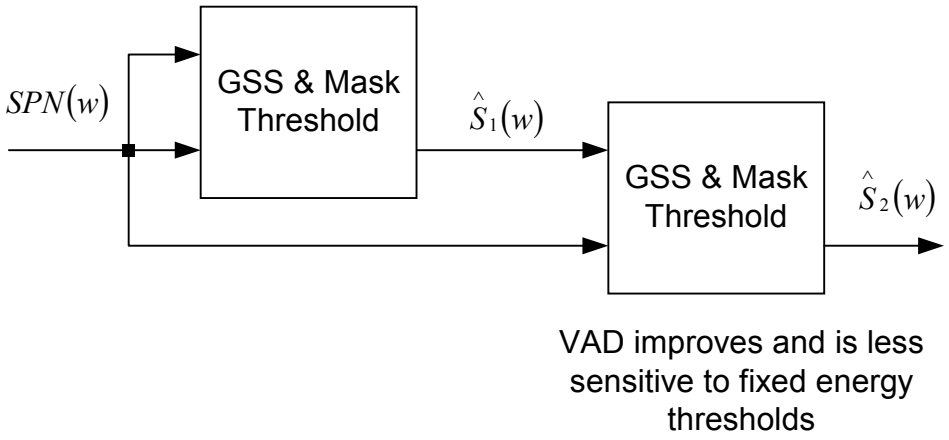


**Figure 5.10: Iteration of GSS**

Table 5.7 shows the speech quality results of iterating the algorithm and the corresponding percentage of voice frames detected. The example in Table 5.7 is for road noise at 5 dB SNR, but similar results were found for all noise types and SNR levels. It was seen that all the speech quality measures improve on the second iteration of the algorithm. The improvement is directly related to the performance of the VAD. The results also show that iterating more than two times starts to degrade the quality of the speech and provide less noise suppression.

| Signal | Road & Engine | GSS iteration 1 | **GSS iteration 2** | GSS iteration 3 |
|---|---|---|---|---|
| %VAD | - | 52 | 60 | 62 |
| SNR ↑ | 5 | 15.5 | 16.8 | 15.4 |
| SSNR↑ | 1.23 | 11.17 | 12.30 | 11.74 |
| AI ↑ | 0.49 | 0.54 | 0.62 | 0.61 |
| IS ↓ | 0.51 | 0.58 | 0.39 | 0.42 |

**Table 5.7: Iterating GSS and VAD**

Another interesting effect of iterating the algorithm is that VAD is less sensitive to the fixed energy thresholds used to determine if speech or noise is present. The threshold for speech in the simulations was fixed to 0.7 of the variance of the noise estimate. This threshold could be moved up or down by as much as 0.3 with little change in % VAD reported in the second iterations. In contrast the first iteration would change the % VAD reported directly corresponding to any variation of the threshold.

The behavior of the VAD, described in section 4.2.1, using the first speech estimate is very logical when the algorithm for voice detection is examined. The speech threshold is a fixed constant, $\alpha_S$, multiplied by the standard deviation of the noise estimate, $\sigma_N$, and added to the mean, $\mu_N$, of the noise estimate as shown in equation (5.14).

$$Thresh_S = \mu_N + \alpha_S * \sigma_N \quad \textbf{(5.14)}$$

Noise variance is significantly smaller when the signal has passed through the first iteration of the perceptually weighted non-linear spectral subtraction. The smaller variance causes the overall value of the speech detection threshold, $Thresh_S$, to be lower,

which naturally detects more speech frames. Lower noise variance also makes the VAD less sensitive to the choice of the fixed threshold constant because the value of the standard deviation, $\sigma_N$, that the fixed constant multiplies is less, so the impact of the constant, $\alpha_S$, on the VAD performance is also less. The lower threshold is less likely to classify a speech frame as noise, thus avoiding attenuation of the speech. If this is taken too far by iterating the algorithm many times, then not enough frames contribute to the noise estimate and the noise is less effectively removed the from the signal.

## 5.5 Beamforming

This section examines speech enhancement using GSC beamforming alone as described in section 4.1.3. Measuring the objective speech quality needs to be done differently for beamforming because there is not a single channel as a reference, so the method chosen here is to send the clean speech through the beamformer and then the noise. The processed signals are compared to analyze the relative speech attenuation versus noise suppression. The results in Table 5.8, Table 5.9 and Table 5.10 are for a Generalized Side-lobe Canceller (GSC) where it is assumed the filters are only adapting during silent frames in order to avoid too much attenuation of the speech. This is an ideal case and the results in Table 5.8, Table 5.9, and Table 5.10 show how well GSC can do when it is only adapting to the noise.

| | Original noise + speech | | | | After Beamforming | | | |
|---|---|---|---|---|---|---|---|---|
| Noise Type | Road & Engine | Fan | Interfering talkers | AWGN | Road & Engine | Fan | Interfering talkers | AWGN |
| SNR ↑ | **0** | **0** | **0** | **0** | 3.45 | 3.11 | 5.21 | 5.76 |
| SSNR↑ | -3.77 | -3.41 | -3.47 | -4.35 | -1.77 | -1.96 | 0.59 | -0.56 |
| AI ↑ | 0.32 | 0.20 | 0.11 | 0 | 0.40 | 0.30 | 0.17 | 0.01 |
| IS ↓ | 0.77 | 0.94 | 0.85 | 11.28 | 0.69 | 0.85 | 0.97 | 10.25 |

**Table 5.8: Beamforming Results at 0 dB SNR**

| | Original noise + speech | | | | After Beamforming | | | |
|---|---|---|---|---|---|---|---|---|
| Noise Type | Road & Engine | Fan | Interfering talkers | AWGN | Road & Engine | Fan | Interfering talkers | AWGN |
| SNR ↑ | **5** | **5** | **5** | **5** | 8.45 | 8.11 | 10.21 | 10.76 |
| SSNR↑ | 1.23 | 1.59 | 1.53 | 0.65 | 3.23 | 3.04 | 5.59 | 4.44 |
| AI ↑ | 0.49 | 0.37 | 0.27 | 0.02 | 0.57 | 0.46 | 0.34 | 0.18 |
| IS ↓ | 0.51 | 0.64 | 0.58 | 8.41 | 0.50 | 0.65 | 0.81 | 5.97 |

**Table 5.9: Beamforming Results at 5 dB SNR**

| | Original noise + speech | | | | After Beamforming | | | |
|---|---|---|---|---|---|---|---|---|
| Noise Type | Road & Engine | Fan | Interfering talkers | AWGN | Road & Engine | Fan | Interfering talkers | AWGN |
| SNR ↑ | **10** | **10** | **10** | **10** | 13.45 | 13.11 | 15.21 | 15.76 |
| SSNR↑ | 6.23 | 6.59 | 6.53 | 5.65 | 8.23 | 8.04 | 10.59 | 9.44 |
| AI ↑ | 0.65 | 0.53 | 0.44 | 0.18 | 0.73 | 0.63 | 0.51 | 0.35 |
| IS ↓ | 0.36 | 0.45 | 0.44 | 5.00 | 0.39 | 0.50 | 0.66 | 3.28 |

**Table 5.10: Beamforming Results at 10 dB SNR**

The beamformer is less effective on the road and fan noise, which are heavily weighted in the low frequency end of the spectrum. The lower frequencies fall below the resolution of the beamformer and the results support this theoretical expectation. The interfering speech and AWGN have more high frequency content and are consequently attenuated more by the beamformer. The gains are very consistent for the different noise sources across the SNR levels of 0, 5, and 10 dB. The SNR gain for the fan noise is 3.11 dB for

all three starting SNRs. This consistency at different noise levels can be attributed to the assumption that the filters are adapting only to the noise and are not dependent on the VAD for these simulations.

The distortion as measured by the IS metric is reduced by the beamformer for AWGN, but changes only slightly for the other noise sources. The AI and SNRs measures show good improvement for all the noise sources. The reduced distortion and good gains in SNR make the beamformer an excellent candidate for up front processing of the noise speech signal.

## 5.6 Spectral subtraction

The results reported here are for noise suppression using spectral subtraction with half-wave rectification with the same VAD that was used for the SSPN algorithm's perceptually weighted spectral subtraction.

| | **Original noise + speech** | | | | **After Spectral Subtraction** | | | |
|---|---|---|---|---|---|---|---|---|
| Noise Type | Road & Engine | Fan | Interfering talkers | AWGN | Road & Engine | Fan | Interfering talkers | AWGN |
| SNR ↑ | **0** | **0** | **0** | **0** | 9.25 | 5.25 | 5.69 | 4.51 |
| SSNR↑ | -3.77 | -3.41 | -3.47 | -4.37 | 5.86 | 2.46 | 3.98 | 1.12 |
| AI ↑ | 0.32 | 0.20 | 0.11 | 0 | 0.41 | 0.27 | 0.16 | 0.01 |
| IS ↓ | 0.77 | 0.94 | 0.85 | 15.80 | 0.86 | 1.12 | 1.47 | 32.41 |

**Table 5.11: Spectral Subtraction Results at 0 dB SNR**

| | Original noise + speech | | | | After Spectral Subtraction | | | |
|---|---|---|---|---|---|---|---|---|
| Noise Type | Road & Engine | Fan | Interfering talkers | AWGN | Road & Engine | Fan | Interfering talkers | AWGN |
| SNR ↑ | **5** | **5** | **5** | **5** | 12.36 | 9.67 | 10.62 | 7.36 |
| SSNR↑ | 1.23 | 1.59 | 1.53 | 0.63 | 9.06 | 6.81 | 8.45 | 4.16 |
| AI ↑ | 0.49 | 0.37 | 0.27 | 0.01 | 0.55 | 0.44 | 0.32 | 0.11 |
| IS ↓ | 0.51 | 0.64 | 0.58 | 11.00 | 0.52 | 0.81 | 1.30 | 23.34 |

**Table 5.12: Spectral Subtraction Results at 5 dB SNR**

| | Original noise + speech | | | | After Spectral Subtraction | | | |
|---|---|---|---|---|---|---|---|---|
| Noise Type | Road & Engine | Fan | Interfering talkers | AWGN | Road & Engine | Fan | Interfering talkers | AWGN |
| SNR ↑ | **10** | **10** | **10** | **10** | 16.04 | 13.93 | 12.5 | 11.80 |
| SSNR↑ | 6.23 | 6.59 | 6.53 | 5.63 | 12.23 | 10.43 | 9.77 | 8.73 |
| AI ↑ | 0.65 | 0.53 | 0.44 | 0.18 | 0.67 | 0.56 | 0.48 | 0.26 |
| IS ↓ | 0.36 | 0.45 | 0.44 | 7.50 | 0.43 | 0.56 | 1.13 | 9.60 |

**Table 5.13: Spectral Subtraction Results at 10 dB SNR**

Spectral subtraction provides significant SNR gains, but the gains are less for lower SNR possibly attributed to the change in VAD accuracy. The results show that spectral subtraction is better at removing the low frequency noise of the road, engine, and fan, but the interfering talkers and AWGN show less of an SNR gain when compared to the other noise sources. The AI measures are generally improved by spectral subtraction while the IS measures indicate there is more distortion after the spectral subtraction then before. Very noticeable musical noise artifacts are present when listening to the processed signal in contrast to the smooth sounding output of the SSPN algorithm. The interfering talkers are not attenuated very much because they can be clearly heard in the processed signal.

## 5.7 Theoretical limit of spectral subtraction

The theoretical limit of a speech enhancement algorithm based on spectral subtraction is perfect estimation and subtraction of the noise's spectral magnitude from the spectral magnitude of the received signal. Thus, the only effect of the noise is to distort the phase of the clean speech signal as seen in Figure 5.11. The theoretical limit can be calculated by multiplying the spectral magnitude of the clean speech times the phase of the noisy speech signal. An algorithm cannot achieve the theoretical perfect performance, but it does serve as an upper bound.

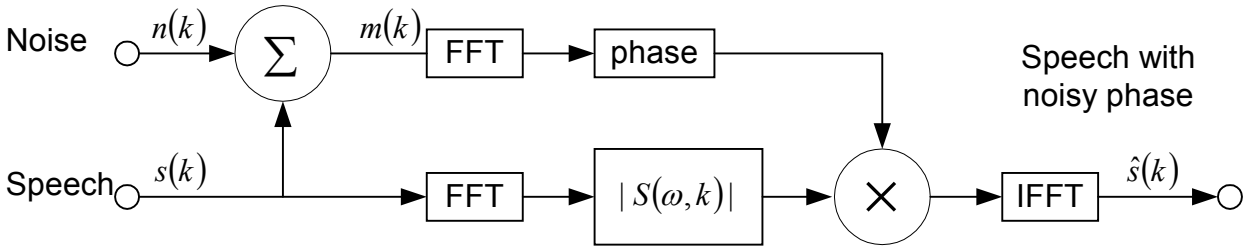


**Figure 5.11: Theoretical limit of spectral subtraction**

| | **Original noise + speech** | | | | **Theoretical limit of SS algorithms** | | | |
|---|---|---|---|---|---|---|---|---|
| Noise Type | Road & Engine | Fan | Interfering talkers | AWGN | Road & Engine | Fan | Interfering talkers | AWGN |
| SNR ↑ | **0** | **0** | **0** | **0** | 17.45 | 13.98 | 17.70 | 12.51 |
| SSNR↑ | -3.77 | -3.41 | -3.47 | -4.37 | 13.18 | 10.95 | 14.13 | 9.04 |
| AI ↑ | 0.32 | 0.20 | 0.11 | 0 | 0.60 | 0.50 | 0.56 | 0.25 |
| IS ↓ | 0.77 | 0.94 | 0.85 | 15.80 | 0.27 | 0.33 | 0.29 | 0.69 |

**Table 5.14: Theoretical limit of spectral subtraction at 0 dB SNR**

| | Original noise + speech | | | | Theoretical limit of SS algorithms | | | |
|---|---|---|---|---|---|---|---|---|
| Noise Type | Road & Engine | Fan | Interfering talkers | AWGN | Road & Engine | Fan | Interfering talkers | AWGN |
| SNR ↑ | **5** | **5** | **5** | **5** | 20.42 | 17.41 | 20.39 | 15.57 |
| SSNR↑ | 1.23 | 1.59 | 1.53 | 0.63 | 15.89 | 13.91 | 16.32 | 11.75 |
| AI ↑ | 0.49 | 0.37 | 0.27 | 0.01 | 0.73 | 0.63 | 0.67 | 0.35 |
| IS ↓ | 0.51 | 0.64 | 0.58 | 11.00 | 0.21 | 0.25 | 0.23 | 0.61 |

**Table 5.15: Theoretical limit of spectral subtraction at 5 dB SNR**

| | Original noise + speech | | | | Theoretical limit of SS algorithms | | | |
|---|---|---|---|---|---|---|---|---|
| Noise Type | Road & Engine | Fan | Interfering talkers | AWGN | Road & Engine | Fan | Interfering talkers | AWGN |
| SNR ↑ | **10** | **10** | **10** | **10** | 22.32 | 20.98 | 22.87 | 18.99 |
| SSNR↑ | 6.23 | 6.59 | 6.53 | 5.63 | 18.06 | 16.84 | 18.60 | 14.95 |
| AI ↑ | 0.65 | 0.53 | 0.44 | 0.18 | 0.84 | 0.77 | 0.77 | 0.46 |
| IS ↓ | 0.36 | 0.45 | 0.44 | 7.50 | 0.17 | 0.19 | 0.19 | 0.50 |

**Table 5.16: Theoretical limit of spectral subtraction at 10 dB SNR**

There are large improvements in all the speech quality measures as expected. The improvements are not perfect because of the noisy phase. For example, the SNR does not even approach 30 dB. The gains are less as the starting SNR increases from 0 to 10 dB because there is less relative noise to remove as SNR increases. The IS measure is very low when compared to any of the other methods, this would suggest that the noisy phase does not contribute much in the way of displacing the all-pole model of the speech.

## 5.8 Comparison of results

### 5.8.1 Speech quality measures

Figure 5.12, Figure 5.13, Figure 5.14, and Figure 5.15 graphically compare the algorithms' performance for each noise condition and speech quality measure. Higher numbers are better for SNR, SSNR, and AI. Lower numbers are better for the IS distance because it represents smaller distortion. The values of the objective speech quality measures vary significantly when the noise type is changed as can be seen when comparing the different groups of subplots. The SSPN algorithm (solid blue line) outperforms the SS (dashed red line) and BF (dash-dotted purple line) in most measures and noise types. The original noisy speech measures are included in each subplot as a dotted black line.

