Real-Time Adaptive Noise Cancellation for Automatic Speech Recognition in a Car Environment

Ziming Qi

2008

Real-Time Adaptive Noise Cancellation for Automatic Speech Recognition in a Car Environment

A thesis presented in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Engineering
at
Massey University
School of Engineering and Advanced Technology
Auckland, New Zealand

Ziming Qi

2008

Table of Contents

Table of figures

List of Tables

Abstract

This research is mainly concerned with a robust method for improving the performance of a real-time speech enhancement and noise cancellation for Automatic Speech Recognition (ASR) in a real-time environment. Therefore, the thesis titled, "Real-time adaptive beamformer for Automatic speech Recognition in a car environment" presents an application technique of a beamforming method and Automatic Speech Recognition (ASR) method. In this thesis, a novel solution is presented to the question as below, namely:

How can the driver's voice control the car using ASR?

The solution in this thesis is an ASR using a hybrid system with acoustic beamforming Voice Activity Detector (VAD) and an Adaptive Wiener Filter.

The beamforming approach is based on a fundamental theory of normalized least-mean squares (NLMS) to improve Signal to Noise Ratio (SNR). The microphone has been implemented with a Voice Activity Detector (VAD) which uses time-delay estimation together with magnitude-squared coherence (MSC). An experiment clearly shows the ability of the composite system to reduce noise outside of a defined active zone. In real-time environments a speech recognition system in a car has to receive the driver's voice only whilst suppressing background noise e.g. voice from radio. Therefore, this research presents a hybrid real-time adaptive filter which operates within a geometrical zone defined around the head of the desired speaker. Any sound outside of this zone is considered to be noise and suppressed. As this defined geometrical zone is small, it is assumed that only driver's speech is incoming from this zone. The technique uses three microphones to define a geometric based voice-activity detector (VAD) to cancel the unwanted speech coming from outside of the zone. In the case of a sole unwanted speech incoming from outside of a desired zone, this speech is muted at the output of the hybrid noise canceller. In case of an unwanted speech and a desired speech are incoming at the same time, the proposed VAD fails to identify the unwanted speech or desired speech. In such a situation an adaptive Wiener filter is switched on for noise reduction, where the SNR is improved by as much as 28dB.

In order to identify the signal quality of the filtered signal from Wiener filter, a template matching speech recognition system that uses a Wiener filter is designed for testing. In this thesis, a commercial speech recognition system is also applied to test the proposed beamforming based noise cancellation and the adaptive Wiener filter.

List of Abbreviations and Acronyms

ANC	Adaptive Noise Canceller
ASR	Automatic Speech Recognition
AWF	Adaptive Wiener Filter
BSS	Blind Source Separation
DOA	Direction of Arrival
DFT	Discrete Fourier Transform
DS	Delay and Sum
DSP	Digital Signal Processing
EOD	Estimation of Direction
EOZ	Estimation of Zone
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
GCC	Generalized Cross Correlation
GPS	Global Positioning System
GSC	Generalized Sidelobe Canceller
IDFT	Inverse Discrete Fourier Transform
IFFT	Inverse Fast Fourier Transform
IIR	Infinite Impulse Response
LabVIEW	Laboratory Virtual Instrument Engineering Workbench
LMS	Least Mean Square
MSC	Magnitude Squared Coherence
NLMS	Normalized Least Mean Squares
NLMF	Normalized Least Mean Forth
RLS	Recursive Least Square
SNR	Signal-to-Noise Ratio
TDOA	Time Difference of Arrival
VAD	Voice Activity Detector

Nomenclature

μ	Step-size parameter for LMS
μ_n	A modified input dependent step size for NLMS
$\mathbf{h}_n$	Tap weight vector at time n of LMS or NLMS
$\hat{\mathbf{h}}_n$	Instantaneous estimate of the tap weight vector at time n
$E[\cdot]$	Expectation operator
$\mathbf{R}_{xx}(k)$	Discrete autocorrelation function of the input signal x_n
$\mathbf{R}_{xd}(k)$	Discrete cross-correlation function between x_n and the desired response d_n
Z	Z-transform operator
$\Phi_{xx}(z)$	Z-transform auto power spectrum of the input signal x_n
$\Phi_{xd}(z)$	Z-transform cross power spectrum between the input signal x_n and a desired response d_n
$\mathbf{R}$	$E[X_n X_n^H]$, autocorrelation vector of tap input vector $\mathbf{x}_n$
$\mathbf{P}$	$E[X_n d_n^*]$, cross-correlation vector between the tap input vector $\mathbf{x}_n$ and the desired response d_n
$\mathbf{x}_n^T$	Transposition input vector $\mathbf{x}_n$ at time n
$\mathbf{x}_n^H$	Hermitian transposition input vector $\mathbf{x}_n$ at time n
$\delta(t)$	Dirac delta function
$\hat{S}_{x_1x_1}(i)$	
$\psi_g(f)$	General frequency weighting function
$\mathbf{R}_{d'x'}^{(g)}(\tau)$	Generalized cross correlation function between $d'(t)$ and $x'(t)$
$\hat{\gamma}_{dx}(f)$	Coherence estimate between $x_d(t)$ and $x_x(t)$
$\lvert\gamma_{dx}(f)\rvert^2$	Magnitude squared coherence function
$\lambda_{\max}$	The largest eigenvalue of the tap input auto correlation matrix $\mathbf{R}$
β	Forgetting factor

$G_{x_1x_2}(f)$	Cross-spectrum of $x_1(t)$ and $x_2(t)$ at frequency f
$G_{x_1x_1}(f)$	Auto spectral density functions of $x_1(t)$ at frequency f
$G_{x_2x_2}(f)$	Auto spectral density functions of $x_2(t)$ at frequency f
$\gamma_{x_1x_2}(f)$	coherence between two zero-mean stationary random processes $x_1(t)$ and $x_2(t)$, at frequency f
$d_{\max}$	Maximum desired time-difference of arrival (TDOA) between two microphones
$C_{\min}$	Minimum desired MSC (with empirical meaning) to prevent reverberant speech from being detected as desired speech

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor, Dr. Tom J. Moir for his invaluable guidance in his position as a top class world researcher in this field. From beginning to end, he has provided many opportunities for me to develop my research interests as well as his solid background in research expertise. Secondly, I would also like to thank to both Massey University and Unitec Institute of Technology for providing an excellent research and study environment, and financial support for the participation in international conferences.

Declaration

I declare that the thesis is based on my own research work under the supervision of Dr. T. J. Moir during the Ph.D. study in School of Engineering and Advanced Technology, Massey University at Albany.

The research work has produced a book chapter, journal papers, conference proceedings and presentations during the Ph.D. study. The content of this thesis therefore contains theory, procedure, application and experimental outputs from the research papers published during the research period as listed below.

Book Chapter

Qi, Z and Moir T (2008), An Adaptive Wiener Filter for Automatic Speech Recognition in a Car Environment with Non-Stationary Noise. In S. Mukhopadhyay & G. S. Gupta (Eds.), *Smart Sensors and Sensing Technology*: Springer-Verlag.

Refereed journal paper

Qi, Z., & Moir, T. J. (2006). Automotive 3-microphone Noise Canceller in a Frequently Moving Noise Source Environment. *International journal of signal processing, 3* (4), 298-304.

Refereed conference proceedings

Qi, Z., & Moir, T. (2008). Automotive speech control in a non-stationary noisy environment. Paper submitted at the 15th International Conference on Mechatronics and Machine Vision in Practice, Auckland, New Zealand.

Qi. Z. & Moir. T. (2007). A Design of Automotive Voice Recognizer Using LabVIEW. Paper was accepted to The 14th Electronics New Zealand Conference (ENZCon). Wellington New Zealand.

Qi, Z., & Moir, T. J. (2007). An Adaptive Wiener Filter for an Automotive Application with Non-Stationary Noise, Paper was accepted to 2nd International Conference on Sensing Technology 2007. Palmerston North, New Zealand.

Qi, Z., & Moir, T. J. (2005). An Automotive three-microphone Voice Activity Detector and noise canceller, *2005 International Conference on Intelligent Sensors, Sensor Networks and Information*. Melbourne, 5 - 8 December, Melbourne.

Qi, Z., & Moir, T. J. (2005). A geometrical active zone voice activity detector in car, *2nd IIMS Post-Graduate Conference*. Auckland, 27 October.

Non-refereed journal paper

Qi, Z., & Moir, T. J. (2005). Automotive three-microphone Voice Activity Detector and noise-canceller. Research Letters in the Information and Mathematical Sciences, 7 (July 2005), 147-156, Institute of Information and Mathematical Sciences, Massey University, Auckland, New Zealand.

Candidate's signature:

Candidate's name: Ziming Qi Date: 18-08-2008

1 Introduction

During the last several decades, a speech signal and array processing technology has been developed in various application areas. For a real-time application in real environments, using the driver's voice to control some devices in the car is one of the most difficult challenges.

1.1 Research objective

The objective of this thesis is to simplify the use of microphones and reduce the real-time environment noise in order to control devices using Automatic Speech Recognition (ASR) successfully. As in Figure 1-1, a three-microphone beamforming is used to cancel the unwanted voice from passenger 1, 2 and 3 as well as from voice from the 4 loud-speakers. In the thesis, a hybrid Adaptive Noise Cancellation (ANC) is proposed, which uses a system with acoustic beamforming based Voice Activity Detector (VAD), Normalize least mean squares Adaptive filter and an Adaptive Wiener Filter as shown in Figure 1-2. Automatic Speech Recognition (ASR) is applied to test the quality of this hybrid ANC.

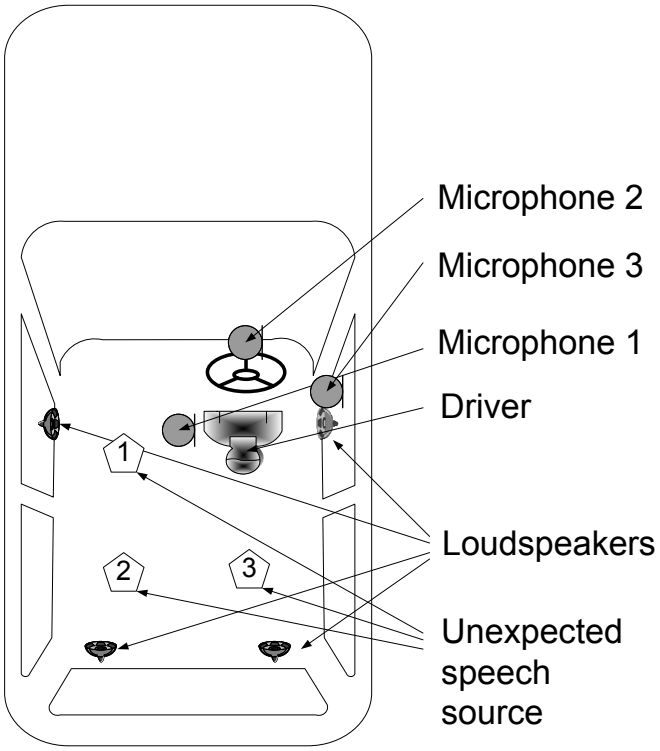

Figure 1-1 Three-microphone beamforming in car

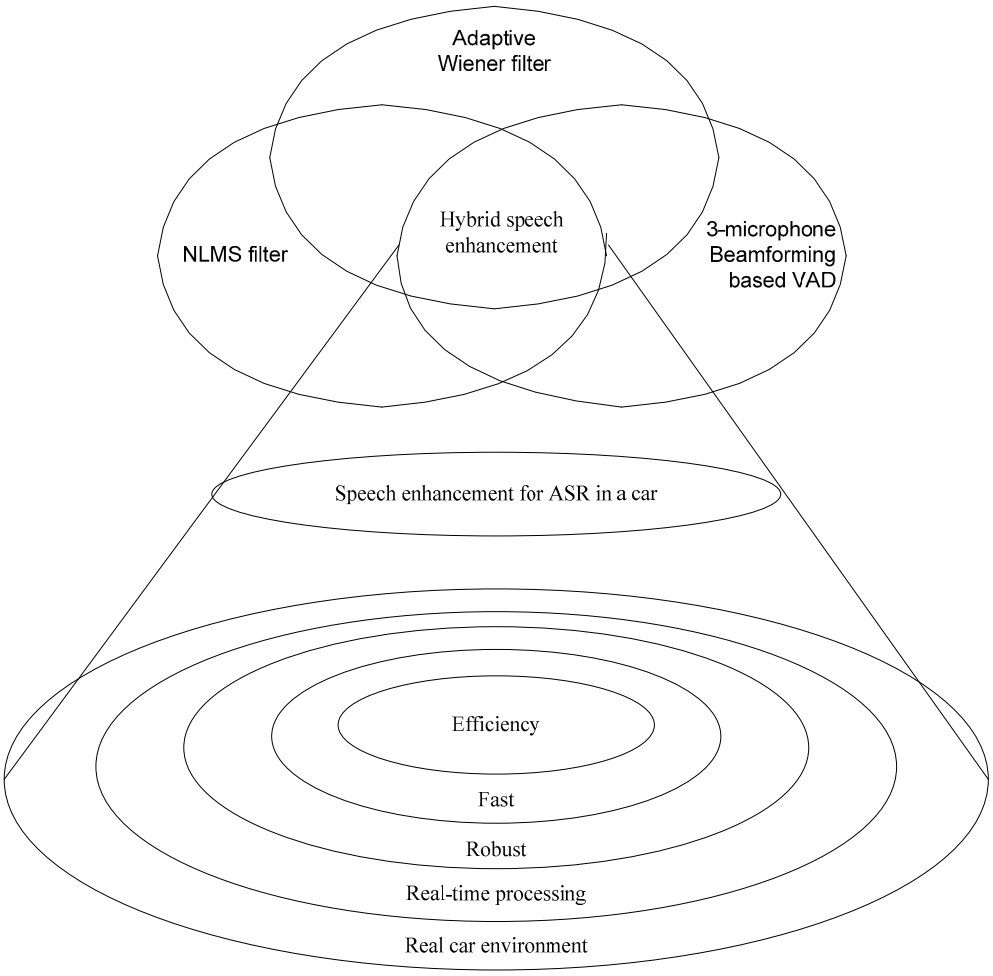

Figure 1-2 Research objective

1.2 Speech enhancement approach

In this thesis, the main approach is 3-microphone beamforming. A 3-microphone beamforming approved will be shown to be able to mute the input to ASR when voice or interference is incoming from outside of desired zone. When a desired command voice is incoming, a 3-microphone adaptive noise cancellation system can suppress the stationary noise in a car e.g. from a car engine.

When non-stationary noise incoming e.g. from a car radio, an Adaptive Wiener filter is an addition to this main approach when the beamforming is "confused" by multi-source noise or interference. ASR is introduced in this thesis to switch the desired devices.

There are three approaches in this thesis as in Figure 1-4: Adaptive noise cancellation, Real-time adaptive speech separation (using an adaptive Wiener filter) and Voice activity detection.

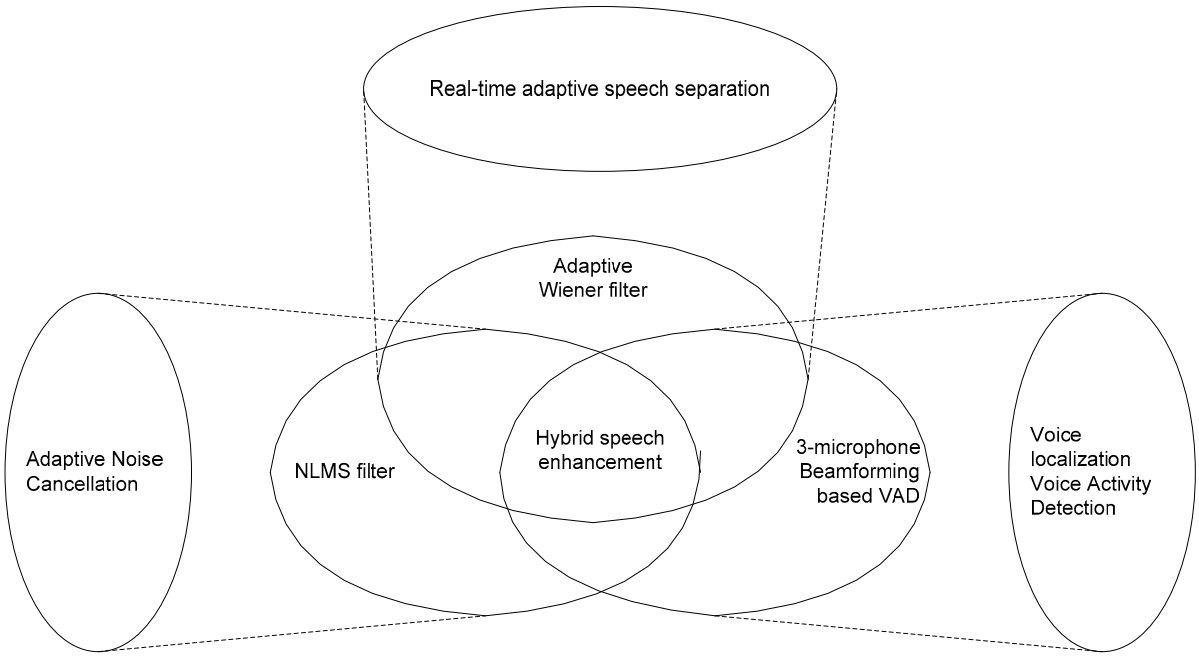

Figure 1-3 Hybrid noise cancellation approach

1.2.1 Voice activity detection approach

The Voice activity detection (VAD) approach is based on a fundamental theory of time-delay estimation with magnitude-squared coherence (MSC). An experiment will clearly show the ability of the composite system to reduce noise outside of a defined active zone. A 3-microphone VAD approach is used in this thesis since it is relatively simple and not too much of a computational overload for real-time system.

1.2.2 Adaptive noise cancellation approach in this thesis

Noise cancellation is based on a fundamental theory of normalized least-mean squares (NLMS) to improve Signal to Noise Ratio (SNR). The signal and noise inputs are shared with a Voice Activity Detector (VAD). A 3-microphone NLMS adaptive noise cancellation is used.

1.2.3 Adaptive Wiener filter approach

Closely related to the NLMS approach is the technique used in this thesis which uses an adaptive Wiener filter. Whereas the NLMS approach uses weight estimation to minimize a mean-square error, this alternative approach constructs the Wiener filter from estimation of covariance matrices (for signal + noise and "noise-alone").

In real-time environments a speech recognition system in a car has to receive the driver's voice only whilst suppressing background noise e.g. voice from radio. Therefore, this research presents a hybrid real-time adaptive filter which operates within a geometrical zone defined around the head of the desired speaker. Any sound outside of this zone is considered to be noise and suppressed. As this defined geometrical zone is small, it is assumed that only driver's speech is incoming from this zone. The technique uses three microphones to define a geometric based voice-activity detector (VAD) to cancel the unwanted speech coming from outside of the zone. However, when unwanted speech and desired speech are incoming at the same time, the VAD fails to identify the unwanted speech or desired speech. In such a situation, the adaptive Wiener filter is switched on for noise reduction. In the case of sole unwanted speech incoming from outside of a desired zone, this speech is muted at the output of the hybrid noise canceller. In the case of desired and unwanted speech incoming together, SNR is improved by as much as 28dB.

1.3 Contributions to knowledge

In the thesis, the following original main contributions have been made:

Firstly, 3-Microphone beamforming is applied in real-time to a car environment. The noise-cancelling is only required when noise is present during desired speech since the VAD will mute any solo noise-source outside of the zone. The experiments in this thesis clearly show the ability of the composite system to reduce noise outside of a defined active zone.

Secondly, a Wiener filter is used in this thesis, similar to the dual microphone method of Widrow et al except here Least-Mean Squares (LMS) is not used but instead is an update of the Wiener filter direct by estimation of the signal + noise and noise covariance matrices and by direct solution of the Wiener-Hopf equation.

Thirdly, in this thesis, I suggest that it is not necessary for the enhanced signal to sound better to the human ear, but only needs to be good enough to provide a Boolean on or off command in all of these kinds of automated 'smart-car' technologies, (in contrast with some other similar technologies) although the filtered signals were clearly recognizable during informal listening tests and sounded vastly improved to the original. As an example, an ASR with Wiener filter has been designed to improve the recognition rate from 7 % up to 95%.

Finally, a completed engineering model is presented in this thesis: a hybrid adaptive noise cancellation system, which employs 3-microphone voice activity detection with NLMS adaptive noise cancellation and an adaptive Wiener filter.

1.4 Performance with favorable effects

The beamforming and ASR approaches have provided an improved performance with some favorable effects to the three-microphone approach, such as the system allows voice controlled equipment available for a car in real-time environment. For example, drivers can control their car in noisy environments e.g. turn off the radio whilst the radio is on a news channel, or control devices whilst passengers are simultaneously speaking. Therefore, a full-voice controlled car is available.

1.5 Structure of this thesis

In chapter 2, a literature review will cover the relevant area of real-time adaptive acoustic noise cancellation for automatic speech recognition in a car environment. Chapter 3 will have the problem definition and research environmental set-up and Chapter 4 proposes the main solutions to such a problem. Chapter 5 has experiments to verify and confirm the solutions in the previous chapters. Finally, a design method of hybrid adaptive noise cancellation is discussed in Chapter 6 Conclusions and Future Work. The structure of this thesis is shown in Figure 1-4.

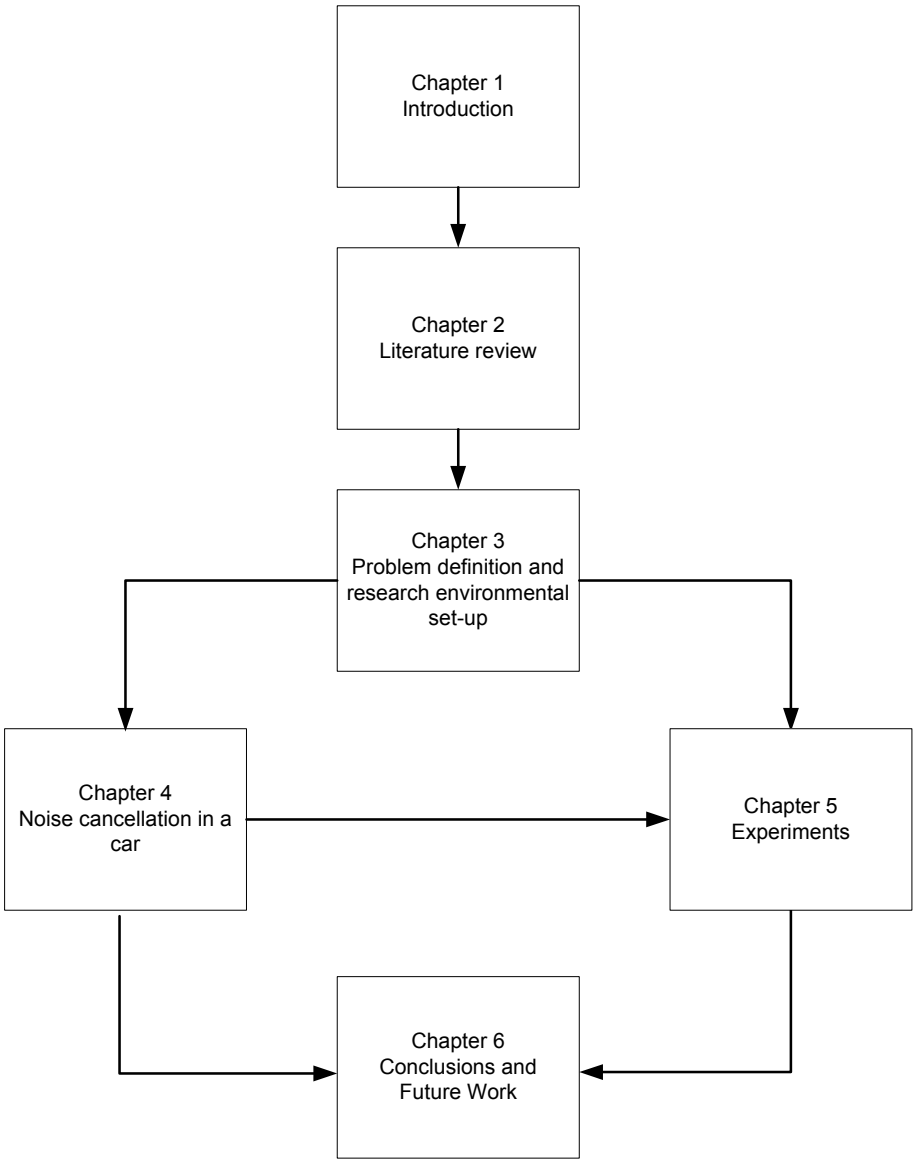

Figure 1-4 Structure of this thesis

2 Literature review

2.1 Introduction

In this chapter, the literature review explores the technical background in an area of microphones, signal conditioning and adaptive digital signal processing to propose a real-time adaptive acoustic noise cancellation for automatic speech recognition in a car environment as showed in Figure 2-1.

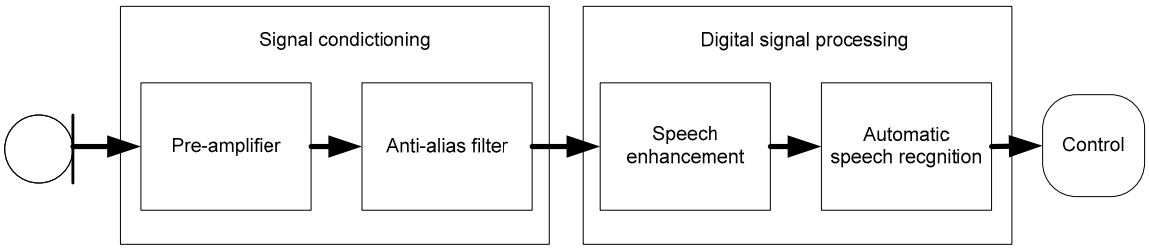

Figure 2-1 Block diagram for a Real-Time Adaptive Acoustic noise cancellation for Automatic Speech Recognition in a Car Environment

In a car environment, a controlled object could be any device e.g. a window, a radio or a GPS. However, the purpose of this thesis is not to discuss such a control object. The focus is the method of a real-time adaptive noise cancellation for automatic speech recognition in a car environment. In Figure 2-1, microphone and signal conditioning are very important hardware support to this research. DSP hardware is the main body of digital signal processing. More importantly, this thesis will discover the best solution to enhance speech in a car. Therefore, the literature review will include acoustic beamforming, voice activity detection, the cocktail party effect and solution, speech enhancement in a car noise environment and also digital signal processing hardware and software.

2.2 Acoustic beamforming

2.2.1 Overview

Acoustic beamforming consists of signal processing using arrays of microphones to control the directionality and sensitivity of sound. Such a beamformer can be either a receive beamformer or transmit beamformer. In a receive beamformer, beamforming can increase the receiver sensitivity in the desired direction and decrease the sensitivity in the unwanted direction. As an example, the human brain uses a form of signal processing on its ears and determines sound localization. In transmit beamformer, beamforming can increase the power in a given direction(*Beamforming*, 2008). As early as 1969, Capon (Capon, 1969) proposed a minimum variance distortionless response (MVDR) beamformer which these results were given to seismic data obtained from the large aperture seismic array located in eastern Montana.

Acoustic beamforming techniques can be divided into two categories: Acoustic conventional beamformer and Acoustic adaptive beamformer. An acoustic conventional beamformer is a fixed beamformer. An acoustic adaptive beamformer is an adaptive array of microphones.

An acoustic conventional beamformer uses a fixed set of weightings and time-delays to combine the signals from the microphones in the array whilst an acoustic adaptive beamformer is able to adapt automatically its response to different weightings or time-delays. In the acoustic adaptive beamformer, a criterion is built up to allow the adaptation to minimize the noise output. In wide band systems, acoustic adaptive beamformer is very often to be considered to process in the frequency domain(*Beamforming*, 2008).

2.2.2 Conventional "delay and sum" acoustic beamformer

A 2-microphone acoustic beamforming is shown in Figure 2-2. Figure 2-2(a) is shown as both of microphones receive incoming sound at the same time. It means sound from an acoustic source reaches the two microphones from an equal distance.

Figure 2-2(b) is shown that the incoming wave front reach the right microphone first and then the left microphone. Since microphones across the array receive signals with differential time delays, the microphones outputs no longer add coherently and cause the output of sum drops.

Figure 2-2(c) is shown that there is a delay between the right microphone and an input of signal processer. While this delay equals to differential time delay of arrival of right microphone and left microphone, beamforming output is assumed as same as at Figure 2-2(a). Therefore a time delay at one of the two signal channels, output will enhance the sound source incoming from a desired direction.

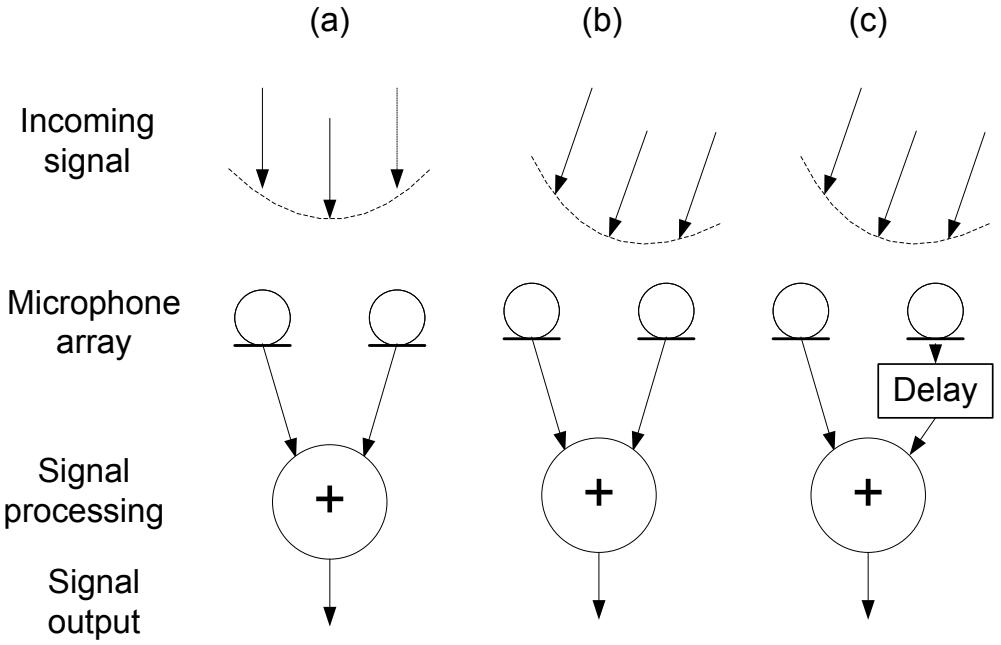

Figure 2-2 A 2-microphone beamforming

If there are more than two microphones, e.g. n microphones in Figure 2-3, time delays with fixed values are set up at each signal processing channel. The received signals from the microphones are then delayed by different values and finally is summed to emit as *Y(k)*. This beamformer is normally called a delay-and-sum beamformer.

In delay–and-sum beamformer, while the incoming signal is $x(k)$ and k is 1, 2 … n,

$$x(k) = [x_1(k),..., x_M(k)]^T \tag{2.1}$$

The output of the beamformer is:

$$y(k) = w^H x(k) \tag{2.2}$$

The w^H denotes the complex conjugate transpose of the weight vector w,

$$w = [w_1,..., w_M]^T \tag{2.3}$$

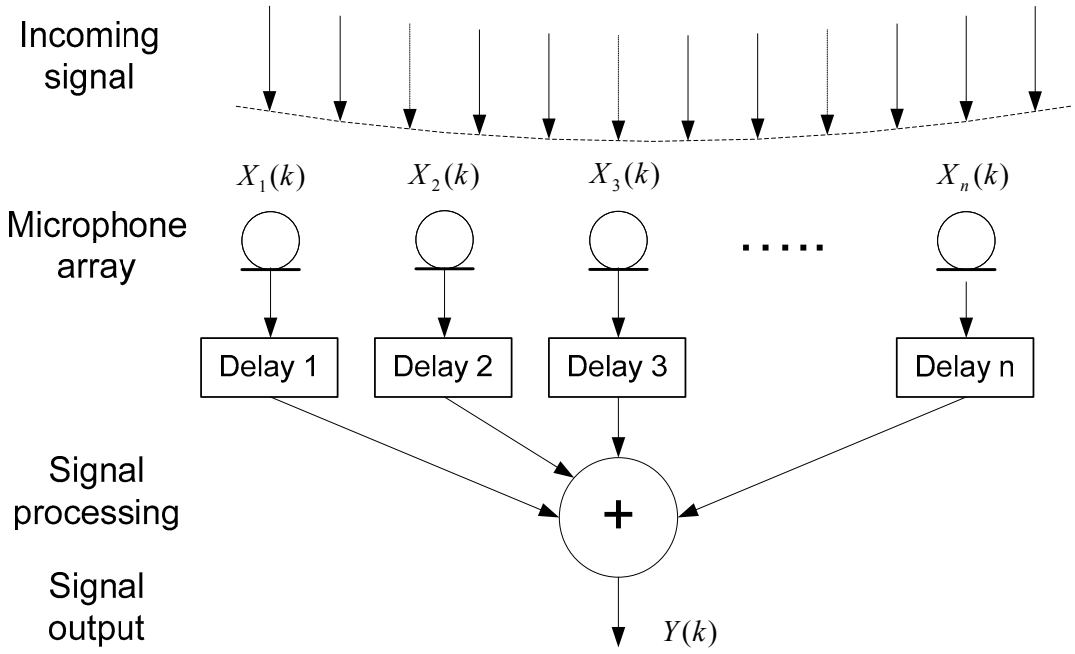

Figure 2-3 Delay-and-sum beamformer

An n-microphone receiving beamformer takes advantage of interference to change the directionality of the array. On the other hand, a transmit beamformer can control the phase and relative amplitude of the signal at each transmitter to create a pattern of constructive and destructive interference in the wavefront. As an example, Hodgkiss (Hodgkiss, 1980) designed a programmable time-delay-and-sum digital beamformer, which permits the incorporation of slow changes in element positions and/or beam steering direction while the beamformer carries out the real-time formation of 1300 beams from 32 input sensors. This thesis only considers receive beamformers.

2.2.3 Far-field and Near-field acoustic wavefronts

Conventional analysis and synthesis techniques for microphone arrays are based on the simplifying assumption that all acoustic sources are located far from the array. In this case wavefront curvature can be neglected and all waves impinging upon the array are assumed to be planar (Ryan & Goubran, 1997). However, when the microphone array receives the sound which is near by, the sound wavefront does not appear as planar.

Mailloux suggested (Mailloux, 1994) the far-field assumptions are no longer valid when

$$r < \frac{d_{dod}^2 f_s}{c} \tag{2.4}$$

with r the distance of the signal source to the centre of the microphone array, $d_{dod} = d_{N-1} - d_0$ the total length of the microphone array, f_s the sampling frequency and c the speed of sound.

In Figure 2-2 and 2-3, the incoming signal wavefronts can be planar in case of far-field assumptions. However, in this thesis, the total length of microphone array in a car is 1 m; the sampling frequency is 11025 Hz. The r from equation 2.4 will be 32 m. Since the distance of signal source (a driver in a car) to the centre of microphone array is less than 0.5 m, we will consider near-field acoustic wavefronts in any cases in this thesis.

2.2.4 Adaptive Acoustic Beamforming

An adaptive beamformer is a signal processing system which applies an algorithm to adjust an array in real-time. A typical application is an array of radar antennas which is to transmit or receive signals in different directions without mechanical steering. An adaptive acoustic beamformer has the ability to adjust its performance to suit differences in its environment e.g. reducing sensitivity to the directions of arrival where unwanted noise or interference is incoming.

Adaptive Beamforming often employs Least Mean Squares (LMS) Algorithm to adjust an array in real-time to enhance the desired signal and cancel the noise or interference.

As early as 1970s, Frost (Frost, 1972) applied a constrained least mean-squares algorithm which is capable of adjusting an array of sensors in real time to respond to a signal coming from a desired direction while unwanted noises coming from other directions. Analysis and simulations confirm that this algorithm is able to iteratively adapt variable weights on the taps of the sensor array to minimize noise power in the array output. A set of linear equality constraints on the weights keeps a chosen frequency characteristic for the array in the desired direction.

Ferrara and Widrow (Ferrara & Widrow, 1981) also applied an adaptive algorithm to enhance a signal against noise. It used two or more input channels containing correlated signal components but uncorrelated noise components. The output is a best least squares estimate of the underlying signal in a chosen input channel. They stated that the more input channels available containing correlated signal components, the better will be the system performance. Excellent performance is obtained when the sum of the filter input signal-to-noise ratios (SNR), is large compared to unity at all frequencies of interest. In this case the output noise is small, the output signal distortion is small and the output SNR is approximately equal to the sum of the filter input SNR. However, this beamformer can improve signal-to-noise ratio (SNR) from a desired direction but cannot cut off the sound or noise which came from outside of this direction. The signal-to-noise ratio (SNR) of the total signal is greater than (or at worst, equal to) that of any individual microphone's signal. This system makes the array pattern more sensitive to sources from a particular desired direction (Sullivan, 1996).

In fact, most applications use acoustic beamforming to enhance desired direction sound source and also use other algorithms to improve signal-to-noise ratio (SNR). The most popular one is a digital filter.

Griffiths and Jim (1982) described an adaptive beamformer as Figure 2-4.

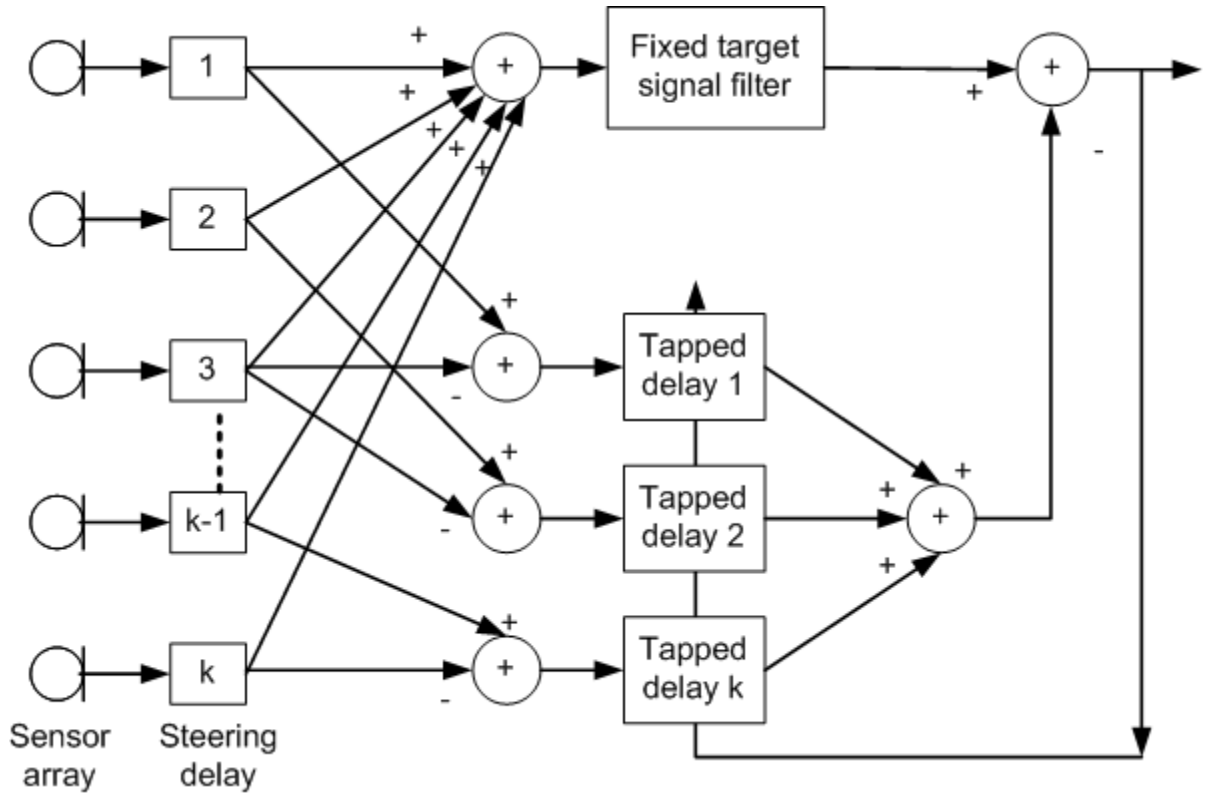

Figure 2-4 Griffiths-Jim beamformer

As in Figure 2-4, this beamformer consists of the summing process from the delay-and-sum beamformer, to provide an enhanced desired signal component, and an adaptive multi-channel section which uses the differences between the time-aligned input signals to provide a representation of the undesired signal components. This simple beamformer shown above consists of a main antenna and one or more auxiliary antennas. The main antenna on the top is highly directional and is pointed in the desired signal direction. It is assumed that the main antenna receives both the desired signal and the interfering signals through its sidelobe (this beamformer is also known as sidelobe noise canceller). The auxiliary antenna primarily receives the noise or interfering signals since it has very low gain in the direction of the desired signal. The auxiliary array weights are chosen such that they cancel the interfering signals that are present in the sidelobe of the main array response.