The road noise plots in Figure 5.12 show BF at 0 dB input has less distortion as shown in the IS metric plot. Spectral subtraction type algorithms can reduce distortion by being less aggressive at reducing the overall noise level in low SNR conditions, but that tradeoff was not explored in this thesis.

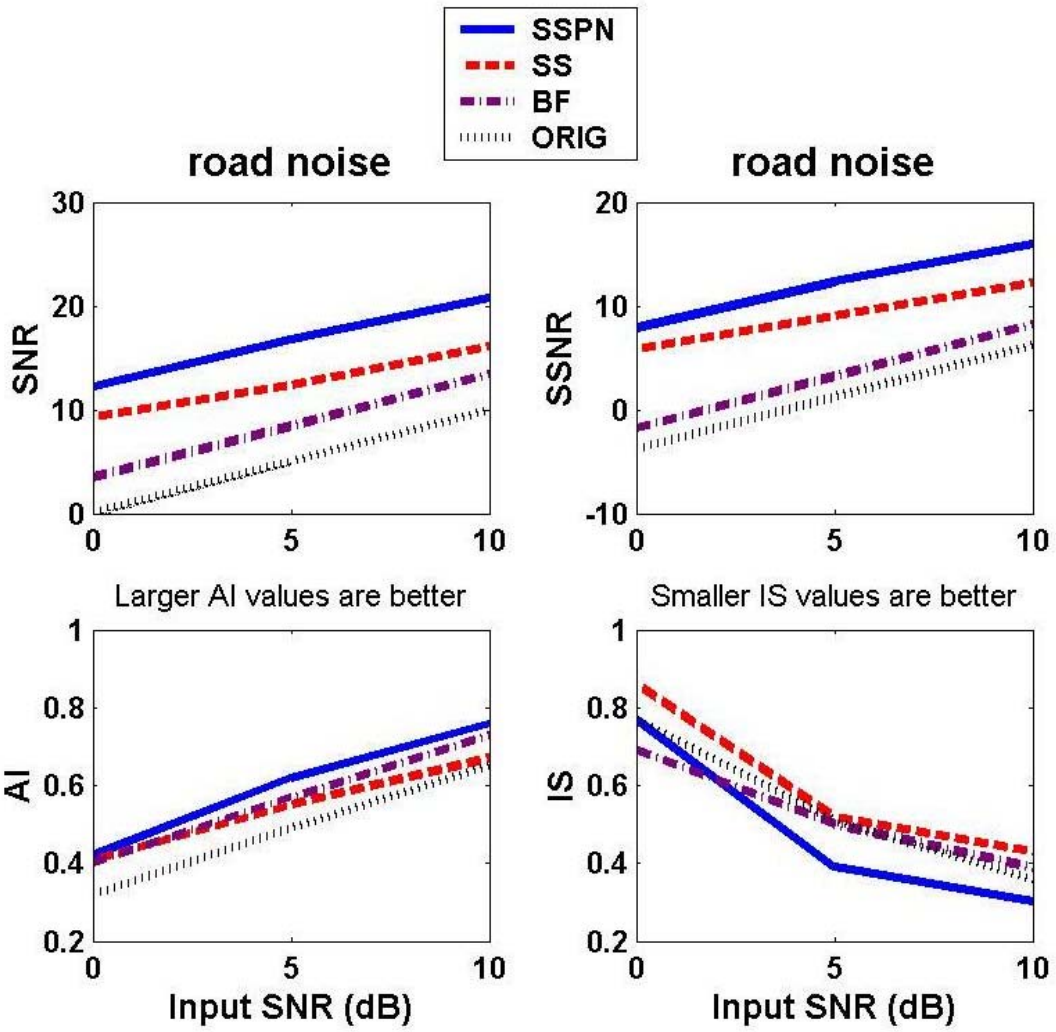


**Figure 5.12: Results for road noise**

All the algorithms show less improvement with the fan noise in Figure 5.13 when compared to the road noise results in Figure 5.12, but SSPN still outperforms the other algorithms on increasing SNR. Again beamforming introduces less distortion as shown in the IS subplot of Figure 5.13.

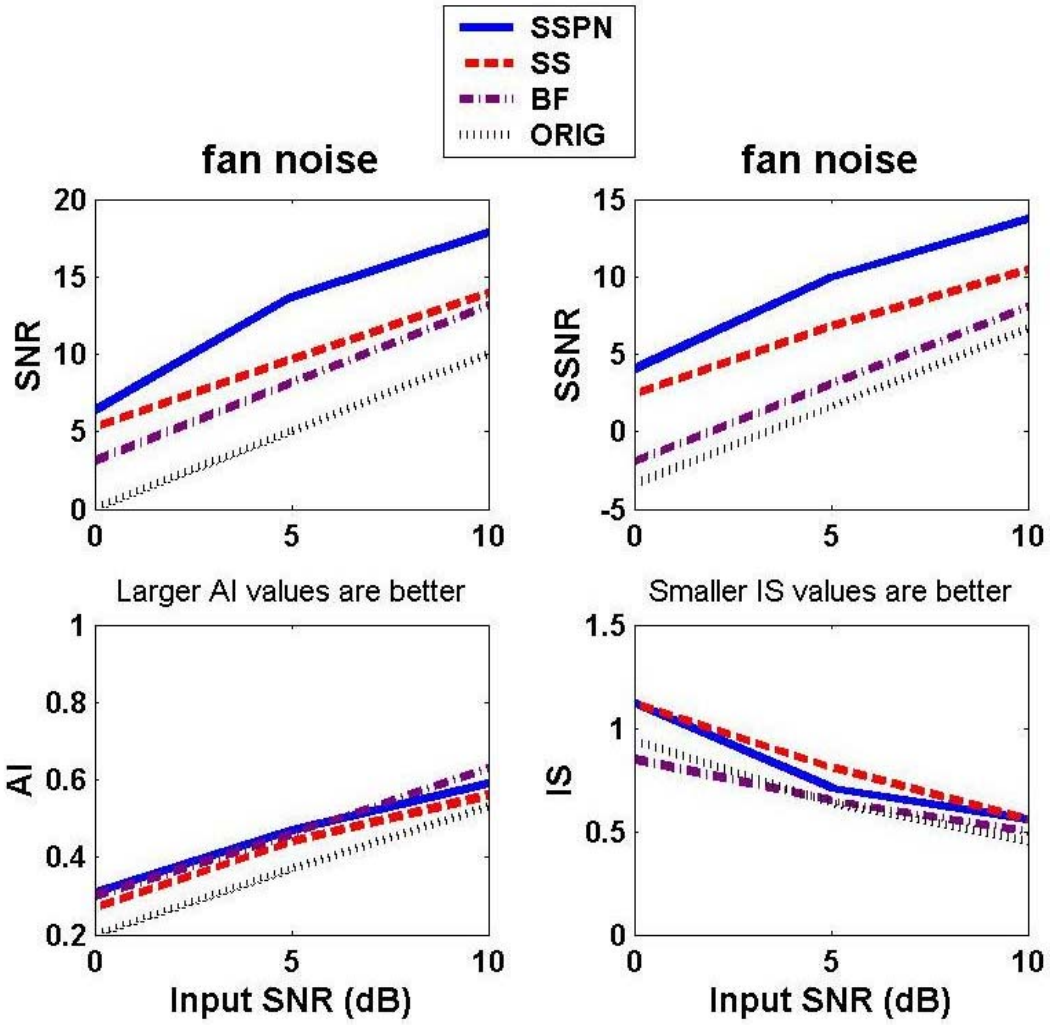


**Figure 5.13: Results for fan noise**

Beamforming effectively suppresses interfering talkers in Figure 5.14 when compared to the spectral subtraction methods because the interfering talker noise is less diffuse than road and engine noise, so the directional gains have a greater effect on the received signal. Poor performance of the spectral subtraction type algorithms on the interfering talker noise can be attributed to the VAD in the SS routines classifying the talker noise as voiced and not including it as part of the noise estimate.

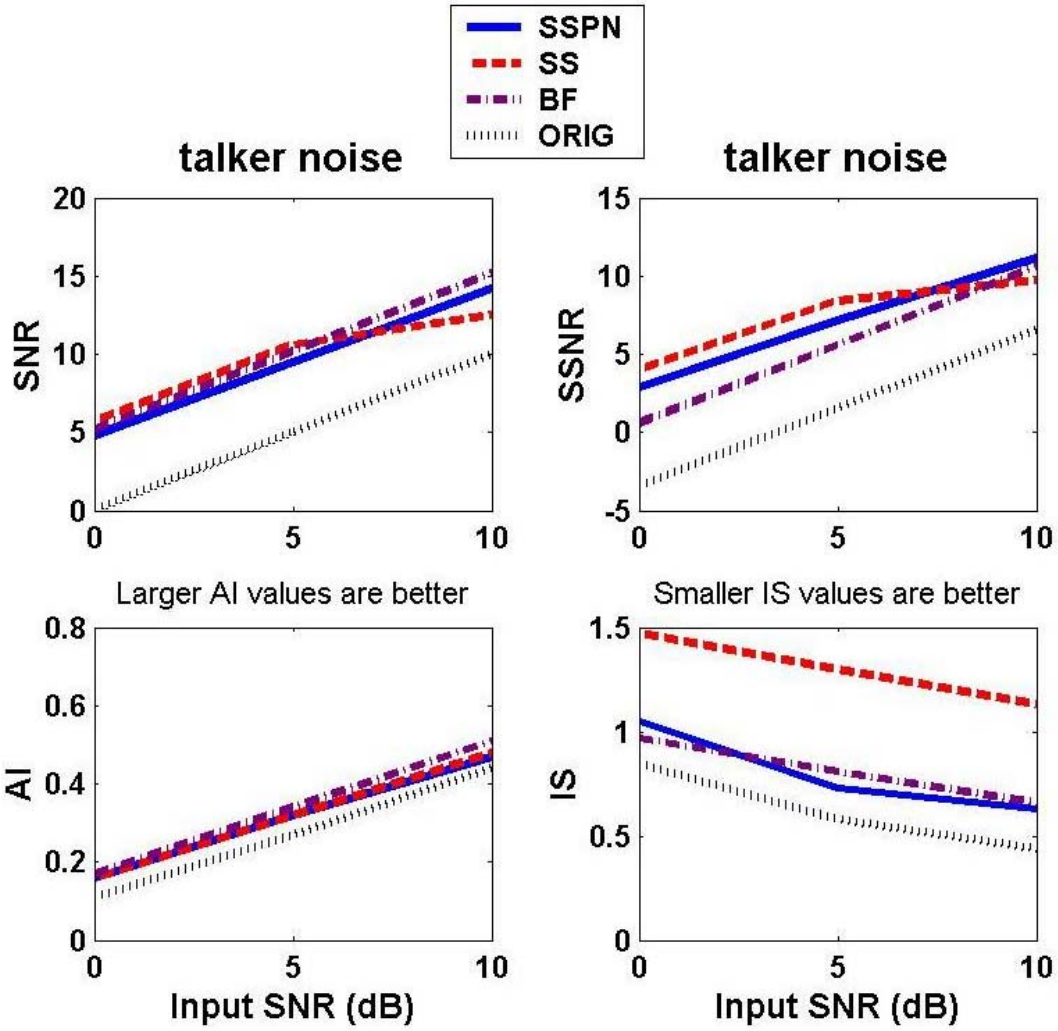


**Figure 5.14: Results for talker noise**

Beamforming does well in AWGN, as shown in Figure 5.15, because there is more noise energy in the higher frequencies than the other noise types. The higher frequencies are better resolved by the 4-element microphone array with 5 cm spacing and an overall aperture of 15 cm. The IS distance is very high for all the algorithms in AWGN when compared to other noise sources.

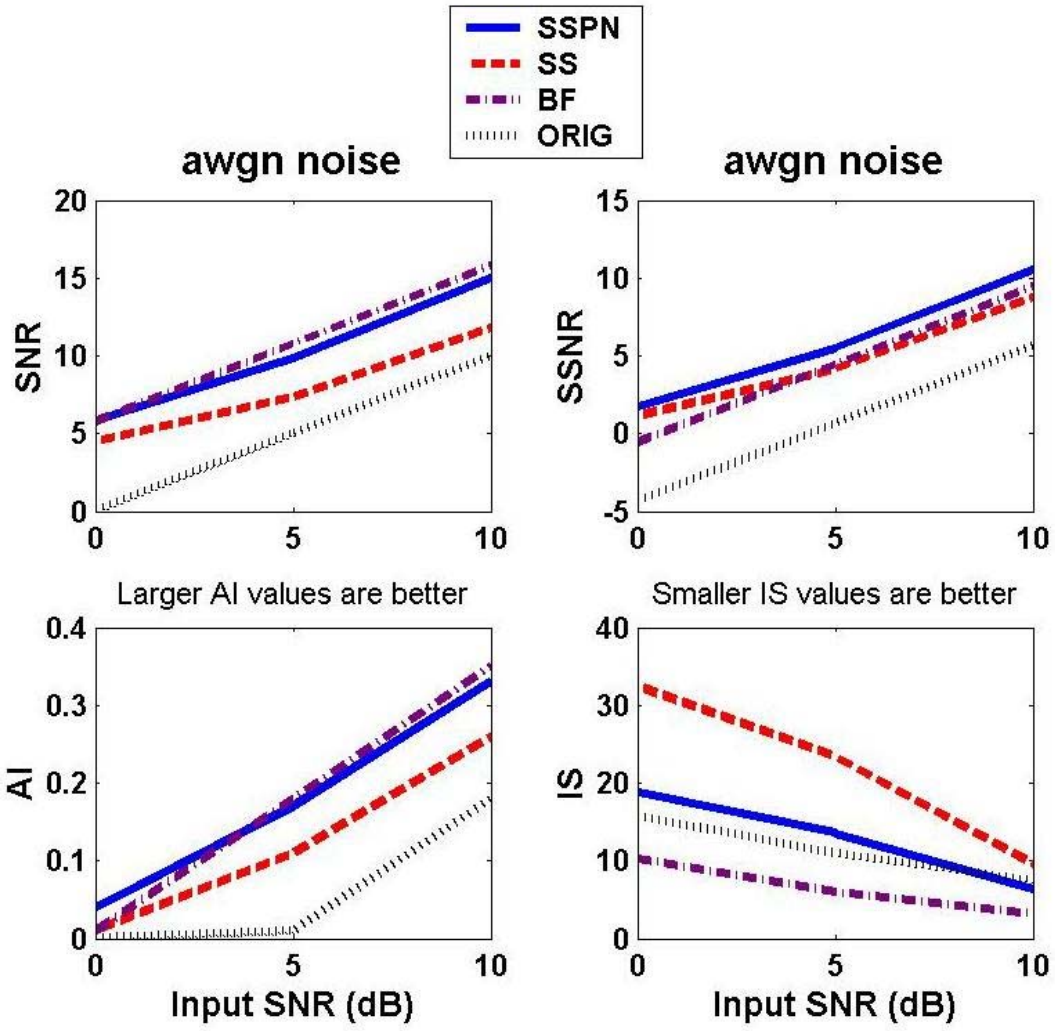


**Figure 5.15: Results for AWGN**

### 5.8.2 Time domain plots

Figure 5.16, Figure 5.17, Figure 5.18, and Figure 5.19 show the clean speech, noise speech, and enhanced signals for different noise types at 0 dB SNR. The female speech shown in the plots is "Turn up the radio".

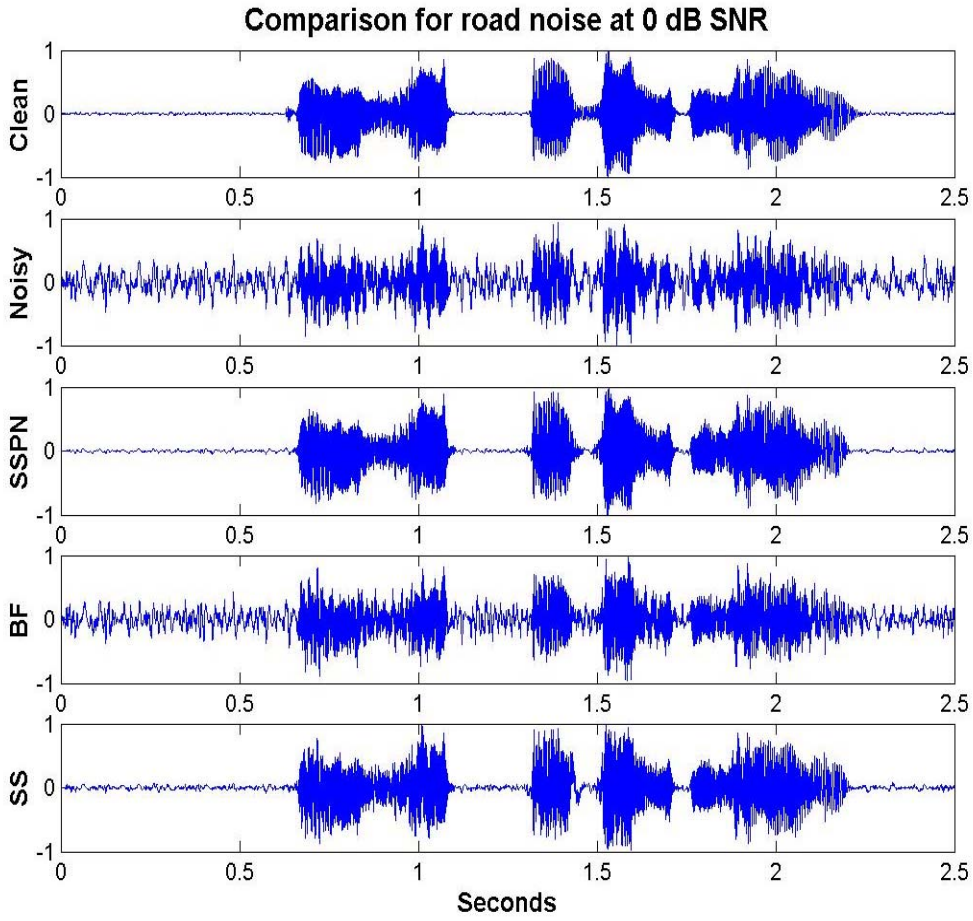


**Figure 5.16: Road noise + speech signals**

The low frequency content of the noise is evident in the noisy speech plot of Figure 5.16. Less residual noise is left in SSPN when compared to BF and SS, as can be seen by looking at the silent periods of the plots. Most of the original speech content remains in the processed signal with the exception of the SSPN and SS removing lower energy speech around 1.4 seconds. The overall envelope of the SSPN plot is much smoother than the SS plot which is expected because the SSPN algorithm removes artifacts that are introduced by SS.

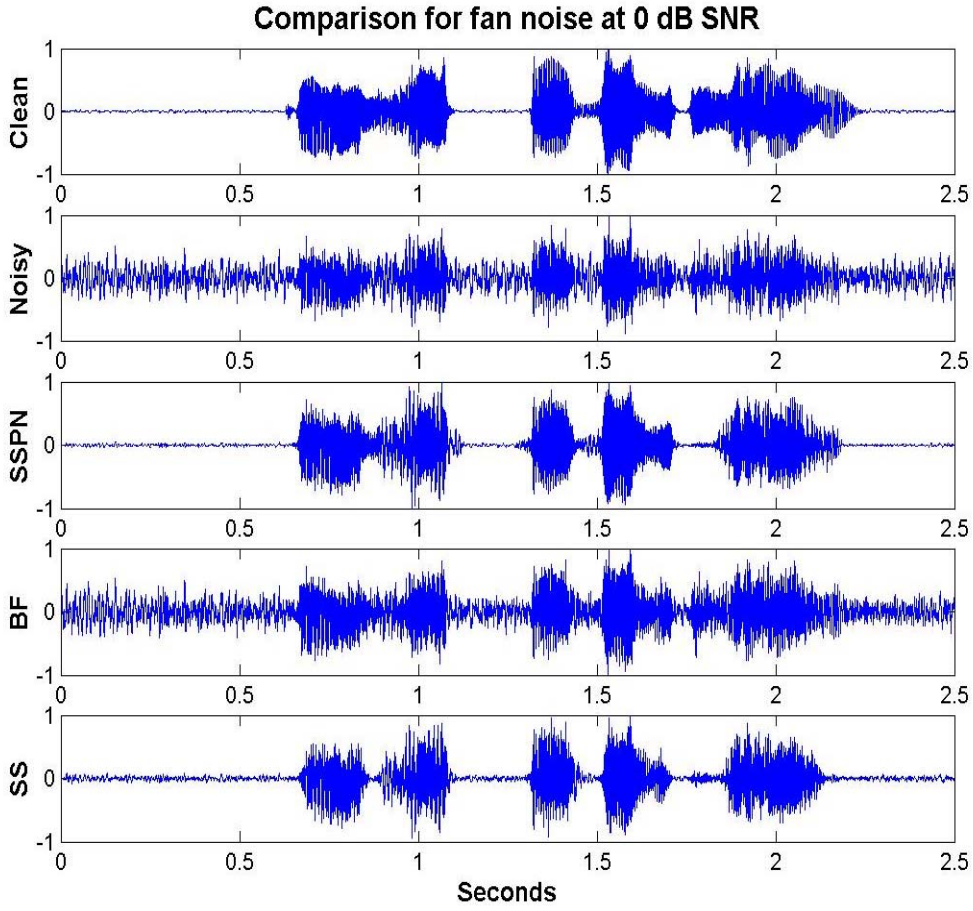


**Figure 5.17: Fan noise + speech signals**

Much more of the low energy speech is attenuated when SS or SSPN removes the fan noise. This speech attenuation is evident by comparing the plots at around 1.8 seconds in Figure 5.17. The fan noise estimate contains large spectral magnitude near the pitch of the female talker, which would explain the speech attenuation of SS and SSPN. The BF does not rely on a noise estimate, so it does not attenuate the low energy speech as seen in the other algorithms.

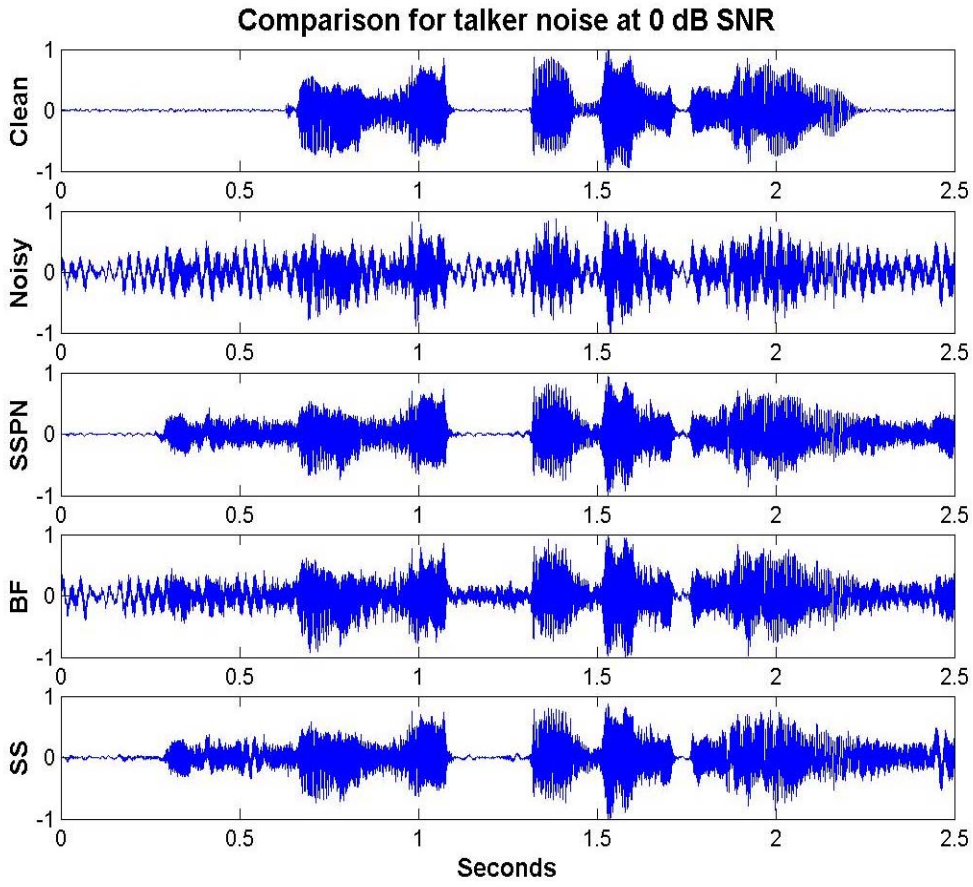


**Figure 5.18: Talker noise + speech signals**

Figure 5.18 shows the problem that occurs when the VAD classifies interfering talkers as a voiced signal and does not update the noise estimate. The desired speech has no energy around 0.5 seconds and 2.4 seconds, but the interfering speech is present. The subsequent plots of the processed signal show that very little interfering speech is removed. This is strong evidence for using talker separation as input to the VAD during a second iteration of processing on the same frame, which was not implemented in this thesis work and can be explored during future research on the SSPN algorithm.

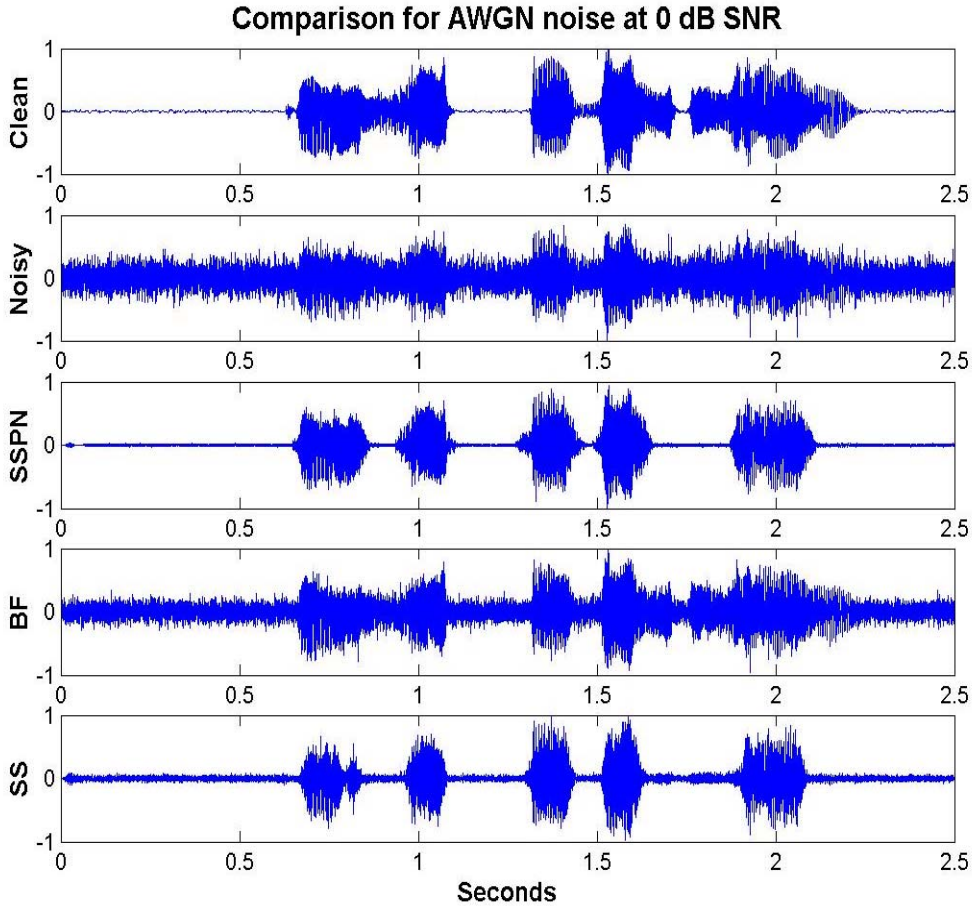


**Figure 5.19: White Gaussian Noise + speech signals**

It can be seen in Figure 5.19 that the SS and the SSPN algorithms do a better job reducing the white noise than beamforming, but at the cost of some distortion. The beamforming retains more detail of the original clean speech signal, which can be seen by looking at the signal between high-energy locations around 0.8 and 1.8 seconds. Again, the SSPN algorithm has less residual noise and a smoother envelope than SS or BF.

### 5.8.3 Spectrograms

The spectrograms in Figure 5.20, Figure 5.21, Figure 5.22, Figure 5.23, and Figure 5.24 show the signal's frequency energy versus time. Musical noise artifacts introduced by spectral subtraction will show up as a localized short-term smudge in the spectrogram.