Later Krolik et al. (Krolik, Eizenman, & Pasupathy, 1986) used adaptive beamforming with generalized correlation (GC) to demonstrate a means of time delay estimate (TDE) bias reduction in interference dominated environments. They gave analysis of GC TDE bias with both minimum variance distortionless response and conventional delay-and-sum

beamforming. At high interference-to-noise ratios, theoretical and simulation results indicated that the adaptive structure can facilitate improved delay estimation performance compared to conventional methods.

In 2004, Dmochowski and Goubran (J. Dmochowski & Goubran, 2004) presented a noise cancellation structure with a fixed beamformer front end. Simulation results showed a noise reduction of 21 dB with a speech source-noise source separation of 1 meter. The experimental results showed a complex noise reduction pattern with a maximum noise reduction of 17 dB.

Dmochowski and Goubran (J. Dmochowski & Goubran, 2005) also examined in 2005 a microphone array based, combined beamformer noise canceller structures. The performance of the structures was evaluated using computer simulation as well as experimental measurements. The inter-operation of the beamformer and noise canceller was measured under the SNR. An experimental procedure for evaluating output SNR was presented in their research: the desired signal was captured from a set location in the recording environment. The noise signal was measured from a second location. Results revealed an SNR improvement of up to 17 dB.

More recently Dmochowski and Goubran (J. P. Dmochowski & Goubran, 2007) informed that the enhancement of noise-corrupted speech acquired by microphones is indispensable to the functioning of a wide variety of digital signal processing algorithms. As many existing products are equipped with steerable, stand-alone fixed beamformers which provide moderate levels of directivity, these applications have long employed the classical adaptive noise canceller configuration with a reference sensor near the noise source to cancel unwanted noise. In their research, the cascading of stand-alone beamformers with back-end adaptive noise cancellers was presented. A decoupled model for signal enhancement using front-end beamformers and cascaded noise cancellers was also presented. The inter-operation of the beamforming and noise cancelling units was studied by defining the signal-to-interference ratio (SIR) gain, directivity index, and white noise gain offered by the beamforming and noise cancelling components. An experimental result shows the SIR improvement as much as to 27 dB.

2.2.5 Adaptive algorithm for beamforming

2.2.5.1 Recursive Least Square algorithm

The RLS algorithm uses an adaptive method to determine the coefficients of an adaptive filter. The method utilizes information from all the previous input data to estimate the inverse of the autocorrelation matrix of the input vector (S. Haykin, 1996). It has been derived independently by several investigators, but the original reference on the RLS algorithm is appeared to be Plackett in 1950 (Plackett, 1950).

RLS adaptive filter block diagram is shown as Figure 2-5.

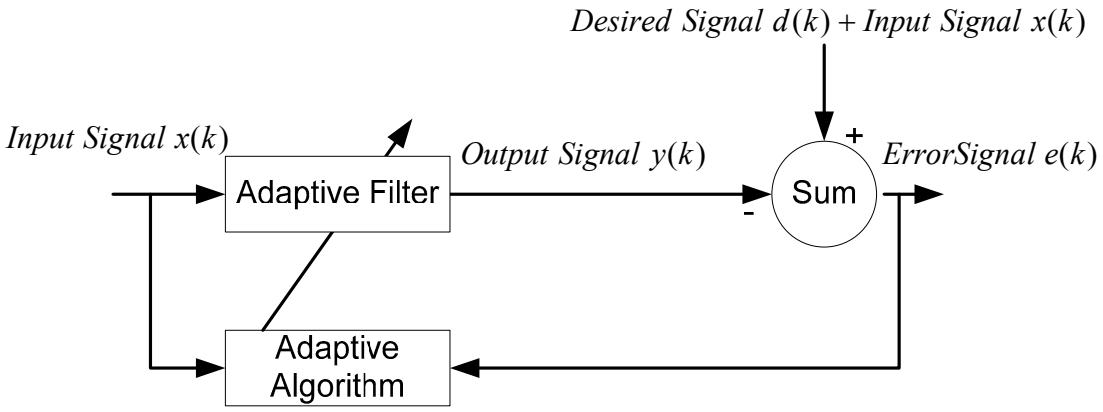

Figure 2-5 RLS adaptive filter

The recursive method of least squares is to minimize the residual sum of squares of the error signal $e(k)$.

To adjust of influence of input samples from the far past, the weighting factor is used in the cost function $J(k)$.

$$J(k) = \sum_{i=1}^{k} \beta^{k-i} e^2(i) \tag{2.5}$$

where β is the exponentially weighted forgetting factor. It is selected between $0 < \beta < 1$.

The resulting equation for the optimum filter coefficients at time k is,

$$\mathbf{h}(k) = \mathbf{R}^{-1}(k)\mathbf{P}(k) \tag{2.6}$$

where $\mathbf{R}(k) = \sum_{i=1}^{k} \beta^{k-i} \mathbf{x}(i)\mathbf{x}^H(i)$

and $\mathbf{P}(k) = \sum_{i=1}^{k} \beta^{k-i} d(i)\mathbf{x}^H(i)$.

Both $\mathbf{R}(k)$ and $\mathbf{P}(k)$ can be computed recursively:

$$\mathbf{R}(k) = \beta\ \mathbf{R}(k-1) + (1-\beta)\mathbf{x}(k)\mathbf{x}^H(k) \tag{2.7}$$

$$\mathbf{P}(k) = \beta\ \mathbf{P}(k-1) + (1-\beta)d(k)\mathbf{x}(k) \tag{2.8}$$

To find the coefficient vector $\mathbf{h}(k)$, we need the inverse matrix $\mathbf{R}^{-1}(k)$. Using a matrix inversion lemma (S. Haykin, 1996), a recursive update equation for $\mathbf{R}^{-1}(k)$ is found as

$$\mathbf{R}^{-1}(k) = \beta^{-1}\mathbf{R}^{-1}((k-1) + \beta^{-1}\mu'(k)\mathbf{x}(k) \tag{2.9}$$

where $\mu'(k) = \dfrac{\beta^{-1}\mathbf{R}_1^{-1}(k-1)\mathbf{x}(k)}{1+\beta^{-1}\mathbf{x}^H(k)\mathbf{R}^{-1}(k-1)\mathbf{x}(k)}$

Therefore, we find the weights update equation as

$$\mathbf{h}(k) = \mathbf{h}(k-1) + \mu'(k)(d(k) - \mathbf{x}(k)\mathbf{h}(k-1)) \tag{2.10}$$

For the computational complexity, RLS requires $5N^2 + 2N + 2$ multiplications (Lim, 1994).

2.2.5.2Least Mean Square algorithm

Least mean squares (LMS) algorithms are used in adaptive filters to find the filter coefficients that relate to producing the least mean squares of the error signal (difference between the desired and the actual signal). It is a stochastic gradient descent method in that the filter is only adapted based on the error at the current time. It was invented in 1960 by Bernard Widrow and Ted Hoff (*Least mean squares filter*, 2008). A LMS filter block diagram is shown as Figure 2-6.

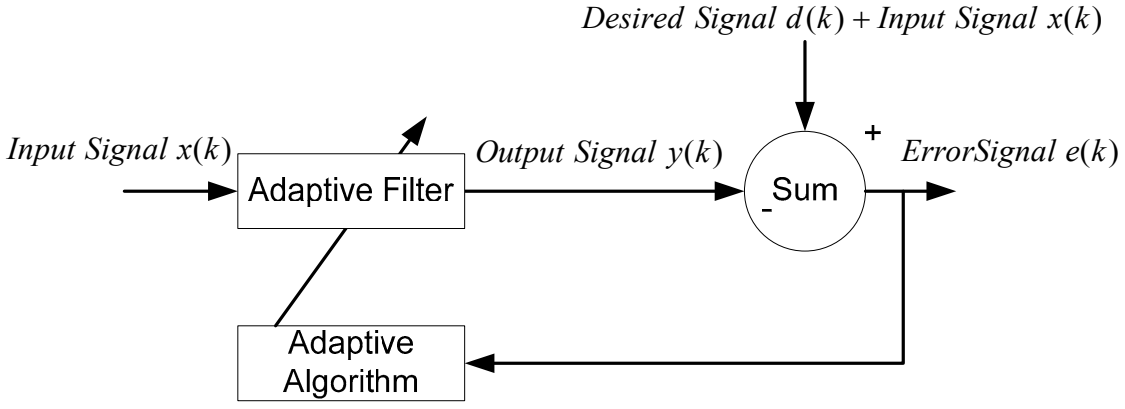

Figure 2-6 LMS adaptive filter as noise canceller block diagram

As in Figure 2-6, LMS adaptive filter equations is shown as below: (Bellanger, 2001; Diniz, 2002; Ifeachor, 1993; Lyons, 2004; Rorabaugh, 1999)

$$y(n) = \sum_{k=0}^{n-1} w_k(n) \bullet x(n-k) \tag{2.11}$$

$$e(n) = d(n) - y(n) \tag{2.12}$$

$$w_k(n+1) = w_k(n) + 2\mu e(n)x(n-k) \tag{2.13}$$

Where k = 1, 2 …, N - 1.

To define the self learning process the filter uses, adaptive algorithm is selected to reduce the error between the output signal *y(k)* and the desired signal *d(k)*. When the LMS performance criteria for *e(k)* has achieved its minimum value through the iterations of the adapting algorithm, the adaptive filter is finished and its coefficients have converged to a solution. Now the output from the adaptive filter matches closely the desired signal *d(k)*. If the input data characteristics are changed, sometimes called the filter environment, the filter adapts to the new environment by generating a new set of coefficients for the new data. Notice that when *e(k)* goes to zero and remains there you achieve perfect adaptation; the ideal result but not likely in the real world(Lyons, 2004).

2.2.5.3Normalized least mean square algorithm

In the LMS algorithm the selection of step size (μ) may affect its stability and the convergence rate. When μ is too large, it has faster convergence but less stability. On the other hand, if μ is smaller, it has slower convergence but greater stability. Therefore a LMS has difficulties in a real-time processing in a real environment since a speech has a wide dynamic range and may give rise to stability in quiet utterances and conversely instability when the utterances are louder or when non stationary noise is added. Haykin (Haykin, 1996) suggested that the LMS algorithm can be convergent or stable in the mean, if and only if $0 < \mu < \frac{2}{\lambda_{\max}}$. Using the analysis that the maximum value of μ depends on the largest eigenvalue $\lambda_{\max}$ of the input autocorrelation R, which can be approximated to $tr(\mathbf{R})$ and then in the same way to $\|\mathbf{x}_n\|^2$ (i.e., $\lambda_{\max} \approx tr(\mathbf{R}) \approx \|\mathbf{x}_n\|^2$), it can be induced that the maximum value of μ depends on the input power signal. Accordingly, the step size for the stable adaptation has to be constrained according to $0 < \mu < \frac{2}{\lambda_{\max}} \approx \frac{2}{tr(\mathbf{R})} \approx \frac{2}{\|\mathbf{x}_n\|^2}$, where, $\lambda_{\max}$ is the largest eigenvalue of the tap input auto correlation matrix $\mathbf{R}$, $tr(\mathbf{R})$ is trace of $\mathbf{R}$, which is the sum of the elements on its diagonal ($\sum_{i=1}^{P} \lambda_i$), and $\|\mathbf{x}_n\|^2$ is the input power.

Based on the above, normalize least mean squares (NLMS) algorithm is used in adaptive filtering algorithms due to its simplicity for real-time applications. (Simon. Haykin, 2002) A NLMS filter block diagram is shown as Figure 2-7. To define the self learning process the filter uses an adaptive algorithm to reduce the NLMS between the output signal *y(k)* and the desired signal *d(k)*. For stationary signals, when the NLMS performance criteria for the NLMS have achieved its minimum value through the iterations of the adapting algorithm, the adaptive filter is finished and its coefficients have converged to a constant solution. Then the output from the adaptive filter matches closely the desired signal *d(k)*. If the input data characteristics are changed, the filter adapts to the new environment by generating a new set of coefficients for the new data. The *e(k)* goes to zero and remains there.

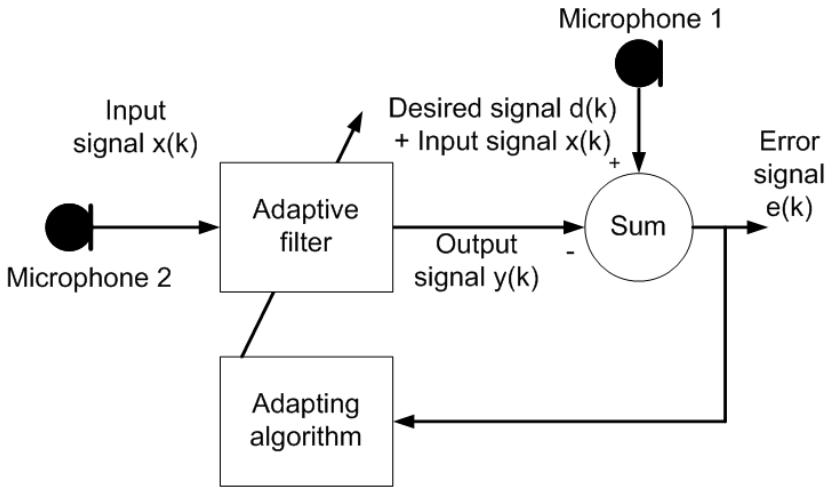

Figure 2-7 NLMS adaptive filter as noise canceller block diagram

The NLMS adaptive filter weights are updated accordingly (Barrault, Costa, Bermudez, & Lenzi, 2005)

$$W_{k+1} = W_k + 2\mu_n e_k X_k \tag{2.14}$$

Where the weight vector

$$W_k = [w_{1,k}\ w_{2,k}\ \ldots w_{N,k}]^T \tag{2.15}$$

are the coefficients of the adaptive filter at time k,

$$X_k = [x_k\ x_{k-1}\ \ldots x_{k-N+1}]^T \tag{2.16}$$

are the N samples of the input data in filter memory at time k,

$$e_k = d_k - W_k^T X_k \tag{2.17}$$

$$\mu_n = \frac{\widetilde{\mu}}{\delta + \|\mathbf{x}_n\|^2} \tag{2.18}$$

Where $0 < \tilde{\mu} < 1$, μ_n is a modified input dependent step size and δ is an infinitesimal positive value added to prevent the possibility of zero division in the event of a very small input value.

where $\delta = 0.0001$ and $\|X_k\|$ is Euclidean norm of X_k and is given by $\|X_k\|^2 = x_k^2 + x_{k-1}^2 + \ldots x_{k-N+1}^2$.

When Microphone 1 is defined as the primary input and Microphone 2 as the reference input, as in Figure 1, experiments (Rulph, 2002) show that voice close to the primary input is enhanced while voice close to the reference input is reduced.

2.2.5.4Normalized Least Mean Forth algorithm

Walach and Widrow (Walach & Widrow, 1984) presented a steepest descent algorithms for adaptive filtering and have been devised which allow error minimization in the mean fourth. Ideally, during adaptation, the weights undergo exponential relaxation toward their optimal solutions. Time constants have been derived, and surprisingly they turn out to be proportional to the time constants that would have been obtained if the steepest descent least mean square (LMS) algorithm of Widrow and Hoff had been used. The gradient algorithms are insignificantly more complicated to program and to compute than the LMS algorithm. Conditions have been derived for weight-vector convergence of the mean and of the variance for the new gradient algorithms. The behavior of the least mean fourth (LMF) algorithm is of special interest. In comparing this algorithm to the LMS algorithm, when both are set to have exactly the same time constants for the weight relaxation process, the LMF algorithm, under some circumstances, will have a substantially lower weight noise than the LMS algorithm. It is possible, therefore, that a minimum mean fourth error algorithm can do a better job of least squares estimation than a mean square error algorithm. This intriguing concept has implications for all forms of adaptive algorithms, whether they are based on steepest descent or otherwise.

Recently Zerguine (Zerguine, 2000) presented a NLMF algorithm. It sounds that NLMF had shown to have potentially faster convergence than NLMS. Unlike the LMF algorithm, the convergence behaviour of the NLMF algorithm is independent of the input data correlation statistics. Sufficient conditions for the NLMF algorithm convergence in the mean were obtained and the analysis of the steady-state performance was carried out using the feedback approach. Simulation results confirmed the performance of the NLMF algorithm.

However, Zerguine (Zerguine, 2000) stated that NLMF algorithm results in the fastest average convergence for a gradient step: this was of course at the expense of higher mis-adjustment values.

Moinuddin et al (Moinuddin, Zerguine, & Sheikh, 2005) had tracking analysis of the normalized least mean fourth (NLMF) algorithm, which is carried out in the presence of two sources of non-stationeries (carrier frequency offset between transmitter and receiver, and random variations in the environment). The concept of energy conservation was used to carry out the analysis. Simulation results agreed very closely with theory. The results showed that, unlike in the stationary case, the steady-state excess mean-square error (MSE) was not a monotonically increasing function of the step size. Moreover, the ability of the adaptive algorithm to track the variations in the environment is shown to degrade with increasing frequency offset.

In summary, LMF algorithm is known for its fast convergence and lower steady state error, especially under sub-Gaussian noise conditions. Meanwhile, the recent work on the normalised versions of LMF algorithm has further enhanced its stability and performance in both Gaussian and sub-Gaussian noise. For example, the XE-NLMF algorithm is normalised by the mixed signal power and error power, and weighted by a fixed mixed-power parameter. Unfortunately, this algorithm depends on the selection of this mixing parameter. Chen et al (2003) introduced a time-varying mixed-power parameter technique to optimise its selection. An enhancement in performance is obtained through the use of this procedure in both the convergence rate and steady-state error. (Chan, Zerguine, & Cowan, 2003)

Nascimento and Bermudez in 2005 described that the least-mean fourth and the least-mean mixed norm algorithms are not mean-square stable when the input is Gaussian-distributed. (Nascimento & Bermudez, 2005)

Zerguine (Zerguine, 2000) described the LMF adaptive filter weights are updated accordingly

$$W_{k+1} = W_k + 2\mu e_k^3 X_k \tag{2.19}$$

Where the weight vector

$$W_k = [w_{1,k}\ w_{2,k}\ \ldots w_{N,k}]^T \tag{2.20}$$

are the coefficients of the adaptive filter at time k,

$$X_k = [x_k\ x_{k-1}\ \ldots x_{k-N+1}]^T \tag{2.21}$$

are the N samples of the input data in filter memory at time k,

$$e_k(n) = d_k(n) - W_k^T X_k \tag{2.22}$$

is the error between the adaptive filter output and the desired signal d_k and μ is a user specified convergence parameter which if chosen to be too small will lead to slow convergence. If chosen to be too large the LMF algorithm will become unstable and the weights will diverge. To overcome this problem in a non-stationary environment we use the NLMF algorithm.

The NLMF algorithm is given by (Zerguine, 2000)

$$W_{k+1} = W_k + 2\tilde{\mu}e_k^3 \frac{X_k}{\|X_k\|^2 + \delta} \tag{2.23}$$

Where $0 < \tilde{\mu} < 1$, $\delta = 0.0001$ and $\|X_k\|$ is Euclidean norm of X_k and is given by

$$\|X_k\|^2 = x_k^2 + x_{k-1}^2 + \ldots x_{k-N+1}^2 \tag{2.24}$$

In this thesis, an experiment in Chapter 5 computes NLMF and NLMS performances for noise cancellation.

2.2.6 Robust acoustic adaptive beamforming

Although in the preceding discussion on the adaptive acoustic beamformers are considered as a good resolution to background noise and interference, the adaptive acoustic beamformers are much more sensitive to errors compared with conventional acoustic beamformers, such as the array steering vector errors caused by imprecise sensor calibrations (J. Li & Stoica, 2005). Therefore, over the past decades, much effort has been devoted to build robust adaptive beamformers. Adaptive beamforming algorithms are sensitive to small errors in array characteristics. In their new book in 2007, Li et al. (J. Li & Stoica, 2005) presented their research developments on robust adaptive beamforming. They had concluded that most of the early methods of making the adaptive beamformers more robust to array steering vector errors are rather ad hoc in that the choice of their parameters is not directly related to the uncertainty of the steering vector, until recently an uncertainty set of the array steering vector had been proposed. They had suggested four areas of current concerns on Robust Adaptive Beamforming: Array steering vector uncertainty, the finite sample size effect, the signal waveform estimation, constant modulus algorithms and robust wideband beamforming.

2.3 Voice Activity Detection

Voice activity detection (VAD) is an algorithm used in detecting the presence of human speech from silence, music, noise or other non-speech signals. In this thesis, VAD is also used to separate the desired speech from unwanted speech (Qi & Moir, 2005). The typical applications of VAD are in speech coding and speech recognition (Moir, 2008).

2.3.1 Time delay estimation

As one of important VAD methods, in this thesis, estimation theory is applied. The goal of VAD is to gather the values of disturbance parameters e.g. noise variance, signal parameters e.g. amplitude or propagation direction, or signal waveforms. Estimation theory assumes that the observations contain an information-bearing quantity, thereby tacitly assuming that detection-based pre-processing has been performed. Conversely, detection theory often requires estimation of unknown parameters: Signal presence is assumed, parameter estimates are incorporated into the detection statistic, and consistency of observations and assumptions tested. Consequently, detection and estimation theory form a symbiotic relationship, each requiring the other to yield high-quality signal processing algorithms. (Johnson, 2003)

At best, estimation theory is less structured than detection theory. Detection is science, estimation art. Inventiveness coupled with an understanding of the problem e.g. what types of errors are critically important, are key elements to deciding which estimation procedure "fits" a given problem well (Johnson, 2003). In speech recognition, word boundary detection is a main method. In spoken language, there are no gaps between words; where to place the word boundary often depends on what choice makes the most sense grammatically and given the context. In written form, languages e.g. Chinese do not have word boundaries either. Therefore, an estimation theory is suitable for the purpose of word boundary detection.

Time delay estimates are used to locate the position of voice source. The main method is time domain cross correlation. The cross correlation between two zero-mean stationary random processes $x_1(t)$ and $x_2(t)$ is defined as (Myers, Erim, & Lowery, 2004):

$$R_{x_1x_2}(\tau) = E[x_1(t)x_2(t+\tau)] \tag{2.25}$$

where E [·] is the estimation operator. Assuming periodicity, for single time-limited realizations of each random process, this is determined using the integral:

$$R_{x_1x_2}(\tau) = \int_{-\infty}^{\infty} x_1^*(t)x_2(t+\tau)dt \tag{2.26}$$

where * denotes complex conjugation and τ is the time lag between the signals. The Fourier transform of the cross correlation function, defines the cross-spectrum, $G_{x_1x_2}(f)$. Cross

correlation functions are unbounded measures and are typically normalized by the values of the autocorrelations at zero lag to bound the estimate between -1 and 1. The autocorrelation functions are the time domain equivalent of the auto power spectra and their value at zero lag represents the total energy in the signal. The normalized and bounded measure is known as the cross correlation coefficient, $\rho_{x_1x_2}(\tau)$, which provides a measure of the linear association between the two signals at a given time lag and is given by:

$$\rho_{x_1x_2}(\tau) = \frac{R_{x_1x_2}(\tau)}{\sqrt{R_{x_1x_1}(0)R_{x_2x_2}(0)}} \tag{2.27}$$

represents the time series as a binary pulse train with ones corresponding to the firing times of the signals. A moving average window is then used to smooth these binary signals, which is analogous to filtering the time-series with a low-pass filter (Myers et al., 2004).

2.3.2 Magnitude squared coherence

Magnitude Squared Coherence (MSC) is a frequency domain method which can be applied to voice activity detection. Carter et al. (1973) (Carter, Knapp, & Nuttall, 1973) describe a method for estimating the magnitude-squared coherence function for two zero-mean wide-sense-stationary random processes is presented. The estimation technique utilizes the weighted overlapped segmentation fast Fourier transform app roach. Analytical and empirical results for statistics of the estimator are presented. The analytical expressions are limited to the non-overlapped case; empirical results show a decrease in bias and variance of the estimator with increasing overlap and suggest a 50-percent overlap as being highly desirable when cosine (Hanning) weighting is used.

The coherence between two zero-mean stationary random processes x_1 (t) and x_2 (t), at frequency f, is defined as:

$$\gamma_{x_1x_2}(f) = \frac{G_{x_1x_2}(f)}{[G_{x_1x_1}(f)G_{x_2x_2}(f)]^{1/2}} \tag{2.28}$$

where $G_{x_1x_2}(f)$ is the cross spectral density and $G_{x_1x_1}(f)$ and $G_{x_2x_2}(f)$ are the auto spectral density functions of $x_1(t)$ and $x_2(t)$ respectively. The coherence function is a complex quantity and its squared magnitude provides a bounded measure of linear association between the two series, taking on a value of 1 for a perfect linear relationship and a value of 0 to indicate that the series are uncorrelated. In practice, it is necessary to estimate the magnitude squared coherence, $C_{x_1x_2}(f) = |\gamma_{x_1x_2}(f)|^2$, by windowing the time series to obtain multiple sections as follows:

$$\hat{C}_{x_1x_2}(f) = \frac{|\sum_{n=1}^{N} X_{1n}(f) \sum_{n=1}^{N} X^{*}{}_{2n}(f)|^2}{\sum_{n=1}^{N} |X_{1n}(f)|^2 \sum_{n=1}^{N} |X_{2n}(f)|^2} \tag{2.29}$$

where * denotes complex conjugation, N is the number of data segments employed and X_{1n} (f) and X_{2n} (f) are the discrete Fourier transforms of the nth data segments of x_1 (t) and x_2 (t). This estimate is biased and its probability density function for non-weighted and non-overlapping windows has been analytically determined. This may be used to derive the value of the estimated coherence, with a particular probability of occurrence, α, that would be obtained when the true value is zero. Any value exceeding this level is considered to be unlikely to be a false indication of coherence with ($\alpha \times 100$) % confidence. This confidence level is given by

$$E_a = 1 - (1-\alpha)^{1/(N-1)} \tag{2.30}$$

The resolution of the coherence estimate is determined from the inverse of the length of the windowed sections, i.e., for a 2 s window, the coherence resolution will be 0.5 Hz. (Myers et al., 2004)

Passive sonar systems using two or more stationary receivers have been used for the detection and localization of moving acoustic targets. In far-field configurations, where the receivers are remotely located from the source, the use of the magnitude-squared coherence function (MSCF) has proved attractive. Patzewitsch (Patzewitsch, Srinath, & Black, 1979) consider near-field configurations, where the inter-receiver spacing is not appreciably greater than the distance from the source. The concept of an extended MSCF (EMSCF) is introduced to take into account the relative Doppler shift and time delay between the signals at two receivers. Asymptotic expressions are derived for the EMSCF as the number of data segments becomes large. The effect on the EMSCF of noise in the received signals is analyzed by obtaining approximate expressions for the mean and variance of the density function of the EMSCF. An example is presented to illustrate the use of these expressions in the evaluation of the performance of passive sonar detectors.

Recently Wang and Tang (2004) (S. Wang & Tang, 2004) described the magnitude-squared coherence function is widely used in many applications. The approximate confidence interval is only reliable for large data segments. In their research, an iterative algorithm wss provided to compute the exact confidence interval from the cumulative distribution function. In order to use the confidence interval conveniently in practice, some libraries were provided using the iterative algorithm and cubic spline interpolation.

2.4 Cocktail party effect and solution

2.4.1 Cocktail party effect

Whilst a single speaker is talking in a mixture of conversations and background noises, a listener could focus on this talk and ignore other conversations. The ability to focus one's listening attention on such a mixture of conversations and background noise environment is called the cocktail party effect(Simon Haykin & Chen, 2006). The effect was first introduced by Colin Cherry in 1953 (Cherry, 1953). Cao et al. (Cao, Sridharan, & Moody, 1997) presented a speech separation structure which simulates the cocktail party effect using a modified iterative Wiener filter and a multi-layer perceptron neural network. The neural network is used as a speaker recognition system to control the iterative Wiener filter. The neural network is a modified perceptron with a hidden layer using feature data extracted from LPC cepstral analysis. The proposed technique has been successfully used for speech separation when the interference is competing speech or broad band noise. Yu et al. (Yu, Boling, Mingyang, & Chongzhi Yu, 2000) presented a improved iterative speech enhancement approach to suppress the undesired speech or noise in cocktail party environments. The key idea in the proposed approach is that a difference phase-correlation-based criterion function (PCF) is introduced to work as the convergence controller for a modified iterative Wiener filter (MIWF). The interference component is attenuated continually until convergence occurs. Evaluation results of objective quality measurement and informal listening test show the proposed system provides speech with good quality and good intelligibility.

In real-time environment, many approaches had been done with digital filters e.g. adaptive filters.

2.4.2 Adaptive digital filter

Adaptive digital filter is a most important part of current applications of noise cancellation. In fact, basic beamforming cannot cancel the noise or undesired sound completely. An adaptive digital filter is employed to improve its SNR.

An adaptive filter can be Finite Impulse Response (FIR) or Infinite Impulse Response (IIR) (Bellanger, 2001). FIR or IIR filter designs itself based on the characteristics of the input signal to the filter and a signal which represent the desired behaviour of the filter on its input. Designing the filter does not require any other frequency response information or specification (Rorabaugh, 1999).

IIR can be represented by

$$y(n) = \sum_{k=0}^{\infty} h(k)x(n-k) \tag{2.31}$$

And FIR can be represented by

$$y(n) = \sum_{k=0}^{N-1} h(k)x(n-k) \tag{2.32}$$

There are two main classes of adaptive filtering algorithm: recursive least squares (RLS) and least mean squares (LMS) algorithms. The LMS algorithm is usually considered to be the low computational cost, low convergence rate option. RLS algorithms theoretically give the least squares optimal output at every time step but are more computationally expensive (Skidmore & Proudler, 2001). The IIR approach also can have problems with stability of the solution.

2.4.2.1Finite impulse response filter

To defines relation between the input signal and the output signal, a difference equation can be presented as (*Finite impulse response*, 2008)

$$y(n) = a_0x(n) + a_1x(n-1) + ... + a_Nx(n-N) \tag{2.33}$$

where $x(n)$ is the input signal, $y(n)$ is the output signal and $a_k, k = 1,2,...N$ are the filter coefficients. N is known as the filter order.

The equation can also be expressed as a convolution of filter coefficients and the input signal.

$$y(n) = \sum_{k=0}^{N} a_k x(n-k) \tag{2.34}$$

To find the impulse response we set

$$x(n) = \delta(n) \tag{2.35}$$

where $\delta(n)$ is the Kronecker delta impulse. The impulse response for an FIR filter is the set of coefficients a_n as follows

$$h(n) = \sum_{k=0}^{N} a_k \delta(n-k) = a_n \tag{2.36}$$

where $n = 0,...,N$.

The Z-transform of the impulse response yields the transfer function of the FIR filter

$$H(z) = Z\{h(n)\} = \sum_{k=0}^{N} h(n) z^{-n} = \sum_{k=0}^{N} a_n z^{-n} \tag{2.37}$$

A FIR filter has a number of useful properties which sometimes make it preferable to an infinite impulse response filter. FIR filters has no feedback require. This means that any rounding errors are not compounded by summed iterations. The same relative error occurs in each calculation. The FIR filter can be designed to be linear phase, which means the phase change is proportional to the frequency e.g. a crossover filters, where transparent filtering is adequate. (*Finite impulse response*, 2008)

2.4.2.2Infinite impulse response filter

To defines relation between the input signal and the output signal, a difference equation can be presented as (*Infinite impulse response*, 2008)

$$y(n) = b_0 x(n) + b_1 x(n-1) + ... + b_P x(n-P) - a_1 y(n-1) - a_2 y(n-2) - ... - a_Q y(y-Q) \tag{2.38}$$

Where P is the feed-forward filter order, bi are the feed-forward filter coefficients, Q is the feedback filter order, ai is the feedback filter coefficients, $x(n)$ is the input signal, $y(n)$ is the output signal.

A more condensed form of the difference equation is:

$$y(n) = \sum_{i=0}^{P} b_i x(n-i) - \sum_{j=1}^{Q} a_j y(n-j) \tag{2.39}$$

if we let $a_0 = 1$ it becomes:

$$\sum_{i=0}^{P} b_i x(n-i) = \sum_{j=1}^{Q} a_j y(n-j) \tag{2.40}$$

.

To find the transfer function of the filter, firstly we take the Z-transform of each side of the above equation, where we use the time-shift property to obtain:

$$\sum_{i=0}^{P} b_i z^{-i} X(z) = \sum_{j=0}^{Q} a_j z^{-j} Y(z) \tag{2.41}$$

We define the transfer function to be:

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{i=0}^{P} b_i z^{-i}}{\sum_{j=0}^{Q} a_j z^{-j}} \tag{2.42}$$

Using the transfer function we can judge whether or not a system is bounded-input, bounded-output (BIBO) stable. The BIBO stability criteria require all poles of the transfer function to have an absolute value smaller than 1. It means all poles must be located within a unit circle in the *z*-plane.

The poles are defined as the values of *z* which make the denominator of *H*(*z*) equal to 0:

$$\sum_{j=1}^{Q} a_j z^{-j} = 0 \tag{2.43}$$

If $a_j \neq 0$, the poles are not located at the origin of the z-plane. It is in contrast to the FIR filter, where all poles are located at the origin, and is always stable (Finite impulse response, 2008).

IIR filters are sometimes preferred over FIR filters because an IIR filter can achieve a much sharper transition region roll-off than an FIR filter of the same order.

2.4.3 Wiener filter

The Wiener filter was introduced by Norbert Wiener in 1949 (Wiener, 1949) and independently for the discrete-time case by Kolmogorov (Kolmogorov, 1941).

The goal of the Wiener filter is to filter out additive noise from a desired signal. It is based on a statistical approach. Typical standard filters are designed for a desired frequency response. An ordinary filter is assumed to have knowledge of the spectral properties of the original signal and the noise, and the Wiener filter would come as close to the original signal as possible. (*Wiener filter*, 2008)

In an adaptive noise cancellation, a discrete Wiener filter is very often applied. With N weights, the primary signal (signal + noise) y_k, the noise x_k, and the error e_k, a Wiener FIR filter output can be

$$e_k = y_k - \mathbf{W}^T X_k = y_k + \sum_{i=1}^{N-1} w(i) x_{k-1} \tag{2.44}$$

where X_k input signal vector is

$$\mathbf{X}_k = \begin{bmatrix} x_k \\ x_{k-1} \\ \vdots \\ x_{k-(N-1)} \end{bmatrix} \tag{2.45}$$

and W the weight vector is

$$\mathbf{W} = \begin{bmatrix} w(0) \\ w(1) \\ \vdots \\ w(N-1) \end{bmatrix} \tag{2.46}$$

The Wiener filter normally uses a cost function of mean-square error which is minimized to give the optimal solution. The square of the error is given as

$$e_k^2 = y_k^2 - 2 y_k \mathbf{X}_k^T \mathbf{W} + \mathbf{W}^T \mathbf{X}_k \mathbf{X}_k^T \mathbf{W} \tag{2.47}$$

When the filter weight vector has its optimum value,

$$\mathbf{W}_{opt} = \mathbf{R}^{-1} \mathbf{P} \tag{2.48}$$

Where

$$\mathbf{R} = E[\mathbf{X}_k \mathbf{X}_k^T], \quad \mathbf{P} = E[y_k \mathbf{X}_k] \tag{2.49}$$

The $E[\bullet]$ is symbolized expectation. $\mathbf{R}$ is the $N \times N$ autocorrelation matrix and $\mathbf{P}$ is the N length cross-correlation matrix.

The equation 2.48 is known as the Wiener-Hopf solution. Ifeachor (Ifeachor, 1993) suggested the Wiener filter has a limited practical usefulness because R and P are not known a priori, and matrix inversion is required. When the signals are non-stationary, R and P will change with time. For real-time application, Ifeachor (Ifeachor, 1993) considered a way of obtaining W_{opt} on a sample-by-sample basis: Adaptive algorithms are used to achieve this without having to compute R and P explicitly or performing a matrix inversion.

2.5 Speech enhancement in car noise environments

2.5.1 Overview

One of the most challenging and important problems in Intelligent Transport Systems (ITS) is to keep the driver's eyes on the road and his hands on the wheel. Speech recognition offers one such solution to this problem. Speech control in a car is a safe solution e.g. to enter a street name in a Global Positioning System (GPS) navigation system by speech is better than to do it by hand. However, speech recognition in a car has the inherent problem of acquiring speech signals in noisy environments. There are two types of additive noise in a car cabin: stationary and non-stationary. Stationary noise in car is from the engine (though it varies with speed), road, wind, air-conditioner etc. Non-stationary noise is from the car stereo, navigation guide, traffic information guide, bumps, wipers, indicators, conversational noise and noise when passing a car travelling in the opposite direction (Shozakai, Nakamura, & Shikano, 1998). To understand car noise environment, Puder and Dreiseitel (Puder & Dreiseitel, 2000) presented an improved method for the spectral estimation of car noise in order to enhance the performance of noise reduction systems. An algorithm is developed that allows us to track changes in the noise spectrum during speech activity. They suggested there are three major components of car noise which are weighted differently according to the type of car and engine, are engine, wind, and tyre noise. These three components are examined in detail in the following and a dependence of these components on the characteristic car parameters is worked out.

Engine Noise - The rotating engine and the movement of the pistons generate a harmonic noise spectrum having narrow band spectral components. The frequency localization of these components depends directly on the rpm of the engine. The harmonic structure of the spectrum can be easily identified. Engines normally exhibit spectral components at multiples of half the rpm frequency.

Wind Noise - For modern cars with low aerodynamic drag, wind noise generally exhibits less power than tyre noise. Remarkable components normally only appear at higher speeds. The higher noise power for the higher speed is obvious. However, it is especially interesting that the characteristic of the power spectral density stays nearly the same.

Tyre / Rolling Noise - For tyre noise, a correlation of the power spectral density of the noise and the speed of the car can also be discerned. Even though the spectral characteristics are not identical, they are very similar. It is important to notice that frequency shifts of spectral components do not appear when the car speed changes.

Most of the researches in car noise environments are applied above noise components and normally are pre-defined.

At the end of the 1980s, Lecomte et al (Lecomte, Lever, Boudy, & Tassy, 1989) addressed the problem of speech enhancement in car environment. After describing the method used to characterize the noisy environment, the main influence on speech processing in two applications: speech recognition and speech transmission.

2.5.2 Voice Activity Detection in a car

2.5.2.1 Beamforming applications in automotives

The most popular application in a car is Griffiths and Jim beamformer as discussed in earlier chapter.