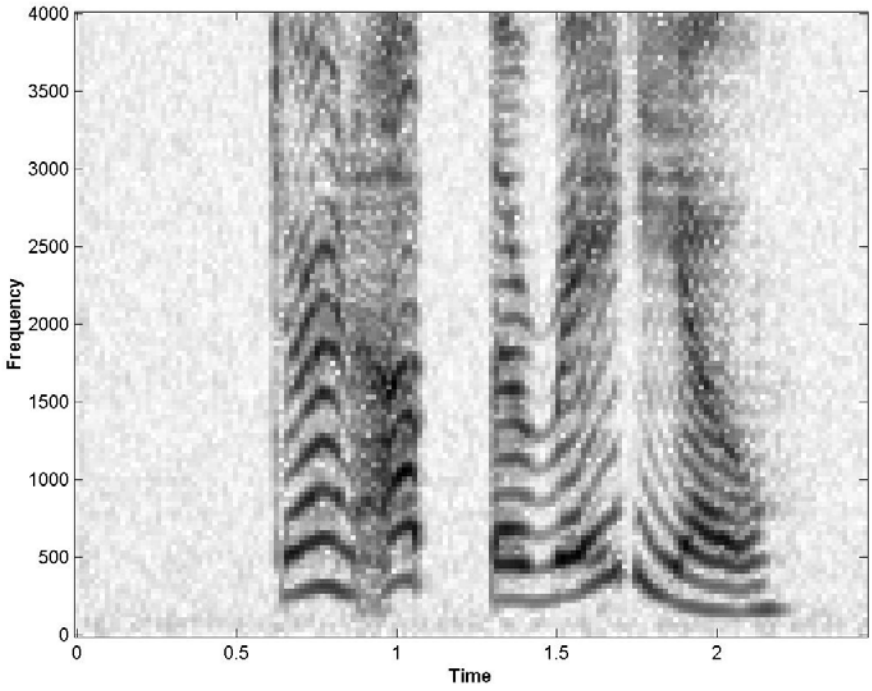


**Figure 5.20: Spectrogram for clean speech**

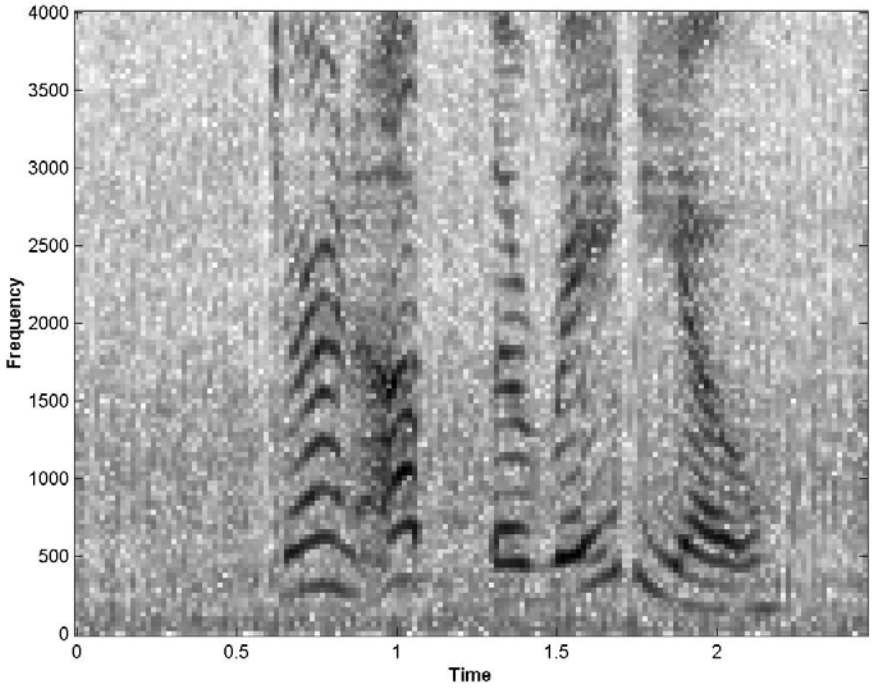


**Figure 5.21: Spectrogram for noisy speech**

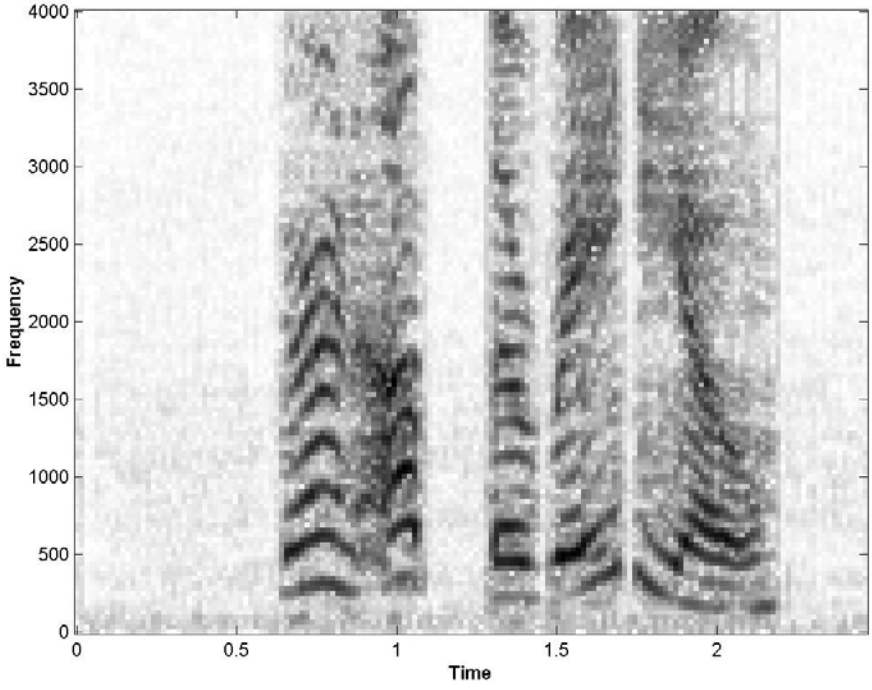


**Figure 5.22: Spectrogram for SSPN enhanced speech**

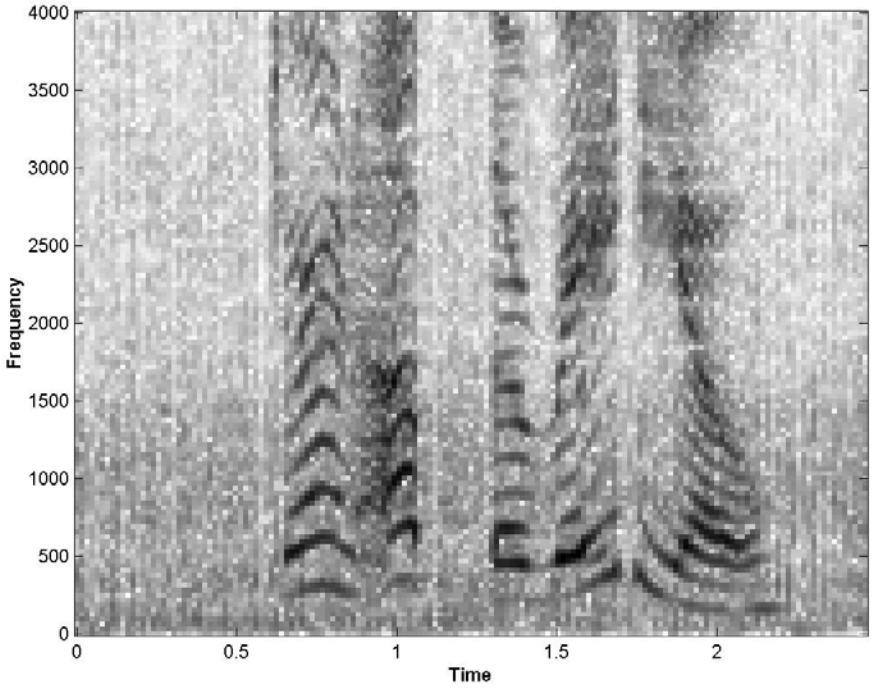


**Figure 5.23: Spectrogram for beamform enhanced speech**

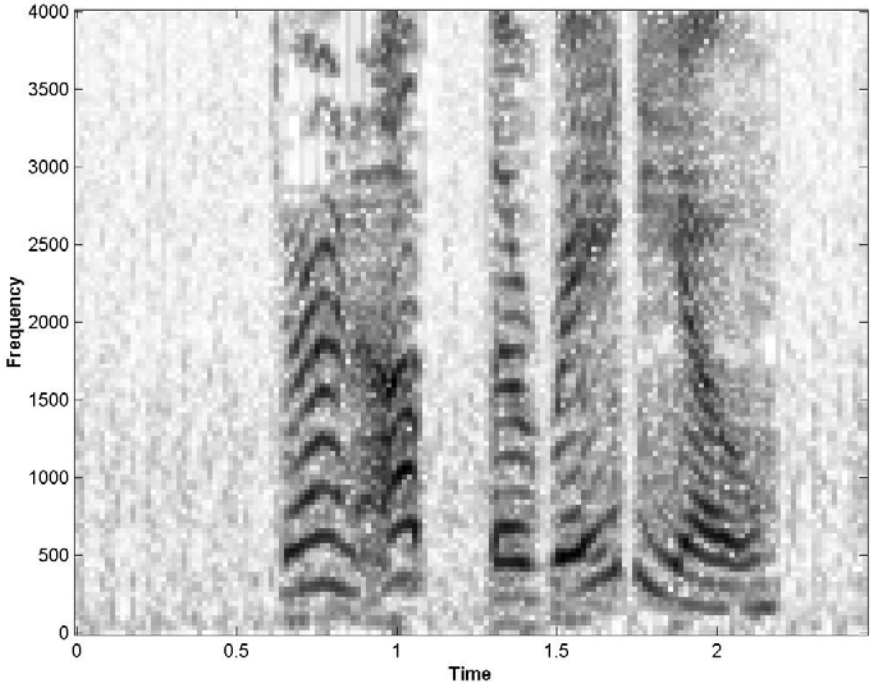


**Figure 5.24: Spectrogram for SS enhanced speech**

## 5.9 VAD performance

A closer analysis of the VAD is warranted because it plays such a critical role in the SSPN algorithm. Correct VAD decisions for each frame are determined by visually inspecting the clean speech signal and used as a reference, $VAD_{ref}$, for comparison to the VAD decisions calculated by the algorithm.

$$VAD\_Accuracy = \frac{correct\_frames}{total\_frames} = \frac{sum(abs(VAD_{ref} - VAD))}{total\_frames} \quad \textbf{(5.15)}$$

Parameters of the test signal:

- 8 kHz sampling rate
- 128 sample frame-size
- 157 frames for signal length
- Length of the female speech used is 2.5 seconds or 20,000 samples

Visual inspection of the clean speech signal found 88 speech frames and 69 silent frames, which corresponds to 56% voice activity. The comparison to this reference for each noise type, SNR level, and SSPN iteration is reported in Table 5.17.

| SNR (dB) | 0 | 0 | 5 | 5 | 10 | 10 |
|---|---|---|---|---|---|---|
| Iteration | 1 | 2 | 1 | 2 | 1 | 2 |
| Noise-free % VAD accuracy | 95 | | | | | |
| Road % VAD accuracy | 91 | 91 | 95 | 94 | 96 | 95 |
| Fan % VAD accuracy | 81 | 89 | 90 | 94 | 94 | 96 |
| Talker % VAD accuracy | 73 | 75 | 76 | 75 | 74 | 74 |
| AWGN % VAD accuracy | 54 | 56 | 82 | 87 | 90 | 95 |

**Table 5.17: SSPN VAD accuracy**

VAD accuracy, using the fan noise and AWGN, was improved by iterating the algorithm a second time. Road noise VAD accuracy stayed about the same for both iterations, but the second iteration tended to classify more frame as speech. Interfering talkers produced

consistently poor VAD accuracy across all SNR levels and iterations because the algorithm does not distinguish between desired and undesired talkers.

## 5.10 Pitch detection

Pitch detection experiments were performed with the SSPN algorithm to evaluate the possibility of using talker isolation to improve VAD accuracy and suppression of interfering talkers. Talker isolation was not tested directly with the SSPN algorithm in this thesis because of the overall complexity of the many approaches, mixed results reported in the research, and the large amount of time required for implementation. Pitch detection is simple to implement and was able to offer some insight into how a talker separation algorithm might perform before and after noise suppression.

As mentioned in section 3.3.7, the pitch detection is not a good candidate for modifying the spectral subtraction parameters. Experimental results consistently showed the spectral subtraction modified by pitch estimate weighting was worse than without it.

It was shown that the pitch detection algorithm could be more accurate using the noise-reduced signal when compared to the noisy signal. However, there are occasions that the enhanced signal causes the pitch detection to deviate from what would have been estimated in the original speech. Thus the noise removal helps pitch detection, but sometimes makes it worse. A more advanced pitch detection algorithm that is designed with knowledge of the noise suppression algorithms would likely have more success.

These results on pitch detection are not conclusive and future research should look closer at the effects of noise suppression on pitch detection.

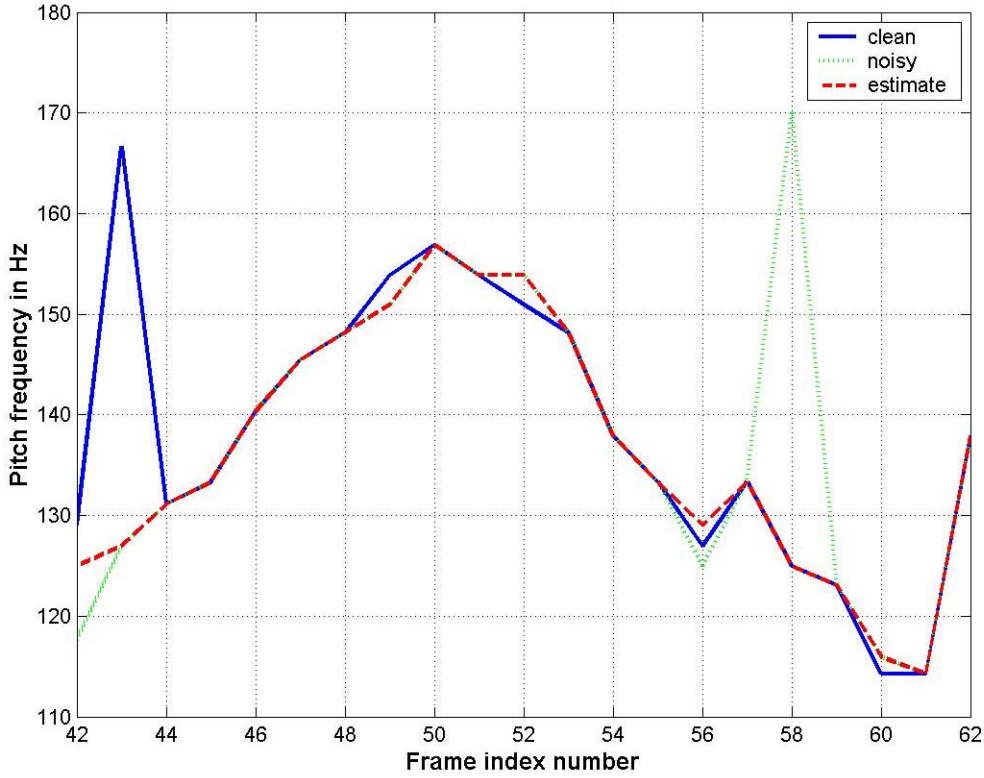


**Figure 5.25: Pitch detection example**

The measure of the pitch detection was done only on frames containing speech where the correct pitch is assumed to be the pitch detected on the original clean speech signal. Figure 5.25 is an example comparing the pitch detection for the clean original (solid blue line), speech + noise (green dotted line), and the speech estimate after noise removal (red dashed line).

## 5.11 Statistical analysis of Segmental SNR results

Segmental SNR (SSNR) is an objective speech quality measure that is fairly well correlated to subjective speech quality tests, so it is a good candidate for further analysis of the algorithms. Twenty-one different noise measurements were made in the car as reported in Table 2.1, which can be used to test the enhancement algorithms over a broad range of conditions. Statistical analysis examining the variances of the SSNR results will demonstrate the overall performance difference between the algorithms and the effect of input SNR.

### 5.11.1 SSNR results for multiple speech + noise data sets

Before doing the statistical analysis, some insights can be gained by examining the SSNR results directly. Figure 5.26, Figure 5.27, and Figure 5.28 compare the difference in improvement between the algorithms for 0, 5, and 10 dB input SNR respectively. It can be seen that SSPN outperforms the other two algorithms in most cases. Spectral subtraction provides more SSNR improvement than beamforming for all 21 data sets used. The SSNR for the original unprocessed noisy speech is included in each figure to show the relative improvement achieved by each algorithm across the 21 data sets.

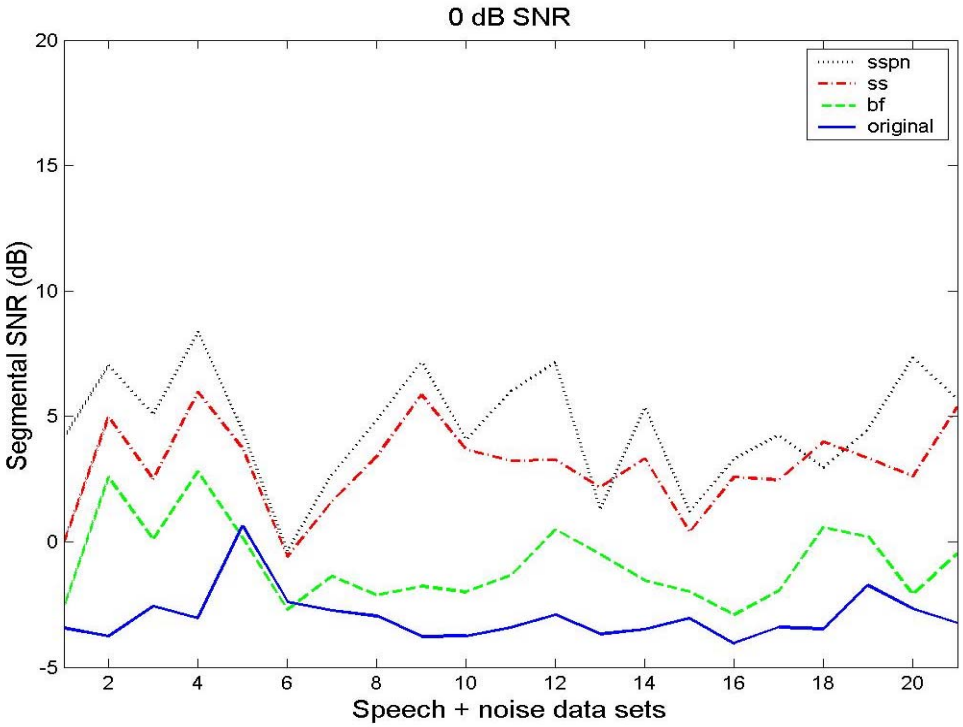


**Figure 5.26: SSNR results at 0dB input SNR**

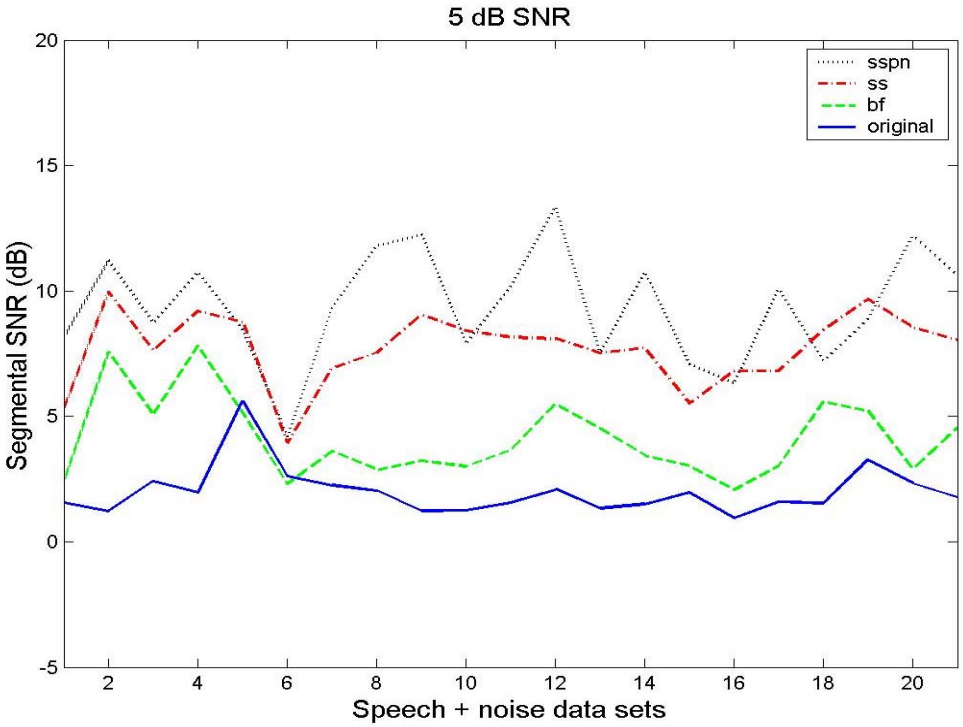


**Figure 5.27: SSNR results at 5dB input SNR**

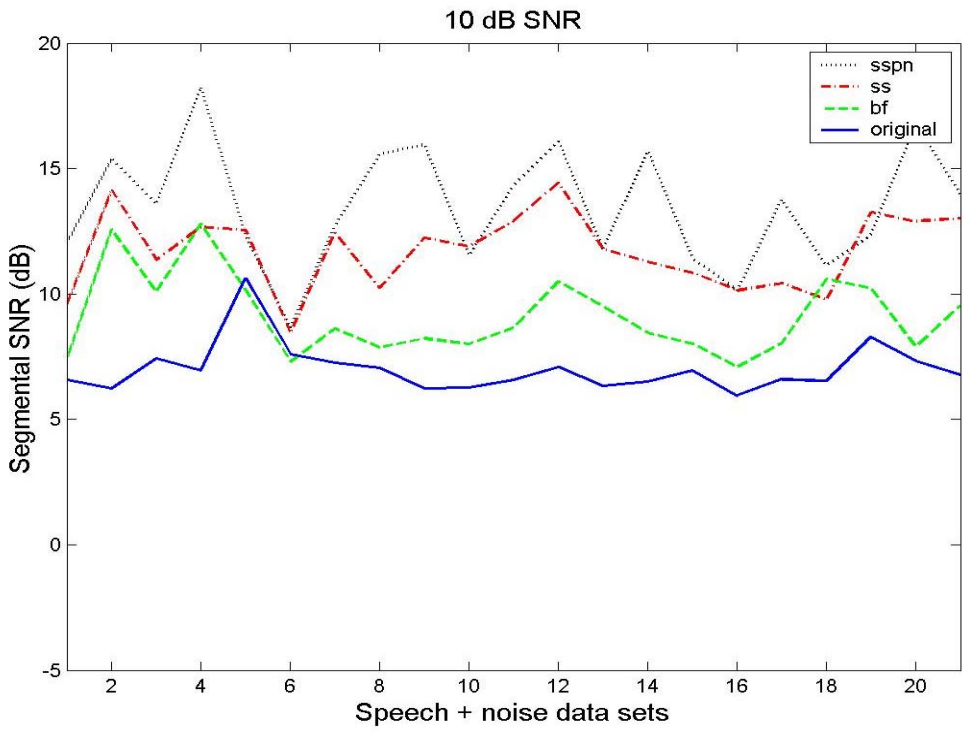


**Figure 5.28: SSNR results at 10dB input SNR**

Taking a different view of the data, the following figures compare the relative effects on SSNR of the input SNR. The SSNR results for input SNR levels of 0, 5, and 10 dB are shown for the original noisy speech in Figure 5.29, the beamforming output in Figure 5.30, the spectral subtraction output in Figure 5.31, and the SSPN algorithm output in Figure 5.32. Not surprisingly, the SSNR increases directly with increases of the input SNR for all four figures. The beamformer results scale linearly with input SNR with little or no variation in the shape of the curve each input SNR level. However, SS and SSPN do have some slight variations between input SNR curves, perhaps indicating the changing performance of the voice activity detection.

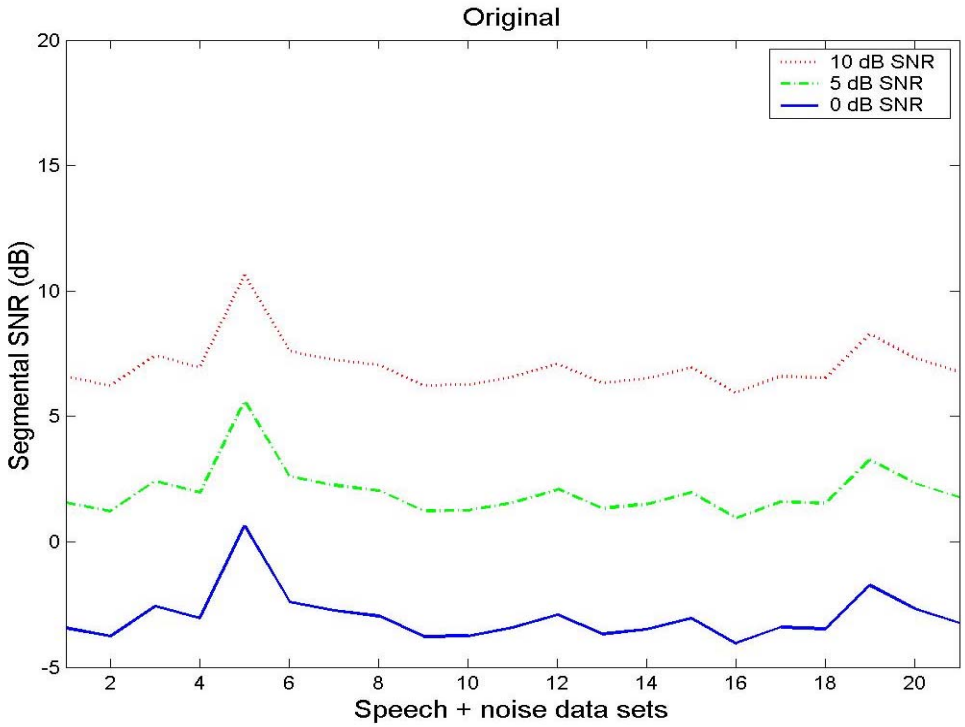


**Figure 5.29: SSNR of original speech + noise**

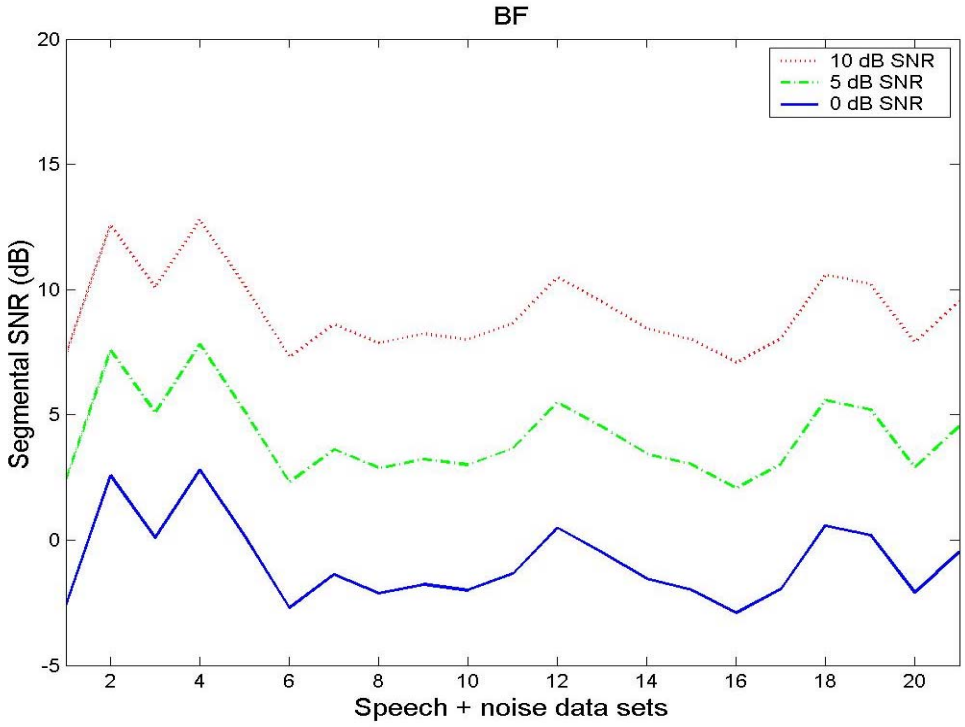


**Figure 5.30: SSNR results for Beamforming**

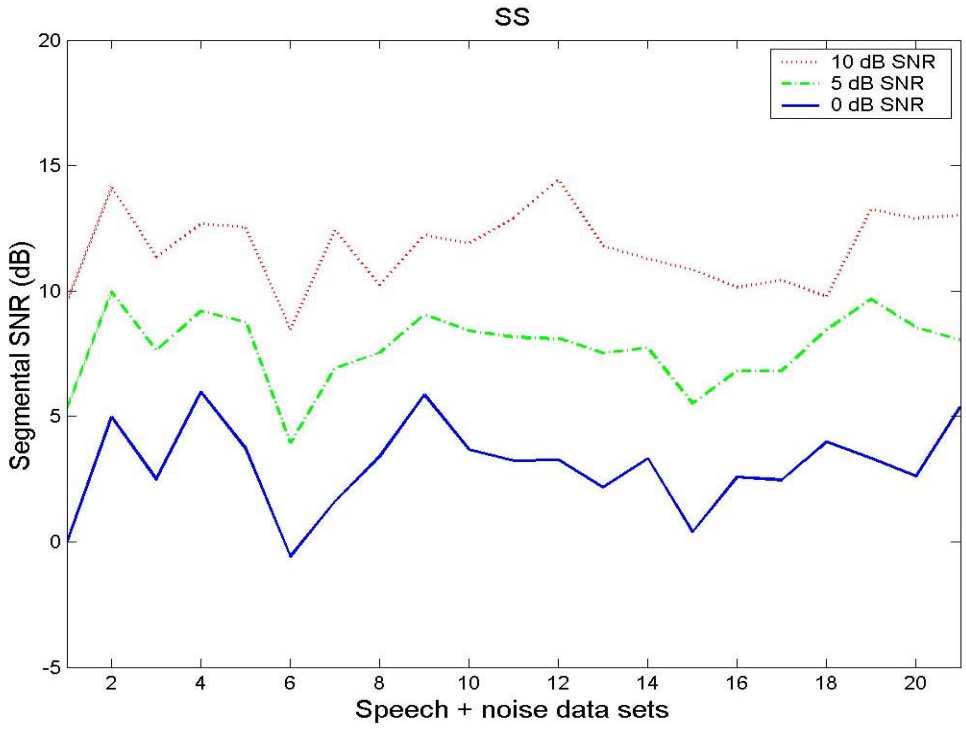


**Figure 5.31: SSNR results for Spectral Subtraction**

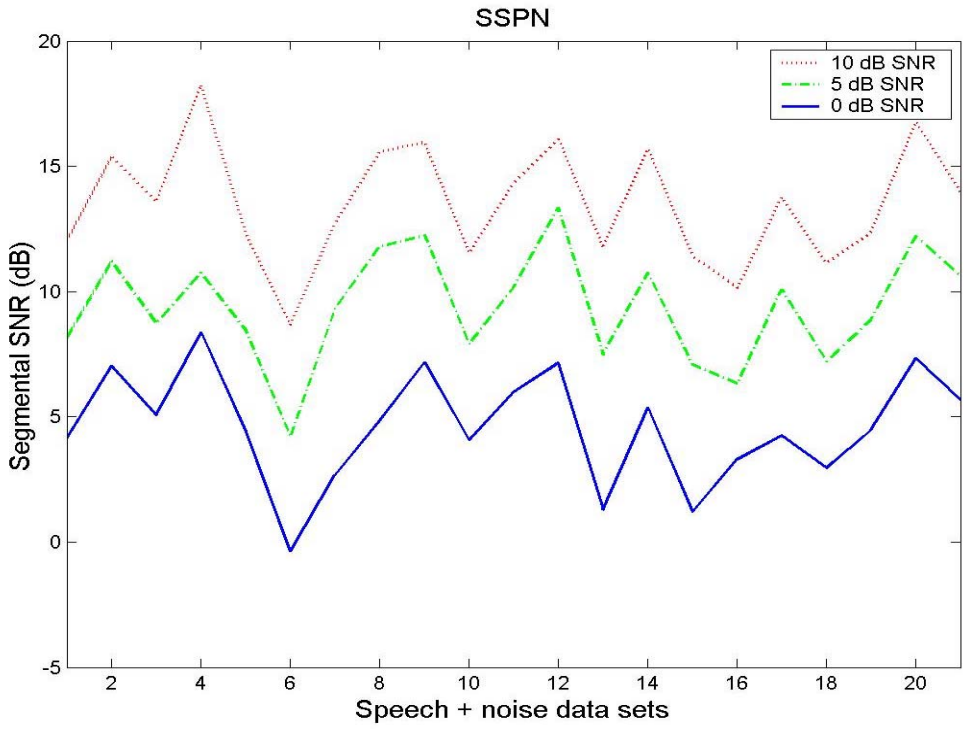


**Figure 5.32: SSNR results for SSPN algorithm**

### 5.11.2 ANOVA analysis of SSNR results

The purpose of the ANOVA analysis is to show that the SSPN algorithm does better than the other two algorithms for a wide set of speech + noise measurements. Two-way ANOVA finds out whether data from several groups have a common mean. One-way ANOVA and two-way ANOVA differ in that the groups in two-way ANOVA have two categories of defining characteristics instead of one. Effects of input SNR and algorithm type are the two characteristics of interest for the SSNR results. Two-way ANOVA can show how much the SSNR results vary with input SNR and with algorithm type. If an algorithm type only works well at a specific input SNR, then this is called an *interaction* between the two characteristics.

The input to the two-way ANOVA analysis function is a 63 x 4 matrix where the columns represent the different algorithm types (original data, beamforming, spectral subtraction, and the SSPN algorithm). The rows are made up of 63 SSNR results divided into groups of 21 for each input SNR level of 0, 5, and 10 dB. If the prediction variances are small compared to the model variances, the model gives a good description of the data. Hence the SSPN algorithm comparison to the other algorithms is accurate and the *means* are significantly different. Table 5.18 contains the results of the two-way ANOVA analysis where the F statistic is calculated as the mean squares divided by the error and used for hypothesis testing. The *mean squares* are equal to the *sum of squares* divided by the *degrees of freedom*. Larger *F* is better because it is an indication that the error is small compared to the effect of the characteristic being analyzed.

| | Sum of squares | Degrees of freedom | Mean squares | F | *p-value* Probability > F |
|---|---|---|---|---|---|
| **Columns** | 1970.84 | 3 | 656.95 | 228.51 | 0.0000 |
| **Rows** | 3711.70 | 2 | 1855.85 | 646.64 | 0.0000 |
| **Interaction** | 16.97 | 6 | 2.83 | 0.98 | 0.4369 |
| **Error** | 689.98 | 240 | 2.87 | | |
| **Total** | 6389.49 | 251 | | | |

**Table 5.18: Two-way ANOVA**

F statistics can be viewed as a measure of significance, so the results in Table 5.18 show that rows (input SNR) have more influence on the data than columns (algorithm type). The F statistics can be used to do hypotheses tests to find out if the SSNR is the same across algorithms, input SNRs, and algorithm-SNR pairs, where the *p-value* (Probability > F) from these tests is given in Table 5.18. The *p-value* for the algorithm effect is zero to four decimal places, which is a strong indication that the SSNR varies from one algorithm to another. An F statistic as extreme as the observed F would occur by chance less than once in 10,000 times if the SSNR were truly equal from algorithm to algorithm. Each pair of the 4 data sets is significantly different. The *p-value* for input SNR effect is also zero, which indicates that the SSNR values are significantly different for each of the 3 data sets. The *p-value*, 0.4369, means that there is high risk (44 out 100 times) of wrongly deciding that there is *interaction*.

Figure 5.33 shows a box-plot of the SSNR result means and variances for the original noisy speech and three enhancement algorithms. Visual inspection of Figure 5.33 shows that the means are different with the closest relationship being between SS and SSPN results. This proximity of SSNR means for SS and SSPN results agrees with the direct view of the SSNR results shown in Figure 5.26, Figure 5.27, and Figure 5.28 from section 5.11.1.