Zhang and Hansen (2003) (Xianxian Zhang & John H.L. Hansen, 2003) have investigated various speech enhancements and processing schemes for in-vehicle speech systems, the delay-and-sum beamforming (DASB) and adaptive beamforming are two typical methods that both have their advantages and disadvantages. A combined fixed/adaptive beamforming solution was presented for speech enhancement and recognition in real moving car environment. The working scheme consists of two steps: source location calibration and target signal enhancement. The first step is to pre-record the transfer functions between speaker and microphone array from different potential source positions using adaptive beamforming under quiet environments; and the second step is to use this pre-recorded information to enhance the desired speech when the car is running on the road.

Fuchs et al. (2004) (Fuchs, Haulick, & Schmidt, 2004) discuss the use of adaptive beamformers in noise suppression systems for automotive applications with two-microphone system. They presented a method of extracting additional spatial information from a conventional beamformer in generalized sidelobe structure. This spatial information can be utilized to control parameters like overestimation or spectral floor of classical noise suppression schemes in a frequency selective manner or to compute a simple attenuation factor for suppressing nonstationary noise.

Zhang et al. (2004) (Zhang et al., 2004) presented hardware prototypes that integrate several heterogeneous sensors into a single headset and describe the underlying DSP techniques for robust speech detection, enhancement and recognition in highly non-stationary noisy environments. However, this hardware prototype with headset is not an ideal input device for an automobile driver.

2.5.2.2 Detect a geometrical zone with three-microphone beamforming

Chen and Moir (1999) (W. Chen, N & Moir, 1999) presented a three-microphone system in lab as Figure 1. The A new active word boundary detection of algorithm of speech is investigated. In this algorithm three microphones are used to detect the desired and undesired periods of speech by defining a geometrical ‘active zone’. With three microphones this word

boundary detector can retrieve the desired speech embedded with noise from varieties of noisy background. Some simulation experiments are produced in this paper to show that the algorithm is an effective speech detecting method that exceeds to an average 80% of success rate.

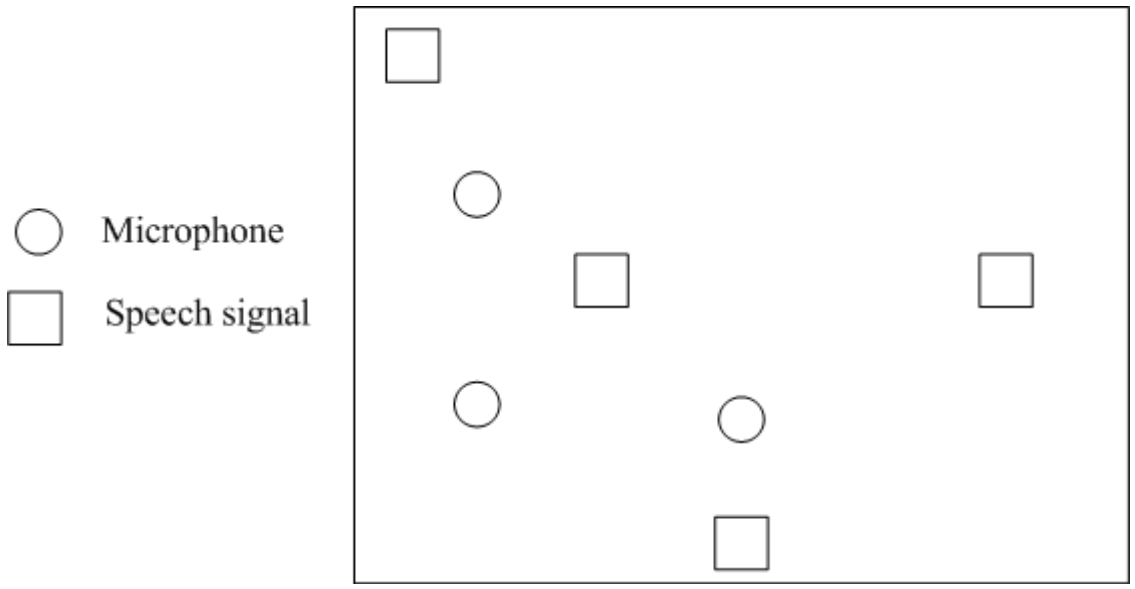

Figure 2-8 Chen and Moir's three-microphone system

2.5.2.3Speech enhancement for automotive speech recognition in a car

From earlier research, Lecomte et al (Lecomte et al., 1989) stated without modifications the results obtained with car noise were poor. With a pre-processing of the noisy signal, improvements were shown especially for speech recognition; the acoustical analysis proved that adaptive noise cancelling method with two inputs is not suitable in a car environment. The performances obtained with white noise may not be generalized directly for other environments. Traditional adaptive noise cancelling methods use two inputs: noisy speech and a noise reference that corrupts the primary input. But these methods do not work in a car because of the low coherence of noise between two microphones: the only coherent noise is the engine noise below 500 Hz. It is not possible to process a SNR because the filter coefficients change. For noisy speech filtered high pass there is no kind of improvement.

Lockwood et al (Lockwood, Boudy, & Blanchet, 1992) applied Nonlinear Spectral Subtraction (NSS) and Hidden Markov Model (HMM) in defined stationary noise environment. Then Alexandre et al (Alexandre, Boudy, & Lockwood, 1993) extended its previous contribution by proposing a robust front-end achieving reliable performance without need for any specific noise compensation. This front-end is based on the Root homomorphic deconvolution schemes. They presented the standard speech analysis schemes with a unified Root-cepstral framework.

Erzin et al (Erzin, Cetin, & Yardimci, 1995) proposed a set of speech feature representations for robust speech recognition in the presence of car noise. These parameters were based on sub-band analysis of the speech signal. In their research, Line Spectral Frequency (LSF) representation of the Linear Prediction (LP) analysis in sub-bands and cepstral coefficients were derived from sub-band analysis, and the performances of the new feature representations were compared to mel scale cepstral coefficients in the presence of car noise. Sub-band analysis based parameters were observed to be more robust than the commonly employed mel scale cepstral coefficients representations.

Yang and Haavisto (Yang, Yang, & Haavisto, 1995) presented a noise compensation algorithm for HMM based speech recognition systems, which utilizes the parallel model combination concept.

Dahl (Dahl, Claesson, & Nordebo, 1997) presented a method to simultaneously perform acoustic echo cancellation and speech enhancement using an adaptive microphone array combined with spectral subtraction. Primarily intended for hands-free telephones in automobiles, the microphone array system simultaneously emphasizes the near-end talker and

suppresses the hands-free loudspeaker and the broadband car noise. The array system was based on a fast and efficient on-site calibration and can be used in other situations such **as** conventional speaker phones.

Jabloun et al (Jabloun, Cetin, & Erzin, 1999; Jabloun & Enis Cetin, 1999) presented a set of speech feature parameters based on multi-rate signal processing and a Teager Energy Operator. The speech signal was first divided into non-uniform sub-bands in mel-scale using a multi-rate filter-bank, and then the Teager energies of the sub-signals were estimated.

Jeong and Hahn (Jeong & Hahn, 2001) applied High-pass filtering and Kalman filtering with a whitening process to car noise corrupted speech signals. The improvement both in speech quality and in speech recognition is achieved in car noise environments.

Ban et al (Ban, Banno, Takeda, & Itakura, 2002) described a method for generation of car noise based on the engine noise and the "friction noise". The engine noise was modeled by composition of a stationary ac ground noise that depends on the size of the engine a non-stationary noise that depends rotational speed of the engine. The friction noise was modeled as a white noise with ranging power. Based on these models, methods for synthesis of these components were developed. Subjective assessment of the car noise synthesis method showed that it is fairly similar to the actual noise.

Wang et al (J. Wang, Yang, & Chang, 2004) proposed a method incorporates a perceptual filter bank which is derived from a psycho-acoustic model for sub-band processing. The experiments were performed using the Taiwan in-CAR speech database TAICAR in-car noisy speech database. Subjective and objective tests showed that its method outperforms other existing signal subspace methods.

Cho and Ko (Cho & Ko, 2004) presented a Griffiths-Jim acoustic beamformer in reducing stationary or non-stationary noise in car cabin.

Li et al (W. Li, Takeda, & Itakura, 2005) addressed issues in improving hands-free speech recognition performance in different car environments. They proposed a speech-enhancement approach based on optimizing regression of the log-spectra, which was used to estimate the log-spectra of speech at a close-talking microphone by using multiple spatially distributed microphones. The regression weights can be adapted automatically for different noise environments. Compared to the nearest distant microphone and adaptive beamformer generalized sidelobe canceller (GSC), the proposed approach shows an advantage in the average relative word error rate (WER) reduction of 58.5 and 10.3%, respectively, for isolated word recognition under 15 real-car environments.

Ding et al (Ding, He, Yan, Zhao, & Hao, 2006) presented a research on robust automatic speech recognition in car noise environments. In this front-end design, speech enhancement technologies were used to suppress the background noise in frequency domain, and then spectrum smoothing is implemented both in time and frequency index to compensate those spectrum components distorted by noise over-reduction. In acoustic model training, they proposed to use an immunity learning scheme, in which pre-recorded car noises were artificially added to clean training utterances with different signal-to-noise ratios (SNR) to imitate the in-car environments. After analyzing the SNR and noise spectrum of real in-car utterances, they refined the immunity training set by adjusting the distribution of SNR and increasing the proportion of training noises that has a similar characteristic. Evaluation results of isolated phrase recognition showed that the ASR system with proposed technologies achieves the average error rate reduction (ERR) of 90.68% and 79.08% for artificial car noisy speech and real in-car speech respectively.

Ma et al (Ma, Shangguan, & Zang, 2007) used a Band Partitioning Spectral Entropy endpoint detection (BPSE) method to get the speech start and end point of speech precisely. Mel Frequency Cepstral Coefficients (MFCC) was extracted from segmented speech signals. The coefficients were recognized by Hidden Markov Model. The results showed that the recognition accuracy was improved from 39.3% to 95.5%.

2.5.3 Speech enhancement for voice communication in car

2.5.3.1 An Adaptive filter in a car

Since telecommunication became popular, and safety issue on using it in a car has arisen indicating, a hands-free voice communication system is needed. In earlier 1990s, Oh et al (Oh, Viswanathan, & Papamichalis, 1992) presented the result of their research on developing a hands-free voice communication system with a microphone array for use in an automobile environment. Their goal of this research was to develop a speech acquisition and enhancement system so that a speech recognizer can reliably be used inside a noise automobile environment, for digital cellular phone application. Speech data have been collected using a microphone array and a digital audio tape (DAT) recorder inside a real car for several idling and driving conditions, and processed using delay-and-sum and adaptive beamforming algorithms. Performance criteria including signal-to-noise ratio and speech recognition error rate have been evaluated for the processed data. Detailed performance results presented show that the microphone array was superior to a single microphone

Nordholm (Nordholm, Claesson, & Dahl, 1999) gave an analytical description of an adaptive microphone array that facilitates a simple built-in calibration to the environment and instrumentation. This method, suggested for use in hands-free mobile telephones and speech recognition systems for cars, provides speech enhancement and acoustic echo-cancellation. The scheme offers several advantages, such as a simple calibration procedure, suppression of directional sources, versatile robust beamforming, and reduced target signal distortion. The analysis employed non-causal Wiener filters yielding compact and effective theoretical suppression limits.

Grbic et al (Grbic, Nordholm, & Johansson, 2001) discussed signal processing methods for speech extraction in use with voice communication applications such as personal digital assistants (PDA), mobile telephone terminals and personal computers. The user was distant from the device and thus the speech signal entering the device may be subject to reverberation and may be disturbed by background noise. The proposed structure consists of a microphone array which allows for techniques of directional processing. Three different optimal beamforming methods were considered in a real world car hands-free environment: an optimal near-field array gain optimization procedure, a theoretical diffuse noise field model for a point source and a least squares solution.

De Haan et al (de Haan, Grbic, Claesson, & Nordholm, 2002) presented a method for the design of non-uniform DFT filter banks for sub-band beamforming. Filter banks designed with the method are evaluated in sub-band beamforming in a real-world microphone array application. The objective of the proposed design was to minimize the magnitude of all aliasing components individually, such that aliasing distortion was minimized although phase alterations occur in the sub-bands. The proposed method was evaluated in a car hands-free mobile telephony environment with real speech signals. The results showed that the performance can be increased by several decibels when using non-uniform filter banks instead of uniform filter banks while maintaining the length of the sub-band filters.

Yermeche et al (Yermeche, Garcia, Grbic, & Claesson, 2002) proposed a calibrated adaptive frequency domain beamformer for speech enhancement. The beamformer is based on the principle of a soft constraint formed from calibration data, rather than pre-calculated from free-field assumptions. The proposed algorithm continuously estimated the spatial information for each frequency band, based on weighting of the received data. The update of the beamforming weights was done recursively where the initial pre-calculated correlation estimates of the speech constitute a soft constraint. The soft constraint secured the spatial-temporal passage of the desired source signal, without the need of any speech detection. The performance was evaluated in real world scenarios, in both car and restaurant environments. Interference and noise suppression was achieved at more than 15 dB with very small distortion.

Cohen (Cohen, 2003) analysed a two-channel generalized sidelobe canceller with post-filtering in nonstationary noise environments. The post-filtering includes detection of transients at the beamformer output and reference signal, a comparison of their transient power, estimation of the signal presence probability, estimation of the noise spectrum, and spectral enhancement for minimizing the mean-square error of the log-spectra. Transients were detected based on a measure of their local non-stationarity, and classified as desired or interfering based on the transient beam-to-reference ratio. The evaluation showed that desired and interfering transients can generally be differentiated within a wide range of frequencies. To further improve the transient noise reduction at low and high frequencies in case the signal is wideband, they estimated for each time frame a global likelihood of signal presence. The global likelihood is associated with the transient beam-to-reference ratios in frequencies, where the transient discrimination quality is high. Experimental results showed the usefulness of the proposed approach in various car environments.

Cohen and Berdugo (Cohen & Berdugo, 2003) designed multichannel systems for spatially filtering interfering signals coming from undesired directions in reverberant and noisy environments. They presented a two-channel post-filtering approach for signal detection and speech enhancement. A mild assumption was made, that a desired signal component was stronger at the beamformer output than at the reference noise signal, and a noise component is stronger at the reference signal. The ratio between the transient power at the beamformer output and the transient power at the reference noise signal was used for indicating whether such a transient is desired or interfering.

Zhang and Hansen (Xianxian Zhang & J.H.L. Hansen, 2003) investigated various speech enhancement and processing schemes for in-vehicle speech systems, little research has been performed using actual voice data collected in noisy car environments. They proposed a constrained switched adaptive beamforming algorithm (CSA-BF) for speech enhancement and recognition in real moving car environments. The proposed algorithm consists of a speech/noise constraint section, a speech adaptive beamformer, and a noise adaptive beamformer. They investigated CSA-BF performance with a comparison to classic delay-and-sum beamforming in realistic car environments using a large quantity of data recorded in various car noise environments from across the United States. After analyzing the experimental results and considering the range of complex noise situations in the car environment using the CU-Move corpus, they formulated the CSA-BF algorithm. This method was shown to decrease word error rate for speech recognition by up to 31% and improve speech quality via the segment signal-to-noise ratio by up to 5.5 dB on the average, simultaneously.

Cornelius et al (Cornelius, Yermeche, Grbic, & Claesson, 2004) discussed speech enhancement in an enclosed environment such as communication in a motorcycle helmet. They proposed a constrained sub-band adaptive beamformer for a hands-free in-car environment. The highly non-stationary nature of the disturbing sound field encountered in a motorcycle helmet and the fact that the source is situated in the extreme near-field of the array, causes the beamformer to produce an unwanted fluctuation in the output level. The proposed spatially constrained beamformer ensured that the output maintains a constant gain, as long as the corresponding source originates from the desired location.

Hai et al (Hai Huyen, Nordhohm, Dam, & Siow Yong Low, 2004) presented an adaptive beamformer employing updates for both the source and interference short term power spectral densities (PSDs). These PSD updates track the variations in the spectral content, thereby yielding a spectrally optimized constraint for each time instant. Consequently,

probability density function models for the source, interference and noise covariance matrices are proposed for the estimation of the PSD parameters. An optimization problem is formulated based on the proposed model. The evaluation in a real car environment showed significant suppression levels for the noise and interference by employing the proposed beamformer whilst maintaining low source signal distortion.

Hai et al (Hai Huyen et al., 2004) presented a sub-band adaptive beamformer equipped with noise statistics updates. These updates were employed to effectively estimate and track the noise statistics continuously in the solution. Additionally, an update on the source power spectral density (PSD) was incorporated to enhance the timbre of the source of interest. Furthermore, the beamformer was also equipped with a space constraint on the source area to provide robustness against steering vector errors and good capture of the target signal spatially. The evaluations on real car data with variations in the car noise level show that the proposed scheme achieves a good noise suppression level up to 20 dB.

Hai et al (Hai Quang, Siow Yong, Hai Huyen, & Nordholm, 2004) presented a space constrained adaptive beamformer employing an updated source power spectral density (PSD). The space constraints were used to capture the target signal spatially and to provide robustness against steering error vectors. The PSD update on the other hand ensures that the spectral information of the desired source is reflected continuously on the space constraints. Therefore, target signal extraction can be achieved with minimum distortion. The beamformer operated in a sub-band structure to allow time-frequency operation for each channel, yielding a combination of weighted spatial and temporal filters. Evaluations on real car data show that the proposed algorithm significantly improves the speech intelligibility with noise suppression level up to 21 dB.

Yermeche et al (Yermeche, Cornelius, Grbic, & Claesson, 2004) presented a spatial filter bank design method for speech enhancement beamforming applications. The aim of this design was to construct a set of different filter banks that would include the constraint of signal passage at one position (and closing in other positions corresponding to known disturbing sources). By performing the directional opening towards the desired location in the fixed filter bank structure, the beamformer was left with the task of tracking and suppressing the continuously emerging noise sources. This algorithm has been implemented in MATLAB and tested on real speech recordings conducted in a car hands-free communication situation. Results show that a reduction of the total complexity can be achieved while maintaining the noise suppression performance and reducing the speech distortion.

Davis et al (Davis, Siow Yong, Nordholm, & Grbic, 2005) introduced a sub-band adaptive space constrained beamforming structure for use in hands-free speech enhancement applications. The scheme incorporates space constrained source model and voice activity information through the integration of a VAD. The VAD information was used to estimate noise covariance information during non-speech periods and to optimally estimate the source power spectral density, which is used to provide a spectrally optimized constraint on the source. The proposed structure was evaluated in a real car environment, the results compared well to the optimal Wiener solution.

Hai et al (Hai Quang, Nordholm, Hai Huyen, & Siow Yong Low, 2005) proposed a adaptive algorithm for speech enhancement in hands-free communication systems. The scheme aims to enhance a target signal corrupted by background noise and by acoustic echo. The main idea was the incorporation of both the source and the interference short-term spectral amplitudes in the solution. These spectral components were estimated iteratively to provide a spectrally optimized constraint while the background noise statistics are estimated and updated continuously. Evaluations in a hands-free car environment showed that the proposed algorithm is superior in suppressing both the noise and the interference even in a double-talk situation.

Li et al (W. Li et al., 2005) addressed the issues in improving hands-free speech recognition performance in different car environments. They proposed a speech-enhancement approach based on optimizing regression of the log-spectra, which was used to estimate the log-spectra of speech at a close-talking microphone by using multiple spatially distributed microphones. Compared to the nearest distant microphone and adaptive beamformer generalized sidelobe canceller (GSC), the proposed approach showed an advantage in the average relative word error rate reduction of 58.5 and 10.3%, respectively, for isolated word recognition under 15 real-car environments.

Hai et al (Dam, Nordholm, Dam, & Low, 2006) investigated the problem of enhancing a desired speech source corrupted by a stationary noise. A beamformer structure was proposed which combines an adaptive constrained beamformer with a post-filtering technique. The adaptive constrained beamformer was designed to spatially enhance the desired source from the observed signal. After the adaptive constrained beamformer, a post-filter is employed to increase noise suppression capability. Experimental results in a real car situation showed that the proposed structure achieves a good noise suppression level with a low distortion.

Kim et al. (Kim, Hasegawa-Johnson, & Koeng-Mo, 2006) presented a theoretical basis for optimal multichannel speech enhancements, sufficient, flexible to be used with any assumed

statistical model and optimality criterion. Any Bayesian optimal one-channel estimator for speech enhancement can be generalized to the multichannel case as a sequentially constructed minimum variance distortionless response (MVDR) beamformer followed by an optimal one-channel post-filter. They presented experimental results using the minimum mean-square error log-spectral amplitude (MMSE-logSA) optimality criterion, applied to a statistical model with simplified channel but realistic inter-microphone noise coherence. Word error rate in the audio-visual speech in a car (AVICAR) corpus (moving car, windows open) was reduced from 18% to 9%.

Yoon et al. (Yoon, Tashev, & Acero, 2007) proposed a adaptive beamforming algorithm with enhanced noise suppression capability. The proposed algorithm incorporates the sound-source presence probability into the adaptive blocking matrix, which is estimated based on the instantaneous direction of arrival of the input signals and voice activity detection. The proposed algorithm guarantees robustness to steering vector errors without imposing ad hoc constraints on the adaptive filter coefficients. It can provide good suppression performance for both directional interference signals as well as isotropic ambient noise. For an in-car environment the proposed beamformer showed SNR improvement up to 12 dB without using an additional noise suppressor.

2.5.3.2 An Wiener filter in a car

Meyer and Simmer (Meyer & Simmer, 1997) presented a multichannel-algorithm for speech enhancement for hands-free telephone systems in cars. This algorithm takes advantage of the special noise characteristics in fast driving cars. The incoherence of the noise allows using adaptive Wiener filtering in the frequencies above a theoretically determined frequency. Below this frequency a smoothed spectral subtraction (SSS) is used to get improved noise suppression. The algorithm yielded better results in noise reduction with significantly less distortions and artificial noise than spectral subtraction or Wiener filtering alone. Chen et al. (A. Chen, Vaseghi, & McCourt, 2000) discussed the performance of Wiener filters in restoring the quality and intelligibility of noisy speech depends on: the accuracy of the estimates of the power spectra or the correlation values of the noise and the speech processes, and on the Wiener filter structure. They proposed a Bayesian method where model combination and model decomposition are employed for the estimation of parameters required to implement sub-band LP Wiener filters. The use of sub-band LP Wiener filters provides advantages in terms of improved parameter estimates and also in restoring the temporal-spectral composition of speech. The method is evaluated, and compared with the parallel model combination, using the TIMIT continuous speech database with BMW and VOLVO car noise databases.

Ortega et al. (Ortega, Lleida, Masgrau, & Gallego, 2002) presented a cabin car communication system (CCCS) to improve the communication among passengers inside a car. Noise, distance between speakers and many other factors make it difficult to maintain a conversation inside a car. The CCCS picks up the speech of each passenger, amplifies it, and uses the car loudspeaker system to return it into the cabin. Two problems arise when designing a CCCS; the electro-acoustic coupling between loudspeakers and microphones, and the amplification of the inside car noise. As a result of the first problem, the system may become unstable. To maintain the stability of the system, the CCCS makes use of a robust acoustic echo cancellation scheme based on system identification and a Wiener echo suppressor. Using a noise reduction system based on Wiener filtering reduces the second problem, noise amplification. Experimental results showing the performance of the system in terms of acoustic echo and noise reduction and speech reinforcement are presented. A system with 2-input/2-output channels has been built on a DSP board for medium size cars and minivan vehicles.

Liang et al. (Liang, Rosca, & Balan, 2003) proposed a single channel algorithm to reduce car noise. The approach employed two phases. First, independent component analysis (ICA) was applied to a large ensemble of clean speech training frames to reveal their underlying statistically independent basis. The distribution of the ICA transformed data was estimated in the training phase. Second, a Wiener filter was applied to estimate the clean speech from the received noisy speech. The Wiener filter minimized the mean-square error between the estimated signal and the clean speech signal in the ICA domain. An inverse transformation from ICA domain back to time domain reconstructed the enhanced signal. Extensive experiments showed considerable noise reduction capability of the proposed algorithm. The evaluation is performed with respect to four objective quality measure criteria.

Davis et al. (Davis et al., 2005) introduced a sub-band adaptive space constrained beamforming structure for use in hands-free speech enhancement applications. The scheme incorporates space constrained source model and voice activity information through the integration of a VAD. The VAD information was used to estimate noise covariance information during non-speech periods and to optimally estimate the source power spectral density, which was used to provide a spectrally optimized constraint on the source.

2.6 Summary

As showed in Figure 2-20, adaptive beamformer has been developed since 1972 as adaptive algorithms have been improved.

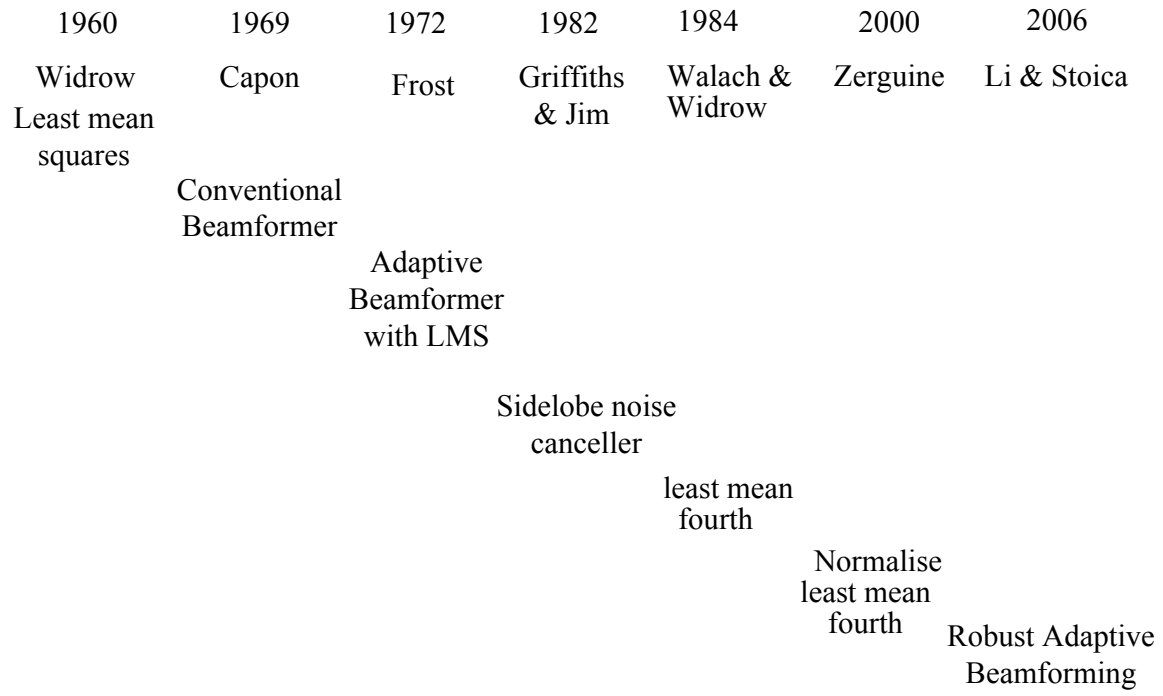

Figure 2-9 Overview of adaptive beamformer history

Literature has been reported in the field of speech enhancement in car environments since the 1970's. Most of these approaches had considered either a general speech enhancement without a link to ASR in a car, or an ASR in a car is assuming the input is a clean voice signal. Since both of above destinations have not been satisfied yet, ASR in a car has not been commercialized.

Therefore, here is a question:

As with all of these kinds of automated 'smart-car' technologies, is it not necessary for the enhanced signal to sound better to the human ear, but only needs to be good enough to provide a Boolean on or off command?

For this question, this thesis is to give a "yes".

Learnt from this literature review, real-time beamforming based VAD is able to identify sounds' direction or area; an adaptive filter with NLMS can reduce unwanted voice and enhance desired voice; an adaptive Wiener filter has a capability of picking up a desired voice from the cocktail party effect environment.

As the need of identifying the quality of signal processed by VAD and digital filter, SNR and ASR successful rate are the key points. Therefore, SNR and ASR successful rate are main figures for experiments in this thesis.

This thesis has the approach of three-microphone VAD switch with a three-microphone NLMS filter and a Wiener filter (discussion in details in following chapters), the novelty for these approaches in this thesis are:

Although multi-microphone VAD has been applied in a laboratory environment and also multi-microphone sidelobe adaptive noise canceller has been employed in stationary noise environment in a car, this thesis is the first approach to a hybrid three-microphone VAD switch with a three-microphone NLMS adaptive noise canceller in a real non-stationary noise environment in a car,.

Before the approach to adaptive Wiener filters in this thesis, an adaptive Wiener filter is normally updated by the LMS algorithm. This thesis updates the Wiener filter directly by estimation of the signal + noise and noise covariance matrices and by direct solution of the Wiener-Hopf equation. This novel solution to adaptive Wiener filter is confirmed by experiments in real non-stationary noise environment in a car.

A hybrid of item 1 and 2 is an engineering solution in a real-time environment in a car.

3 Problem definition and research environmental set-up

3.1 Problems of speech enhancement in a car

As the discussion in Chapter 2 literature review, the most challenging task of speech enhancement in a car is to deal with non-stationary noise and interference e.g. other passenger's speech or voice from radio loudspeakers as showed in Figure 3-1.

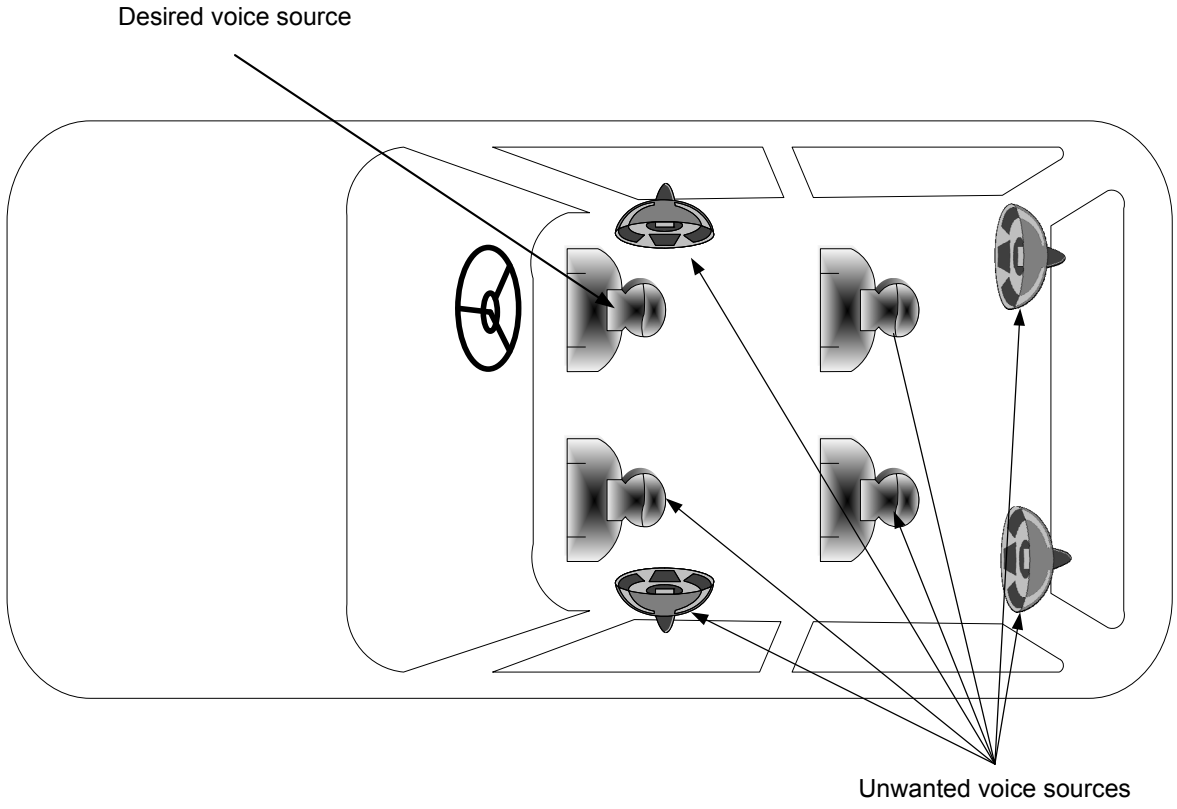

Figure 3-1 non- stationary noise and interference

3.2 Approaches on speech enhancement in a car

This thesis approaches the best solution to this problem as described in Figure 3-1. A hybrid speech enhancement in a car is presented as showed in Figure 3-2. In this hybrid system, the main approach is a 3-microphone beamforming. Since 3-microphone beamforming can mute the output to ASR while voice or interference is incoming from outside of desired zone, ASR does not receive noise or interference in this case.

Adaptive Wiener filter is an addition to this main approach when beamforming is confused by multi-source noise or interference. ASR is introduced in this thesis to test the result from the Adaptive Wiener filter.

In this thesis 3-microphone beamforming based Voice Activity Detection (VAD) is the main approach for the solution to non-stationary noise and interference. As beamforming based VAD identifies where the voice sources are incoming, the output of Digital Signal Processing (DSP) is muted while an incoming voice is not incoming from a desired area. Whilst a desired voice is picked up, a 3-microphone NLMS filter enhances this voice and reduces the background noise e.g. from the engine.

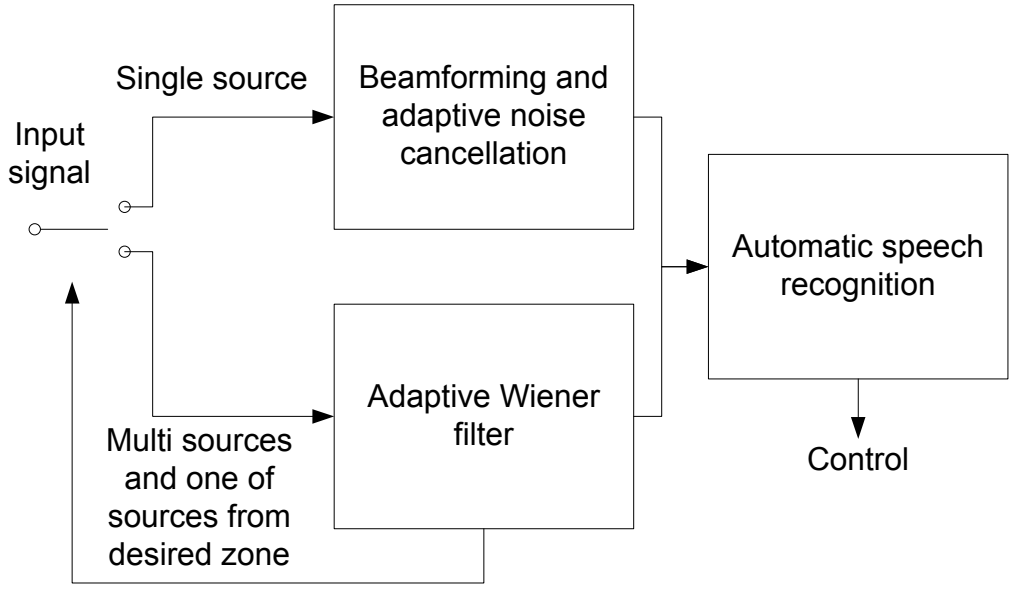

Figure 3-2 A hybrid system with acoustic beamforming VAD and an AWF

3.2.1 Three-microphone VAD switch and NLMS filter

Carter et al.(Carter et al., 1973) describe a method for estimating the magnitude-squared coherence (MSC) function for two zero-mean wide-sense-stationary random processes. The estimation technique utilizes the weighted overlapped segmentation fast Fourier transform (FFT). Analytical and empirical results for statistics of the estimator are presented. The analytical expressions are limited to the non-overlapped case. Empirical results show a decrease in bias and variance of the estimator with increasing overlap and suggest a 50-percent overlap as being highly desirable when cosine (Hanning) weighting is used. Once the MSC is found the Generalized Cross-Correlation (GCC) method is used to give a robust

estimate of time-delay. A microphone array as shown in Figure 3-3 is currently located to ensure that there is an intersection. Clearly, a linear array cannot have such an intersection. In Figure 3.3 three microphones are located as shown and there is 50 cm distance between these microphones. A desired speech source is located 35.4 cm away from Microphone 1, 2 and 3. Therefore, when speech travels to microphones 1, 2 and 3, it has the same distance to travel. The sample rate of Microphone 1, 2 and 3 is 11025 Hz, and the speed of sound in air is around 34600cm/second. Therefore during every sample the speech travels 3.1 cm so that the wave-front of speech arriving at microphones 1, 2 and 3 have no time difference of arrival (TDOA) with respect to one another. When the speaker is away from the desired position, a finite TDOA between Microphone 1, 2 and 3 is expected. For any point on a hyperbolic curve as shown at Figure 3-3, the difference between distances to a pair of microphones (as foci) is constant e.g. speech source from the star point on the hyperbolic curve travels to Microphone 1 has 5 samples intervals delayed with respect to the microphone 2.

The technique can be summarized as follows for three microphones and two estimated time-delays. When the VAD is set to be within some defined number of samples (e.g. 5 samples is typically used), then the estimation of time delay (TDOA)'s from each microphone pair is estimated and compared with some threshold value $d_{\max}$.Therefore an "VAD valid zone" is defined as in Figure 3-3 and Figure 3-4 (Qi & Moir, 2005). In Figure 3-3, the plan view of three microphones 50cm apart is shown. The distance between Microphone 1 and Microphone 3 is 70.7 cm. The microphones need not be at right-angles but are positioned in such a way that the intersection of the two hyperboloids (in 3-D space Figure 3-4) form an VAD valid zone around the drivers head (Agaiby & Moir, 1997). A 3-microphone VAD valid zone can be steered by pre-defined time-difference of arrival (TDOA). For example, the VAD valid zone can be moved towards microphone 1.

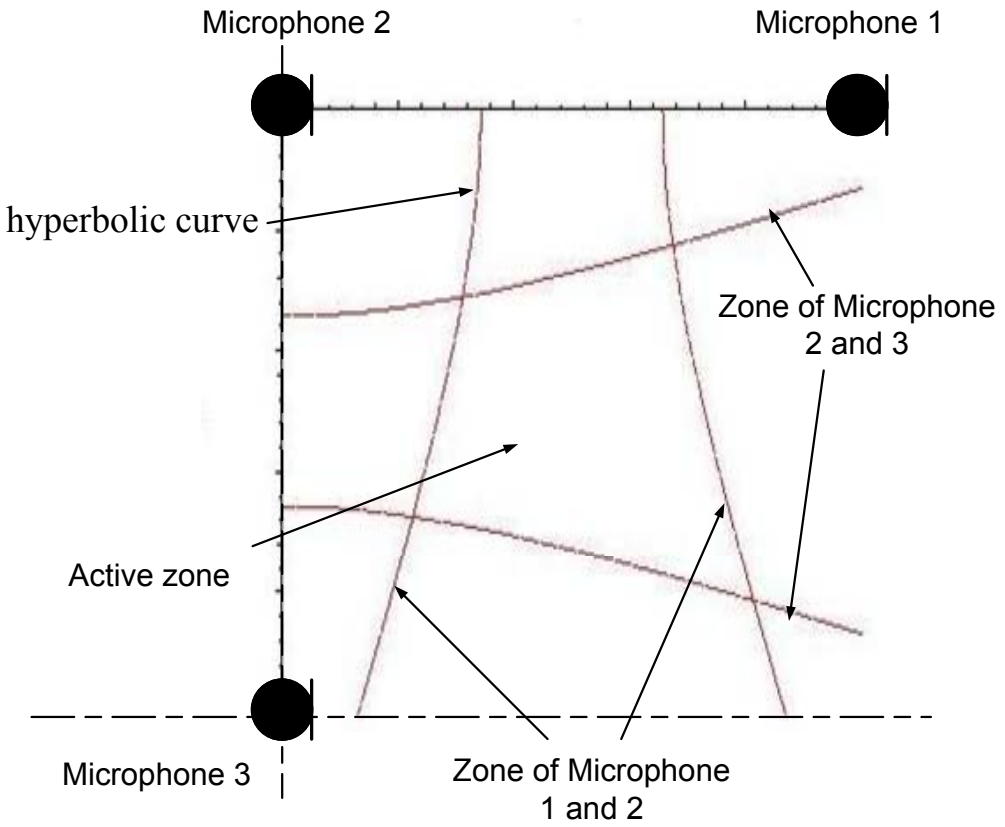

Figure 3-3 A desired zone is defined with 3 microphone

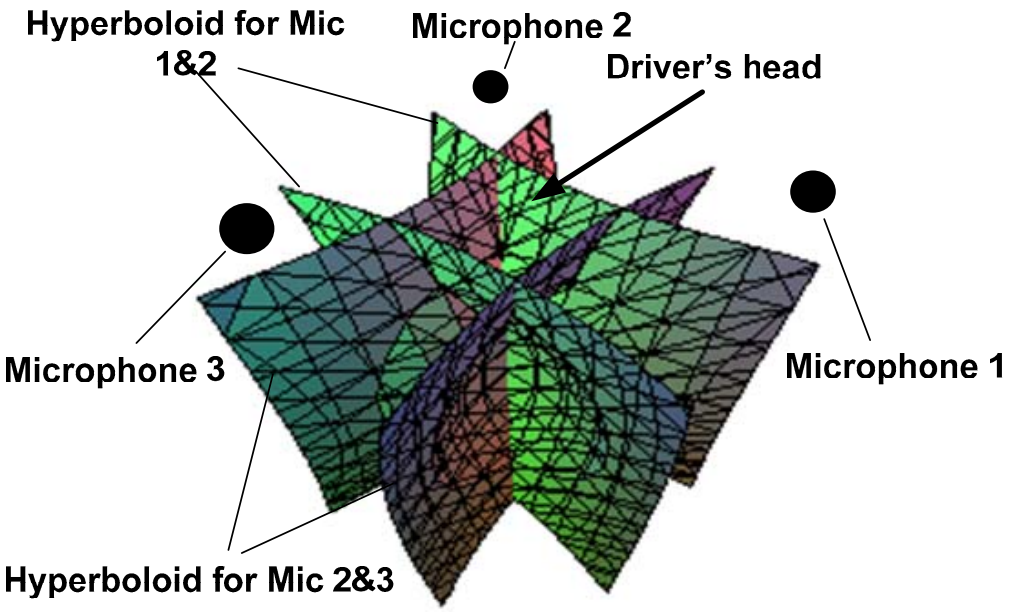

Figure 3-4 Plan view of 3-microphone VAD valid zone in 3D

These 3 microphones are also acted as data acquisition for a 3-microphone NLMS. Whilst a desired voice is picked up, this voice is filtered by a 3-microphone NLMS filter. The details will be described in chapter 4.

3.2.2 Wiener filter in 3 microphone array

From the literature review in Chapter 2, the Wiener filter is a good solution to the Cocktail party problem. As in Figure 3-4, the desired signal with noise/interference is inputted from microphone 1 and is filtered by a Wiener filter matrix. The Wiener filter matrix is updated with inputs from microphone 1 and 2. Ideally, Microphone 2 is only receiving noise/interference without the desired signal. Therefore, Microphone 2 should be located as closer to the noise/interference source as possible.

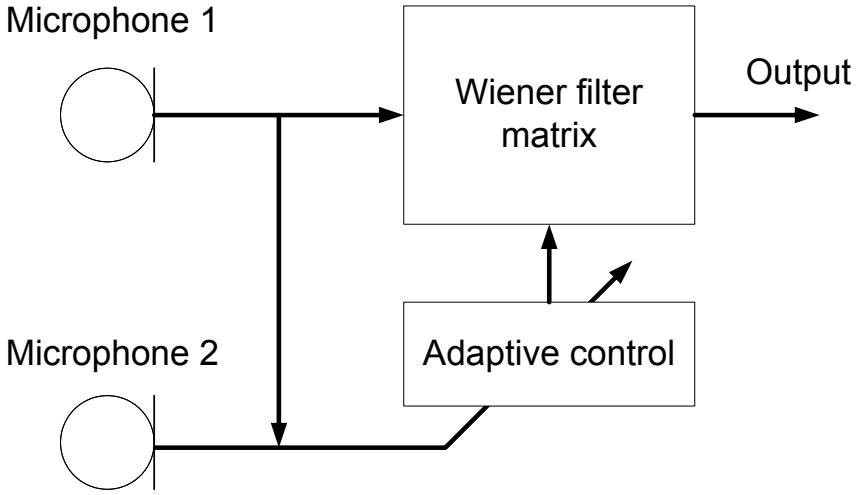

Figure 3-5 A modified adaptive Wiener filter with two microphone

In a 3-microphone array in this thesis, only 2 of 3 microphones are used in the Wiener filter as in Figure 3-5.

3.3 Overview of system build-up

In Chapter 4, this beamforming based VAD is confirmed as good performance in sole voice source during test time. However, once unwanted voice is challenging the desired voice during test time, the 3-microphone beamforming based VAD need an additional real-time speech separation to pick up the desired voice e.g. driver's voice. Therefore, a novel real-time adaptive Wiener based filter is discussed in Chapter 4. Finally, an engineering application is discussed in Chapter 6.

3.4 Research environmental set-up

Research environmental set-up is to verify the system design with experiments. 3 microphones are located in car and signal conditioning and Digital Signal Processing (DSP) are designed to suit the requirement of experimental test.

3.4.1 Three-microphone data acquisition in car

A 3-microphone system is set-up in a car (BMW 318i) as in Figure 3-6 and Figure 3-7. Three microphones are located as shown and there is 50 cm distance between these microphones. A desired speech source is located 50 cm away from Microphone 1 and Microphone 3. The distance between the speaker and Microphone 2 is 70.7 cm.

Therefore, when speech travels to microphone 2 it has 20.7 cm more distance from to microphone 1 and also has 20.7 cm more than from microphone 3.

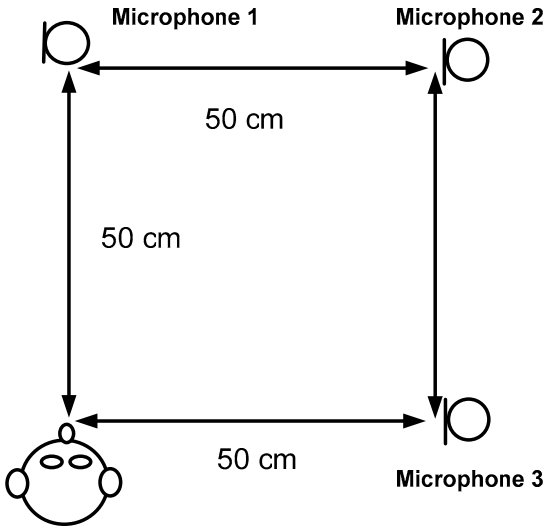

Figure 3-6 Automobile environment layout

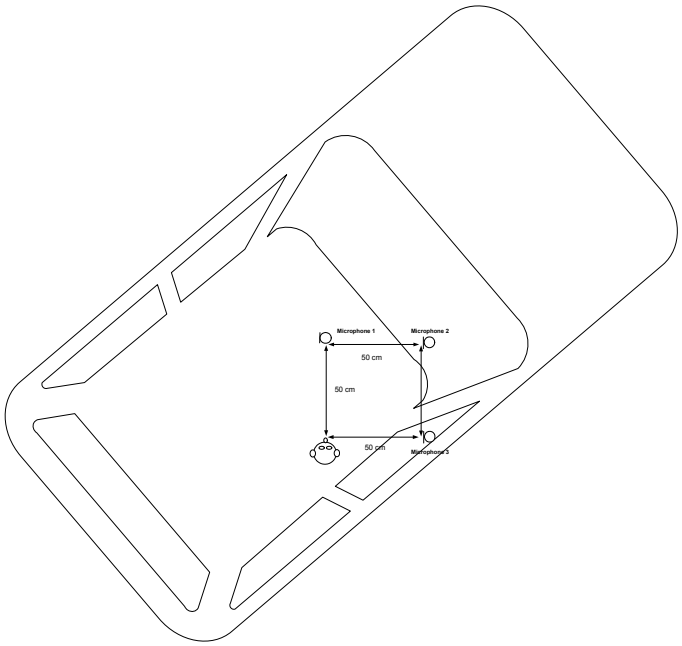

Figure 3-7 Three microphones in a car

The sample rate of Microphone 1, 2 and 3 is 11025 Hz, and the speed of sound in air is 34600cm/second. Therefore during every sample the speech travels 3.1 cm so that the wave-front of speech arrives at microphone 2 delayed by 7 sample intervals with respect to the other two microphones.

3.4.2 Signal conditioning in a car

Microphones 1, 2 and 3 are omni-directional electret condenser microphones with the following specifications:

Sensitivity: -62 +/- 3 dB

Impedance: <2K Ohm

Frequency Range: 50-12, 500Hz

A pre-amplifier and an anti-aliasing filter with a 5 kHz cutoff are used. The sample rate of Microphone 1, 2 and 3 is 11025 Hz. A desired source and a noise source were used in the test which was taken from loud-speakers. The enclosures are 21cm x 10 cm x 11 cm. The loudspeakers in the enclosure have a specification:

0.5W 3"; External Diameter: 3 inches;

Frequency response, lower limit: 200Hz; Frequency response, upper limit: 6kHz;

Impedance: 8 Ohms;

A PC sound card line output is used to drive this speaker. All results were recorded in a car in real time.

Alias protection is critical to accurate signal analysis. Aliasing is an effect associated with digitizing an analogue signal. Aliasing can cause gross errors in measurement and testing. The only way to protect against aliasing errors is by using high quality anti-alias filters on the input channels. The input signal is processed with an electronic low-pass filter to remove all frequencies above the Nyquist frequency (one-half the sampling rate). This is done to prevent aliasing during sampling, and is correspondingly called an anti-alias filter. (Smith, 1999)

A pre-amplifier was designed to enlarge microphone signal as Figure 3-8.

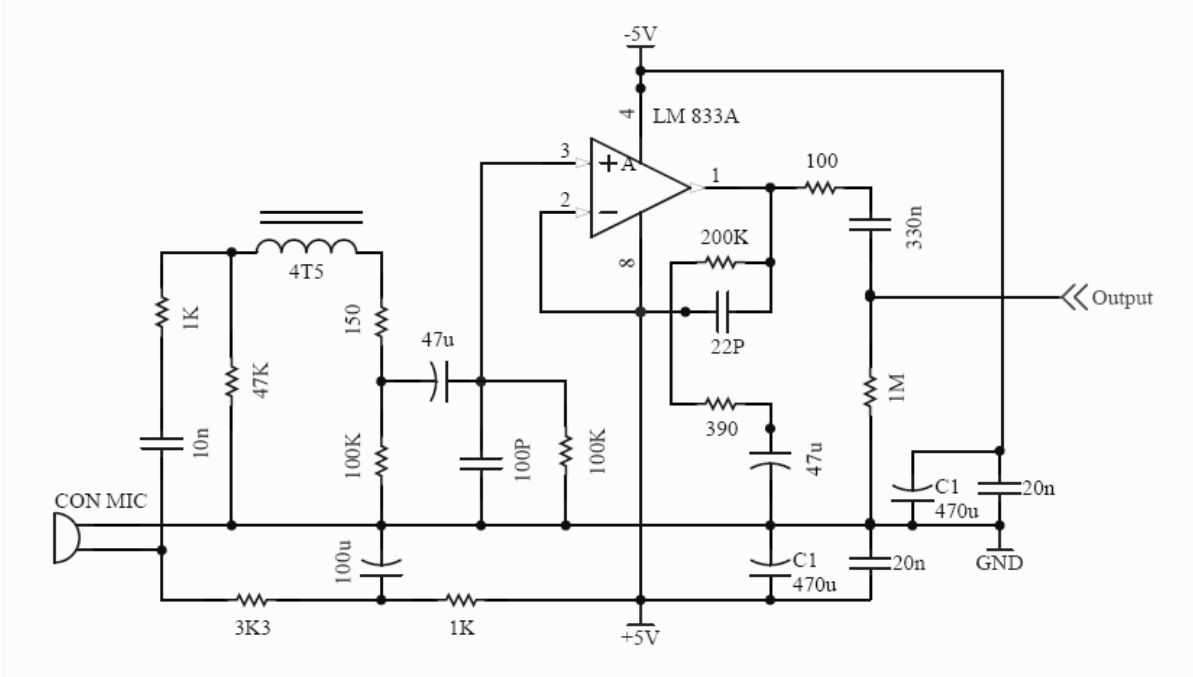

Figure 3-8 Pre-amplifier

A 6 order low pass Butterworth filter was designed as Figure 3-9. As the sample rate is 11026 Hz, the pass band frequency was set to 5000 Hz. The design frequency response is shown as Figure 3-10.

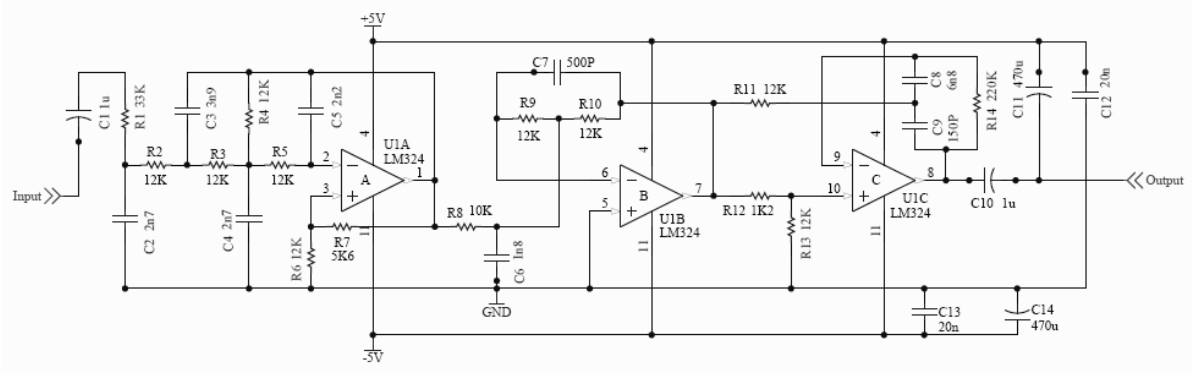

Figure 3-9 Anti-alias filter

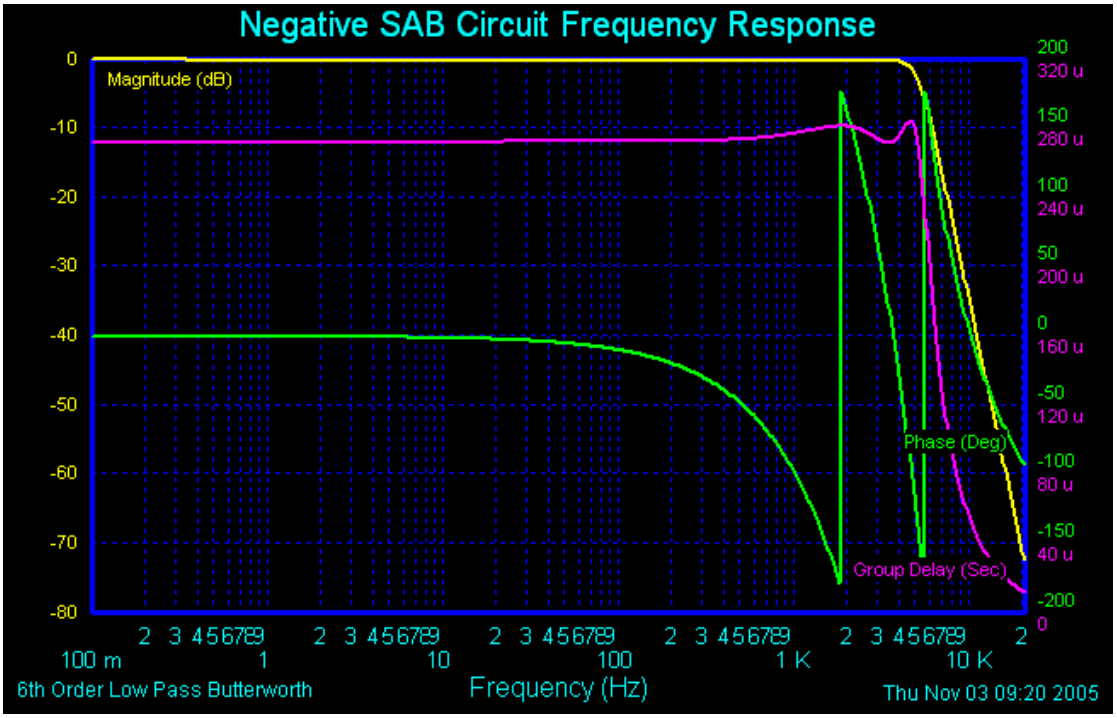

Figure 3-10 frequency respond

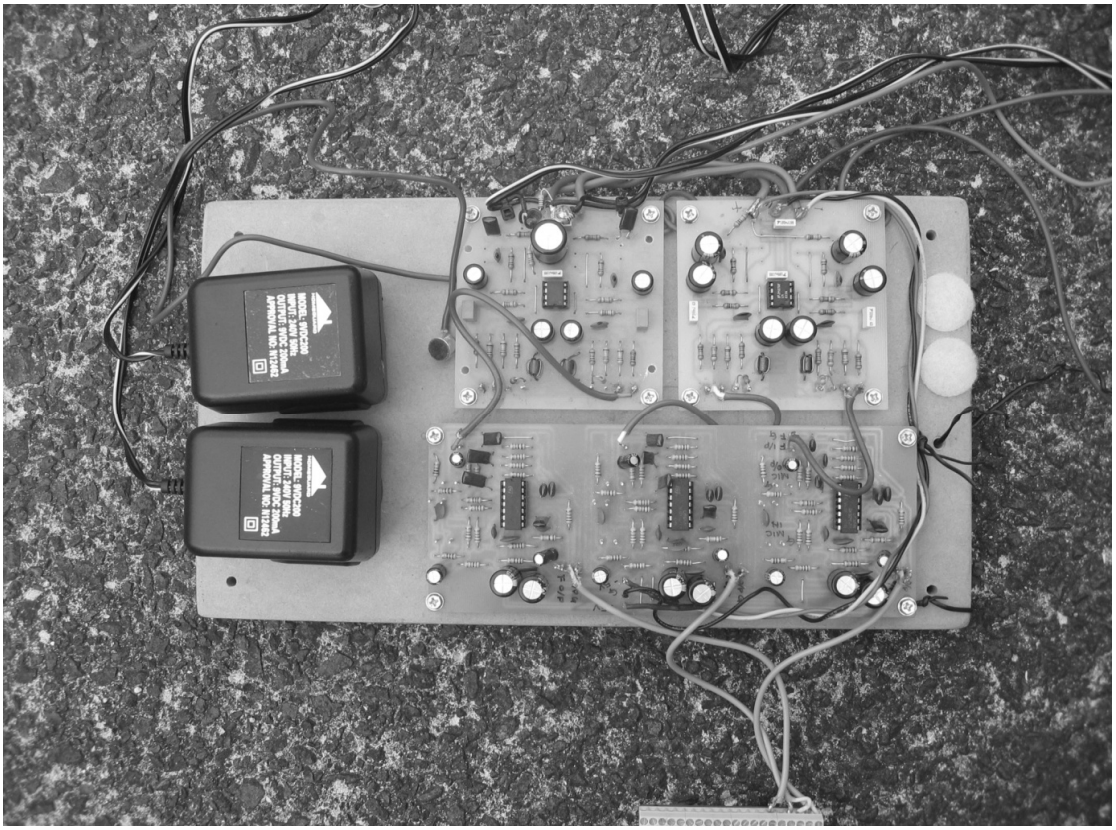

Figure 3-11 Pre-amplifier and anti-alias filter

The hardware assembles of pre-amplifier and anti-alias filter are shown in Figure 3-11.

3.4.3 Digital Signal Processing hardware and software

Many Data Acquisition (DAQ) Systems are available as details in Appendix. PMD-1608FS is a low-cost DAQ but offered a software interface with LabVIEW program (see details in Appendix). LabVIEW program has been used in the Digital Signal Processing group in Massey University at Albany for a long time and many creditable samples of applications are available e.g. NLMS filter.

PMD-1608FS with LabVIEW 8.0 is applied as showed in Figure 3-12.

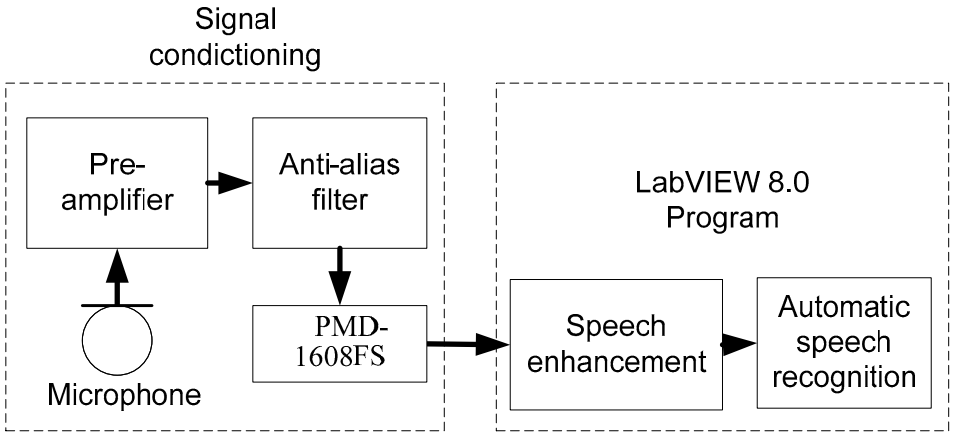

Figure 3-12 A block diagram of DSP hardware and software

3.5 Summary

This chapter highlights the most challenging task of speech enhancement in a car, which is to deal with non-stationary noise and interference, e.g. to turn off the playing radio.

As a solution, 3-microphone system with signal conditioning is introduced.

Therefore, the discussion on digital signal processing is ready at following chapters.

4 Real-time Adaptive Noise Cancellation in a car

4.1 Three-microphone beamformer in a car

4.1.1 Introduction

This thesis focuses on a typical real environment such as found in a car. There are two parts in this system:

Part 1: A three-microphone beamformer with normalised least-mean squares (NLMS).

Part 2: A three-microphone Voice Activity Detection (VAD) algorithm.

The VAD acts as a switch on a double-acting Griffiths-Jim adaptive beamformer. Van Compernolle (Van Compernolle, 1990) introduced this switching adaptive filter with a 4 microphone array in a highly reverberant room with both music and fan type noise as jammers. SNR improvements of l0 dB were typical with no audible distortion. Zhang and Hansen (Xianxian Zhang & J.H.L. Hansen, 2003) introduced a constrained switched adaptive beamforming algorithm for speech enhancement and recognition in real moving car environments. In a frequently moving noise sources environment (noise sources are coming from different locations but not always presented at the same time), the 3-microphone noise canceller with geometric VAD has the effect of canceling un-wanted speech or noise from outside of a VAD valid zone. As the same time, there is a 3-microphone noise canceller valid zone defined. In order to enhance desired speech and reduce noise(s), a desired voice should be in the intersection of the noise canceller valid zone and the VAD valid zone. Thus all noise is suppressed outside this intersected area. Experiments performed have verified the improvements given by this method in a real environment(Qi & Moir, 2005, 2006).

4.1.2 Three microphone beamforming Voice Activity Detection with Adaptive noise cancellation

A block diagram of the three-microphone VAD-controlled three-microphone adaptive noise canceller (ANC) is shown in Figure 4-1. The noise canceller (three-microphone adaptive digital filter) is detailed in Figure 3. The VAD switches various NLMS filters on or off depending if the desired speech is presented. Moreover, the VAD allows signal output only when desired speech presented i.e. it mutes the output when there is noise present outside the desired zone but only if simultaneously there is no desired speech.

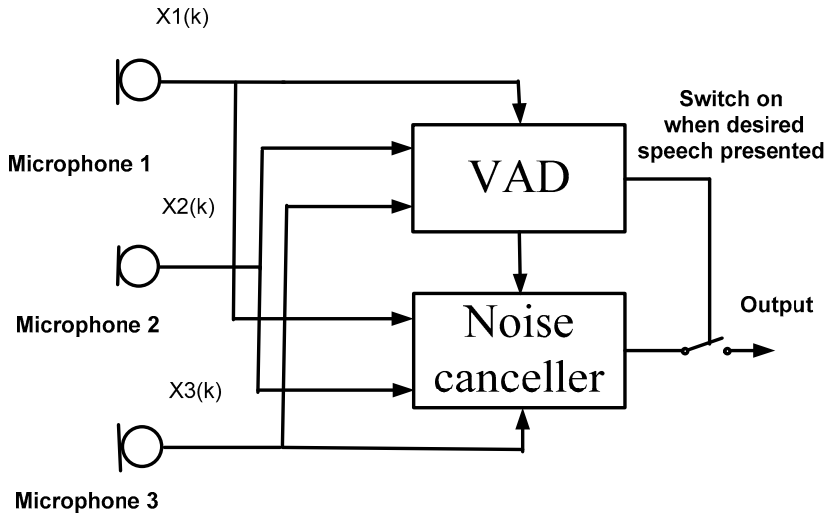

Figure 4-1 Overview of three-microphone VAD controlled three-microphone noise canceller

4.1.3 Three-microphone Adaptive noise cancellation

A three-microphone noise canceller based on Van Compernolle's work is showed as Figure 4-2. There are four NLMS algorithms in a three-microphone noise canceller. The top path of the beamformer has a summation term which forms the primary input whilst both of the bottom paths have a difference term which forms the reference input. The three microphone signals contain speech as well as noise. The left section of the system serves at improving the noise reference by eliminating speech so that the VAD switches this part on when speech energy is dominant. The right section consists of NLMS 2 and NLMS 4, which are only switched on to adapt during the absence of speech (i.e. during noise periods). For these experiments the number of weights used in W1 and W3 were 100 and in W2 and W4, 450.

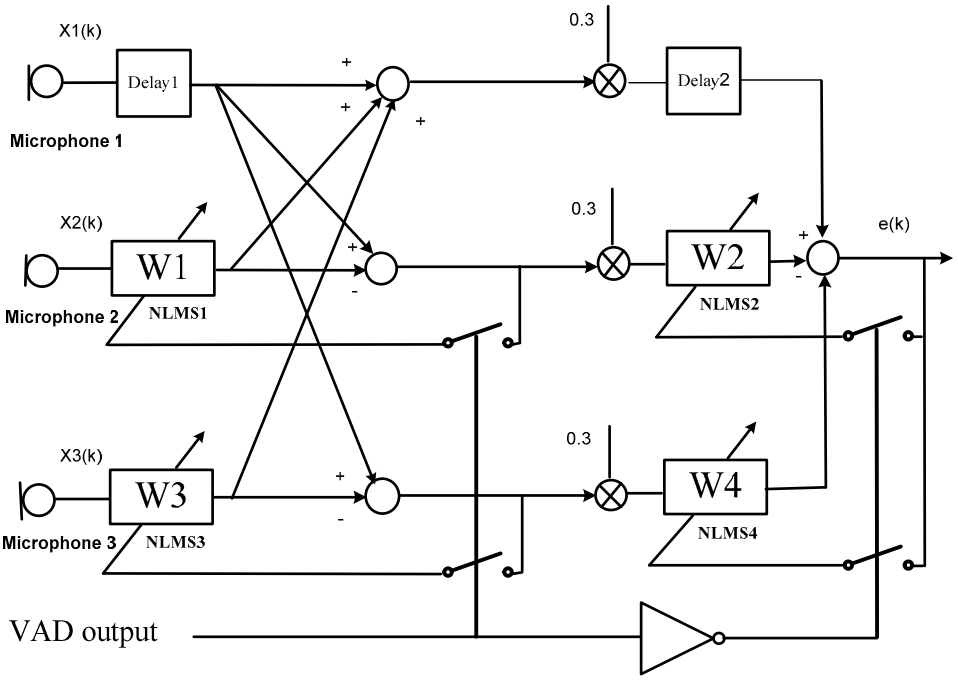

Figure 4-2 Three-microphone noise canceller block diagram

4.1.4 A three-microphone VAD

Carter et al. (Carter et al., 1973) describe a method for estimating the magnitude-squared coherence (MSC) function for two zero-mean wide-sense-stationary random processes. The estimation technique utilizes the weighted overlapped segmentation fast Fourier transform (FFT). Analytical and empirical results for statistics of the estimator are presented. The analytical expressions are limited to the non-overlapped case. Empirical results show a decrease in bias and variance of the estimator with increasing overlap and suggest a 50-percent overlap as being highly desirable when cosine (Hanning) weighting is used. Once the MSC is found the Generalized Cross-Correlation (GCC) method is used to give a robust estimate of time-delay. The technique can be summarized as follows for three microphones and two estimated time-delays.

At each FFT frame index $i = 1,2,3,...$ assign the three vectors

$$\mathbf{x}_1 = [n_0, n_1, ...n_{N-1}]^T \tag{4.1}$$

$$\mathbf{x}_2 = [m_0, m_1, ...m_{N-1}]^T \tag{4.2}$$

$$\mathbf{x}_3 = [l_0, l_1, ...l_{N-1}]^T \tag{4.3}$$

which are composed of N samples of the three microphone inputs and have been suitably windowed with their corresponding frequency vectors corresponding to $\mathbf{X_1}$, $\mathbf{X_2}$ and $\mathbf{X_3}$ respectively.

The periodograms of the signals, which are estimates of power spectral density can be estimated from each of the three microphones according to

$$\hat{\mathbf{S}}_{x_1x_1}(i) = \beta \hat{S}(i-1) + (1-\beta)\mathbf{X_1}\mathbf{X_1}^* \tag{4.4}$$

$$\hat{\mathbf{S}}_{x_2x_2}(i) = \beta \hat{S}(i-1) + (1-\beta)\mathbf{X_2}\mathbf{X_2}^* \tag{4.5}$$

$$\hat{\mathbf{S}}_{x_3x_3}(i) = \beta \hat{S}(i-1) + (1-\beta)\mathbf{X_3}\mathbf{X_3}^* \tag{4.6}$$

where (4.4), (4.5) and (4.6) is a method of smoothly updating the spectrum recursively at each FFT frame. In the above equation '*' represents complex conjugate and $0 \le \beta \le 1$ is a forgetting factor. For the results used in this paper $\beta = 0.5$ was used as a compromise between fast tracking and smoothing. If chosen to be too large then the tracking ability of the GCC time-delay estimator is severely compromised. Some experimentation is required depending on the application. Two cross-spectrums (cross-periodograms) are found in a similar manner.

$$\hat{S}_{x_1x_2}(i) = \beta \hat{S}(i-1) + (1-\beta)\mathbf{X_1}\mathbf{X_2}^* \tag{4.7}$$

$$\hat{S}_{x_2x_3}(i) = \beta \hat{S}(i-1) + (1-\beta)\mathbf{X_2}\mathbf{X_3}^* \tag{4.8}$$

The magnitude-squared coherence (MSC) (Carter et al., 1973) gives a measure of dependency of one signal upon another as a function of frequency and can be computed at each FFT frame from

$$\left|\hat{\boldsymbol{\gamma}}_{x_1x_2}(i)\right|^2 = \frac{\left|\hat{\mathbf{S}}_{x_1x_2}(i)\right|^2}{\hat{\mathbf{S}}_{x_1x_1}(i)\hat{\mathbf{S}}_{x_2x_2}(i)} \tag{4.9}$$

$$\left|\hat{\boldsymbol{\gamma}}_{x_2x_3}(i)\right|^2 = \frac{\left|\hat{\mathbf{S}}_{x_2x_3}(i)\right|^2}{\hat{\mathbf{S}}_{x_2x_2}(i)\hat{\mathbf{S}}_{x_3x_3}(i)} \tag{4.10}$$

and at each frame i, average over frequency k the MSC. Thus for an FFT with N points, since the FFT is symmetric s

$$\left|\bar{\gamma}_{x_1x_2}(\mathrm{i})\right|^2 = \frac{1}{\mathrm{N}/2-1}\sum_{\mathrm{k}=0}^{\mathrm{N}/2-1}\left|\hat{\gamma}_{x_1x_2}(\mathrm{k})\right|^2 \tag{4.11}$$

$$\left|\bar{\gamma}_{x_2x_3}(\mathrm{i})\right|^2 = \frac{1}{\mathrm{N}/2-1}\sum_{\mathrm{k}=0}^{\mathrm{N}/2-1}\left|\hat{\gamma}_{x_2x_3}(\mathrm{k})\right|^2 \tag{4.12}$$

Estimate the weighting terms for the GCC $\psi_{g_1}(i)$ and $\psi_{g_2}(i)$ from

$$\psi_{g_1}(i) = \frac{\left|\hat{\gamma}_{x_1x_2}(i)\right|^2}{\left|\hat{S}_{x_1x_1}(i)\right|\left[1-\left|\hat{\gamma}_{x_1x_2}(i)\right|^2\right]} \tag{4.13}$$

$$\psi_{g_2}(i) = \frac{\left|\hat{\gamma}_{x_2x_3}(i)\right|^2}{\left|\hat{S}_{x_2x_3}(i)\right|\left[1-\left|\hat{\gamma}_{x_2x_3}(i)\right|^2\right]} \tag{4.14}$$

Estimate the time-delays of arrival d_1 (between microphone 1 and 2) and d_2 (between microphone 2 and 3) from the generalized cross-correlations.

$$\hat{R}^{g_1}_{x_1x_2}(d_1) = \arg\max F^{-1}\{\psi(i)\hat{S}_{x_1x_2}(i)\} \tag{4.15}$$

$$\hat{R}^{g_2}_{x_1x_2}(d_2) = \arg\max F^{-1}\{\psi(i)\hat{S}_{x_1x_2}(i)\} \tag{4.16}$$

That is the maximum of the inverse FFT of $\psi(i)\hat{S}_{x_1x_2}(i)$ and $\psi(i)\hat{S}_{x_2x_3}(i)$. A positive delay can be inferred if the maximum occurs in the region $0 \le d \le N/2-1$ i.e. the first half of the inverse FFT and a negative delay if the maximum occurs in the upper half of the inverse FFT.

Valid speech is then assumed when an active zone is defined according to

$$d_{1,2} \le d_{\max} \tag{4.17}$$

$$d_{2,3} \le d_{\max} \tag{4.18}$$

where $d_{\max}$ is some maximum desired time-difference of arrival (TDOA) between microphone 1 and 2, or between microphone 2 and 3.

Also we require that both

$$\left|\hat{\gamma}_{x_1x_2}(i)\right|^2 \ge C_{\min} \tag{4.19}$$

and

$$\left|\hat{\gamma}_{x_2x_3}(i)\right|^2 \ge C_{\min} \tag{4.20}$$

where $C_{\min}$ is some minimum desired MSC (with empirical meaning) to prevent reverberant speech from being detected as desired speech e.g. when a reflection of a nearby undesired noise finds its way into the active zone. It is well established that reverberant speech has a higher MSC than non-reverberant speech and this gives rise to (4.19) and (4.20).

For the experiments carried out in this paper a sampling interval of 11025Hz was used so that each sample interval corresponds to 90.7 μs. Typically $\mathrm{d}_{\max}$ was chosen to be no more

than 5 samples and $C_{\min}$ was chosen as 0.5, which has empirical meaning. A three-microphone VAD block diagram is presented at Figure 4-3.

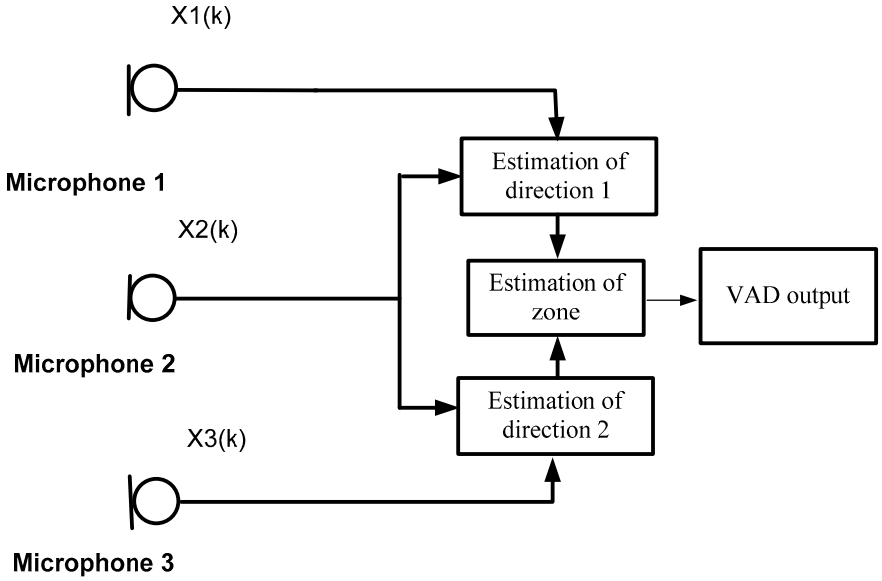

Figure 4-3 Three-microphone VAD Block diagram

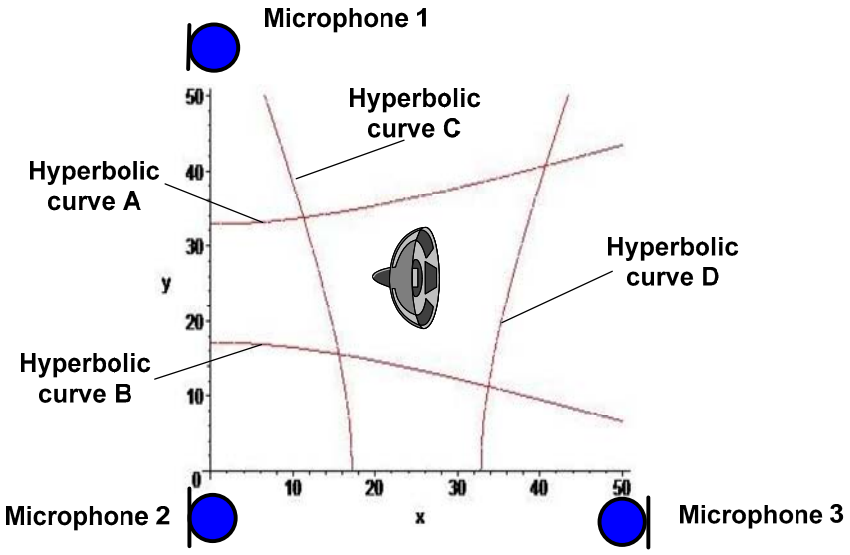

Figure 4-4 A defined active zone

When d_{max} was chosen using microphone 1 and microphone 2 e.g. to be no more than 5 samples, an estimation of time delay (time-difference of arrival TDOA) defines Estimation of Direction 1 (EOD 1) located on the area between the hyperbolic curve A and B as in Figure 4-4. The hyperbolic curve A and B indicate the calculation result of $d_{1,2} = d_{\max}$. Another estimation of TDOA between microphones 2 and 3 defines Estimation of Direction 2 (EOD 2) on the area between the hyperbolic curve C and D. The hyperbolic curve C and D indicate

the calculation result of $d_{2,3} = d_{\max}$. The intersection of EOD 1 and 2 will be the Estimation of Zone (EOZ).

When the VAD is set to be within an EOZ, it means some defined number of samples e.g. 5 samples is typically used, then the TDOA's from each microphone pair (the pair of microphone 1 and 2, or the pair of microphone 2 and 3) is estimated and compared with some threshold value $d_{\max}$. The larger the value of $d_{\max}$, the active zone will be bigger. The VAD works so as to switch to freeze or enable the various LMS algorithms. Also, the VAD switches off (mutes) the signal output when speech does not come from the desired zone. Therefore, only the driver's voice will activate the system whereas a passenger's voice, or for instance the car radio will have no effect.