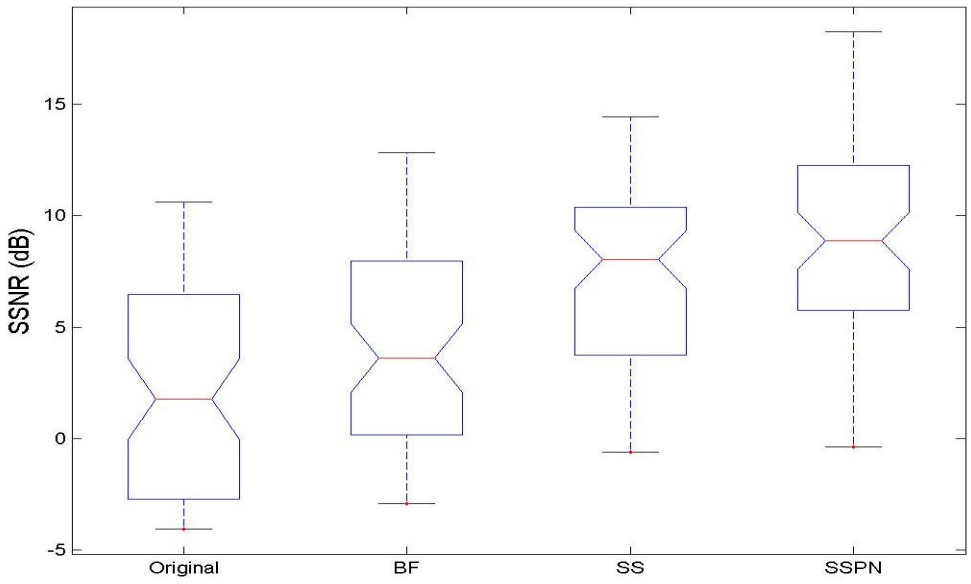


**Figure 5.33: ANOVA box plot for algorithm comparison**

Sometimes it is necessary to determine not just if there are any differences among the means, but specifically which pairs of means are significantly different. The output from the multi-comparison shown in Figure 5.34 indicates the results grouped according to algorithm type are significantly different. The only exception is that SS and SSPN results do appear to be only slightly different. It is interesting to note that the slight SSNR differences between SS and SSPN do not reflect the importance of that gain. SSPN

removes the annoying musical artifacts contained in SS, which is perceptually significant. The statistical analysis of SSNR results is only one view of the enhancement algorithms' performance and must always be considered in the larger context of overall speech quality.

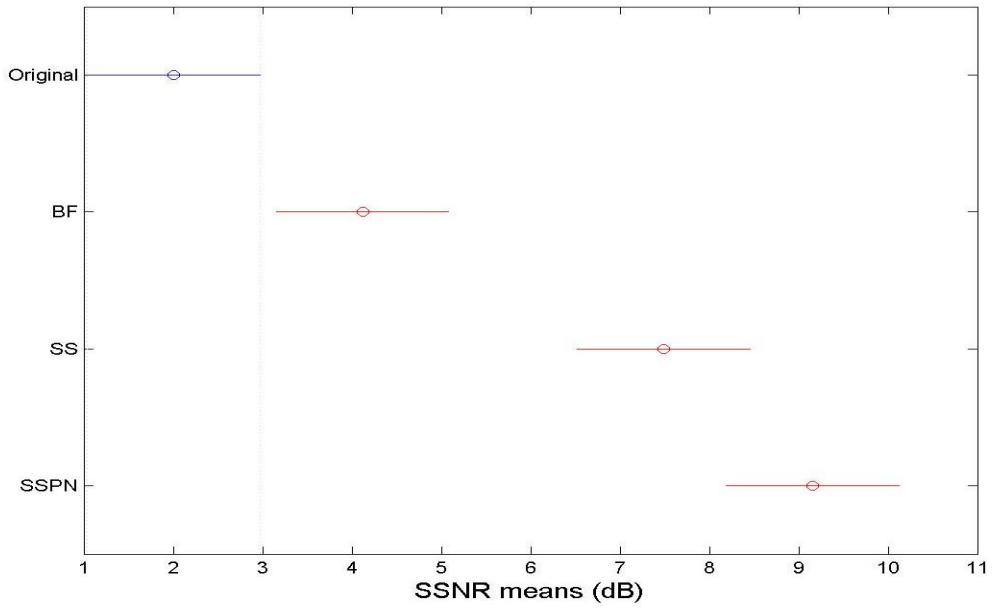


**Figure 5.34: Multi-compare for algorithm type**

Results of the ANOVA for SNR comparison are shown in the plots of Figure 5.35 and Figure 5.36. The SSNR results according to input SNR are significantly different according to *p-value* of 0 and well separated means shown in both figures.

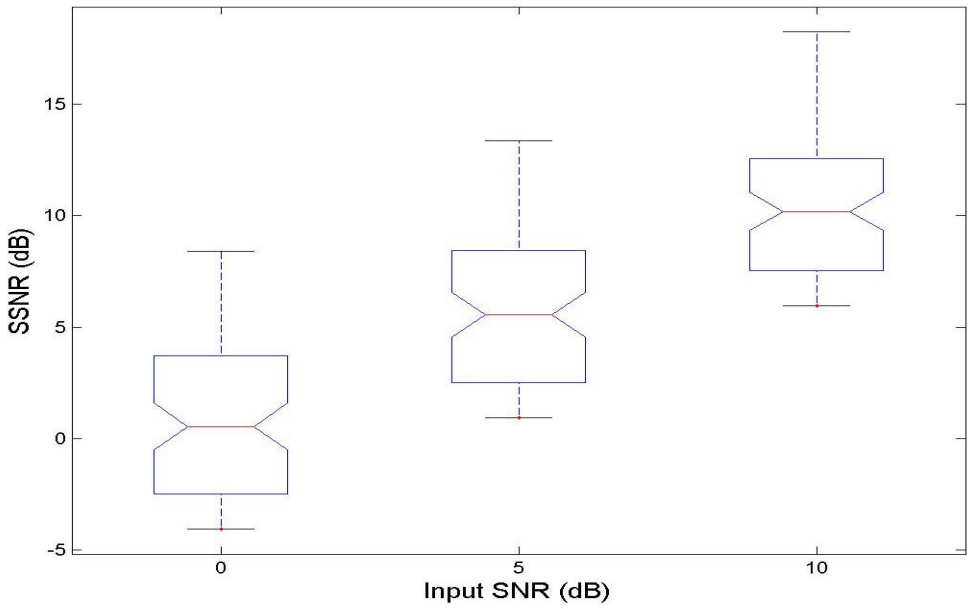


**Figure 5.35: ANOVA box plot for input SNR comparison**

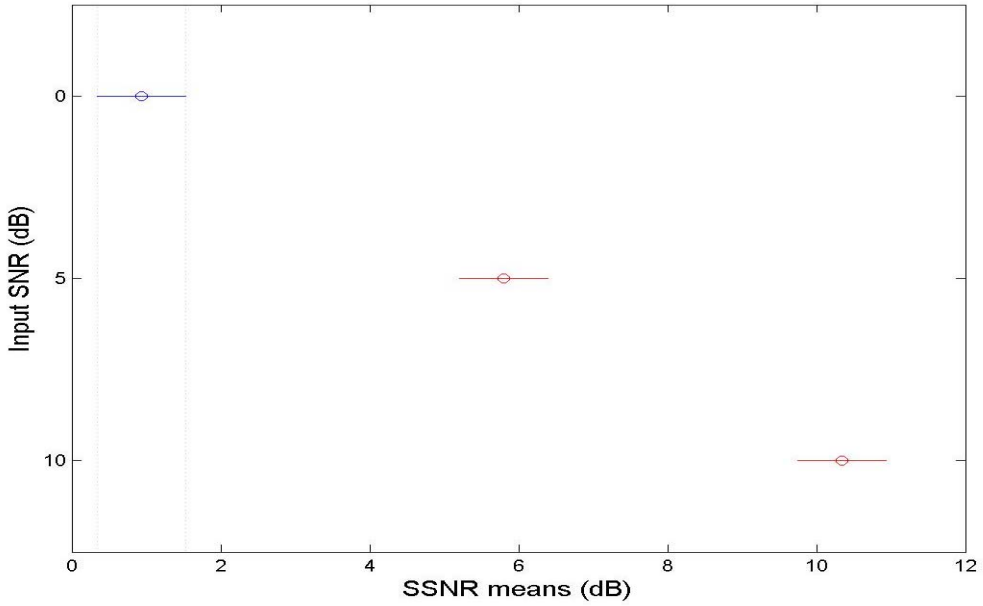


**Figure 5.36: Multi-compare for input SNR**

# Chapter 6

# Conclusions and future work

## 6.1 Conclusions

The SSPN algorithm contained a unique combination of complementary signal processing that was presented and evaluated in this thesis. The combinations of signal processing, in most cases, exceeded the performance of simpler techniques. High-level speech enhancement algorithms must be carefully designed and not simply cascaded one after another expecting the results to improve proportionally. Each algorithm benefits the other in a unique way beginning with beamforming in the SSPN algorithm.

**Beamforming** is used first because the other algorithms would not help its performance greatly and the multi-channel processing collapses to a less computationally intensive single channel problem. GSC beamforming takes advantage of the directional information in the signal to attenuate the noise while introducing only minimal distortion of the desired speech. An important aspect of the SSPN algorithm is that it uses the VAD to instruct the GSC to update its filter coefficients only during frames not containing speech, which avoids attenuating the desired speech. Experiments suggest that the beamformer is not effective at creating a separate noise reference channel to be used in

subsequent spectral subtraction because the noise in the separate channel does not correlate well with the noise in the main speech channel. The effective frequency range of the beamformer is approximately 800 to 3400 Hz because of the fixed spacing of the microphones and length of the array, as explained in 4.1.2. The far field assumption is only valid for a subset of frequencies, so including spherical spreading effects will improve performance at frequencies where the far field assumption is not valid.

**Spectral subtraction** supplements the GSC beamformer's limited frequency range by attenuating noise in the lower frequencies subtraction as shown by the results in Chapter 5. Additional noise reduction in the lower frequencies is important in the car because most of the noise sources are heavily weighted below 1 kHz. Spectral subtraction operates on the single channel output of the beamformer and acts independently of the beamformer. Enhancing the signal before spectral subtraction will improve the accuracy of the VAD, which is crucial for SS. This arrangement of beamforming and spectral subtraction also allows the enhancement gains to be linearly combined. However, simple spectral subtraction can introduce undesirable audible artifacts.

**Perceptual masked threshold frequency weighting** is very effective at both minimizing musical noise artifacts and attenuating the noise further than simple half-wave rectified SS. The perceptual frequency weighting is so effective that any further modification to the generalized spectral subtraction parameters did not help. Using an approximate pitch estimate to modify the gains in the corresponding critical band did not improve speech quality because the perceptual mask threshold already accounted for high signal energy in

that band. This experiment with pitch detection was done using the clean speech to ensure accurate pitch information, but the lack of improvement was the same. Another experiment was to double the amount of frequency bands used below 1 kHz in an attempt to further decrease the noise in narrower bands while leaving behind more of the desired signal. However the higher resolution frequency bands did not offer any improvement over dividing the spectrum into critical bands using the Bark scale. Despite the noise attenuation and good signal quality achieved by using perceptually weighted spectral subtraction based on the Bark scale, non-stationary noises such as interfering talkers were not removed.

**Talker separation** is necessary because beamforming and spectral subtraction do a very poor job at removing this type of noise as shown in Chapter 5. Talker separation and pitch tracking require detailed feature analysis and benefit from the removal of the stationary noise and directionally strengthened speech performed by the previous processing. Talker separation also has potential for improving the accuracy of the VAD, which is so important to the SSPN algorithm.

**The accuracy of the VAD plays a central role** for both the spectral subtraction noise estimate and the decision to adapt the filters in the GSC beamformer. Therefore, a substantial improvement is obtained by feeding back the initial speech estimate from a first pass through the system and using it to drive the VAD as the original input is processed again. Talker separation will avoid falsely classifying data frames as voiced when interfering talkers are speaking while the desired talker is silent. Even without the

talker separation the VAD is improved by limited iteration with spectral subtraction. In doing this, the VAD also becomes less sensitive to the value of the fixed constant multiplier used to determine the speech threshold. The reduced sensitivity will enable the algorithm to perform better over a wider range of noise conditions. The VAD and the overall algorithm performance will be better if the desired speech signal is stronger at the inputs because the microphones were installed in optimal locations.

**Microphone placement** in the car is a very important part of the system design. Placing the microphone(s) over the head and slightly forward of the desired talker is the optimal location in an automobile. The evidence that supports the theory behind this assumption is shown in Appendix B – Microphone Location.

Finally, this unique combination of signal processing algorithms, SSPN, also has the advantage of a modular design. The beamformer, VAD, noise estimate, spectral subtraction, and talker separation can each be modified independently to incrementally introduce advanced techniques and improve the overall performance of the system.

## 6.2 Future work

Ideally the noise suppression system for hands-free phones and speech recognition would remove all the interfering noise in the car. Stationary noise sources are readily removed as demonstrated in this thesis. However, there are various noise sources that remain and continued research is required for their removal. Some of these challenging noise sources and suggested research are listed below.

1. Talker isolation to remove interfering speech
2. Speaker dependent voice extraction from the noise
3. Noise suppression specific to the make and model of automobile
4. Removing the non-stationary noise of passing cars
5. Removal of music played by the stereo
6. Scaling the system up to include multiple talkers in the car
7. Optimization of the SSPN algorithm with respect to the VAD

Further work on talker isolation in the car is required. The current techniques are inadequate at removing interfering talkers. Interfering speech is most prominent in the millions of min-vans full of children out there.

Speaker dependent enhancement is another area well suited to the automobile. Typically, the same few people will be driving a certain automobile, so the opportunity to obtain training data is excellent. Speaker dependent processing would help retain valuable speech energy in the lower part of the spectrum normally swamped by road and engine noise. If the approximate pitch of the driver was 180Hz, then a bandpass filter around that frequency could be used to attenuate the other low frequencies, which are likely due to noise.

Signal processing that is specific to a particular make of automobile could also offer significant improvements over generic noise reduction algorithms. Each car has slightly

different noise characteristics due the engine, tires, interior design, materials used to construct the car, positions of the vents, fans, and many other considerations. Attenuating noise frequencies for each make of car, for example, would allow optimal use of the spectrum. Measuring the source to microphone transfer functions in each car could help de-reverberate the signal.

Future research into removing the non-stationary noise of passing cars is required. The removal of passing car noise and other short transient noise sources is still an unsolved problem. The non-stationary noise sources require continuous update of the noise estimate.

Future work should include Echo Cancellation and removal of music from the received signal, which will be required of any practical speech acquisition system in a car. Investigation into scaling the system up to include multiple pairs of microphones would allow optimal processing for multiple talkers while at the same time improve each individual speech signal. One could envision persons in the back seat of a car wanting to participate in a hands-free phone call.

# Chapter 7

# Appendix A – Psychoacoustics

It is useful to have some understanding of how the human hearing system works when discussing speech enhancements based on perception. The pictures, diagrams, and plots in this chapter where copied from Hyper Physics web pages at Georgia State University[107].

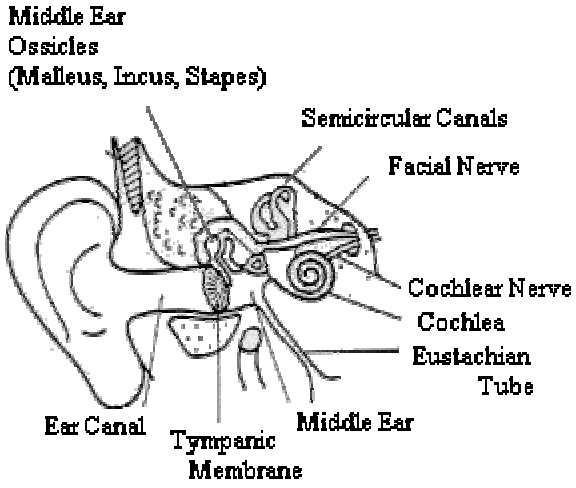


**Figure 7.1: Human Ear**

As seen in Figure 7.1 the ear has three major parts: the outer, middle, and inner ear. The pinna or outer ear-the part of the ear attached to the head, funnels sound waves through

the outer ear. The sound waves pass down the auditory canal to the middle ear, where they strike the tympanic membrane, or eardrum, causing it to vibrate. The vibrations are received by three small bones (ossicles) in the middle ear; named for their shapes: the malleus (hammer), incus (anvil), and stapes (stirrup). The stirrup is attached to a thin membrane called the oval window, which is much smaller than the eardrum and consequently receives more pressure. As the oval window vibrates from the increased pressure, the fluid in the coiled, tubular cochlea (inner ear) begins to vibrate the membrane of the cochlea (basilar membrane), which bends fine hair-like cells on its surface. These auditory receptors generate miniature electrical forces, which trigger nerve impulses that then travel via the auditory nerve, first to the thalamus and then to the primary auditory cortex in the temporal lobe of the brain. The impulses are relayed to association areas of the brain, which convert them into meaningful sounds by examining the activity patterns of the neurons, or nerve cells, to determine sound frequencies. Although the ear changes sound waves into neural impulses, it is the brain that actually "hears," or perceives the sound as meaningful.