4.1.5 Summary and discussion

In the chapter 4.1, 3-microphone voice activity detection and adaptive noise cancellation are presented. The 3-microphone voice activity detection (VAD) approach is based on a fundamental theory of time-delay estimation with magnitude-squared coherence (MSC). This VAD has the ability of the composite system to reduce noise outside of a defined active zone. The 3-microphone adaptive noise cancellation (ANC) is based on a fundamental theory of normalized least-mean squares (NLMS) to improve Signal to Noise Ratio (SNR). The signal and noise inputs are shared with a Voice Activity Detector (VAD). A 3-microphone NLMS adaptive noise cancellation is used.

The VAD in this thesis is based on Chen and Moir's 3-microphone VAD (Chen and Moir, 1999). Chen and Moir derived their VAD which was using equation 4.13 and 4.14. And they also built a test plan in lab using 3 microphones. The sampling frequency was 25.6 kHz and the separation of each pair of microphones is 20 cm. Some simulation experiments had been conducted to present the performance of a word boundary detection algorithm using three microphones. All the investigations showed the accuracy of the word boundary detecting percentage success rate to be more than 80% and therefore prove the algorithm an effective pre-processor for further applications in noise cancellation.

This thesis contributes a method of smoothly updating the spectrum recursively at each FFT frame as equations 4.4, 4.5, 4.6, 4.7 and 4.8 for equation 4.13 and 4.14. The sampling frequency in this thesis is 11025Hz and the separation of each pair of microphones is 50 cm. This thesis also presents experiments in a car in a real-time environment. This 3-microphone VAD approach is used in this thesis since it is relatively simple and not too much of a computational overload for real-time system.

The investigations in this thesis show 93% successful rate using a commercial Automatic Speech Recognition kit as reported in Chapter 5 experiment.

However, the 3-microphone VAD and ANC only work well for a single sound source. In real-time environments a speech recognition system in a car has to receive the driver's voice only whilst suppressing background noise. Therefore, next chapter is to discuss a solution using an adaptive Wiener filter in a multi-source environment.

4.2 Adaptive Wiener filter in a car

4.2.1 Introduction

In real-time environments an Automatic Speech Recognition (ASR) system in a car has to receive the driver's voice only whilst suppressing background noise. Previous work (Qi & Moir, 2005) uses a real-time Voice Activity Detector (VAD) which operates within a geometrical zone defined around the head of the desired speaker. However, VAD only works well for a single sound source. During the period of desired speech the VAD does not work well if the unwanted speech is incoming at the same time as the desired speech. As a solution to the above issue, this thesis presents an adaptive wiener filter for automatic speech recognition in a car environment with non-stationary noise (Qi & Moir, 2008), improving the SNR by slightly better than 28 dB in a noisy car environment (Qi & Moir, 2007a). As an experiment, a template matching ASR was designed to recognize the signal filtered by such a real-time Wiener filter (Qi & Moir, 2007b).

4.2.2 Adaptive Wiener filter

The Wiener filter was introduced by Norbert Wiener in 1949 (Wiener, 1949) and independently for the discrete-time case by Kolmogorov (Kolmogorov, 1941). Wiener-Kolmogorov filters have the following assumptions: a) signal and (additive) noise are stochastic processes with known spectral characteristics or known autocorrelation and cross-correlation, b) the performance criterion minimizes the mean-square error. An optimal filter can be found from a solution based on scalar or multivariable methods.(Barrett & Moir, 1987) The goal of the Wiener filter is to filter out noise that has corrupted a signal by statistical means.(Brown & Hwang, 1996) A Wiener filter block diagram as described by Haykin (Simon. Haykin, 2002) is shown in Figure 5-1. From Figure 5-1, a microphone signal y (an M-dimensional vector) is filtered by the Wiener filter W (an M x M filter matrix) and the output z (an M-dimensional vector) has to estimate the desired signal with some residual error.

$$e = d - z = d - \mathbf{W}^T y \tag{4.21}$$

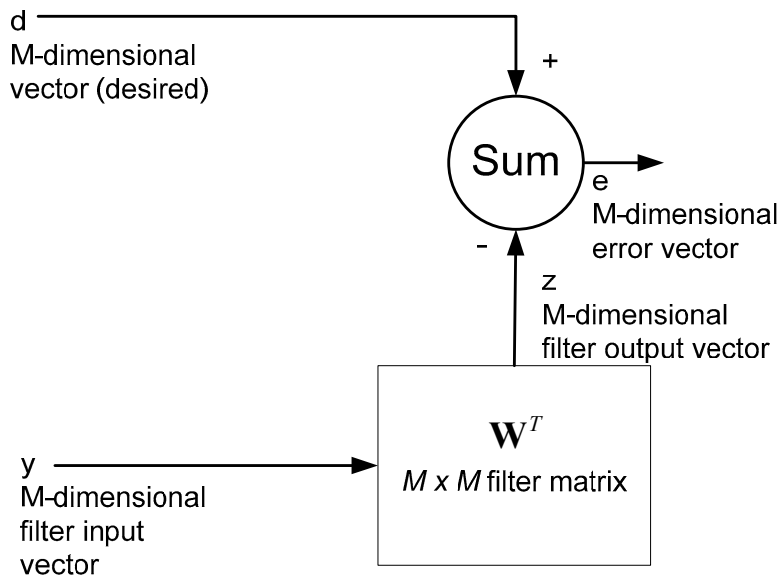

Figure 4-5 A Wiener filter

Figure 4-5 shows that when $z = d$ then $e = 0$. This means that when e = 0, z is the estimated value of d. Therefore, when a desired signal comes with noise (white or coloured), a selected W matrix is available for estimating the desired signal. Thus the goal of speech enhancement is to compute the W matrix. Wiener filtering is commonly used in the time domain or the frequency domain to reduce noise for single-channel and multi-channel (multivariable) signals in speech or image enhancement. In order to compute and update the W matrix, the standard Wiener filter minimizes the mean-square error $E[e^T(k)e(k)]$ between the filter output signal and desired speech output.(Doclo & Moonen, 2005) This method is a so called adaptive filter if the signal and noise model have time-varying statistics.(Dabeer & Masry, 2002; Griffiths, 1978; Simon. Haykin, 2002; Resende, Romano, & Bellanger, 2004) Doclo & Moonen (Doclo & Moonen, 2002) introduced a Wiener filter with N microphones, with each microphone signal defined as $y_n[k], n = 0,1,..., N-1$, at some sample k, the speech as $x_n[k]$ and noise component $v_n[k]$ all received at the nth microphone. Thus the mixture

$$y_n[k] = x_n[k] + v_n[k], n = 0,1,...N-1 \tag{4.22}$$

In single-microphone speech enhancement, the number of microphones is $N = 1$ such that the equation (4.22) simplifies to

$$y_0[k] = x_0[k] + v_0[k] \tag{4.23}$$

Let the filter $w_n[k], n = 0,1,..., N-1$ have length L

$$\mathbf{w}_n[k] = [w_n^0[k] \quad w_n^1[k] \quad ... \quad w_n^{L-1}[k]]^T \tag{4.24}$$

and consider the L-dimensional data vectors $y_n[k]$, the M-dimensional stacked filter $w[k]$ (with $M = LN$), and the M-dimensional stacked data vector $y[k]$, defined as

$$\mathbf{y}_n[k] = [y_n[k] \quad y_n[k-1] \quad \ldots \quad y_n[k-L+1]]^T \tag{4.25}$$

Then we have

$$\mathbf{w}[k] = [w_0^T[k] \quad w_1^T[k] \quad \ldots \quad w_{N-1}^T[k]]^T \tag{4.26}$$

$$\mathbf{y}[k] = [y_0^T[k] \quad y_1^T[k] \quad \ldots \quad y_{N-1}^T[k]]^T \tag{4.27}$$

Therefore, the output signal is the convolution

$$\mathbf{z}[k] = \sum_{n=0}^{N-1} \mathbf{w}_n^T[k]\mathbf{y}_n[k] = \mathbf{w}^T[k]\mathbf{y}[k] \tag{4.28}$$

And also,

$$\mathbf{e}[k] = \mathbf{d}[k] - \mathbf{W}[k]^T \mathbf{y}[k] \tag{4.29}$$

In order to compute $\mathbf{W}[k]$, the Wiener-Hopf equation $\mathbf{W}_{opt} = \mathbf{R}^{-1}\mathbf{P}$ had been applied as equation 2.48. $\mathbf{R}$ is the $N \times N$ autocorrelation matrix and $\mathbf{P}$ is the N length cross-correlation matrix.

Under two assumptions: short-term stationarity of the noise, and statistical independence of the speech and noise signals, Doclo & Moonen (Doclo & Moonen, 2002) derived a solution of the Wiener-Hopf equation

$$\mathbf{W}[k] = \mathbf{R}_{yy}^{-1}[k]\left(\mathbf{R}_{yy}[k] - \mathbf{R}_{vv}[k]\right) \tag{4.30}$$

where $\mathbf{R}_{yy}[k]$ is the *M x M* auto-correlation matrix of the input signal and it is estimated during speech plus additive noise periods. $\mathbf{R}_{vv}[k]$ is the *M x M* auto-correlation matrix of the input signal which is estimated during noise-alone periods.

From Equation 4-30, when the W matrix is known, $z[k]$ is estimated. Therefore, when the W matrix is pre-computed, a block diagram can be illustrated as in Figure 4-6.

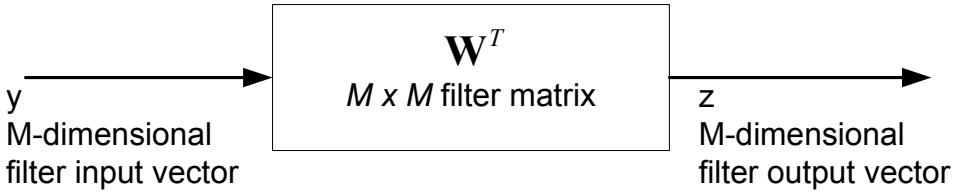

Figure 4-6 A Wiener filter with pre-computed W matrix

Therefore, a schematic for a single microphone was built as showed in Figure 4-7. Microphone signal *y* (an M-dimensional vector) is the input signal and z (an M-dimensional vector) is the filtered signal, which is an estimate of the desired signal.

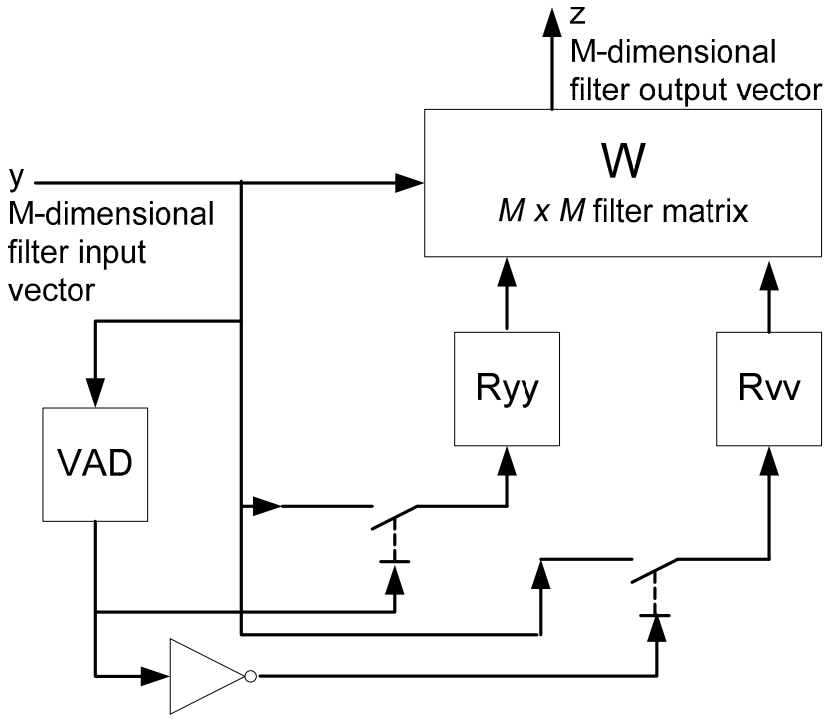

Figure 4-7 W Matrix Calculation for single microphone case

For a single microphone, we have one filter. In Figure 4-7, the **W** matrix can be compute by collecting the input M-dimensional vector in the single microphone y under noise periods and the same microphone vector y under signal + noise periods. Therefore, $\mathbf{R}_{yy}[k]$ is only updated at signal + noise periods and $\mathbf{R}_{vv}[k]$ is only updated during noise alone periods. From (4.24) and (4.26), we have

$$w_0[k] = [w_0^0[k] \quad w_0^1[k]]^T \tag{4.31}$$

From (4.25) and (4.27), we have

$$y_0[k] = [y_0[k] \quad y_0[k-1]]^T \tag{4.32}$$

From (4.28), we have

$$z[k] = w^T[k]y[k] \tag{4.33}$$

During noise-alone periods, the noise input vector

$$\mathbf{Y}(k) = [y(k), y(k-1), \ldots y(k-M+1)]^T \tag{4.34}$$

where n defines the order of the noise process. The Toeplitz measurement correlation matrix is

$$\mathbf{R}_{vv} = E[y(k)y^T(k)] \tag{4.35}$$

$$\mathbf{R}_{vv} = \begin{bmatrix} r_{vv}(0) & r_{vv}(1) & \dots & r_{vv}(M-1) \\ r_{vv}(-1) & r_{vv}(0) & \dots & r_{vv}(M-2) \\ \dots & \dots & \dots & \dots \\ r_{vv}(-M+1) & r_{vv}(-M+2) & \dots & r_{vv}(0) \end{bmatrix} \tag{4.36}$$

The term $r_{vv}(0)$ is always real valued and

$$r_{vv}(k, k-n) = E\left[y(k)y^*(k-n)\right] \tag{4.37}$$

During speech and noise periods, a noise + signal vector as primary input is defined as

$$\mathbf{D}(k) = [d(k), d(k-1), \dots d(k-M+1)]^T \tag{4.38}$$

where n defines the order of the noise process and the Toeplitz signal + noise correlation matrix is

$$\mathbf{R}_{yy} = E\left[d(k)d^*(k)\right] \tag{4.39}$$

$$\mathbf{R}_{yy} = \begin{bmatrix} r_{yy}(0) & r_{yy}(1) & \dots & r_{yy}(M-1) \\ r_{yy}(-1) & r_{yy}(0) & \dots & r_{yy}(M-2) \\ \dots & \dots & \dots & \dots \\ r_{yy}(-M+1) & r_{yy}(-M+2) & \dots & r_{yy}(0) \end{bmatrix} \tag{4.40}$$

The term $r_{yy}(0)$ is always real valued and

$$r_{vv}(k, k-n) = E\left[y(k)y^*(k-n)\right] \tag{4.41}$$

Experiments showed good results when the arrangement of Figure 5-3 was used. That is when the noise source comes from a permanent location and the noise source model is not changing with time. The block diagram in Figure 4-7 can be used in 1-microphone application e.g. mobile phone applications. However, for a real-time environment e.g. in a car the noise sources do not come from a permanent location and the noise source model is non-stationary. If the noise characteristic changes between noise and speech periods then the **W** matrix will be in error.

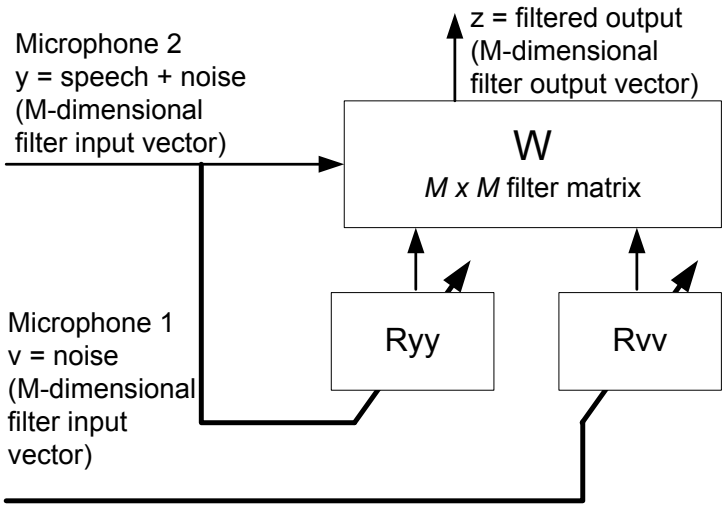

Figure 4-8 matrix updated during speech or noise period

Therefore, a new Wiener filter is built as in Figure 4-8 with test conditions as shown in Figure 4-9. Since Microphone 1 is close to the noise source, Microphone 1 is set as the noise input. Microphone 2 acts as a speech and noise input. Microphones 1 and 2 record desired speech from an equal distance whilst an unwanted speech or noise source is close to Microphone 1.

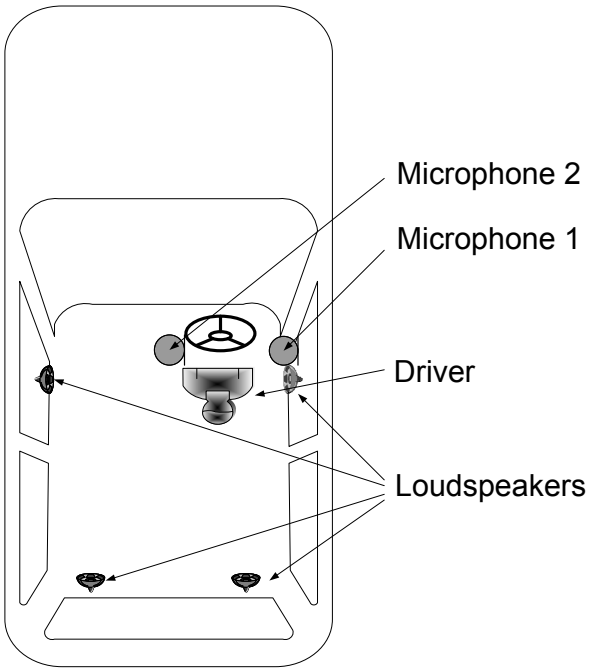

Figure 4-9 In-car test plan

Essentially this technique is similar to the dual microphone method of Widrow et al except here we will not use least-mean squares(LMS) but instead update the Wiener filter directly by estimation of the signal + noise and noise covariance matrices and by direct solution of the Wiener-Hopf equation (4.30).

4.2.3 Matrix inversion method

The matrix inversion method (or DMI method) is important part in this thesis. The method arises out of the following exponential smoothing method to recursively estimate the covariance matrix:

$$\mathbf{R}_{k+1} = \lambda \mathbf{R}_k + (1-\lambda)\mathbf{\Phi}_\mathbf{k}\mathbf{\Phi}_\mathbf{k}^\mathbf{T} \tag{4.42}$$

where $0 < \lambda < 1$ is a scalar forgetting factor chosen in a similar way to RLS. In order to find the inverse we use the matrix inversion lemma. For a matrix

$$\mathbf{A} = \mathbf{B}^{-1} + \mathbf{H}\mathbf{Q}^{-1}\mathbf{H}^\mathbf{T} \tag{4.43}$$

then its inverse is given by

$$\mathbf{A}^{-1} = \mathbf{B} - \mathbf{B}\mathbf{H}\left[\mathbf{Q} + \mathbf{H}^\mathbf{T}\mathbf{B}\mathbf{H}\right]^{-1}\mathbf{H}^\mathbf{T}\mathbf{B} \tag{4.44}$$

Appling to (1) gives the following SMI algorithm update

$$\mathbf{R}_{k+1}^{-1} = \frac{\mathbf{R}_k^{-1}}{\lambda}\left[I_m - \frac{\mathbf{\Phi}_k \mathbf{\Phi}_k^T \mathbf{R}_k^{-1}}{\frac{\lambda}{1+\lambda} + \mathbf{\Phi}_k^T \mathbf{R}_k^{-1} \mathbf{\Phi}_k} \right] \tag{4.45}$$

The method is used for example in smart antennas(Kaiser, 2005) or CDMA for mobile communication networks (Swarts, 1999).

4.2.4 Automatic Speech Recognition

The enhanced signal from the noise canceller is fed to a speech recognizer. At the automated speech recognizer (ASR) we compute features that represent the spectral-domain content of the speech. These features are computed with a "frame", which is the number is of samples within a defined period of, for example 10ms. A neural network can be used to classify a set of these features into phonetic-based categories at each frame. A search is used to match the neural-network output scores to the target words (the words that are assumed to be in the input speech), in order to determine the word that was most likely uttered. It is also possible to analyze the results by looking at the confidence we have in the top-scoring word. The word can be rejected as out-of-vocabulary if the confidence falls below a pre-determined threshold. Using the *M square* W matrix in the Wiener filter, the samples of input speech are filtered and outputted as a *1D* array with size *M*. This *1D* array is considered as a frame, which would be used in a typical neural network. For the W matrix, the larger *M* means more computing resource cost. In the meantime, the Frame size in a neural network is considered by the accuracy of speech recognition. Normally a smaller frame size means larger error.

Since filtered signals from a Wiener filter has as much as 25 dB SNR, word boundaries are easily identified by energy density. In a Wiener filter, the incoming voice is sampled at an 11025 Hz sample rate. The W matrix is 100x100 and is updated by the 100 samples from Microphone 1 and another 100 samples from Microphone 2. These samples from Microphone 2 are also filtered by the W matrix which is from the output as the filtered signal. Since the spectrogram is essentially a "fingerprint" of a person's voice, it shows particular features specific to the subject themselves. Therefore, dictionary fingerprint and sampled fingerprint are built on 100 samples. Based on the voice fingerprinting, different subjects will have different parallel bands as well as other features such as delays and time shifts. In this paper, the spectrograms of two different subjects (dictionary fingerprint $(d_1, d_2, \ldots d_n)$ and sampled word fingerprint $(w_1, w_2, \ldots w_n)$) are compared from each other with Cross-correlation. The dictionary fingerprint and the word fingerprint have to be amplitude normalized. This processing step simply amplifies dictionary fingerprint and the incoming fingerprint by positive or negative gain so that they have same level of peak. When we calculate the cross-correlation of the dictionary fingerprint and the word fingerprint, the average RMS (Root Mean Square) power of the result of cross-correlation shows difference between the same word or not. A process flowchart is showed in Figure 5-6.

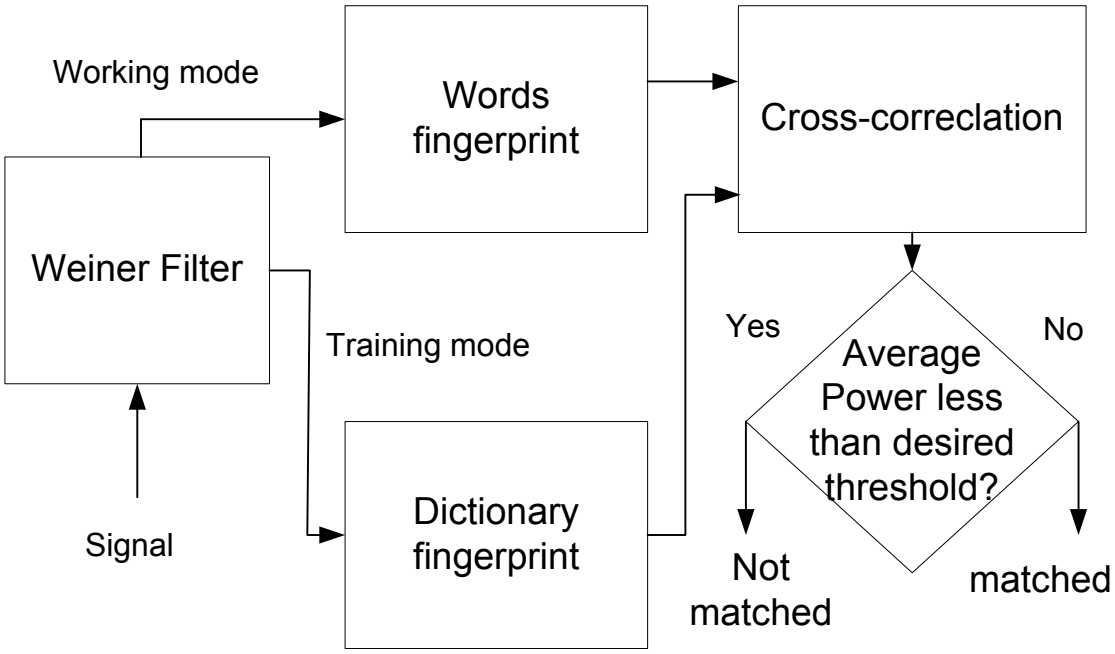

Figure 4-10 A simple ASR block diagram

This template matching ASR was designed using LabVIEW 8.0 (Qi & Moir, 2007b).

4.2.5 Summary and discussion

This thesis applied the Wiener-Hopf equation derived by Doclo & Moonen (Doclo & Moonen, 2002). Doclo & Moonen assumed two assumptions: short-term stationarity of the noise, and statistical independence of the speech and noise signals. Normally the NLMS approach is the technique used in updating an adaptive Wiener filter. Whereas the NLMS approach uses weight estimation to minimize a mean-square error, therefore, it is not suitable the applications of this solution in non-stationary noise environment. This thesis contributes a method of updating the **W** matrix in real-time. This alternative approach constructs the Wiener filter from estimation of covariance matrices (for signal + noise and "noise-alone").

In real-time environments a speech recognition system in a car has to receive the driver's voice only whilst suppressing background noise e.g. voice from radio. Therefore, this research presents a hybrid real-time adaptive filter which operates within a geometrical zone defined around the head of the desired speaker. Any sound outside of this zone is considered to be noise and suppressed. As this defined geometrical zone is small, it is assumed that only driver's speech is incoming from this zone. The technique uses three microphones to define a geometric based voice-activity detector (VAD) to cancel the unwanted speech coming from outside of the zone. However, when unwanted speech and desired speech are incoming at the same time, the VAD fails to identify the unwanted speech or desired speech. In such a situation, the adaptive Wiener filter is switched on for noise reduction. In the case of sole unwanted speech incoming from outside of a desired zone, this speech is muted at the output of the hybrid noise canceller. In the case of desired and unwanted speech incoming together, SNR enhance the desired speech and reduce the unwanted signal.

In next chapter, the experiments are to compare the ANC with a NLMS and the ANC with an adaptive Wiener filter.

5 Experiments

5.1 Overview

In experiments, DAQ device is an important part. Although commercialized products can be employed, pre-amplifier and anti-alias filter had been designed in this thesis to meet the need of DAQ PMD-1608FS. Details can be found in chapter 3.3 Overview of system build-up and the Appendix.

The LabVIEW program is employed in this thesis to fulfil a task to design a real-time beamforming based VAD with an adaptive filter with NLMS, and an adaptive Wiener filter, and also a template matching ASR. The CoolEdit program is used to store and replay the sample voices since we need to analysis the waveform of these voices. Details can be found in the Appendix.

Chapter 4 presents a novel 3-microphone beamformer (with a 3-microphone VAD and a 3-microphone NLMS ANC) and a novel adaptive Wiener filter in a car. Therefore, there are two parts of experiments in this thesis: Three-microphone beamforming in a car and adaptive Wiener filter within 3-microphone beamforming in a car. The first is to confirm the capability of the 3-microphone beamforming speech enhancement for ASR in a car. The second is to verify if the adaptive Wiener filter within 3-microphone beamforming in a car is able to be recognized by an ASR in a car.

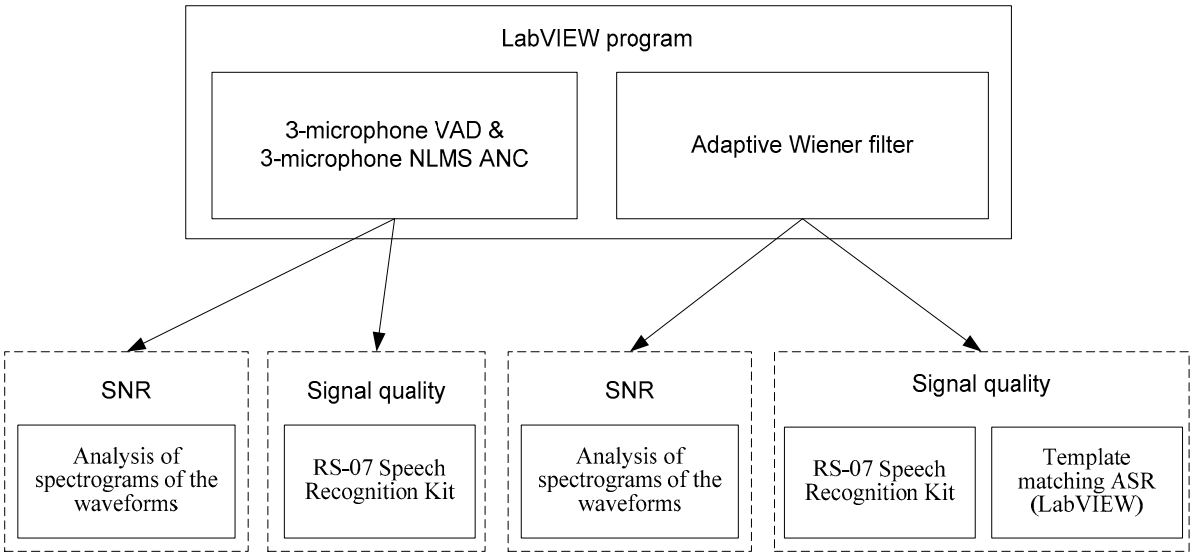

Figure 5-1 Overview of experiments

In order to identify the improvements in this thesis contrasted of other previous works, as showed in Figure 5-1, Signal-to-Noise ratio is a key figure for a noise canceller as it means the capability of picking up desired signal from noise environment. In order to identify the quality of the output signal, successful rate is an important result when the output signal is fed to an ASR. As described in Chapter 4, this thesis is to investigate a novel method of noise cancellation for ASR rather than produce a hi-fi quality of sound, a double blind aural test is not necessary.

In this thesis, chapter 5.2.4 “Comparison of NLMF and NLMS ANC” is to confirm NLMS has advantage of NLMF in noise cancellation as NLMS is much more stable and ability of SNR.

In chapter 5.3.5, an experiment for the size of Wiener filter matrix in noise cancellation is to confirm that possibility of an Adaptive Wiener filter running in real-time.

In chapter 5.3.6, an experiment is to suggest an engineering consideration for the location of the microphones.

More importantly, the experiment is also to compare the improvement of SNR under adaptive Wiener filter and NLMS filter in 3-microphne noise cancellation. Whilst a valid voice incoming from a desired zone an unwanted voice incoming from outside of desired zone, a noise cancellation is to enhance the desired voice e.g. a driver’s voice and reduce the unwanted voice e.g. voice from in-car radio. At an earlier stage, this thesis applied NLMS filter to do this, however, the final decision was made to use Wiener filter on our result of experiment.

5.2 Experiments – Three microphone beamforming in a car

This part is to answer two questions:

Can the 3-microphone VAD mute the voice which is incoming from outside of a desired zone?

Is the 3-microphone NLMS filter able to enhance a desired voice once we know it is incoming from a desired zone?

5.2.1 Three-microphone VAD in a car

Eight testing points have been set as in Figure 5-2. Figure 5-3 shows where these testing points located in a car (in this thesis, the test was in a BMW 318i, 1996 model). As defined in Chapter 4.1.4, around the test point 1 is the estimation of zone (EOZ) where the desired speech (the driver's head) is coming from. These tests were carried out in a stationary automobile with the engine running. While speaking at test point 1, microphone 1, 2 and 3 pick up the signal and output the enhanced signal for test point 1 by using the discussed algorithms. However, noise cancellation takes place at test points 2, 3, 4, 5, 6, 7 and 8 which are outside of the desired zone. For this experiment, a sampling interval of 11025Hz was used so that each sample interval corresponds to 90.7 μs. Typically d_{max} was chosen 5 samples and C_{min} was chosen as 0.5.

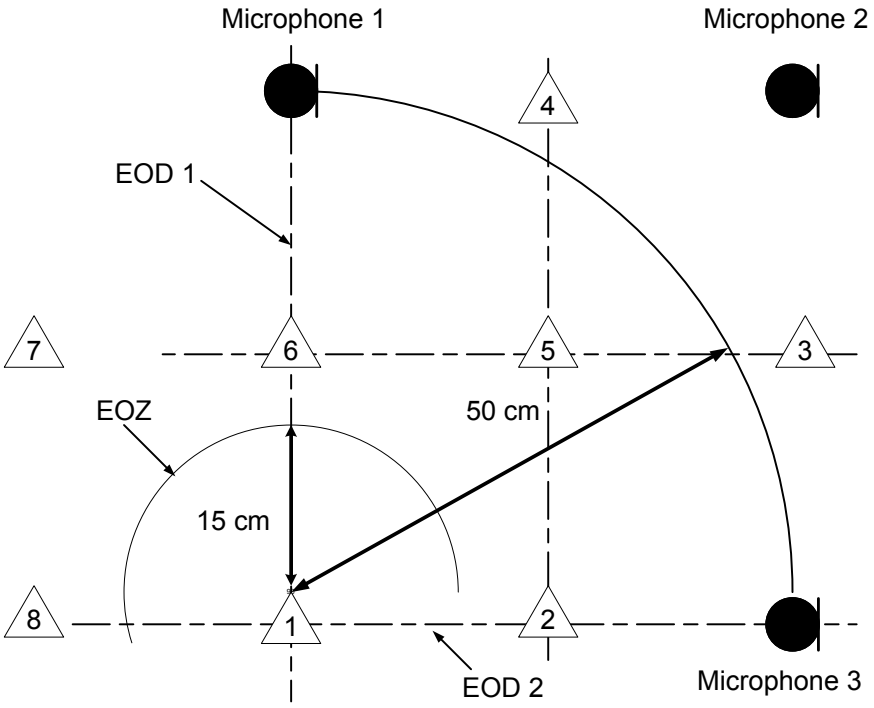

Figure 5-2 Eight testing points

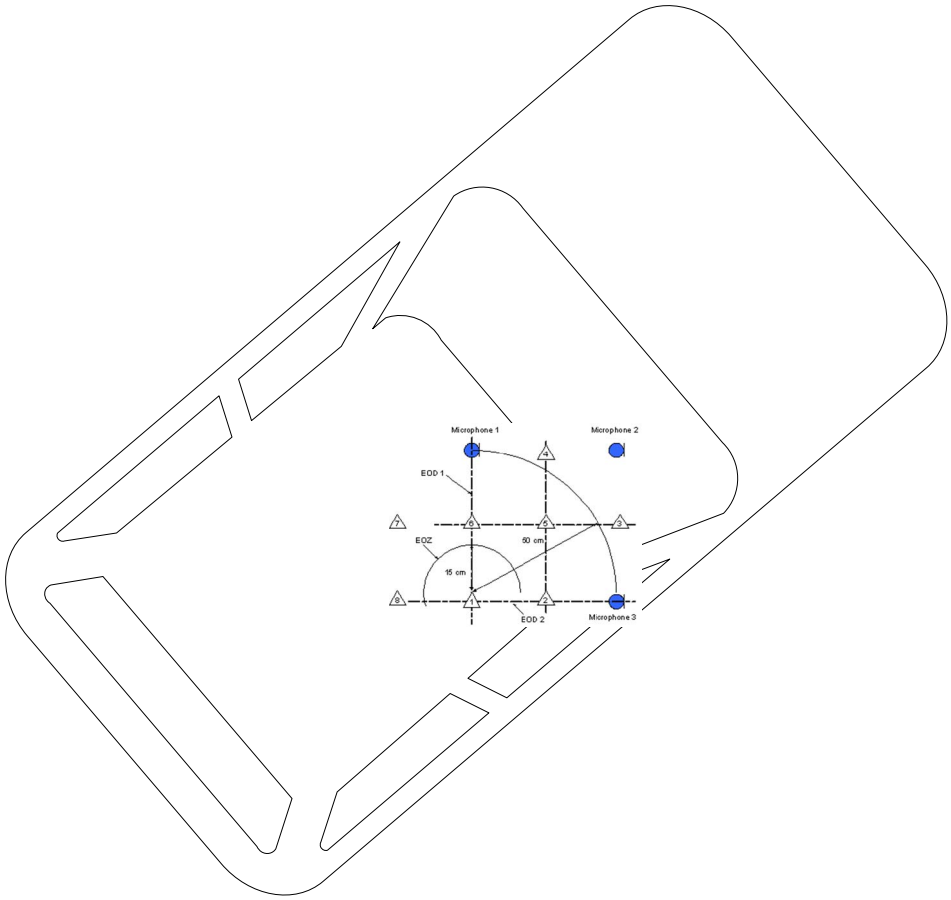

Figure 5-3 An overview of testing unit in a car

The experiment was conducted as follows: a loud-speaker outputs a pre-recorded phrase "Open the door" once at test point 1, then repeats this for test point 2 and so on to test point 8. Therefore Microphone 1, 2 and 3 pick up the phrase "Open the door" eight times with differing strength as shown in Figure 5-4. Waveform "Output A" in Figure 5-4 shows the output at the error *e(k)* from Figure 4-2 in Chapter 4.1. It indicates that speech from point 1 is enhanced but the speech picked up from points 2-8 are attenuated. The VAD can be programmed to switch off (mute) when the speech is not from point 1 so in effect the only noise cancelling that needs to be done is when speech is detected in the active zone. This is shown as "Output B" in Figure 5-4.

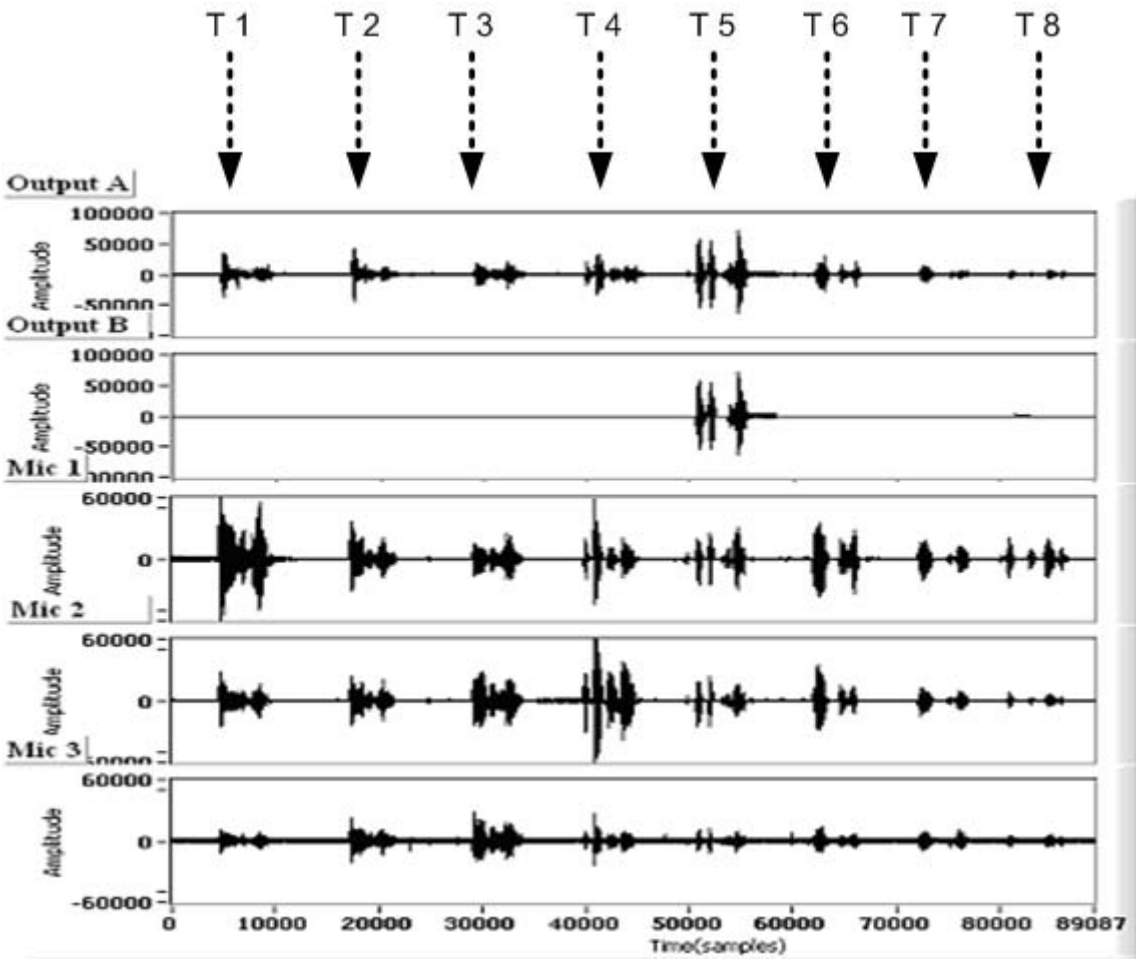

Figure 5-4 Speech waveforms

Since the waveforms in Figure 5-4 have the same sources acting as pre-recorded phrases 1 or 2 and so on SNR can be compared directly from

$$SNR_i = 10\log_{10}\frac{OutputPower}{Mic_i InputPower} \tag{5.1}$$

Where i=1,2,3….

The SNR results are presented at Table 5-1. For T1 in Table 6-1 the SNR should be as high as possible as this is desired speech whilst for the other test-points the SNR should be as small as possible indicating an attenuation in the speech as it appears outside the desired zone. At "Output A" in Figure 5-4, the un-desired speech cannot be cancelled completely. However, points 2 – 8 are very close to microphones indicating that much effort has to be done to reduce their power. Since we have a robust VAD it makes little difference whether there is in fact any residual speech after noise-cancellation since this can easily be muted as shown as Output B in Figure 5-4.

Table 5-1 SNR improvement in different test zones

	SNR_1 dB	SNR_2 dB	SNR_3 dB
T1	7.35	6.58	3.9
T2	0.93	-1.95	-10.76
T3	-1.3	-7.67	-9.04
T4	-4.96	-10.21	-4.82
T5	-7.1	-9.46	-8.76
T6	-8.48	0.58	0.65
T7	-9.62	-0.43	-2.56
T8	-10.17	-4.07	-5.64

5.2.2 NLMS adaptive noise cancellation in 3-microphone beamforming in a car

When all mechanical noise and undesired speech come from unknown directions, a microphone array beamformer is used to enhance speech from a geometrical zone and reduce any other speech or noise outside of this zone.(Agaiby & Moir, 1997) In order to improve hands-free speech recognition performance in car environments, a microphone beamforming array has been implemented with a Voice Activity Detector (VAD) which uses time-delay estimation together with magnitude-squared coherence (MSC). (Qi & Moir, 2005) This microphone array has been used to form a beamformer with normalized least-mean squares (NLMS) to improve Signal to Noise Ratio (SNR). The experiment clearly shows the ability of the composite system to reduce noise outside of a defined zone. Experiments have been conducted in real-time on a combined three-microphone VAD and noise-canceling system. The VAD assumes that the desired speech falls within a desired geometric zone in free-space which is most appropriate for an automobile environment as it can be defined around the drivers head. The noise-canceling is only required when noise is present during desired speech as the VAD will mute any solo noise-source outside of the zone. The experiment used only pre-recorded phrases. This work clearly demonstrates the ability of the algorithm to cancel speech outside of the zone.

However, in a frequently moving noise sources environment, the noise cancellation needs to suppress the unwanted noise when desired speech is also present. This paper investigates this problem in some detail with real-time experiments clearly showing the performance of the canceller.

The well known Normalized least mean squares (NLMS) algorithm is used extensively in adaptive filtering algorithms due to its simplicity for real-time applications. (Simon. Haykin, 2002) A NLMS filter block diagram is shown as Figure 2. To define the self learning process the filter uses an adaptive algorithm to reduce the NLMS between the output signal *y(k)* and the desired signal *d(k)*. For stationary (in the statistical sense) signals, when the NLMS performance criteria for the NLMS have achieved its minimum value through the iterations of the adapting algorithm, the adaptive filter is finished and its coefficients have converged to a constant solution. Then the output from the adaptive filter matches closely the desired signal *d(k)*. If the input data characteristics are changed, (sometimes called the filter environment) the filter adapts to the new environment by generating a new set of coefficients for the new data. Notice that when *e(k)* goes to zero and remains there.

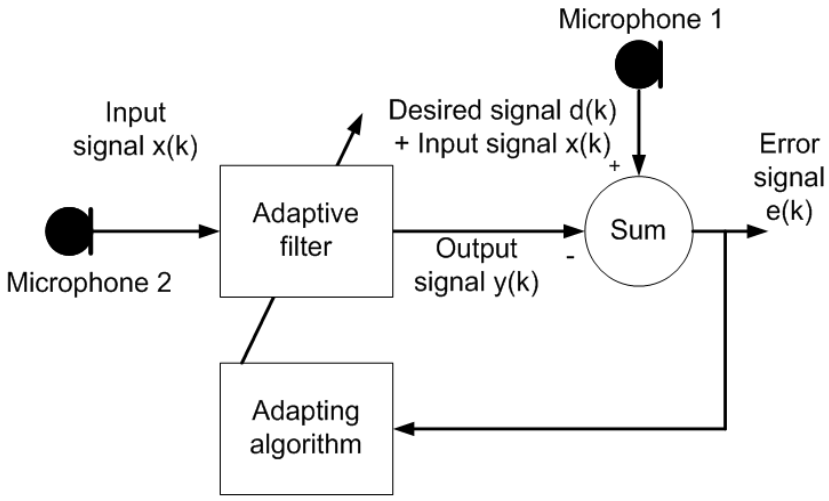

Figure 5-5 NLMS adaptive filter as noise canceller block diagram

As in Figure 5-5, we define Microphone 1 as the primary input and Microphone 2 as the reference input. In Figure 1, experiments (Rulph, 2002) show that voice close to the primary input is enhanced while voice close to the reference input is reduced.

A three-microphone noise canceller (Qi & Moir, 2005) based on Van Compernolle's work (Van Compernolle, 1990) is showed as Figure 5-6. There are four NLMS algorithms in a three-microphone noise canceller. The top path of the beamformer has a summation term which forms the primary input whilst both of the bottom paths have a difference term which forms the reference input. The three microphone signals contain speech as well as noise. The left section of the system serves at improving the noise reference by eliminating speech so that the Voice Activity Detection (VAD) switches this part on when speech energy is dominant. The right section consists of NLMS 2 and NLMS 4, which are only switched on to adapt during the absence of speech (i.e. during noise periods). For these experiments the number of weights used in W1 and W3 were 100 and in W2 and W4, 450. The 3-microphone VAD works so as to switch to freeze or enable the various NLMS algorithms. Also, the VAD switches off (mutes) the signal output when speech does not come from the desired zone. Therefore, while the driver's voice activates the 3-microphone VAD, the 3-microphone noise canceller is expected to reduce a passenger's voice, or for instance the car radio. For this to be successful the noise canceller valid zone must be positioned around the drivers head and be large enough to provide some movement.

In fact, whilst the Enable line in Figure 5-6 is enabled by the 3-microphone VAD (E = 1) and hence NLMS 1 and 3 are enabled, there are two pairs of microphones acting as NLMS filters: the first pair comprises Microphone 1 and 2 and the second pair comprises

microphone 1 and 3. As in Figure 1, experiments (Rulph, 2002) show that a voice close to the primary input (e.g. Microphone 1 in Figure 5-6) is enhanced and a voice close to the reference input (e.g. Microphone 2 or 3) is reduced. Therefore the driver's voice should be close to microphone 1 to be enhanced. In Figure 5-7, the circled area around microphone 1 is called the "noise canceller valid zone", where a desired voice is treated as a desired source but not noise. While the 3-microphone VAD active zone is in the square area as shown in Figure 5-7, a 3-microphone noise canceller valid zone should has an intersection with the VAD valid zone.

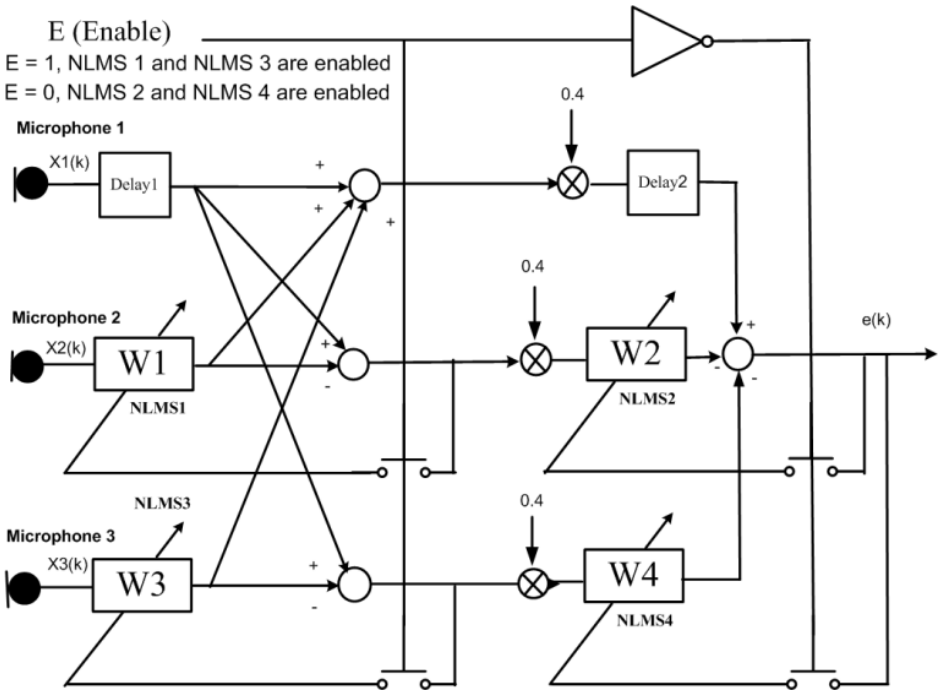

Figure 5-6 Three-microphone noise canceller block diagram

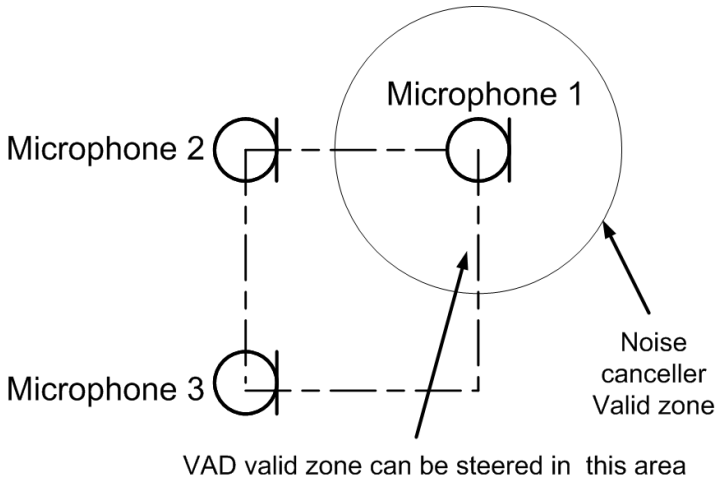

Figure 5-7 Definition of a noise canceller valid zone around microphone 1

When there are two or more voices present simultaneously it becomes difficult to measure definitively the improvement in SNR of the desired voice. Therefore these series of experiments are specially designed to measure this improvement by timing the second unwanted voice appropriately. For all of the following experiments the distance between microphone 1 and microphone 2 is 50 cm and the distance between microphone 1 and microphone 3 is 70.7 cm.

This chapter is to identify the capability of a NLMS filter in 3-microphone system.

Chapter 5.2.2.1 “The 3-microphone noise canceller in a 2 speech environment” is to test a NLMS filter in 3-microphone system when a voice incoming from the active zone and a voice from outside of the active zone.

Chapter 5.2.2.2 “Definition of Noise canceller valid zone” is to suggest a working model when applying a NLMS filter in 3-microphone system. The conclusion is the desired voice should be coming from nearby the primary microphone.

Chapter 5.2.2.3 “A single noise environment: Driver’s voice and a stationary white noise” is to test a NLMS filter in 3-microphone system when a desired voice with a stationary white noise (outside of the active zone) incoming. It also confirms the conclusion in chapter 5.2.2.2: the desired voice should be coming from nearby the primary microphone.

Chapter 5.2.2.4 “A single noise environment: Driver’s voice and a second speech” is to identify the capability of a NLMS filter in 3-microphone system when the desired voice and other speech (outside of the active zone) incoming at the same time. It also confirms the conclusion in chapter 5.2.2.2: the desired voice should be coming from nearby the primary microphone.

5.2.2.1 The 3-microphone noise canceller in a 2 speech environment

This experiment is to test a NLMS filter in 3-microphone system when a voice incoming from the active zone and a voice from outside of the active zone.

A 3-microphone noise canceller was tested as shown in Figure 5-8. When the driver's voice is presented the second voice is also present and is present after the driver's voice has stopped. In this way we can measure by how much the second voice has reduced (otherwise its reduced form overlaps the drivers voice and is difficult to measure). Since the driver's voice is located in the VAD valid zone, the 3-microphone VAD measures a desired voice and enables E (E = 1) as shown in Figure 5-6. When the driver's voice disappears, the 3-microphone VAD disables E (E = 0).

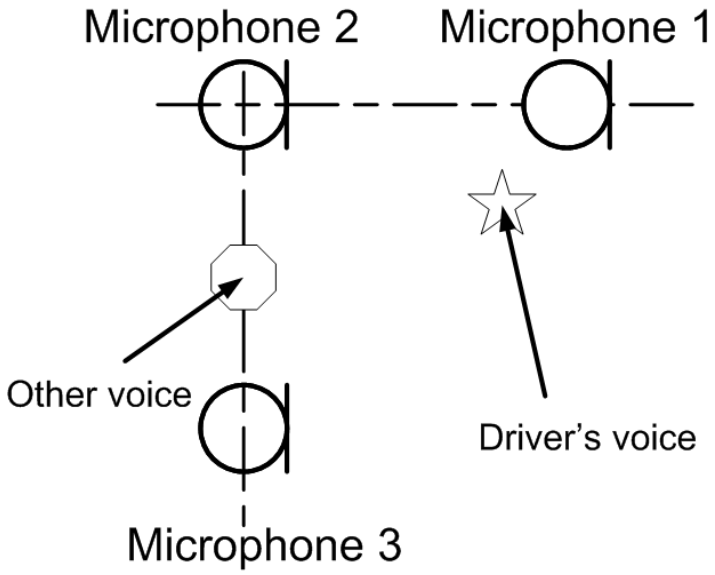

Figure 5-8 The 3-microphone noise canceller in a speech and unwanted speech environment

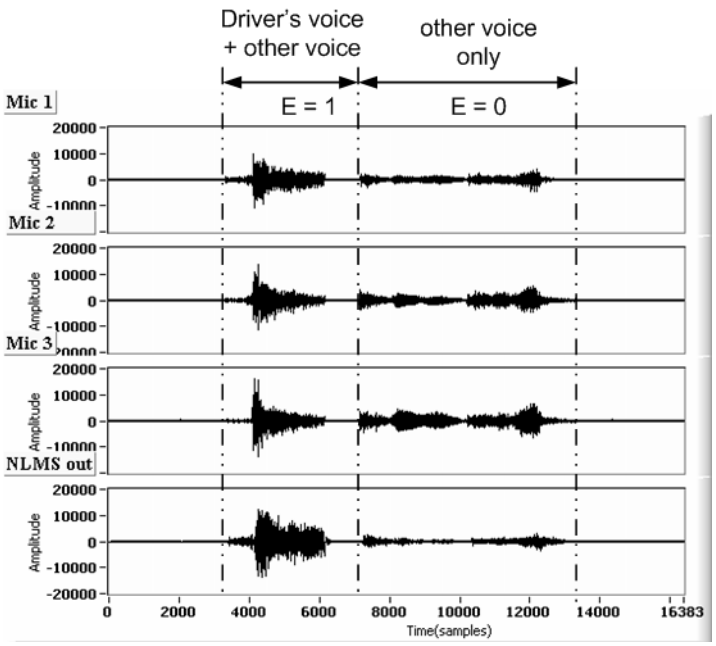

Figure 5-9 Refers to Figure 5-6, driver's voice enabled VAD (E = 1) and then disabled VAD (E = 0)

The result as shown in Figure 5-9 clearly indicates that when the VAD is enabled (E=1) the driver's voice is enhanced and otherwise (when E=0) the other voice is reduced. A previous experiment (Qi & Moir, 2005) clearly shows the Signal to Noise Ratio (SNR) improvement of the 3-microphone noise canceller is more than 6 dB. When both voices appear simultaneously and when E = 1, the driver's voice and other voice are overlapping and the result as shown in Figure 5-9 cannot clearly show (other than by listening tests) if the other voice is reduced significantly. Therefore, the following experiments investigate this in more details.

5.2.2.2 Definition of Noise canceller valid zone

This experiment is to suggest a working model when applying a NLMS filter in 3-microphone system. The conclusion is the desired voice should be coming from nearby the primary microphone.

Whilst E = 1 as shown in Figure 5-6, NLMS 1 and NLMS 3 are enabled by the VAD. In this experiment white noise comes via test points 7, 6, 5, 4, 3, 2 and down to 1. The output waveform of Microphones 1, 2, 3 and the 3-microphone noise canceller error (output) are shown in Figure 5-10. The waveforms in Figure 5-11 are clearly indicating that the output of the 3-microphone noise canceller followed and enhanced voice from microphone 1 but microphone 2 and 3 had no effect as each microphone has a valid sensitivity distance of 25 cm. This result indicates a "noise canceller valid zone" of this 3-microphone noise canceller as shown in Figure 5-10.

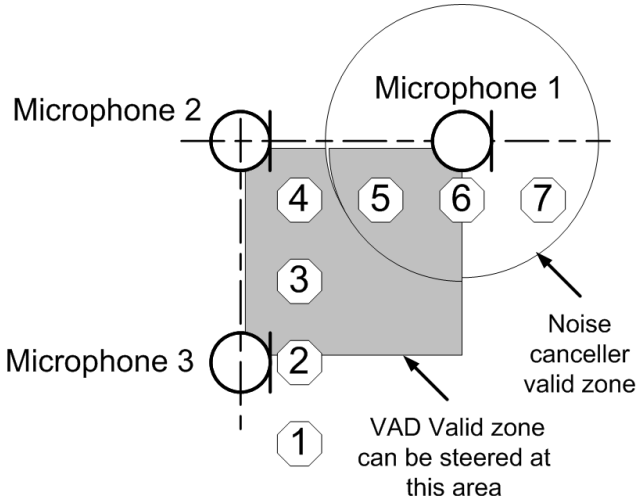

Figure 5-10 A test for the noise canceller valid zone

The SNR calculation is defined here as:

$$SNR_i = 10\log_{10}\frac{Power\ of\ noise\ canceller\ output}{Power\ at\ microphone\ i} \quad i = 1,2,3 \tag{5.2}$$

where the power of the noise at the noise canceller output and the power of the noise at microphone 1, 2 or 3 are total average power during the periods of block samples 1, 2 or 3 as shown in Figure 5-11. All results at the sample blocks are shown in Table 5-2. All block samples are 96ms second duration. The calculated SNR results are shown in Table 5-3. As shown in Table 5-3, only sample block 1 at microphone 1 is amplified and has 2.41 dB power. All other SNR(s) are attenuated since they are from outside of the noise canceller valid zone

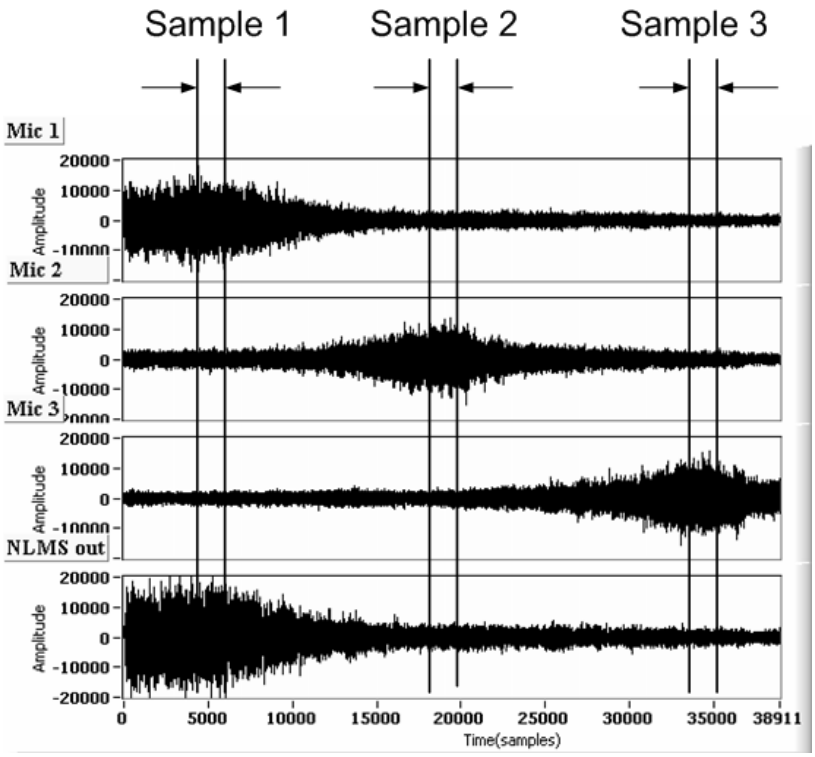

Figure 5-11 White noise source testing waveforms

Table 5-2 Power of microphones inputs and noise canceller's output

	Sample 1 (dB)	Sample 2 (dB)	Sample 3 (dB)
Microphone 1	-13.88	-26.95	-30.03
Microphone 2	-25.91	-15.64	-27.18
Microphone 3	-29.4	-28.04	-14.13
Output	-11.47	-24.15	-26.81

Table 5-3 SNR results of white noise test

	SNR_1 (dB)	SNR_2 (dB)	SNR_3 (dB)
Microphone 1	2.41		
Microphone 2		-8.51	
Microphone 3			-12.6

5.2.2.3A single noise environment: Driver's voice and a stationary white noise

This experiment is to test a NLMS filter in 3-microphone system when a desired voice with a stationary white noise (outside of the active zone) incoming. It also confirms the conclusion in chapter 5.2.2.2: the desired voice should be coming from nearby the primary microphone.

As speech is easily investigated in a mixed graphic of speech plus white noise, a white noise signal is used as a second voice to test the 3-microphone noise canceller in a frequently moving noise sources environment. A 3-microphone VAD valid zone was steered by pre-defined time-difference of arrival (TDOA). So VAD valid zone can be moved towards microphone 1. A driver's voice shown in Figure 5-12 turns on the 3-microphone VAD whilst at the same time, a white noise source jumps from test points 1, 2, 3 and so on up to test point 18.

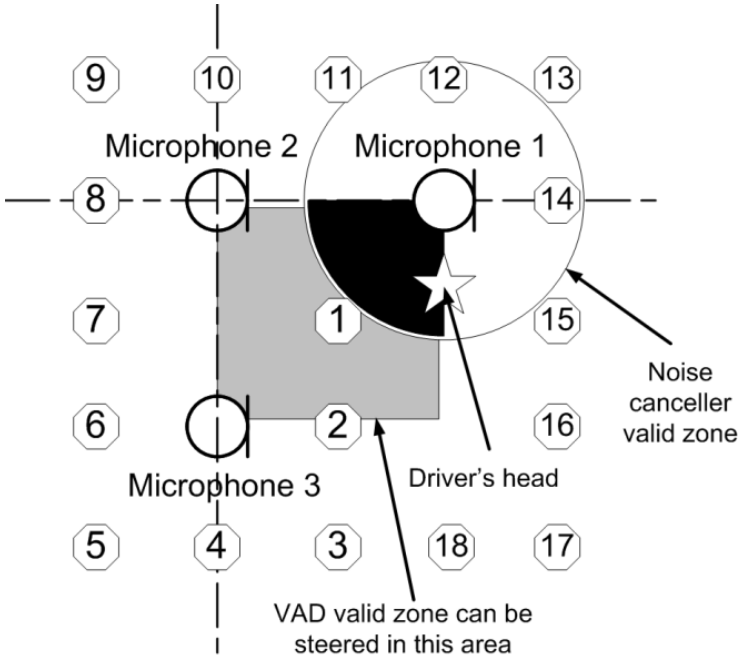

Figure 5-12 Numbering the test points in a frequently moving noise sources environment (Driver's voice and a second voice)

Figure 5-13 shows a test result when a driver's voice is in the VAD valid zone whilst E (shown in Figure 5-6) is held at unity by the VAD. A white noise source comes via test points 1 – 18 as depicted in Figure 5-12. Note from Figure 5-12 that the driver's voice has been enhanced but it is difficult to measure the reduction in noise. Therefore a test result with the Enable pin in Figure 5-6 manually enabled whilst white noise comes via test points 1, 2 ... and finally to test point 18 but with no desired speech present. The measurement of the dB noise reduction is now far easier than with Figure 5-12. As shown in Figure 5-14, when the

driver's (desired) voice is detected and a second voice is presented from test point 1 - 18, the signal output of 3-microphone noise canceller follows the signal from Microphone 1 only.

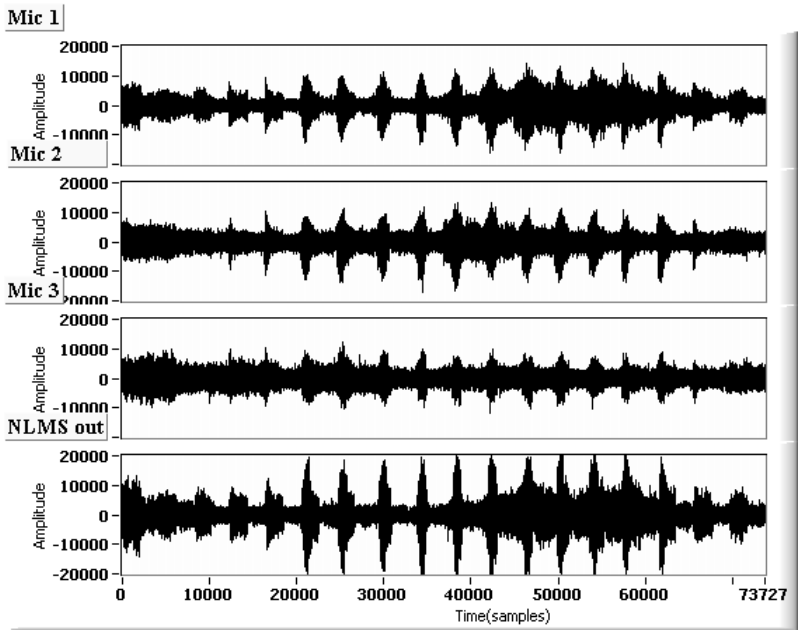

Figure 5-13 Voice in the VAD valid zone activates the VAD, whilst a white noise source comes from test points 1, 2 and so on

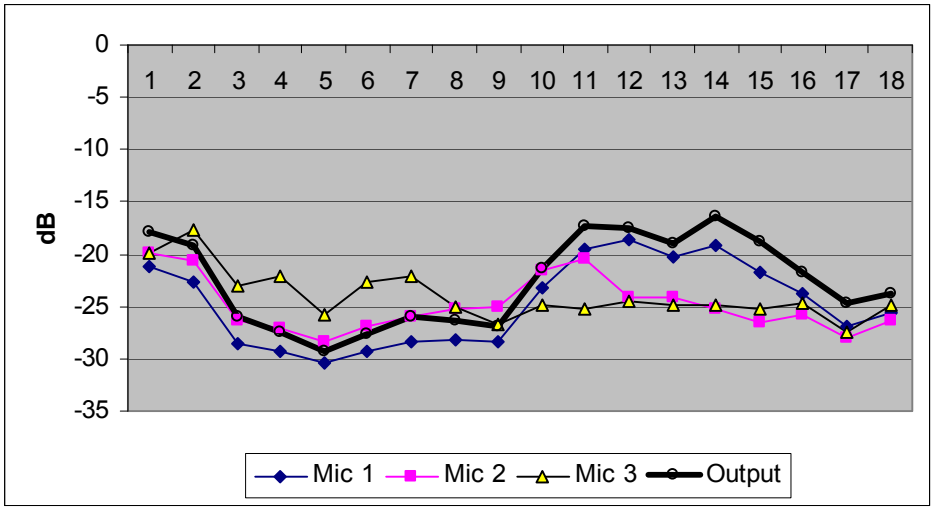

Figure 5-14 Moving second voice (white noise) results

5.2.2.4A single noise environment: Driver's voice and a second speech

This experiment is to identify the capability of a NLMS filter in 3-microphone system when the desired voice and other speech (outside of the active zone) incoming at the same time. It also confirms the conclusion in chapter 5.2.2.2: the desired voice should be coming from nearby the primary microphone.

A 3-microphone VAD valid zone was steered by pre-defined time-difference of arrival (TDOA). So this VAD valid zone can be moved towards microphone 1. A driver's voice shown in Figure 5-15 turns on the 3-microphone VAD whilst at the same time, the second speech comes via test points 1, 2 and 3. In Figure 5-16, from the top, the waveforms are Microphone 1, 2and 3 respectively. The waveform of the 3-microphone noise canceller output is shown at the bottom. Comparing input waveforms of Microphone 1 and the 3-microphone noise canceller output, we can clearly see that the 3-microphone noise canceller output follows the output of Microphone 1 and not microphone 2 and 3. Of course the second voice is still present but can easily be nullified by using the VAD.

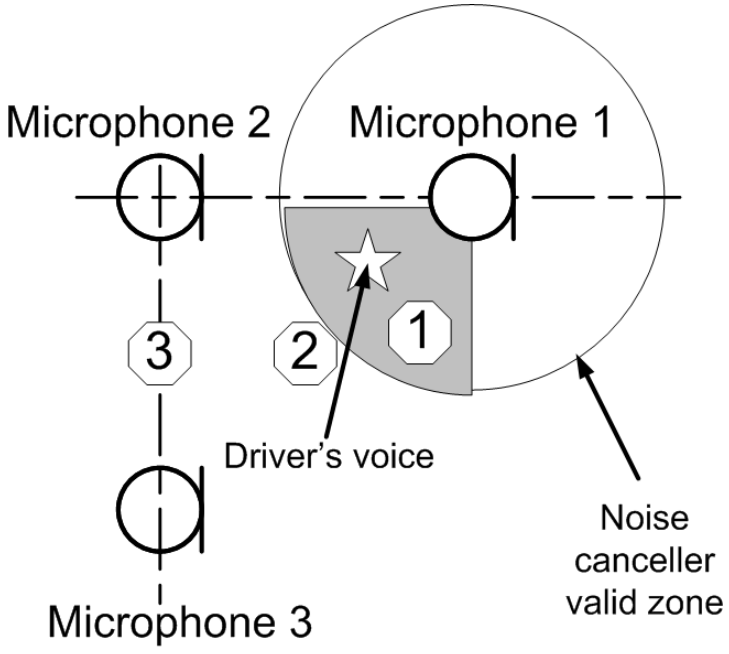

Figure 5-15 Numbering the test points in a frequently moving noise sources environment (Driver's voice and a second voice)

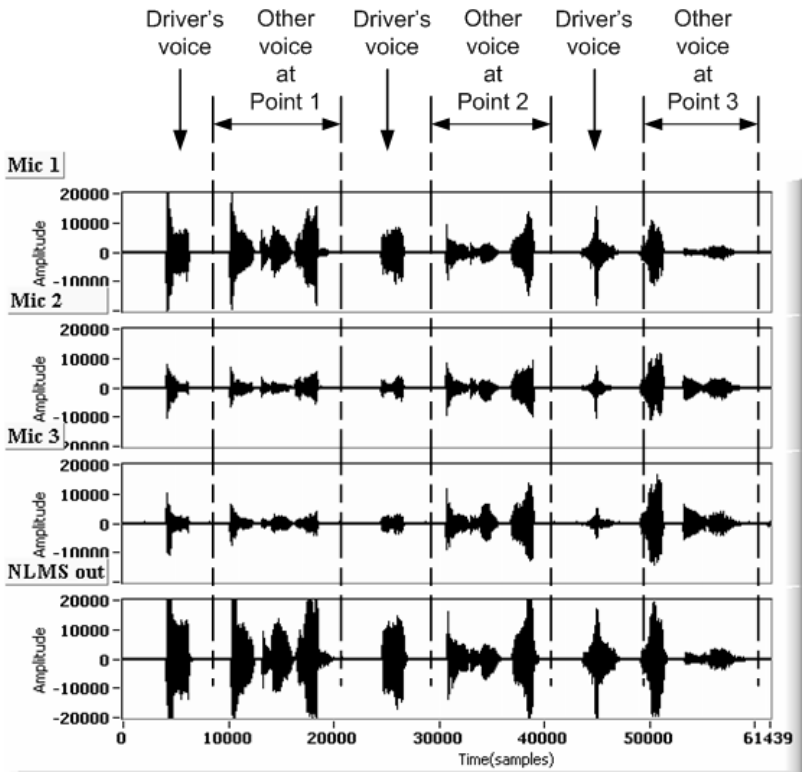

Figure 5-16 E = 1 (as in Figure 6-6), other speech arrives via test points 1, 2 and 3

In a frequently moving noise sources environment (noise sources are coming from different locations but not always presented at the same time), the 3-microphone noise canceller with geometric VAD has the effect of canceling un-wanted speech or noise from outside of a VAD valid zone. As the same time, there is a 3-microphone noise canceller valid zone defined. In order to enhance desired speech and reduce noise(s), a desired voice should be in the intersection of the noise canceller valid zone and the VAD valid zone. Thus all noise is suppressed outside this intersected area. Experiments performed have verified the improvements given by this method in a real environment.

5.2.3 An ASR with 3-microphone VAD and NLMS ANC

This experiment is to identify the signal quality from a 3-microphone VAD and NLMS ANC. As the purpose of the 3-microphone VAD and NLMS ANC in this thesis is to improve the ASR in a real-time environment, a test with an ASR is the best experiment rather than a double blind aural test.

As shown in Figure 5-17, a Speech Recognition Kit is employed to identify the quality of the outcome from 3-microphone VAD and NLMS ANC.

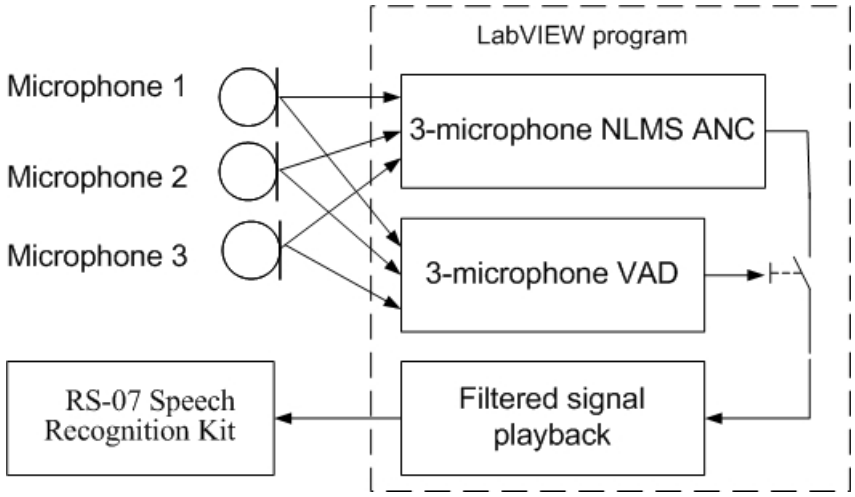

Figure 5-17 An ASR with 3-microphone VAD and NLMS ANC

A test plan as shown in Figure 5-2, a voice command is incoming from position 1, 2, … or 8. The speech recognition kit only recognizes a voice command from position 1 with 93% hit-rate (in 100 tests) as shown in Table 5-4.

Table 5-4 Result of AST successful rate in different source positions in 100 tests

Position	VAD switch on/off	successful rate
1	on	93%
2	off	0
3	off	0
4	off	0
5	off	0
6	off	0
7	off	0
8	off	0

5.2.4 Comparison of NLMF and NLMS ANC

As an argument of the application of NLMS filter or NLMF filter: which is the better noise cancellation, this experiment is to explore the fact. It is to confirm a NLMS filter in a 3-microphone system in this thesis.

A testing plan for Comparison of NLMF and NLMS ANC is shown in Figure 5-18. Microphone 1 is set as Primary input and Microphone 2 is acted as secondary input. Speech phrase "one two three" is presented while white noise is always existing.

From chapter 2.2.4.3, NLMS ANC is updated in:

$$W_{k+1} = W_k + 2\mu e_k^2 \frac{X_k}{\|X_k\|^2 + \delta} \tag{5.3}$$

For NLMF ANC, a new parameter A is introduced then

$$W_{k+1} = W_k + 2\mu A e_k^3 \frac{X_k}{\|X_k\|^2 + \delta} \tag{5.4}$$

μ is a specified convergence parameter. If μ is too large, NLMS or NLMF ANC become non-stable. Here μ is 0.001. When A = 0.1, 0.2,…0.9, the error output of NLMF is shown in Figure 5-19 to Figure 5-27.

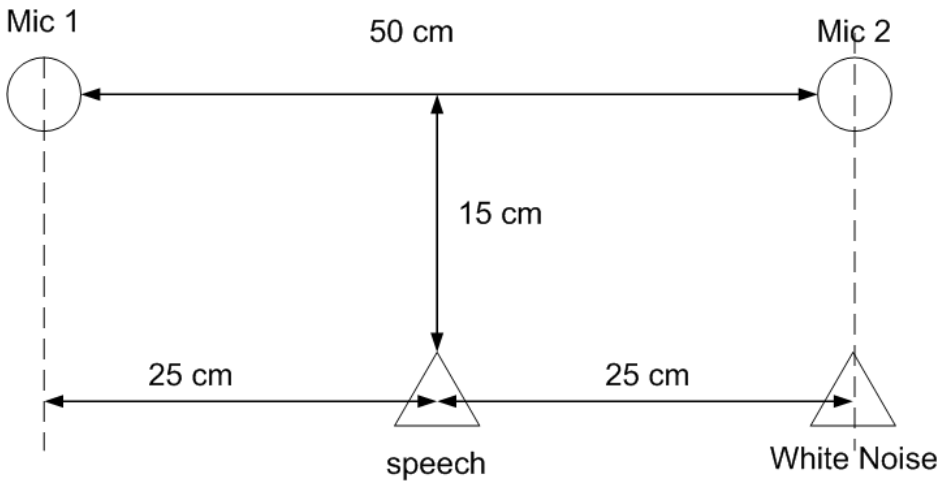

Figure 5-18 testing plan

In the Figure 5-19 to Figure 5-27, an English phrase "open the door" is recorded in a white noise environment. As the phrase "open the door" had constant volume output and the white noise source had also constant volume, the amplitude of the white noise in these figures shows the levels of SNR e.g. Figure 5-20 has a higher SNR than Figure 5-19.