The auditory system contains about 25,000 cochlear neurons that can process a wide range of sounds. The sounds humans hear are determined by two characteristics of sound waves: their amplitude (the difference in air pressure between the peak and baseline of a wave) and their frequency (the number of waves that pass by a given point every second). Loudness of sound is influenced by a complex relationship between the wavelength and amplitude of the wave; the greater the amplitude, the faster the neurons fire impulses to the brain, and the louder the sound that is heard. Loudness of sound is usually expressed

in decibels (dB). A whisper is about 30 dB, normal conversation is about 60 dB, and a subway train is about 90 dB. Sounds above 120 dB are generally painful to the human ear. The loudest rock band on record was measured at 160 dB.

The normal frequency range of human hearing is 20 to 20,000 Hz. Frequencies of some commonly heard sounds include the human voice (120 to approximately 1,100 Hz), middle C on the piano (256 Hz), and the highest note on the piano (4,100 Hz). Differences in frequency are discerned, or coded, by the human ear in two ways, frequency matching and place. The lowest sound frequencies are coded by frequency matching, duplicating the frequency with the firing rate of auditory nerve fibers. Frequencies in the low to moderate range are coded both by frequency matching and by the place on the basilar membrane where the sound wave peaks. High frequencies are coded solely by the placement of the wave peak.

The organ of Corti is the sensitive element in the inner ear and can be thought of as the body's microphone as displayed in Figure 7.2: Hair cells on basilar membrane. It is situated on the basilar membrane in one of the three compartments of the Cochlea. It contains four rows of hair cells, which protrude from its surface. Above them is the tectoral membrane which can move in response to pressure variations in the fluid- filled tympanic and vestibular canals. There are some 16,000 -25,000 of the hair cells distributed along the basilar membrane which follows the spiral of the cochlea.

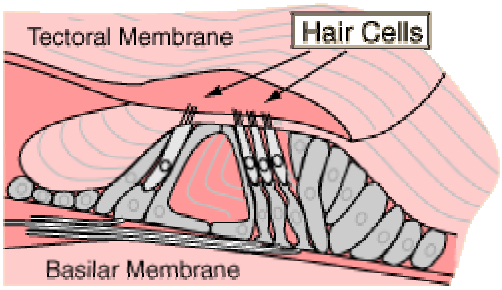


**Figure 7.2: Hair cells on basilar membrane**

The place along the basilar membrane where maximum excitation of the hair cells occurs determines the perception of pitch according to the place theory as shown in Figure 7.3. The perception of loudness is also connected with this organ. High frequency sounds selectively vibrate the basilar membrane of the inner ear near the entrance port (the oval window). Lower frequencies travel further along the membrane before causing appreciable excitation of the membrane. The basic pitch determining mechanism is based on the location along the membrane where the hair cells are stimulated.

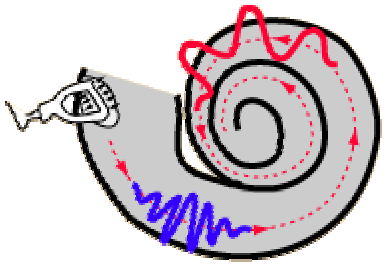

**Figure 7.3: Cochlea**

A schematic view, in Figure 7.4, of the place theory unrolls the cochlea and represents the distribution of sensitive hair cells on the organ of Corti. Pressure waves are sent through the fluid of the inner ear by force from the stirrup.

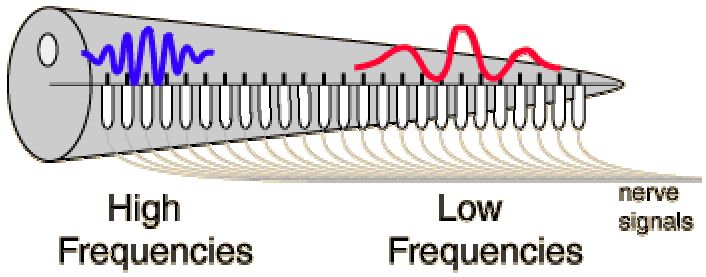


**Figure 7.4: Cochlea Frequency Selectivity**

Whether a sound can be heard depends on its intensity and spectrum, and perceptibility is discussed in terms of hearing thresholds. The inner ear's vibration and neural firings are highly nonlinear with the result that perception of sound energy at one frequency is dependent on the distribution of sound energy at other frequencies. The amount of energy before and after a sound on the time scale also effects it's perception. The phenomenon of masking is when the perception of one sound is obscured by the presence of another. *Frequency masking* occurs when sounds occur simultaneously and *temporal masking* happens when there is a delay between the sounds. The auditory system can be considered as a bank of band-pass filters. Frequency resolution of the ear's filtering mechanism is known as *critical bands*. The widths of the critical bands increase with increasing center frequency. The frequency resolution can be measured in terms of the minimum frequency separation at which two tones can be distinguished. [92]

# Chapter 8

# Appendix B – Microphone Location

Location of the microphones is extremely important because it determines how much noise vs. speech will be received. In a free field, the strength of the source is proportional to the inverse square of its distance from the microphone, corresponding to a 6 dB decrease in intensity for each doubling of distance. The noise suppression algorithms must work much harder as the distance of the talker from the microphone increases because of the weaker signal as shown using rough approximations in Table 8.1.

| Location of microphone | Distance to talker in (cm) | Equivalent intensity in a free field (dB) | Equivalent noise suppression (dB) |
|---|---|---|---|
| Headset | 2 | 32 | 4 |
| Car ceiling above drivers head | 8 | 8 | 28 |
| Conference phone table | 32 | 2 | 34 |
| On rear view mirror in car | 64 | 1 | 35 |

**Table 8.1: Microphone distance and required noise suppression**

Microphones inside an automobile are not in a free field and some desired signal energy would reach them from reflections, so the free field approximations in Table 8.1 should be taken in that context. However, the table does not show how much the noise energy can increase as microphones are moved closer to the noise, which will also be a big factor in determining the required noise suppression. Suggested by results in this thesis and prior research, the optimal location for the microphone(s) in the car is on the ceiling above the driver's head and slightly forward. The reasons for this location are listed below.

1. It is the closest place to the driver's mouth other than the steering wheel.
2. The steering wheel is not a practical location because the air bag deployment needs to be un-obstructed. Considerable vibration from the engine and road is also transmitted along the steering column.
3. The microphone needs to be away from wind currents coming from the window.
4. Keeping the microphone away from the windshield avoids the wind from the defroster, windshield wiper noise, and rain noise.
5. The directionality of the talker's voice does not require the microphone to very far in front of the talker. Distance is the dominating factor.
6. The microphone can easily be insulated against noise from the roof and wind inside the car.
7. It does not impair the driver's vision.

The biggest drawback to locating the microphone on the ceiling of the car is that automakers have to complicate their manufacturing process to install it. Automakers prefer the center console or the rear view mirror because it is easier to install.

An experiment was done to compare two pairs of microphones at different locations. The signals used were 312.5 ms of speech sampled at 16 kHz. There were only two microphones available, so only two positions were compared simultaneously. Comparisons were made between microphones A & B and between A & C.

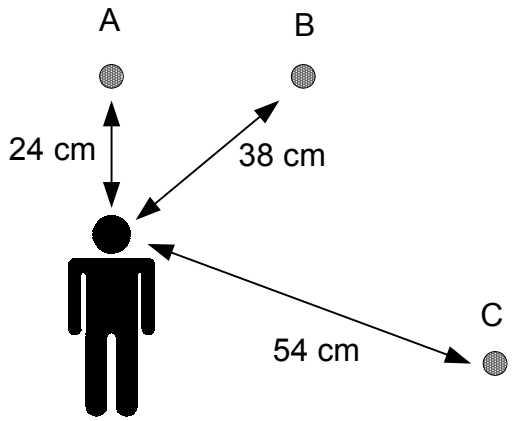


**Figure 8.1: Microphone positions**

A. Microphone A is on the ceiling of the car over the talker's head and slightly forward. The distance between the talker and microphone is 10 cm.
B. Microphone B is on the ceiling of the car close to the windshield in front of the talker. The distance between the talker and microphone is 24 cm.

C. Microphone C is on the rear view mirror in the center of the car. The distance between the talker and microphone is 38 cm.

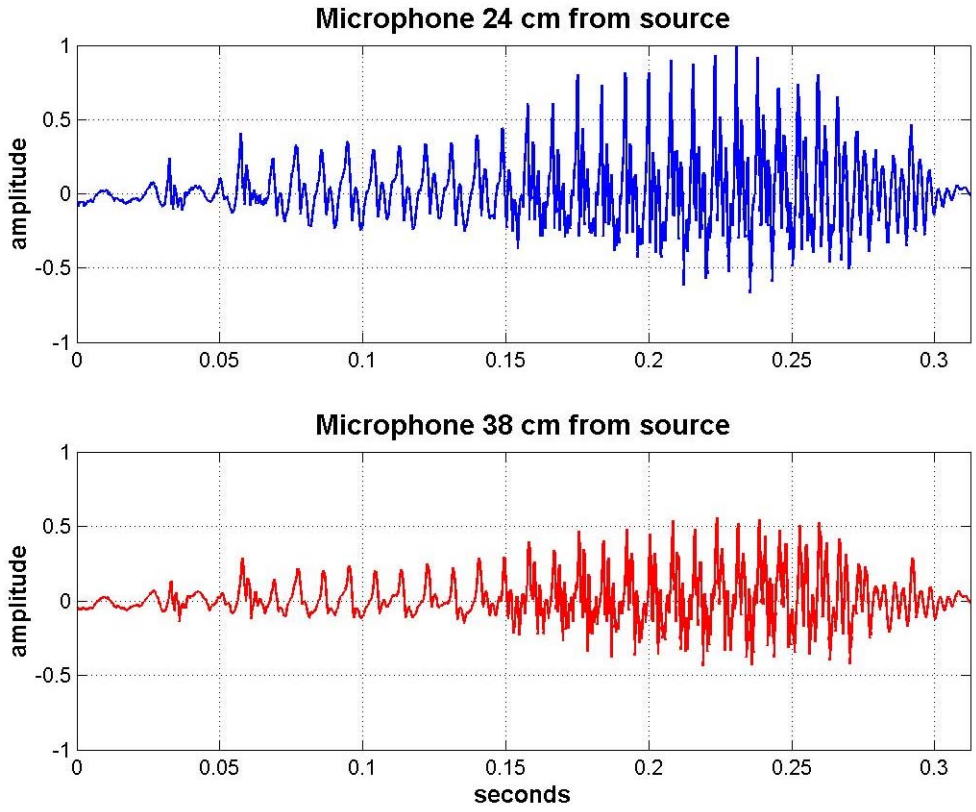


**Figure 8.2: Comparison between 24cm and 38 cm microphone distances**

The source signal strength difference in Figure 8.2 clearly shows the advantage of the closer microphone. The ratio of power between the 24 cm and 38 cm position is 2.21. The inverse ratio of the distances is 38 / 24 = 1.58.

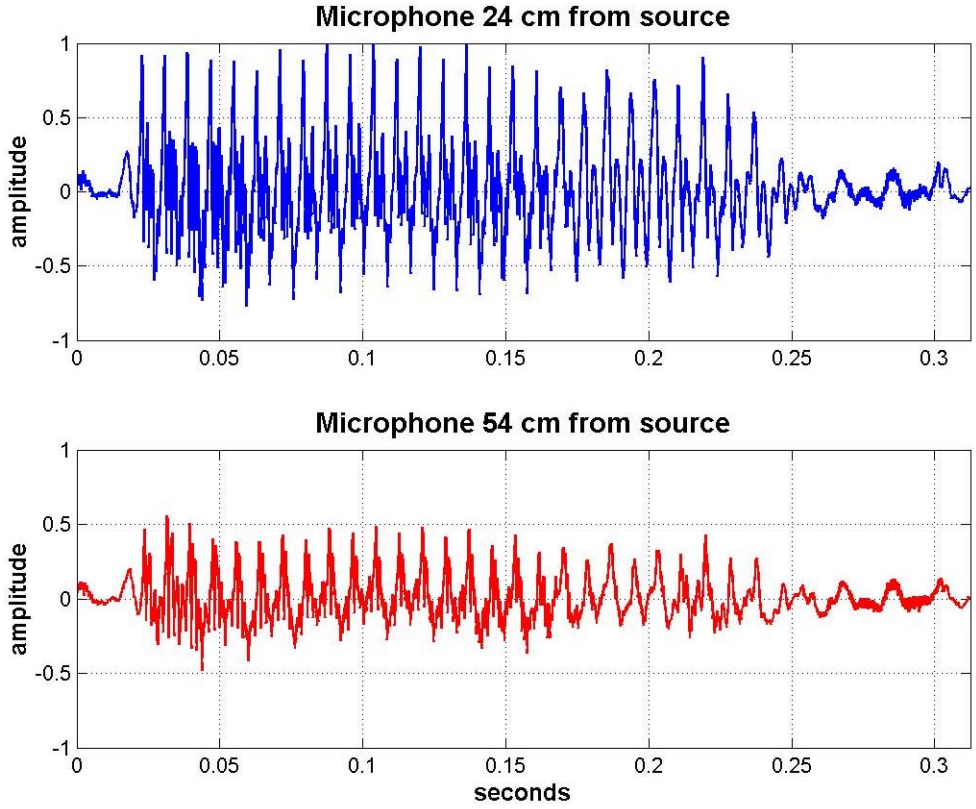


**Figure 8.3: Comparison between 24cm and 54 cm microphone distances**

Figure 8.3 shows how the signal strength scale with the farther distance at 54 cm. The ratio of power between the 24 cm and 54 cm position is 4.12. The inverse ratio of the distances is 54 / 24 = 2.25.

A similar conclusion was reached by research done in 1994 about the best microphone positions in an automobile for speech recognition. After analyzing 7 different locations the conclusion stated, "The position at the ceiling right in front of the speaker gave the best results." [108]

# References

The following references are formatted using the style defined by the Chicago Manual of Style.[109] [110]


[1]Brian Demers, "The Wireless Renaissance Of Speech Recognition," Communications Solutions, November 2001, < http://www.tmcnet.com/comsol/1101/1101demers.htm> (May 2002).

[2] S. Hayki, Adaptive Filter Theory, 3rd edition, (Upper Saddle River, N.J.: Prentice Hall, 1996), 194-196.

[3] B. Widrow, "Adaptive noise canceling: Principles and Application," Proceedings of IEEE 63(1975):1692-1716.

[4] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Transactions on Acoustics, Speech, and Signal Processing, 27 (1979):113-120.

[5] B. D. Van Veen and K.M. Buckley, "Beamforming: a versatile approach to spatial filtering," IEEE Transactions on Acoustics, Speech, and Signal Processing 5, no. 2, (April 1988):4-24.

[6] Lloyds J. Griffiths and Charles Jim, "An alternative approach to linearly constrained adaptive beamforming," IEEE Transactions on Antennas and Propagation AP-30, no. 1, (January 1982):27-34.

[7] Weinstein, Feder, and Oppenheim, "MultiChannel Signal Separation by Decorrelation", IEEE Transactions on Speech and Audio Processing 1, no. 4, (October 1993):405-413.

[8] Attias and Schriener, "Blind Source Separation and Deconvolution by Dynamic Component Analysis," Sloan Center for Theoretical Neurobilogy, University of California, San Francisco, CA, Proceedings of IEEE, (1997).

[9] John R. Deller Jr., John H.L. Hansen, and John G. Proakis, Discrete-Time Processing of Speech Signals, (New York, NY: IEEE Press, 2000):266.

[10] Tsoukalas, Mourjopoulos, Kokkinakis, "Audio Noise Cancellation Using Subjective Signal Representation", IEEE Signal Processing, (1997):613-616.

[11] Tsoukalas, Mourjopoulos, and Kokkinakis, "Speech Enhancement Based on Audible Noise Suppression," IEEE Transactions on Speech and Audio Processing 5, no. 6, (Nov 1997).

[12] John R. Deller Jr., John H.L. Hansen and John G. Proakis, Discrete-Time Processing of Speech Signals, (New York, NY: IEEE Press, 2000):266.

[13] Gail Erten, "Voice Signal Extraction for Enhanced Speech Quality in Noisy Vehicle Environments," Proceedings of IEEE, (1999).

[14] Meyer and Simmer, "Multi-Channel Speech Enhancement in a Car Environment Using Wiener Filtering and Spectral Subtraction", Proceedings of IEEE, (1997):1167-1170.

[15] Dahl, Claesson, and Nordebo, "Simultaneous Echo Cancellation and Car Noise Suppression Employing a Microphone Array", Proceedings of the IEEE, (1997):239-242.

[16] Hardwick, Yoo, and Lim, "Speech Enhancement Using the Dual Excitation Speech Model, "Proceedings of the IEEE, (1993).

[17] Rolf Vetter, "Single Channel Speech Enhancement Using MDL-Based Subspace Approach in Bark Domain," Proceedings of the IEEE, (2001):641-644.

[18] S. R. Quackenbush, T. P. Barnwell, M. A.Clements, Objective Measures of Speech Quality, (Englewood Cliffs, NJ: Prentice Hall, 1988): 37-50.

[19] Iain A. McCowan, C. Marro, and L. Mauuary, "Robust speech recognition using near-field superdirective beamforming with post-filtering", ICASSP Proceedings, (2000): 1723 –1726.

[20] B. Scharf, Foundations of Modern Auditory Theory, (New York: Academic, 1970):ch. 5.

[21] E. Zwicker and H. Fastl, Psychoacoustics: Facts and Models, (Berlin, Germany: Springer-Verlag, 1990).

[22] http://www.minidisc.org/MaskingPaper.html

[23] http://people.cs.uchicago.edu/~odonnell/OData/Courses/CS295/perception.html

[24] E. Zwicker, H. Fastl, and Frater, Psychoacoustics: Facts and Models, 2nd Edition, (Springer, April 1999):149-173.

[25] D. Rabinkin, "Optimum Sensor Placement for Microphone Arrays," (PhD Thesis, Rutgers, New Brunswick, New Jersey, May 1998).

[26] Smolders, Claes, Sablon, and Van Compernolle, "On the Importance of the Microphone Position for Speech Recognition in the Car," Proceedings of the IEEE, (1994).

[27] Nathanial A Whitmal, Janet C. Rutledge, "Noise Reduction in Hearing Aids: A Case For Wavelet-based Methods," Proceedings of the 20th Annual International Conference of the IEE Engineering in Medicine and Biology Society 20, no. 3, (1998).

[28] Wahab and Chong, "Intelligent dashboard with Speech Enhancement," Proceedings of the IEEE, (September 1997).

[29] Petr Pollak, Pavel Sovka, and Jan Uhlir, "Noise Suppression System for a Car," Department of Circuits Theory, Czech Technical University in Proague, Prague, Czech Republic, Proceedings of the IEEE, (1994).

[30] W.B. Kleijn and K.K. Paliwal, Speech Coding and Synthesis, (Amsterdam: The Netherlands, Elsevier, 1995):529-530.

[31] Gustafsson, Claesson, Norholm, & Lindgren, "Dual Microphone Spectral Subtraction," Research Report, (University of Karlskrona Ronneby, Ronneby, Sweden, February 2000).

[32] Abdul Wahab and Eng Chong Tan, "CMAC Spectral Subtraction for Speech Enhancement," Proceedings of the IEEE, (2001).

[33] Martin Rainer, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," IEEE Transactions on Speech and Audio Processing 9, no. 5, (July 2001).

[34] Kim Asano, Suzuki, and Sone, "Speech Enhancement Base on Short-Time Spectral Amplitude Estimation with Two-Channel Beamformer," EICE Trans. Fundamentals E79-A, no. 12, (December, 1996).

[35] Hwai-Tsu Hu, Fang-Jang Kuo, Hsin-Jen Wang, "Supplementary schemes to spectral substation for speech enhancement," Speech Communication, (Elsevier Science, 2002):205-218.

[36] R. Sankar, "Pitch Extraction Algorithm for Voice Recognition Applications," Proceedings of the IEEE, (1998).

[37] Ching and Toh, "Enhancement of Speech Signal Corrupted by High Acoustic Noise," Proceedings of the IEEE, (1993).

[38] M. S. Brandstein and H. F. Silverman, "A Practical Methodology for Speech Source Localization With Microphone Arrays," Harvard University, Cambridge, MA, (November 13, 1996).

[39] Nagai, Kondo, Kaneko, and Kurematsu, "Estimation of Source Localization Based on 2-D MUSIC and its Application to Speech Recognition in Cars," Proceedings of the IEEE, (2001).

[40] J. C. Chen, K. Yao, and R. E. Hudson, "Source Localization and Beamforming in a Distributed Sensor Network", DARPA-ITO contrant N66001-001-8937, UCLA, (August, 4, 2001).

[41] Nishiura, Yamada, Nakamura, and Shikano, "Localization of Multiple Sound Sources Based on a CSP Analysis with a Microphone Array," Proceedings of the IEEE, (2000).

[42] D. Rabinkin, R. Renomeron, A. Dahl, J. French, J. Flanagan, and M. Bianchi, "A DSP Implementation of Source Location Using Microphone Arrays," Rutgers University, (1996).

[43] I. McCowan, D. Moore, S. Sridharam, "Near-field Adaptive Beamformer for Robust Speech Recognition", *In Proceedings of ICASSP 2000*, no. 3 (2000): 1723–1726.

[44] J. Bitzer, K. Simmer, and K D. Kammeyer, Speech Communication, (Germany, Elsiver, 2001):8.

[45] Dimitris G. Manolakis, Vinay K. Ingle and Stephen M. Kogon, Statistical and Adaptive Signal Processing, (Boston: McGraw-Hill Higher Education, 2000):633.

[46] Frost, O. L., "An Algorithm for Linearly Constrained Adaptive Array Processing," Proceedings of the IEEE 60, no. 8 (1972):926-935.

[47] L. J. Griffiths and C. W. Jim, "An Alternative Approach to Linearly Constrained Adaptive Beamforming," IEEE Transactions on Antennas and Propagation 30, no. 1 (January 1982):27-34.

[48] D. Van Compernolle, S. Van Gerven, W. Broos, and L. Weynants, "A Real-Time Griffiths-Jim Beamformer for Speech Applications," Proceedings of the IEEE & ProRISC Symposium on Circuits, Systems, and Signal Processing, (Veldhoven:The Netherlands, April 3-4, 1991):147-150.

[49] S. Hayki, Adaptive Filter Theory, 3rd edition, (Upper Saddle River, N.J.: Prentice Hall, 1996), 366.

[50] S. Hayki, Adaptive Filter Theory, 3rd edition, (Upper Saddle River, N.J.: Prentice Hall, 1996), 433-437.

[51] W.B. Kleijn and K.K. Paliwal, Speech Coding and Synthesis, (Amsterdam: The Netherlands, Elsevier, 1995):267.

[52] Lee, Stern, Mahmoud, "A Voice Activity Detection Algorithm for Communication Systems with Dynamically Varying Background Acoustic Noise," Proceedings of the IEEE, (1998).

[53] Benyassine, Shlomot, Su, Yuen, "A Robust Low Complexity Voice Activity Detection Algorithm for Speech Communication Systems," Proceedings of the IEEE, (1997).

[54] El-Maleh, Kabal, "Comparison of Voice Activity Detection Algorithms for Wireless Personal Communications Systems," Proceedings of the IEEE, (1997).

[55] Vahatalo, Johansson, "Voice Activity Detection for GSM Adaptive Multi-rate CODEC," Proceedings of the IEEE, (1999).

[56] Garner, Barret, Howard, and Tyrrell, "Robust Noise Detection for Speech Detection and Enhancement," Electronics Letters 33, no. 4, (February 13, 1997).

[57] Tanyer, Ozer, "Voice Activity Detection in Non-stationary Noise," IEEE Transactions on Speech and Audio Processing 8, no. 4, (July 2000).

[58] Rabiner and Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances," The Bell System Technical Journal, (February 1975):297-315.

[59] Woo, Yang, Park, Lee, "Robst Voice Activity Algorithm for Estimating Nose Spectrum," Electronics Letters 36, no. 2, (January 20, 2000).

[60] Junqua, Reaves, and Mak, "A Study of Endpoint Detection Algorithms in Adverse Conditions: Incidence on a DTW and HMM recognize," Proceedings Eurospeech '91, (1991):1371-1374.

[61] Tucker, "Voice Activity Detection Using a Periodicity Measure," IEEE Proceedings-I 139, no. 4, (August 1992).

[62] Irwin, M.J., "Periodicity Estimation in the Presence of Noise", Inst. Acoustics Conference, (Windemere: United Kingdom 1979), and JSRU Report 1009, (1980).

[63] W.B. Kleijn and K.K. Paliwal, Speech Coding and Synthesis, (Amsterdam: The Netherlands, Elsevier, 1995):268.

[64] D. L. Wang, and J. S. Lim, "The unimportance of phase in speech enhancement," IEEE Transactions on Acoustics, Speech, and Signal Processing 30, (Aug1982):679-681.

[65] Oliver Cappe, "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor," Proceedings of the IEEE, (1994).

[66] Ghoreishie and Sheikhzadeh, "A Hybrid Speech Enhancement System Based on HMM and Spectral Subtraction," Proceedings of the IEEE, (2000).

[67] Goh, Tan, and Tan, "Post-processing Method for Suppressing Musical Noise Generated by Spectral Subtraction," IEEE Transactions on Speech and Audio Processing 6, no. 3, (May 1998).

[68] Nathanial A Whitmal, Janet C. Rutledge, and Johnathan Cohen, "Reducing Correlated Nose in Digital Hearing Aids," IEEE Engineering in Medicine and Biology, (September/October, 1996).

[69] Wu and Chen, "Efficient Speech Enhancement using Spectral Subtraction for Car Hands-free Applications," Proceedings of the IEEE, (2001).

[70] Seok and Bae, "Reduction of musical noise in spectral subtraction method using sub-frame randomization," Electronics Letters 35, no. 2, (January 1999).

[71] Ogata and Shimamura, "Reinforced Spectral Subtraction Method to Enhance Speech Signal," Proceedings of the IEEE, (2001).

[72] Y. Ephraim and D. Malah, "Speech enhancement using optimal non-linear spectral amplitude estimation," Proceedings IEEE Int. Conference on Acoustics Speech Signal Processing, (Boston, 1983):1118-1121.

[73] Y.Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-32, no. 6, (1984):1109-1121,

[74] Sen Kuo, "Adaptive Acoustic Noise Cancellation Microphone," Proceedings of the IEEE, (1996).

[75] Widrow and Stearns, Adaptive signal processing, (Englewood Cliffs, NJ: Prentice-Hall, 1985).

[76] E. Weinstein, Feder, and Oppenheim, "Multi-channel Signal Separation by Decorrelation," IEEE Transactions on Speech and Audio Processing 1, no. 4, (October 1993).

[77] Yellin and Weinstein, "Multi-channel Signal Separation: Methods and Analysis," IEEE Transactions on Signal Processing 44, no. 1, (January 1996).

[78] Kuan-Chieh Yen, and Yunxin Zhao, "Adaptive C-Channel Speech Separation and Recognition," IEEE Transactions on Speech and Audio Processing 7, no. 2, (March 1999).

[79] R. Goubran, R. Hebert, and H. Hafez, "Acoustic Noise Suppression Using Regressive Adaptive Filtering," Proceedings of the IEEE, (1990).

[80] Itoh and Mizushima, "Environmental Noise Reduction Based on Speech/Non-speech Identification for Hearing Aids," Proceedings of the IEEE, (1997).

[81] Shields and Campbell, "Multi-Microphone Noise Cancellation For Improvement of Hearing Aid Performance," Proceedings of the IEEE, (1998).

[82] BohhPoh NG Zhang and Khoon Seong Lim, "A Practical Noise Suppression Method Using Microphone Array," Proceedings of the IEEE, (1985).

[83] H.G. Hirsch, "Noise Estimation Techniques for Robust Speech Recognition," Proceedings of the IEEE, (1995).

[84] Kim, Asano, Susuki, and Sone, "Speech Enhancement Based on Short-Time Spectral Amplitude Estimation with a Two-Channel Beamfomer," IEICE Trans. Fundamentals E79, no. 12, (December 1996).

[85] Daniel Rabinkin, "Optimum Sensor Placement for Microphone Arrays," (PhD Thesis Dissertation, Rutgers University, New Brunswick, New Jersey, May, 1998).

[86] Rainer Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," IEEE Transactions on Speech and Audio Processing 9, no. 5, (July 2001).

[87] Rainer Martin, "Spectral Subtraction Based on Minimum Statistics," Proceedings of EUSIPCO-94, (Seventh European Signal Processing Conference, Edinburgh: Scotland, U.K, September 13-16, 1994).

[88] Christophe Ris, and Stephane Dupont, "Assessing Local Noise Level Estimation Methods: Application to Noise Robust ASR," Faculte Polytechnique de Mons, Multitel, (Parc Initialis, B-7000 Mons: Belgium, 2000).

[89] Tuffy and Laurenson, "Estimating Clean Speech Thresholds for Perceptual Based Speech Enhancement," Proceedings of 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, (New Paltz, New York, October17-20, 1999).

[90] D. Tsoukalas, Mourjopoulos, Kokkinakis, "Audio Noise Cancellation Using A Subjective Signal Representation", Proceedings of the IEEE, (1997).

[91] Nathalie Virag, "Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System," IEEE Transactions on Speech and Audio Processing 7, no. 2, (March 1999).

[92] Li, McAllister, Black, and De Perez, "Perceptual Time-Frequency Subtraction Algorithm for Noise Reduction in Hearing Aids," IEEE Transactions on Biomedical Engineering 48, no. 9 (September 2001).

[93] M. Lorber and R. Hoeldrich, "A Combined Approach for Broadband Noise Reduction," Institute of Electronic Music Graz, (Graz: Austria, 1997).

[94] Vetter, "Single Channel Speech Enhancement Using MDL-Based Subspace Approach in Bark Domain," Proceedings of the IEEE, (2001).

[95] Kim, Kang, and Ko, "Spectral Subtraction Based on Phonetic Dependency and Masking Effects," IEEE Proceedings on Visual Image and Signal Processing 147, no. 5, (October 2000).

[96] Nandkumar and Hansen, "Dual-Channel Iterative Speech Enhancement with Constraints on an Auditory-Based Spectrum," IEEE Transactions on Speech and Audio Processing, 3, no. 1, (January 1995).

[97] Davidson, Fielder, and Link, "Parametric Bit Allocation in Perceptual Audio Coder," Audio Engineering Society, 97th Convention, (November 10-13th, 1994).

[98] James D. Johnson, "Transform Coding of Audio Signals Using Perceptual Noise Criteria," IEEE Journal on Selected Areas in Communications 6, no. 2, (February 1988):314-323,.

[99] H. Fletcher, "Auditory patterns," Rev. Modern Phys. 12, (1940):47-65.

[100] H. Fletcher, "Auditory patterns," Rev. Modern Phys. 12, (1940):47-65.

[101] Luo and Denbigh, "A Speech Separation System That is Robust to Reverberation," IEEE International Symposium on Speech, Image Processing, and Neural Networks, Hong Kong, (April 13-16, 1994).

[102] S. R. Quackenbush, T. P. Barnwell, M. A.Clements, Objective Measures of Speech Quality, (Englewood Cliffs, NJ: Prentice Hall, 1988): 9.

[103] Jont B. Allen, "How Do Humans Process and Recognize Speech?", IEEE Transactions on Speech and Audio Processing. Vol. 12, No. 4, (October 1994):567-577.

[104] K. D. Kryter, "Methods for the Calculation and Use of the Articulation Index," J. Acoust. Soc. Amer. 34, no. 11, (November 1962):1689-97.

[105] John R. Deller Jr., John H.L. Hansen, and John G. Proakis, Discrete-Time Processing of Speech Signals, (New York, NY: IEEE Press, 2000):582-584.

[106] B. Wei and J. Gibson, "Comparison of Distance Measures in Discrete Spectral Modeling", Proceedings of IEEE DSP Workshop, Hunt, TX. (October, 2000).

[107] Carl R. Nave, "Hyper Physics Sound and Hearing", Department of Physics and Astronomy, (Georgia State University: Atlanta, Georgia, 2001) <http://hyperphysics.phyastr.gsu.edu/hbase/sound/hearcon.html> (May 2002)

[108] J. Smolders, T. Claes, G. Sablon, and D. Van Compernolle, "On the Importance of the Microphone Position for Speech Recognition in the Car", Proceedings of the IEEE, (1994).

[109] Joan I. Miller and Bruce J. Taylor, "The Thesis Writer's Handbook," (West Linn, Oregon: Alcove Publishing Company, 1987).

[110]Andrew Harnack and Eugene Kleppinger, "Online!: a reference guide to using internet sources," 2001, <http://www.bedfordstmartins.com/online/cite7.html >(July 1 2002).

# Biography

**Jeff Faneuff** received a B.S. degree in electrical engineering from Worcester Polytechnic Institute, Worcester, MA in 1991. He is currently employed at Bose Corporation in Framingham, MA.