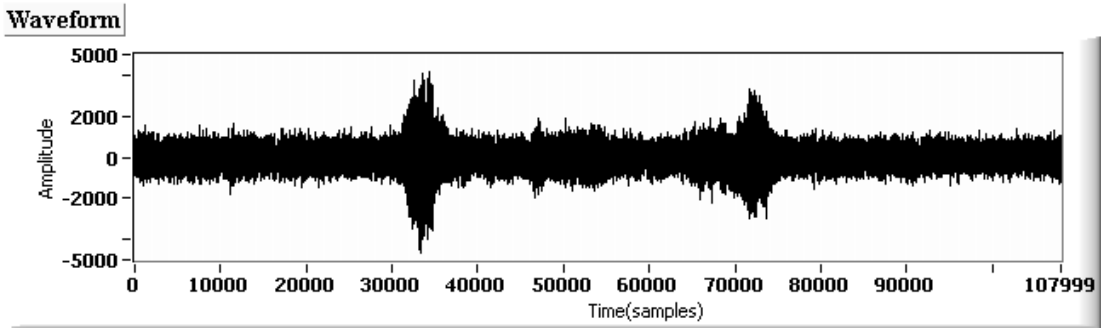

Figure 5-19 Testing of NLMF when A=0.1

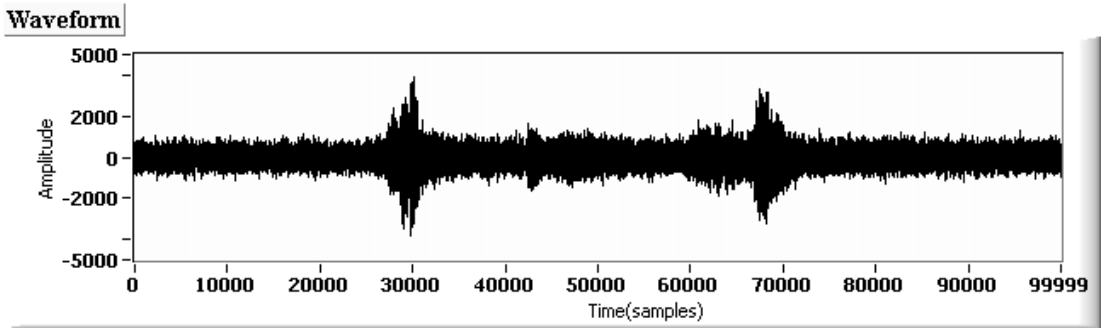

Figure 5-20 Testing of NLMF when A=0.2

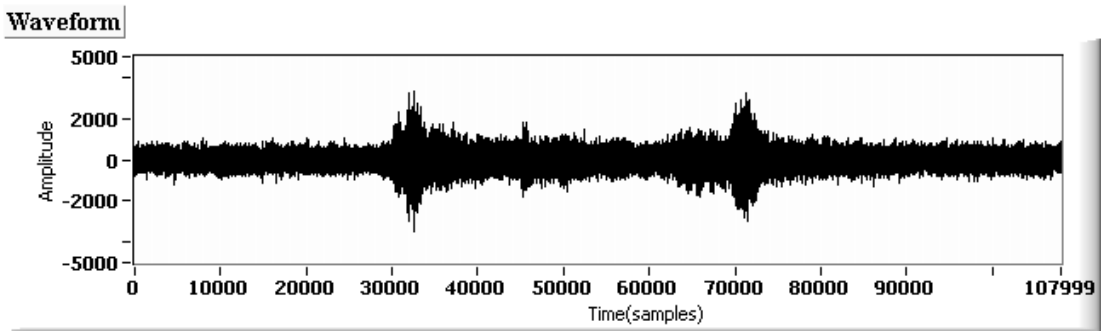

Figure 5-21 Testing of NLMF when A=0.3

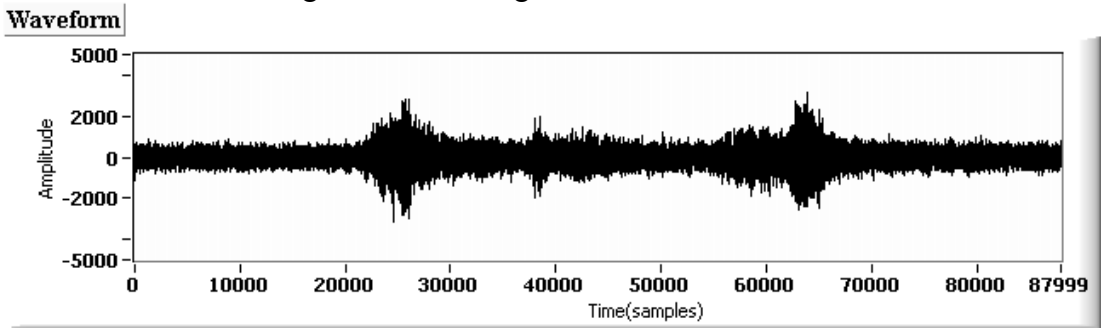

Figure 5-22 Testing of NLMF when A=0.4

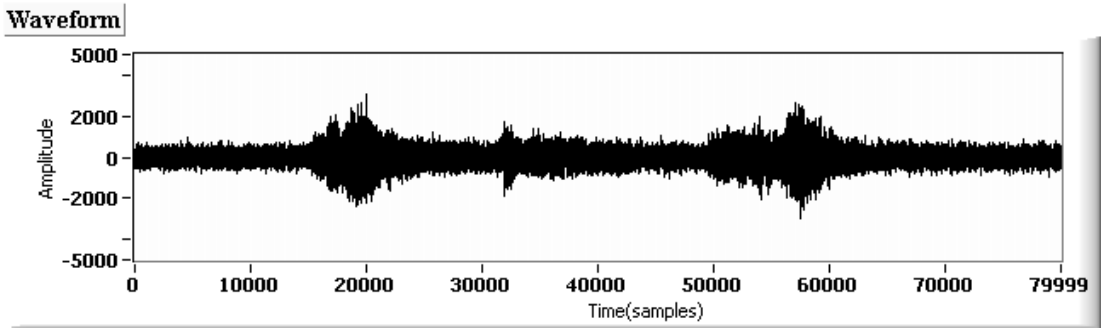

Figure 5-23 Testing of NLMF when A=0.5

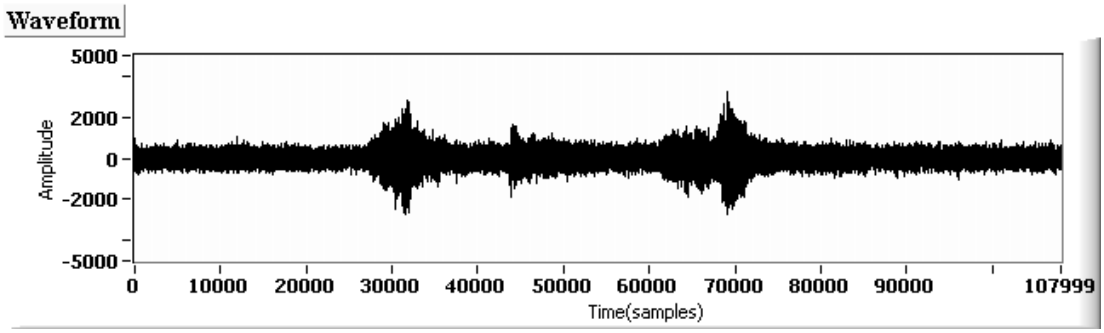

Figure 5-24 Testing of NLMF when A=0.6

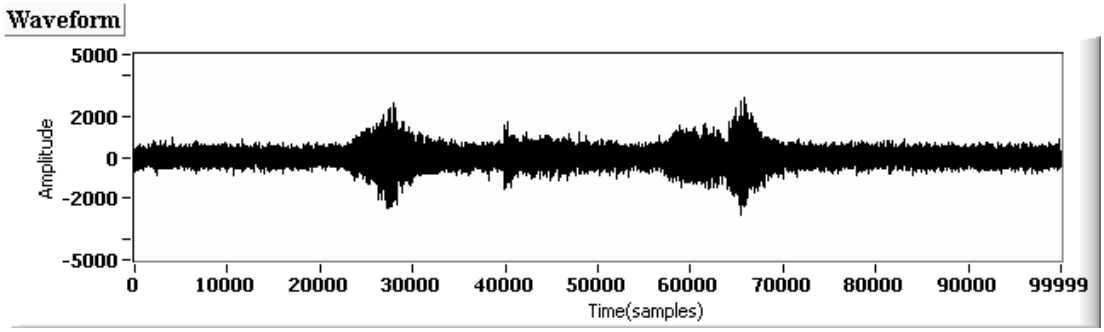

Figure 5-25 Testing of NLMF when A=0.7

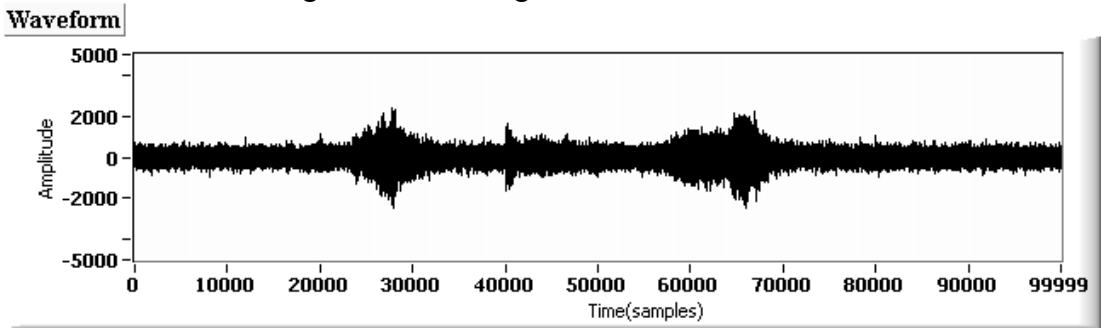

Figure 5-26 Testing of NLMF when A=0.8

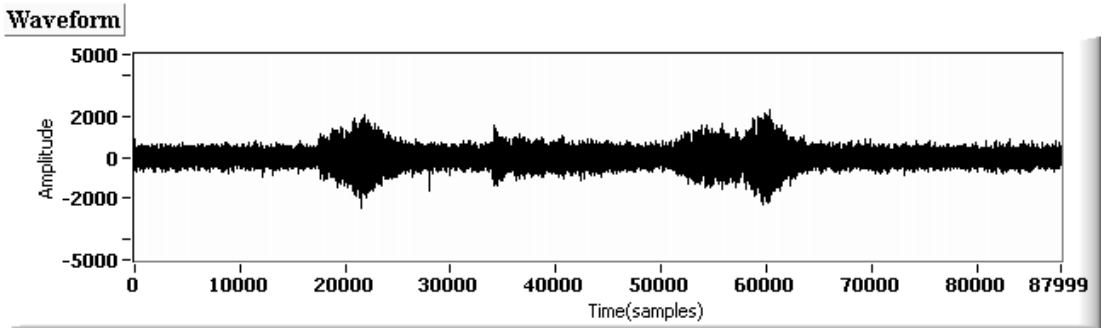

Figure 5-27 Testing of NLMF when A=0.9

When A = 1.0, NLMF become non-stable. Figure 5-27 is the best result of SNR from Figure 5-19 to Figure 5-27. The testing result of NLMS (as the same testing plan as shown in Figure 5-18) is shown in Figure 5-28.

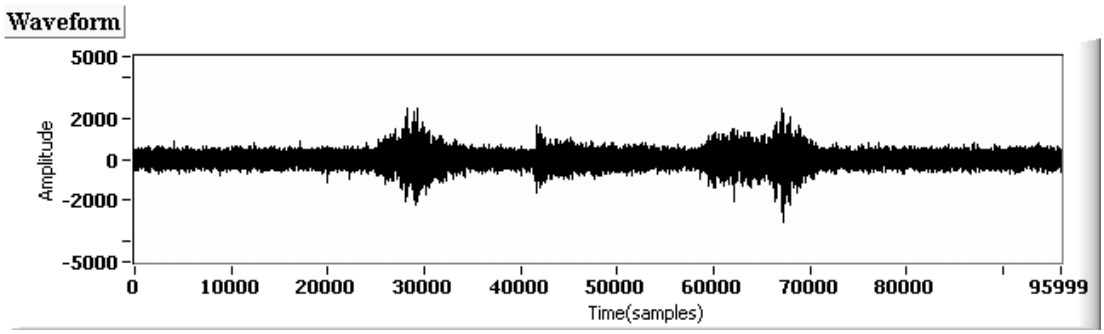

Figure 5-28 Testing of NLMS

As the phrase “open the door” had constant volume output and the white noise source had also constant volume, the strength of the white noise in these figures shows the levels of SNR, comparing Figure 27 and 28, NLMS has better SNR than NLMF.

5.3 Experiments – Adaptive Wiener filter in a car

As discussed in Chapter 4.2.2, this thesis applied the Wiener-Hopf equation derived by Doclo & Moonen (Doclo & Moonen, 2002). Doclo & Moonen had two assumptions: short-term stationarity of the noise, and statistical independence of the speech and noise signals. Normally the NLMS approach is the technique used in updating this adaptive Wiener filter. Whereas the NLMS approach uses weight estimation to minimize a mean-square error, therefore, it is not suitable the applications of this solution in non-stationary noise environment. This thesis contributes a method of updating the **W** matrix in real-time. This alternative approach constructs the Wiener filter from estimation of covariance matrices. This chapter is to prove a novel adaptive Wiener filter in a car.

The first part in this chapter is to compare the Wiener filter updated methods between the new in this thesis and previous applications by Doclo & Moonen (Doclo & Moonen, 2002).

The second part is to analyze the average power in the spectrograms at the filtered signal from adaptive Wiener filter to prove the SNR of this filter.

The third part is employed a commercialized product – RS-07 speech recognition kit to identify the quality of filtered signal from adaptive Wiener filter at speech recognition successful rate.

Fourthly, a LabVIEW program is designed to build a template matching speech recognition to test the quality of filtered signal from Wiener filter.

Finally, experiments show the performance of approached Wiener filter:

- The best selection for the size of the Wiener filter matrix in a car in a real-time environment.
- The best location for the microphone which represents the unwanted speech in a car in a real-time environment.

5.3.1 Case study one: comparison of adaptive Wiener filter update methods

This chapter is to compare the Wiener filter updated methods between the new in this thesis and previous applications by Doclo & Moonen (Doclo & Moonen, 2002).

In previous application, adaptive Wiener filter updates as Figure 4-7 by Doclo & Moonen (Doclo & Moonen, 2002), which is updated during the noise periods or speech periods respectively. However, this thesis is to prove a method, as Figure 4-8, that Wiener filter is updating in both noise periods and speech periods.

A simulation was built to test these assumptions using a sample rate of 11025 Hz as in Figure 5-29. The anti-alias filters had a cut-off frequency of 5 kHz.

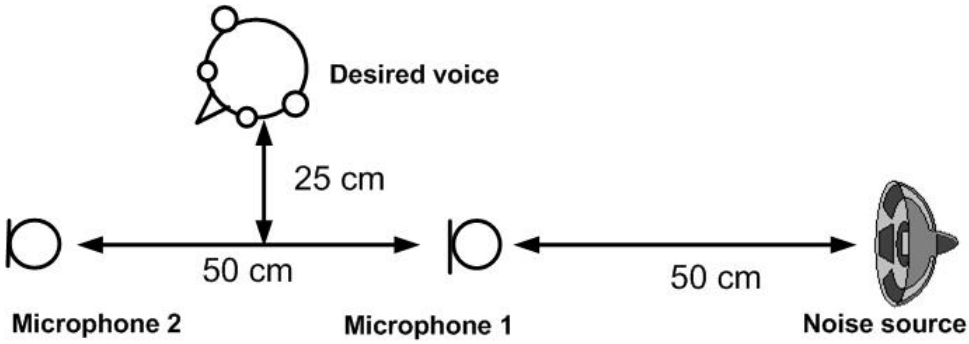

Figure 5-29 Experimental setup

In Figure 5-29, pre-recorded files from Microphone 1 or 2 were recorded at a sample rate of 11025 Hz. The numbers of elements in the array for these pre-recorded files have length *M*. As the method of Figure 4-7 suggests, during active speech periods the elements of the array from Microphone 1 is not updated and keeps the previous record of noise periods, as in Figure 5-30 (c). Therefore, the Wiener filter matrix W is updated during the noise periods only, or during active speech periods. The Wiener filter output is showed in Figure 5-30 (a) whilst Microphone 2 Figure 5-30 (b) is incoming speech + noise (input y). Since the noise statistics change between noise periods and speech periods then the W matrix cannot be updated correctly.

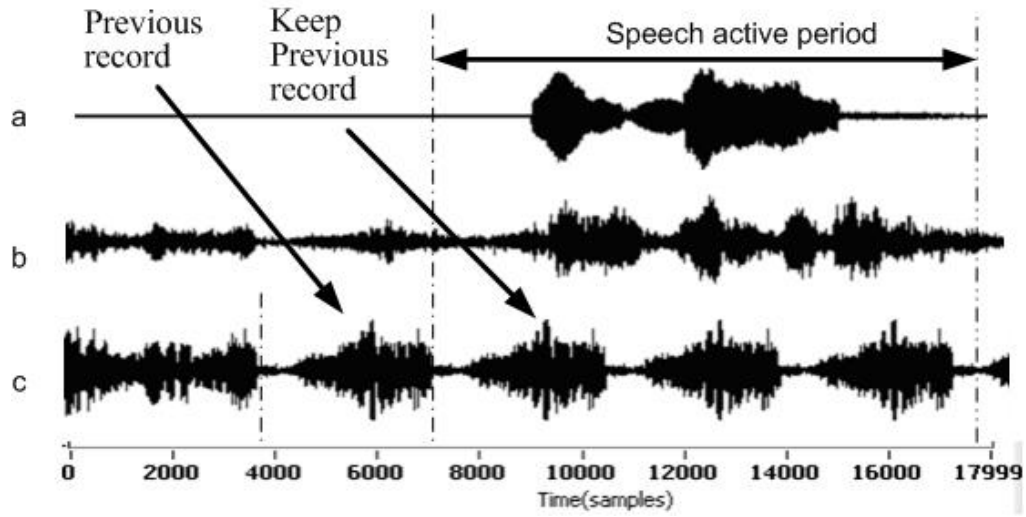

Figure 5-30 The W matrix is updated during a noise period only or a speech activity period. (a) Wiener filter output (b) Waveform of Signal + noise from Microphone 2 (c) Waveform of noise from Microphone 1

In Figure 5-31, using the method shown in Figure 4-8, the Wiener filter output is showed in Figure 5-31 (a). Figure 5-31 (b) is microphone 2 - incoming speech + noise (input y). Microphone 1 shown in Figure 5-31 (c) is changed during the desired speech period. Comparing Figure 5-30 and Figure 5-31, in order to achieve a cleaner estimate from the original speech, the method of Figure 4-8 (which is suggested in this thesis) is clearly an improved method.

Using waveform statistics in CoolEdit program (see Appendix for details), the results show this filter reduces the unwanted speech by approximately 30dB.

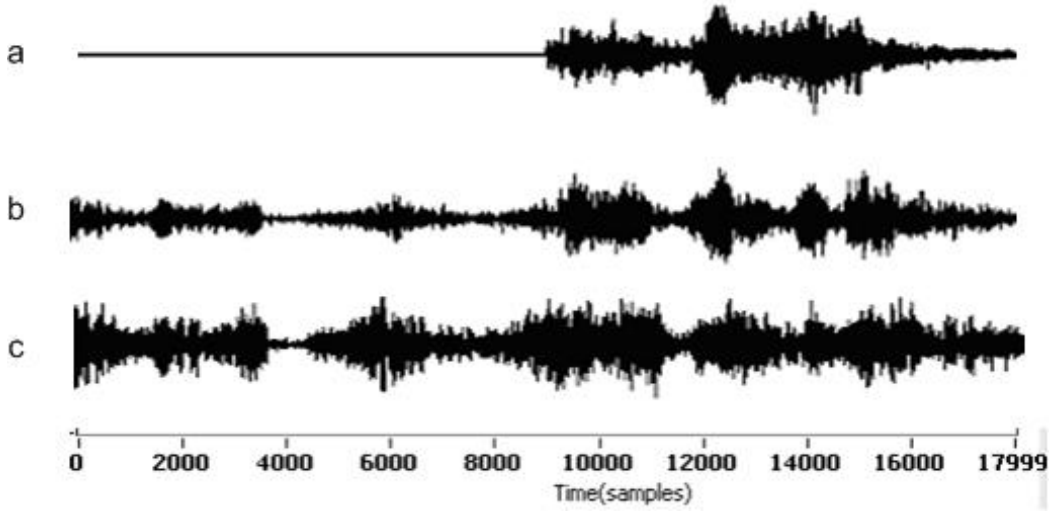

Figure 5-31 W matrix is updated in real-time. (a) Wiener filter output (b) Waveform of signal + noise from Microphone (c) Waveform of noise from Microphone 1

Figure 5-32 shows the spectrograms of the original speech phrase “open the door” (Figure 5-32 (a)). From Microphone 1 in Figure 5-32 (b), the spectrogram shows the desired speech phase “open the door” + simultaneous unwanted speech babble from a Chinese radio channel. Figure 5-32 (d) shows the spectrogram at Microphone 2, the desired speech phase “open the door” + simultaneous unwanted speech from a Chinese radio channel. Figure 5-32 (c) is the spectrogram of the filtered signal using 500 samples from microphone 2 and filtered by a 500 x 500 Weiner filter Matrix. Comparing with Figure 5-32 (a) and (b), it is clear that spectrogram of “open the door” of the filtered signal is similar to the original signal. The 500 x 500 Weiner matrix is updated by the 500 samples from Microphone 1 and another 500 samples from Microphone 2. The samples from Microphone 2 are also filtered by the Wiener matrix whose output is the filtered signal.

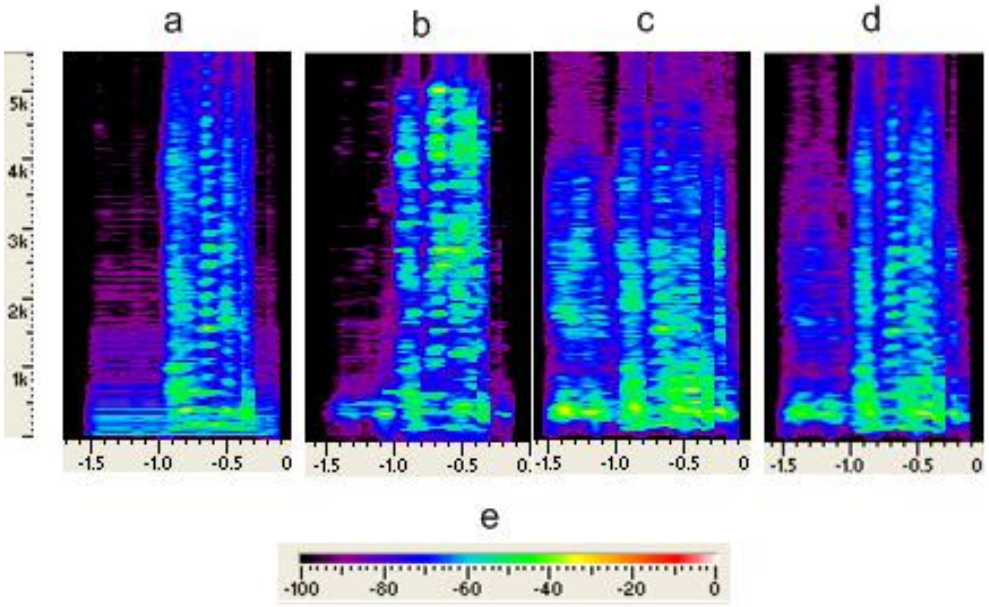

Figure 5-32 Spectrograms of the filtering process. The horizontal axis represents time (in seconds) and the vertical axis is frequency (in Hz) (a) clean speech “open the door” (b) Filtered speech (c) recording of Microphone1 (d) recording of Microphone2 (e) Intensity scale in dB

5.3.2 Case study two: Analysis of average power in the spectrograms

This part is to analyze the average power in the spectrograms at the filtered signal from adaptive Wiener filter to prove the SNR of this filter. The tool for analysis of average power in the spectrograms is CoolEdit program (see the Appendix for details).

An in-car test was performed with layout as in Figure 5-33. Microphone 1 is in the central area of the car and Microphone 2 is in the front of the driver. Whilst the driver is speaking "open the door", the radio in the car is simultaneously switched on to a Chinese radio channel. Although the car is stationary, the car engine is idling during the test making the overall problem quite challenging. The noise volume was typical of a car radio - but not deafeningly loud. The absolute dB level was not measured but rather dB is just used here as a relative measure of signal or noise strength.

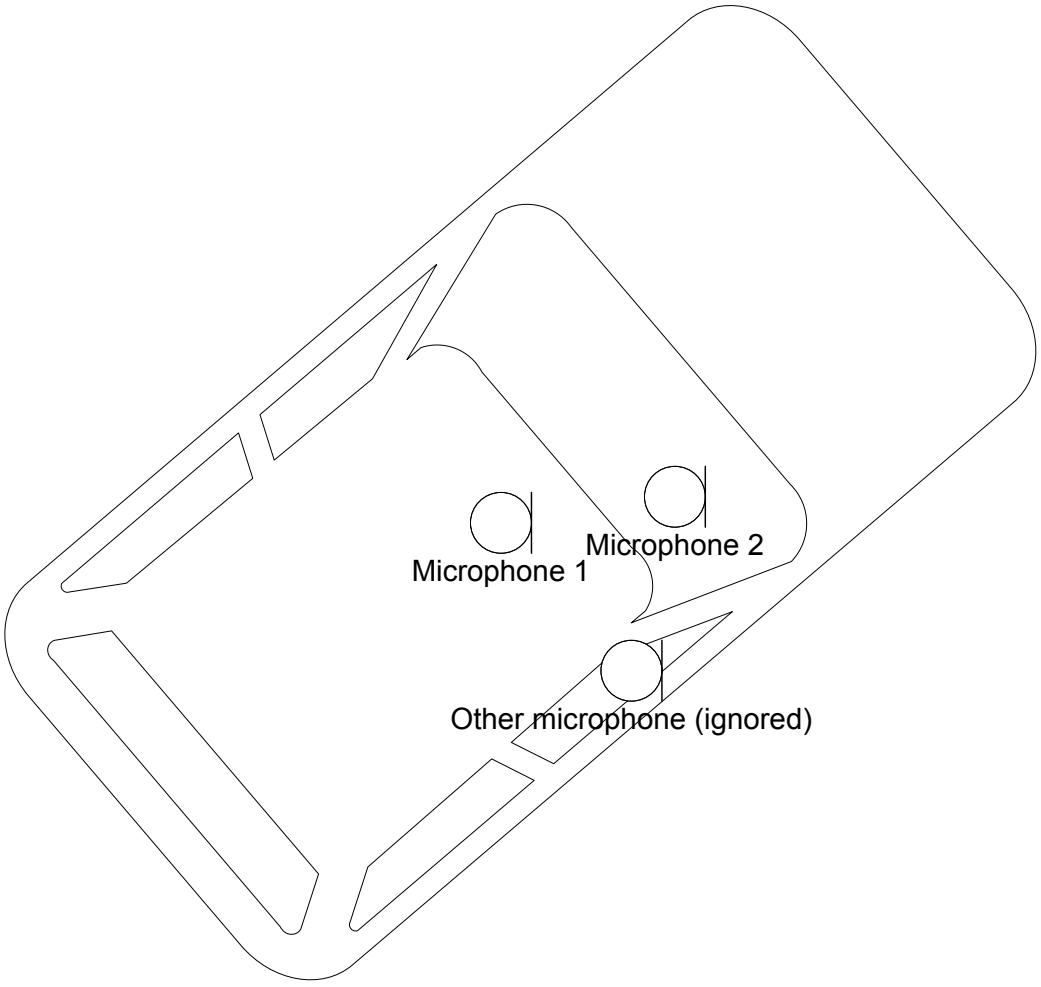

Figure 5-33 In-car test plan

Test results are shown as Figure 5-34. Figure 5-34 (a) is the incoming signal at Microphone 1, which is receiving the voice from all radio loud-speakers as well as the driver's voice as positioned in Figure 5-33. Figure 5-34 (b) is the incoming signal at Microphone 2, which is receiving the voice from all radio loud-speakers as well as simultaneously the driver's voice. Figure 5-34 (c) is the filtered voice signal, which is keeping the driver's voice whilst reducing the voice from the radio loud-speakers. Table 6-5 shows the waveform statistics using Cool Edit Pro regarding the average power during the driver's speech period + speech from the radio period and also the speech only from the radio period. In Table 6-5, SNR is computed as follows:

$$\text{SNR} = \text{Average Signal Power} - \text{Average Noise Power (in dB)} \quad (5.5)$$

The results show Signal-to-Noise Ratio (SNR) is improved by 28.26 dB between the filtered signal and the noisy signal received at Microphone 2. This result was performed at an 11025 Hz sample rate. The W matrix is 1000x1000 and is updated by the 1000 samples from Microphone 1 and another 1000 samples from Microphone 2. These samples from Microphone 2 are also filtered by the W matrix which forms the output as the filtered signal as in Figure 5-34 (c).

Figure 5-35 shows the spectrograms of the waveforms of Figure 5-34. The horizontal axis represents time (in seconds) and the vertical axis is frequency (in Hz). Figure 5-35 (a) is the spectrogram at Microphone1. Figure 5-35 (b) is the spectrogram at Microphone2. Figure 5-35 (c) is the spectrogram of the filtered signal. Figure 5-35 (d) shows the intensity scale in dB.

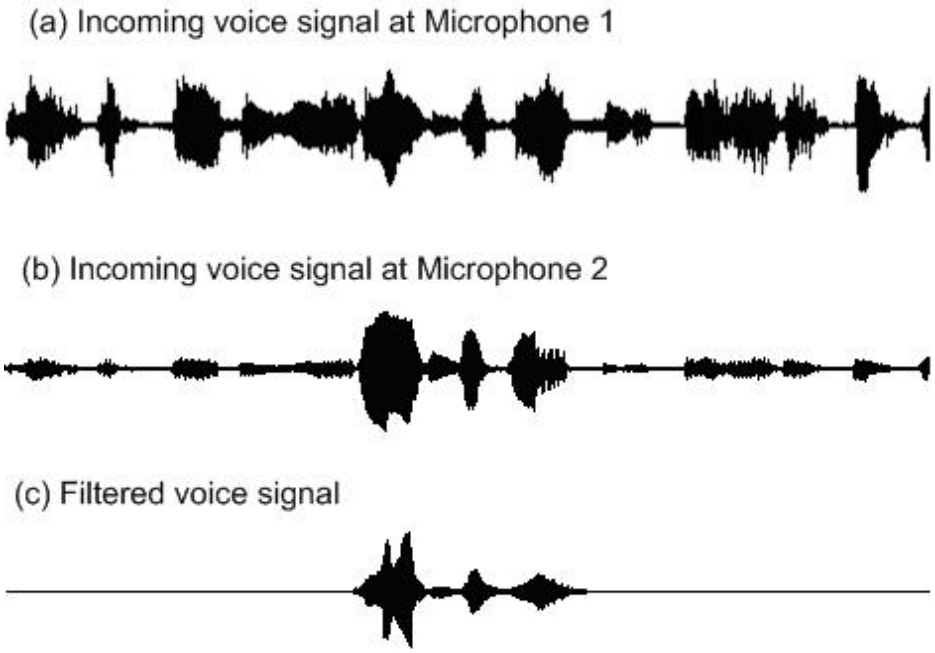

Figure 5-34 Test results in a stationary car with engine and car radio on.

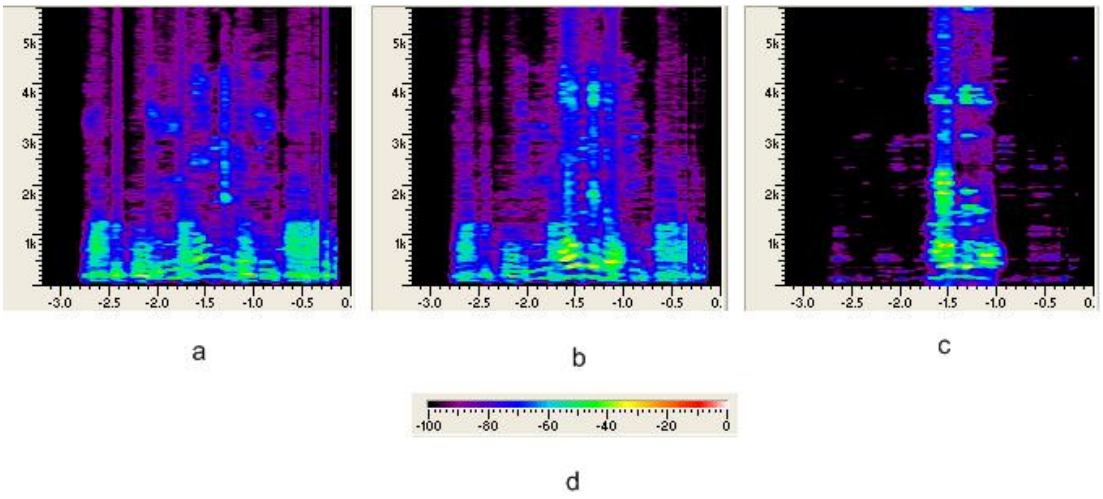

Figure 5-35 The spectrograms of the waveforms in Figure 6-34. The horizontal axis represents time (in second) and the vertical axis is frequency (in Hz). (a) Spectrogram at Microphone1. (b) Spectrogram at Microphone2. (c) Spectrogram of filtered signal. (d) Intensity scale in dB

Table 5-5 SNR Analysis at average power

	Average power in only speech from radio period (dB)	Average power in driver's speech + speech from radio period (dB)	SNR (dB)
Filtered signal	-60.17	-17.44	42.73
Signal at Microphone 2	-31.91	-17.39	14.52
SNR improved			28.21

Table 5-5 shows that the SNR is improved by 28.21 dB.

5.3.3 Case study Three: Test in RS-07 Speech recognition kit

The purpose of this chapter is to investigate an adaptive Wiener filter for ASR. Therefore, the criteria for testing are a SNR and the successful-rate of ASR. A commercial ASR in this experiment is to identify the quality of the filtered signal. The select criteria for a commercial ASR are reliable, stable and repeatable. Currently, there are two major speech recognition kits in the world (see Appendix for details). This study employs a speech recognition kit RS-07 (see Appendix for details).

An in-car test was performed with layout as in Figure 5-35. Microphone 1 is in the central area of the car, in such a position where it will in all probability pick up the radio more than the drivers voice. Microphone 2 is in the front of the driver, so it is in a position where it will more likely pick up the drivers voice than the radio signal. This process was performed at an 11025 Hz sample rate. The W matrix is 100x100 and is updated by the 100 samples from Microphone 1 and another 100 samples from Microphone 2. These samples from Microphone 2 are also filtered by the W matrix which forms the output as the filtered signal.

The test is in two parts: the first test is to identify the speech recognition successful rate without an adaptive Wiener filter. The second test is to identify recognition successful rate with an adaptive Wiener filter.

5.3.3.1 The inputs to the speech recognition successful rate without an adaptive Wiener filter

A LabVIEW program in a laptop is designed to record speech signal from microphone 2 and then playback to RS-07 speech recognition kit as in Figure 5-36.

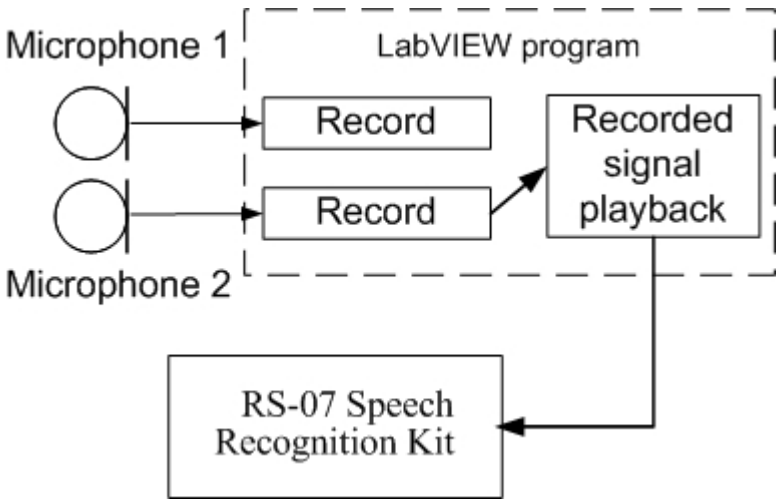

Figure 5-36 Test without an adaptive Wiener filter

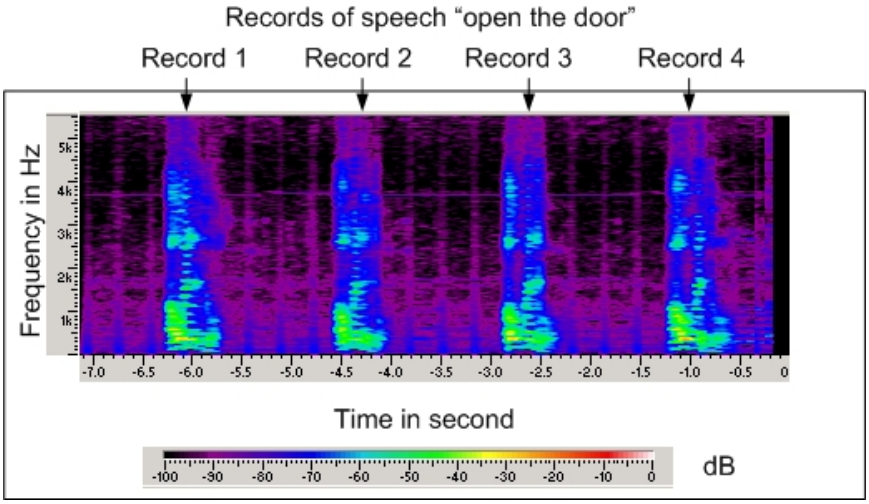

Figure 5-37 records of "open the door" in low noise environment when the car radio and engine does not turn on

As an example, there are 4 records of "open the door" in Figure 5-37. In this case the car radio and engine were not turned on. In this test, 100 records with "open the door" were made. Once the car radio and car engine were turned on, the records were from microphone 1 and 2. As examples, Figure 5-38 shows 4 records of "open the door" from microphone 1 while the car radio is playing classical music. And Figure 5-39 shows 4 records of "open the door" in noisy environment from microphone 2 while the car radio is playing classical music. In this test, 100 records of "open the door" with music were made.

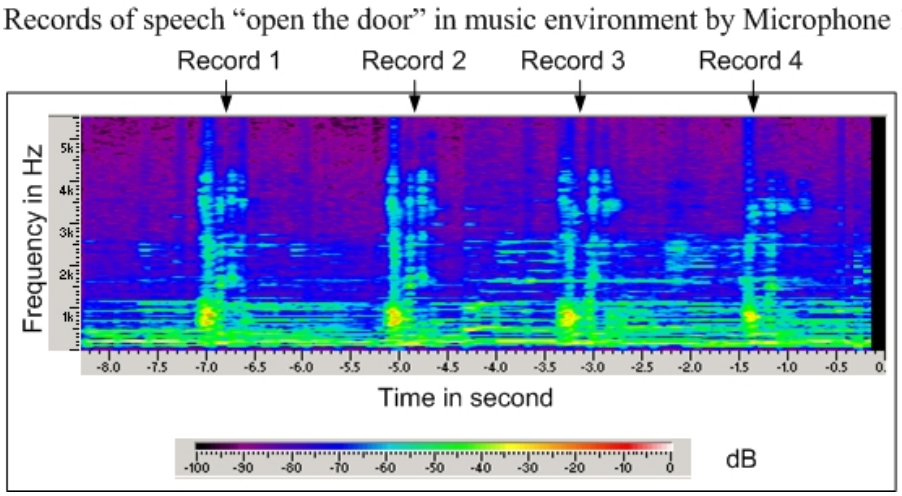

Figure 5-38 Records of speech "open the door" in music environment by Microphone 1

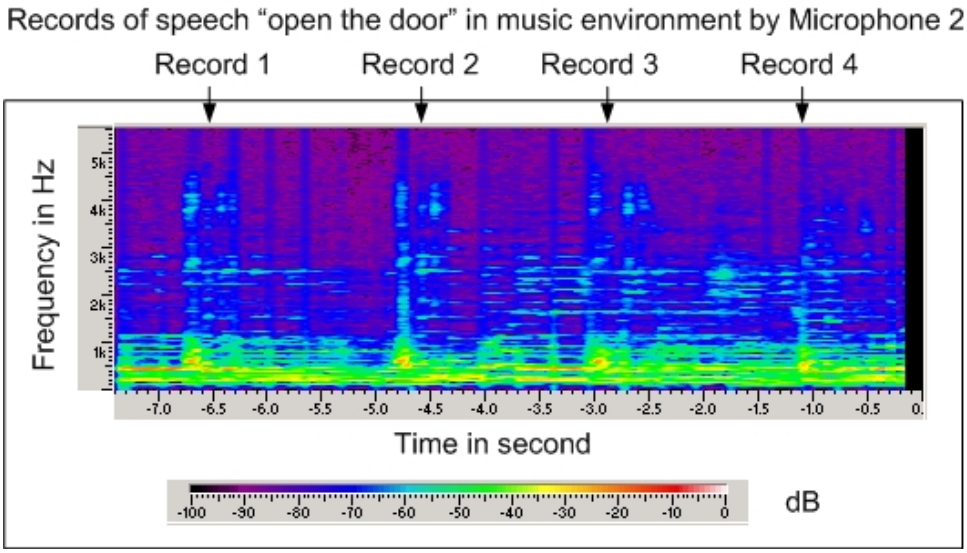

Figure 5-39 Records of speech "open the door" in music environment by Microphone 2

As examples, Figure 5-40 shows 4 records of "open the door" from microphone 1 while the car radio is playing news. And Figure 5-41 shows 4 records of "open the door" in noisy environment from microphone 2 while the car radio is playing news. In this test, 100 records of "open the door" with news were made.

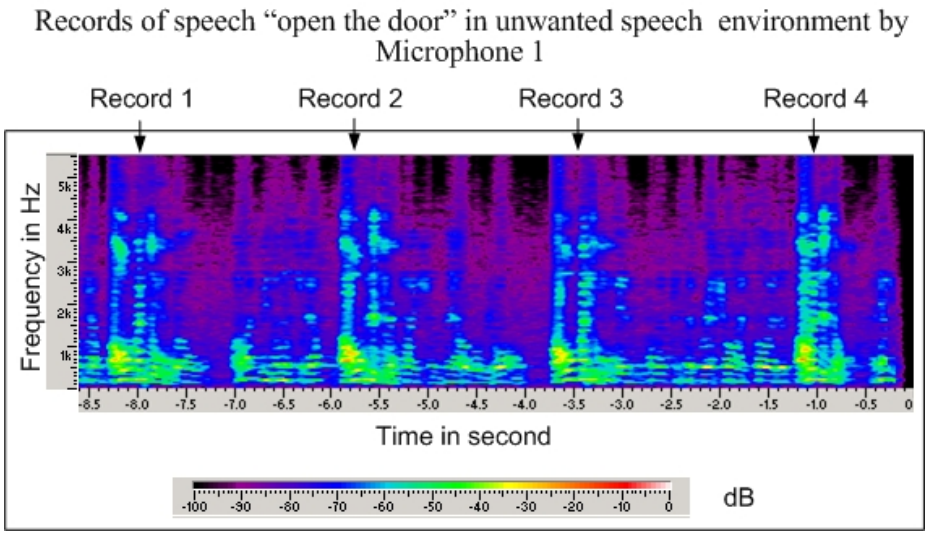

Figure 5-40 Records of speech "open the door" in unwanted speech environment by Microphone 1

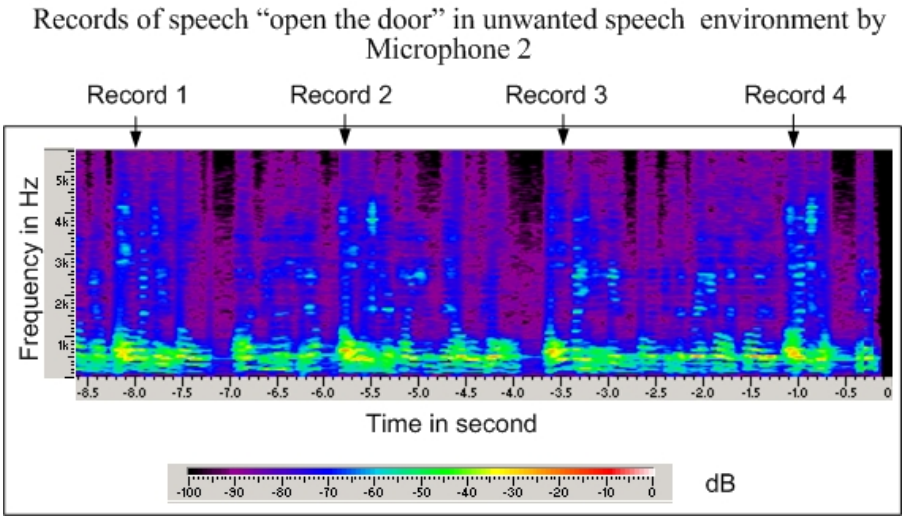

Figure 5-41 Records of speech "open the door" in unwanted speech environment by Microphone 2

5.3.3.2 The inputs to the speech recognition successful rate with an adaptive Wiener filter

In last test, at first 100 records of "open the door" without noise were made, at second 100 records of "open the door" with music were made, and at third 100 records of "open the door" with news were made. These second and third records were used to be filtered by adaptive Wiener filter as Figure 5-42.

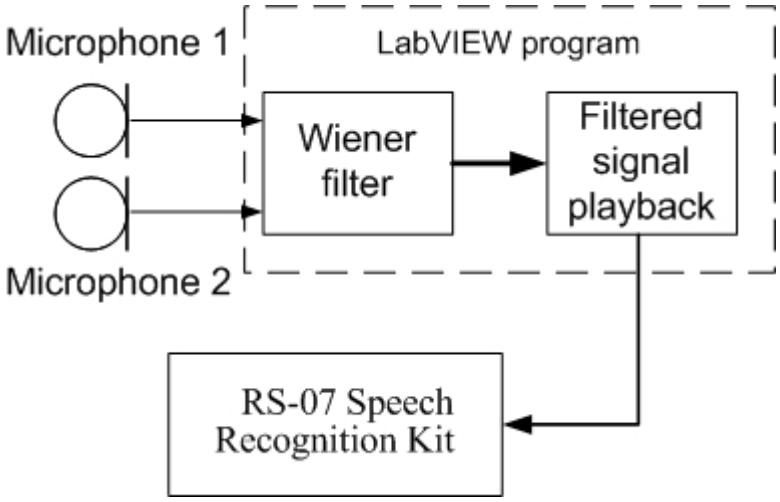

Figure 5-42 Test with AWF

Figure 5-43 shows the spectrograms of the waveforms of filtered signal by adaptive Wiener filter. As an example, there are 4 records of "open the door" with music in Figure 5-43. And there are 4 records of "open the door" with news in Figure 5-44.

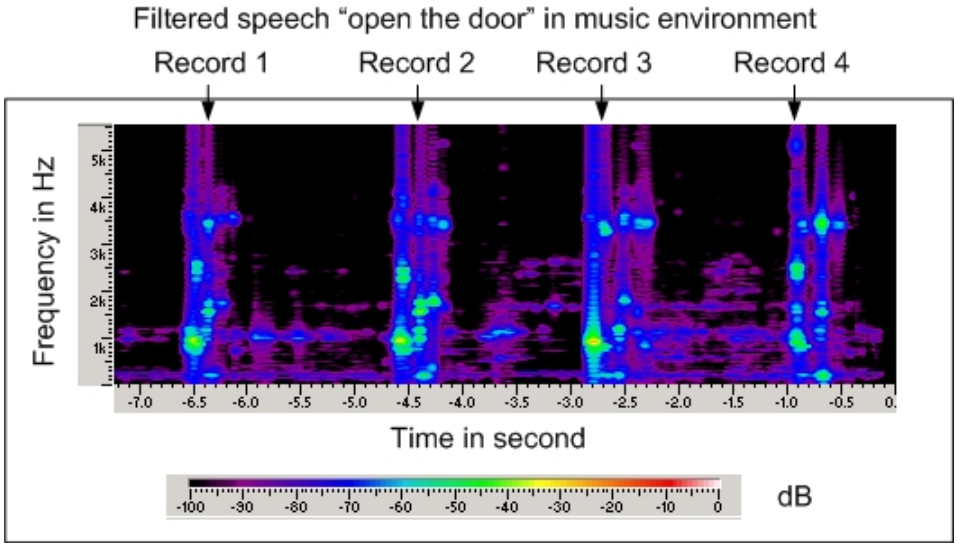

Figure 5-43 Filtered speech “open the door” in music environment

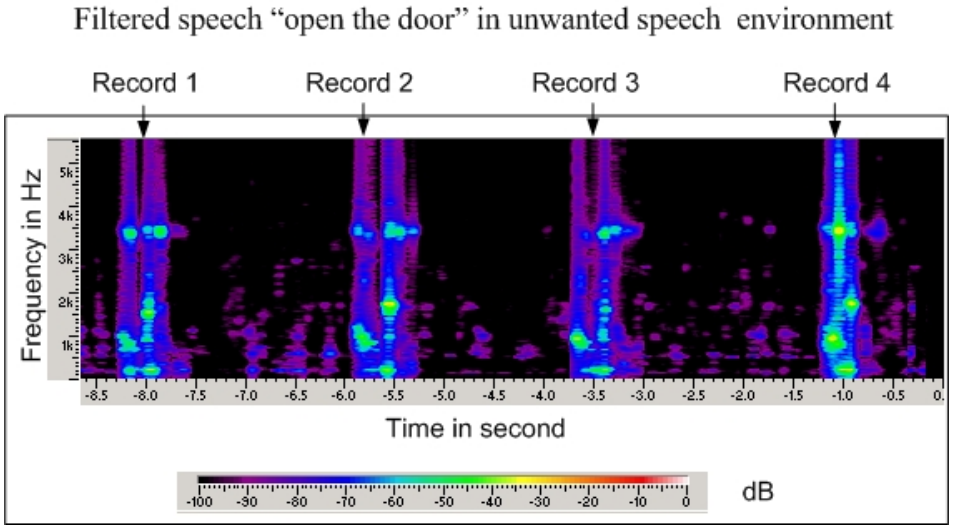

Figure 5-44 Filtered speech “open the door” in unwanted speech environment

5.3.3.3Results of ASR successful rate with an AWF or without an AWF

When 100 records of "open the door" with music, and 100 records of "open the door" with news were applied to Figure 5-36 and Figure 5-42, the ASR successful rate is showed in Table 5-6.

Table 5-6 Results of test with AWF or without AWF

Background noise	Successful rate as Figure 5-36	Successful rate as Figure 5-42
Speech	9.7%	73%
Music	3.5%	75%

5.3.4 Case study Four: Test in Template matching ASR in LabVIEW

Last chapter test the proposed Wiener filter using a commercial project ASR. However, if An ASR integrated with a Wiener filter, the successful-rate can be better. In the ASR, the finger-print training can be directly done using filtered signal from adaptive Wiener filter e.g. when the Wiener filter outputs a m-dimensional vector, such a m-dimensional vector can directly form the frame of a neural network in the ASR system.

An in-car test was performed with layout as in Figure 5-33. Microphone 1 is in the central area of the car, in such a position where it will in all probability pick up the radio more than the drivers voice. Microphone 2 is in the front of the driver, so it is in a position where it will more likely pick up the drivers voice than the radio signal. This process was performed at an 11025 Hz sample rate. The W matrix is 100x100 and is updated by the 100 samples from Microphone 1 and another 100 samples from Microphone 2. These samples from Microphone 2 are also filtered by the W matrix which forms the output as the filtered signal.

A LabVIEW program (see the discussion in chapter 4.2.4 for details) in a laptop is designed to create and update the Wiener matrix. The speech signal is filtered by this Wiener matrix and then fed to a cross-correlation algorithm for a measure of similarity with one of the fingerprints.

In order to test the repeatability and speaker-dependence, two groups of data are collected: a male driver's speech "left" and "right"; a female driver's speech "left" and "right". Whilst the driver is speaking, the radio in the car is simultaneously switched on to a radio channel. The voice from the car radio is different from time to time as is the case with a normal radio transmission.

Figure 5-45 shows a sample of spectrograms of the waveforms of a female driver's speech "right" and "left" when the in-car radio was on and the engine was idling. Figure 5-45 (c) shows the driver's speech "right" and "left" was clearly separated from that of the radio.

Figure 5-46 shows a sample of spectrograms of the waveforms of a male driver's speech saying "left" and "right" when in-car radio was on and the engine was idling. Figure 5-46 (c) shows the driver's speech "left" and "right" was clearly separated from that of the radio noise. There are 6 records of a Female driver's speech "right" and "left" and 6 records of a male driver's speech "left" and "right" in this test.

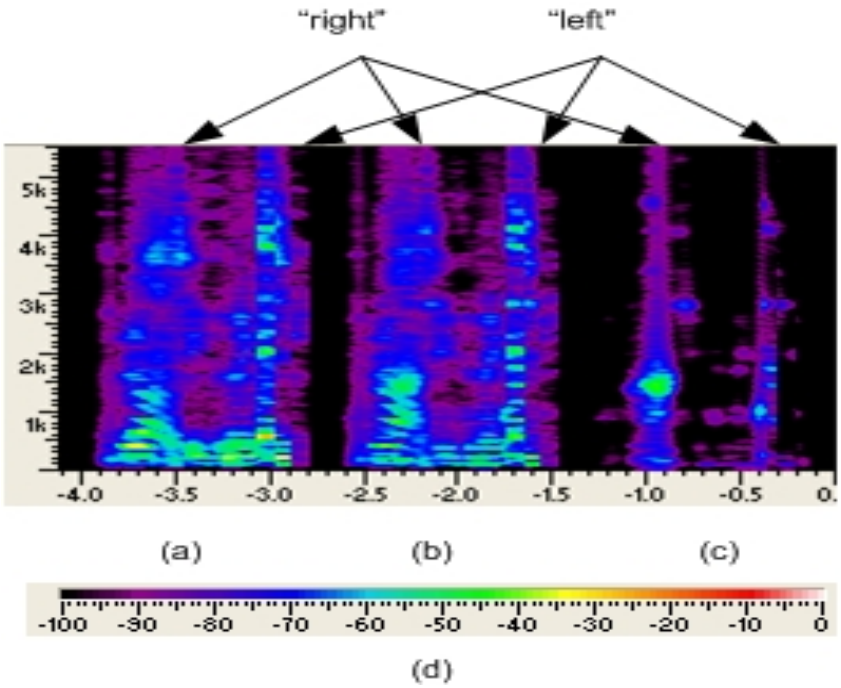

Figure 5-45 The spectrograms of the waveforms of a female driver's speech "right" and "left". The horizontal axis represents time (in second) and the vertical axis is frequency (in Hz). (a) Spectrogram at Microphone 1. (b) Spectrogram at Microphone 2. (c) Spectrogram of filtered signal. (d) Intensity scale in dB

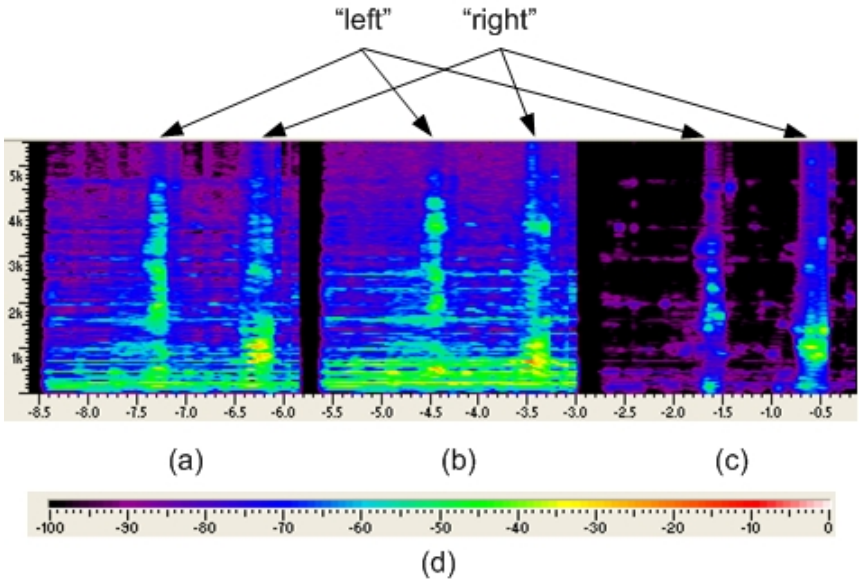

Figure 5-46 The spectrograms of the waveforms of a male driver's speech "left" and "right". The horizontal axis represents time (in second) and the vertical axis is frequency (in Hz). (a) Spectrogram at Microphone 1. (b) Spectrogram at Microphone 2. (c) Spectrogram of filtered signal. (d) Intensity scale in dB

More than 100 experimental records of speech saying “left” and “right” have been performed to calculate the ASR success rate. We present 5 of these results at random for clarity.

Firstly we used a recording to create a fingerprint of a female driver’s speech saying “right”. Then we collected 5 tests of the female driver’s speech “left” and 5 samples of the female driver’s speech “right”. Fingerprint “right” was cross-correlated with 5 records of “left” and 5 records of “right”. Then the average power was calculated. The results are shown in Table 5-7. If a threshold is set at the pre-set value (between -45.48dB and -43.41dB), then the incoming signal “right” will be recognized with a hit-rate of 100%.

Table 5-7 Average power in dB result of cross-correlation between fingerprint "right" and 5 records of “right” or 5 records of “left” from a female driver

	“left”	“right”
Sample 1	-49.38 dB	-39.52 dB
Sample 2	-47.82 dB	-40.68 dB
Sample 3	-45.48 dB	-41.49 dB
Sample 4	-51.10dB	-41.52 dB
Sample 5	-47.99 dB	-43.41 dB

We trained and created a fingerprint of a male driver’s speech “right”. Then we collected 5 records of the male driver’s speech “left” and 5 records of the male driver’s speech “right”. Fingerprint “left” was cross-correlated with 5 records of “left” and 5 records of “right”. Then the average power was calculated. The results are showed in Table 5-8. As shown in Table 5-8, if a threshold is set at the pre-set value (between -48.81dB and -56.89) then the incoming signal “right” will be recognized with a hit-rate of 100%.

Table 5-8 Average power in dB at result of cross-correlation between fingerprint "right" and 5 records of “right” or 5 records of “left” from a male driver

	“left”	“right”
Sample 1	-59.65 dB	-40.56 dB
Sample 2	-67.65 dB	-40.81 dB
Sample 3	-62.99 dB	-46.44 dB
Sample 4	-56.89 dB	-42.27 dB
Sample 5	-63.97 dB	-48.81 dB

We trained and created a fingerprint of a female driver’s speech “right”. Then we collected 5 records of a female driver’s speech “left” and 5 records of the female driver’s speech “right”. The results as shown in Table 5-9 show that there are no successes since there no threshold can be selected to recognize the female driver speech “right”.

Table 5-9 Average power in dB at result of cross-correlation between fingerprint "right" from a female driver and 5 records of “right” or 5 records of “left” from a male driver

	“left”	“right”
Sample 1	-53.04 dB	-51.85 dB
Sample 2	-63.75 dB	-56.23 dB
Sample 3	-47.84 dB	-55.78 dB
Sample 4	-57.14 dB	-55.65 dB
Sample 5	-67.72 dB	-53.73 dB

We trained and created a fingerprint of a male driver’s speech “right”. Then we collected 5 records of a female driver’s speech “left” and 5 records of the female driver’s speech “right”. The results as Table 5-10 showed that there are no successes since no threshold can be selected to recognize the male driver speech “right”.

Table 5-10 Average power in dB at result of cross-correlation between fingerprint "right" from a male driver and 5 records of “right” or 5 records of “left” from a female driver

	“left”	“right”
Sample 1	-60.40 dB	-55.00 dB
Sample 2	-45.75 dB	-48.05 dB
Sample 3	-56.23 dB	-53.85 dB
Sample 4	-57.10 dB	-50.95 dB
Sample 5	-49.04 dB	-42.64 dB

The above testes have been done for a limited number of records. However, as with the cases of Table 5-7 and Table 5-8, more than 100 records have been performed for each threshold with 100% success rate for these cases (as in Table 5-7 and Table 5-9). For the cases in Table 5-9 and Table 5-10, when we used the thresholds (as in Table 5-9 and Table 5-10), a success rate of 40-60% in 100 records was found.

In conclusion, these result shows a high hit-rate rate speaker-dependent ASR in a noisy environment.

5.3.5 Case study Five: The size of the Wiener filter matrix

This experiment is to discuss the best selection for the size of the Wiener filter matrix in a car in a real-time environment.

A real-time hybrid adaptive filter operates within a geometrical zone defined around the head of the desired speaker. From Figure 5-47, Microphone 1, 2 and 3 are applied to a 3-microphone geometrical VAD. These 3 microphones are employed in an adaptive Wiener filter which is switched on for noise reduction.

Any sound outside of this zone is considered to be noise and suppressed e.g. the incoming speech from loud-speakers driven by radio (1, 4, 5 and 8 in Figure 5-47), as well as passengers (2, 3 and 7 in Figure 5-47). The technique uses three microphones to define a geometric based voice-activity detector (VAD) to cancel the unwanted speech coming from outside of the zone. When unwanted speech and desired speech are incoming at the same time, the VAD falls to identify the unwanted speech and desired speech and so an adaptive Wiener filter is switched on for noise reduction.

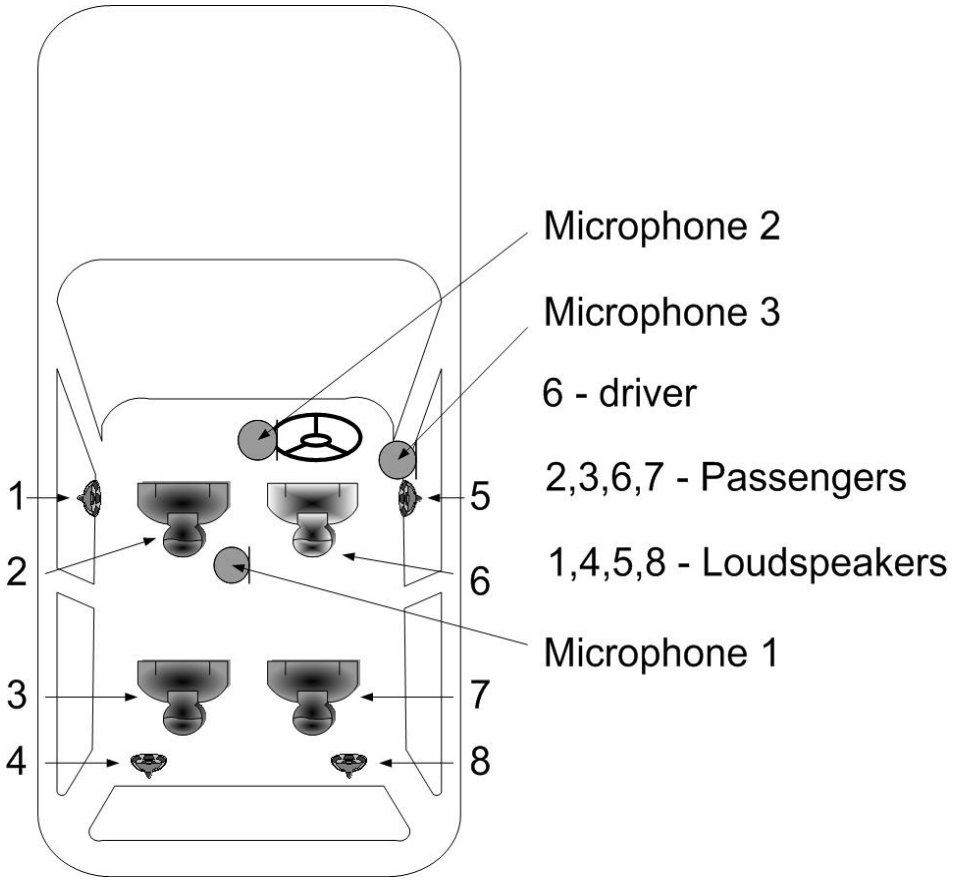

Figure 5-47 Design layout of hybrid noise canceller

Using an *M square* **W** matrix in the Wiener filter, the samples of input speech are filtered and outputted as a *1D* array with size *M*. This *1D* array is considered as a frame, which would be used in a typical neural network.

For the **W** matrix, the larger *M* means more computing resource cost. In the meantime, the Frame size in a neural network is considered under the accuracy of speech recognition. Normally a smaller frame size means larger error.

Figure 5-48 is only a group of data from many repeatable data groups. It is in grey scale intensity.

Figure 5-48 shows the spectral of Wiener filter results at different size of *M,* the dimension of the W matrix. The horizontal axis represents time and the vertical axis is frequency. All experiments in this paper use an English phase "Open the door" as desired speech. The unwanted speech is recorded from a Chinese radio station. Figure 5-48 (a) shows a result for an English phase "Open the door" with background noise (unwanted speech) which is filtered using a 10 X 10 W matrix. This matrix is updated by data from Microphones 1 and 2. Figure 5-48(b) uses a 50 X 50 W matrix, Figure 5-48(c) uses a 100 X 100 W matrix, Figure 5-48(d) uses a 200 X 200 W matrix, Figure 5-48(e) uses a 500 X 500 W matrix, Figure 5-48 (f) uses 1000 X 1000 W matrix and Figure 5-48(g) uses a 3000 X 3000 W matrix. Figure 5-48(i) shows unwanted speech plus desired speech at Microphone 1. Figure 5-48(j) shows unwanted speech plus desired speech at Microphone 2.

From Figure 5-48, Signal to Noise Ratio (SNR) is calculated using CoolEdit (see Appendix for details) and the results are showed in Table 6-12. The SNR is computed as follows.

$$\text{SNR}=\text{Average Signal Power (dB)} - \text{Average Noise Power (dB)} \tag{5.6}$$

In this experiment the Signal is known exactly for measurement purposes. Of course in an automobile environment the signal will be corrupted with noise and not be directly measurable.

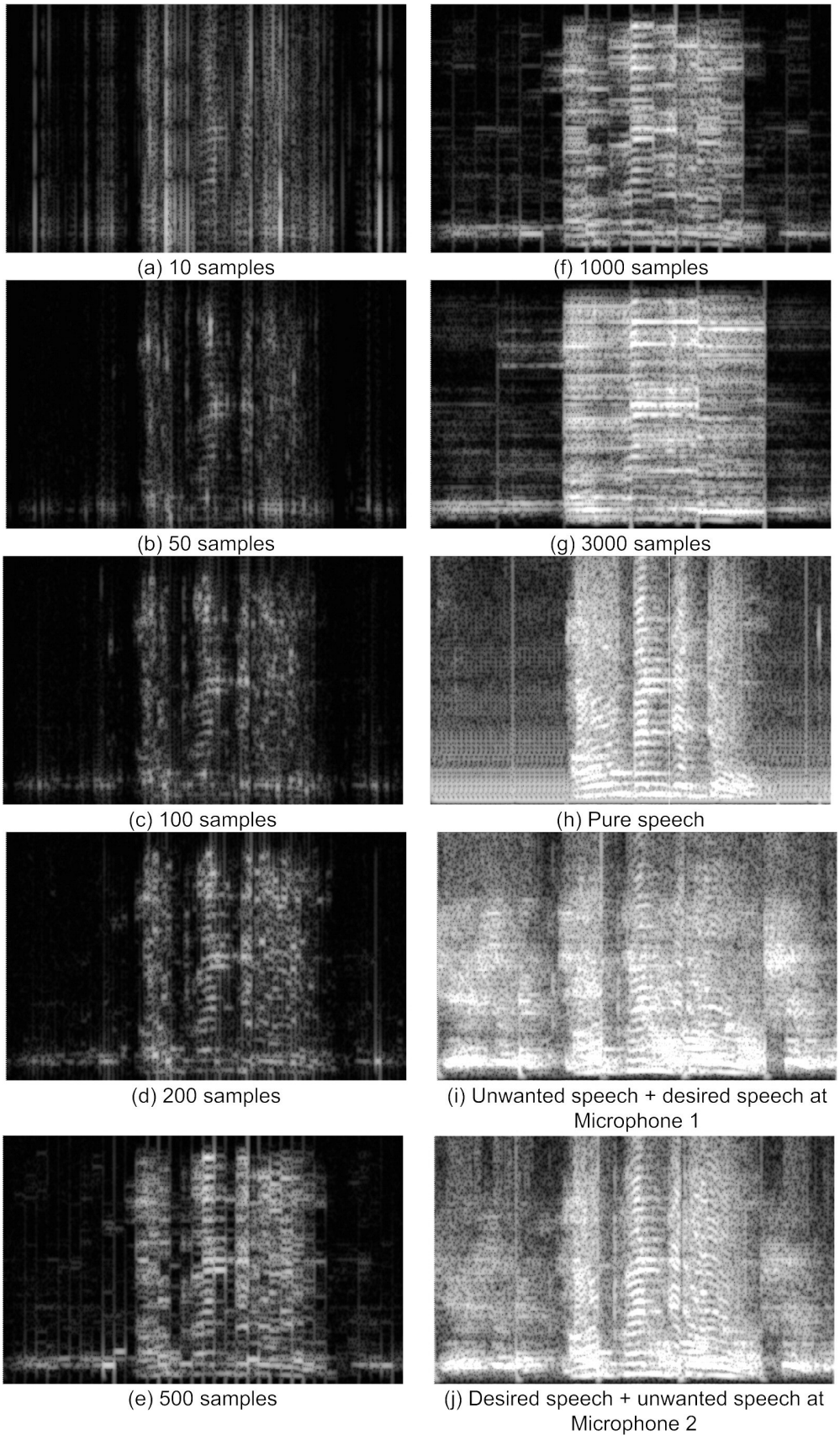

Figure 5-48 Comparing of Frame size at spectrum of filtered speech in grey scale intensity

Table 5-12 shows the Average Power for a W matrix of dimension 10, 50, 100, 200, 500, 1000 and3000 respectively. This was performed at an 11025 Hz sample rate. We are to find a best frame for speech recognition and as well as a best SNR for signal enhancement. From Figure 5-48, it is clear that a 100-sample frame is more representative of the original speech since it resembles the spectrogram of the clean speech more closely than do the others. From Table 5-12, we have high SNR at a 100-sample frame (9.07 ms).

Table 5-11 Average Power for different orders of W matrix

Order of W matrix	Speech Average Power at filter output during periods of desired speech (dB)	Speech Average Power at filter output during periods of only unwanted speech (dB)	Average Power during periods of desired speech incoming to Microphone 1 (dB)	Average Power during periods of only unwanted speech incoming to Microphone 1 (dB)
10	-56.3	-74.4	-21.56	-26.48
50	-64.5	-85.3	-21.56	-26.48
100	-59.3	-82.2	-21.56	-26.48
200	-53.7	-79.1	-21.56	-26.48
500	-46.1	-68.3	-21.56	-26.48
1000	-39.3	-62.5	-21.56	-26.48
3000	-28.3	-53.3	-21.56	-26.48

From Figure 5-48(i), we have a SNR between pre-filtered desired speech and unwanted speech. From Figure 5-48(a)-(g), we have SNRs between pre-filtered desired speech and unwanted speech. And improved SNR is calculated by the differential SNR between pre-filtered and filtered signals:

$$SNR_{improved} = SNR_{filtered} - SNR_{pre-filtered} \tag{5.7}$$

Table 5-13 shows SNRs for a W matrix of dimension 10, 50, 100, 200, 500, 1000 and 3000 samples obtained at an 11025 Hz sample rate. From Table 6-13, when the frame is 9.07 ms (M=100 samples at an 11025 Hz sample rate), the SNR is improved by up to 18.83dB.

Table 5-12 SNRs for different W matrix dimension obtained at an 11025 Hz sample rate

Order of W matrix	$SNR_{filtered}$ dB	$SNR_{pre-filtered}$ dB	$SNR_{improved}$ dB
10	18.1	3.92	14.18
50	20.8	3.92	16.88
100	22.9	3.92	18.98
200	25.4	3.92	21.48
500	22.2	3.92	18.28
1000	23.2	3.92	19.28
3000	25	3.92	21.08

Table 5-12 is to lead to a conclusion from Figure 5-48 that 100 x 100 W matrix has best balance between frame length and also clear filtered output. And Table 5-13 is to identify if the 100 x 100 W matrix has acceptable SNR.

5.3.6 Case study Six: Wiener filter in unwanted speech from different locations

This experiment is to discuss to the best location for the microphone which represents the unwanted speech in a car in a real-time environment.

This case study is based on the geometry of Figure 5-48. The spectrograms of filtered speech are collected whilst driver's speech is incoming with simultaneous unwanted speech from radio loud-speakers or a passenger.

Figure 5-49 (in grey scale intensity) shows the test results of the driver's speech with unwanted speech from different locations. The horizontal axis represents time and the vertical axis is frequency. For Figure 5-49 (a) and Figure 5-49 (b), two front loud-speakers are presented at the same time. For Figure 5-49 (c) and Figure 5-49 (d), two rear loud-speakers are presented at the same time. Figure 5-49 is only a group of data from many repeatable data groups.

In Figure 5-49,

(a) the spectrogram of the filtered speech is shown. Microphone 1 (as reference) and Microphone 2 (as speech) as input to the Wiener Filter whilst noise comes from front loud-speakers(#1 and 5 in Figure 5-49).

(b) is the spectrogram of filtered speech using Microphone 3 (as a reference) and Microphone 2 (as speech) as input to the Wiener Filter whilst noise comes from the front loud-speakers(#1 and 5 in Figure 5-49).

(c) is the spectrogram of filtered speech using Microphone 1 (as a reference) and Microphone 2 (as speech) as input to the Wiener Filter whilst noise comes from rear loud-speakers(#4 and 7 in Figure 5-49).

(d) shows the spectrogram of filtered speech using Microphone 3 (as a reference) and Microphone 2 (as speech) as input to the Wiener Filter whilst noise comes from rear loud-speakers (#4 and 7 in Figure 5-49).

(e) shows the spectrogram of filtered speech using Microphone 1 (as a reference) and Microphone 2 (as speech) as input to the Wiener Filter whilst noise comes from #2 passenger as in Figure 5-49.

(f) shows the spectrogram of filtered speech using Microphone 3 (as reference) and Microphone 2 (as speech) as input to the Wiener Filter whilst noise comes from #2 passenger as in Figure 5-49.

(g) shows the spectrogram of filtered speech using Microphone 1 (as a reference) and Microphone 2 (as speech) as input to the Wiener Filter whilst noise comes from #3 passenger as in Figure 5-49.

(h) shows the spectrogram of filtered speech using Microphone 3 (as a reference) and Microphone 2 (as speech)as input to the Wiener Filter whilst noise from #3 passenger as in Figure 5-49.

(i) shows the spectrogram of filtered speech using Microphone 1 (as a reference) and Microphone 2 (as speech) as input to the Wiener Filter whilst noise from #7 passenger as in Figure 5-49.

(j) shows the spectrogram of filtered speech using Microphone 3 (as a reference) and Microphone 2 (as speech) as input to the Wiener Filter whilst noise from #7 passenger as in Figure 5-49.

Table 6-13 shows the Average RMS Power as per the spectrograms of Figure 5-49. Signal to Noise Ratio (SNR) is calculated and the results are showed in Table 6-14. From Table 6-14, it is clear that we only need microphone 1 and 2 to apply a Wiener filter to separate the desired speech from unwanted simultaneous speech of the radio and passengers. Table 6-14 is to extend our understanding from Figure 5-49 that if there is clear filtered output when voices are incoming from different location in a car.

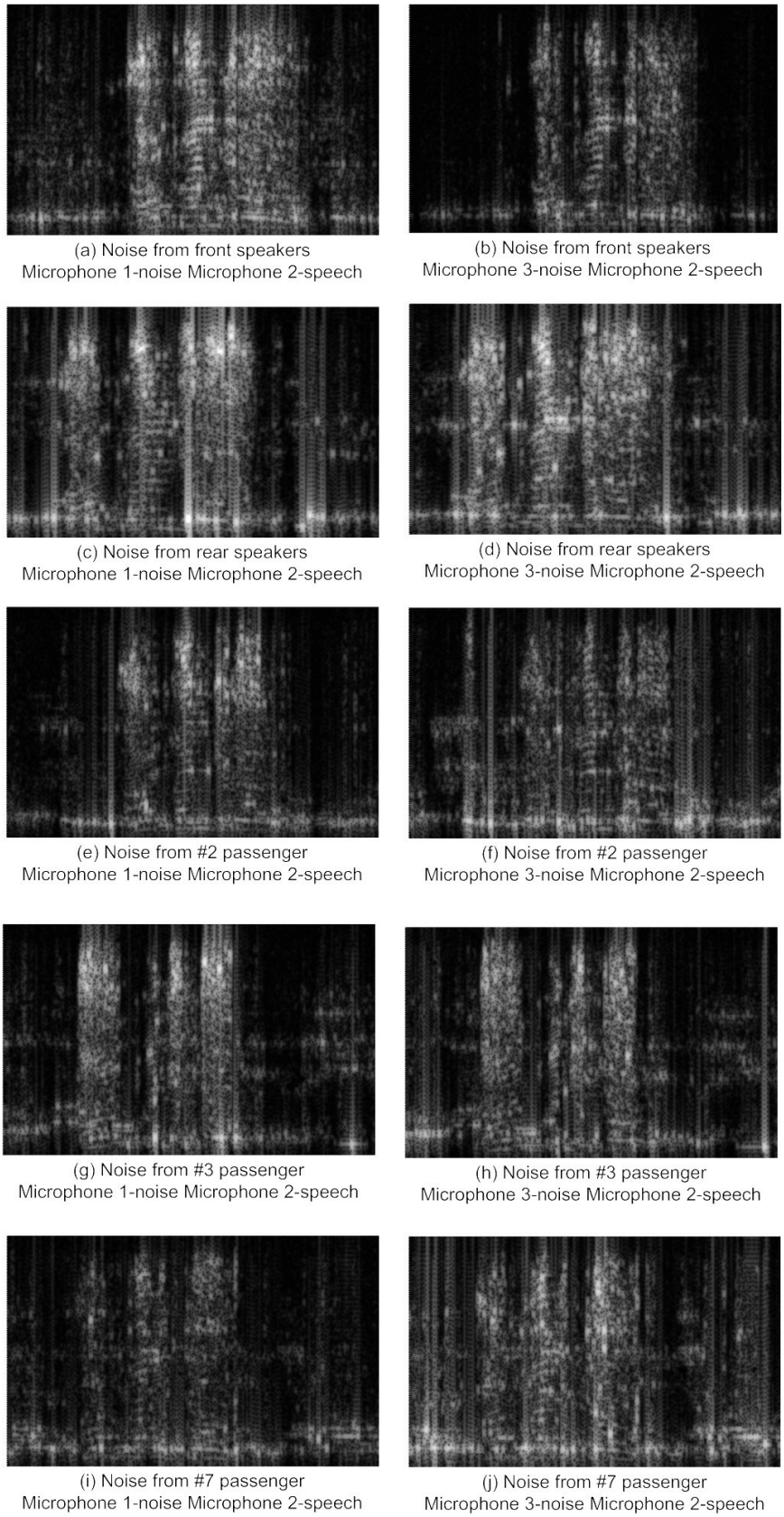

Figure 5-49 The spectrograms of filtered speech whilst driver's speech is incoming with simultaneous unwanted speech from radio loud-speakers or a passenger, in grey scale intensity

Table 5-13 Average Signal Power Samples with reference to the seating positions referred to in Figure 5-49

Spectrogram	Filtered Desired Speech Average Power (dB)	Filtered noise (Unwanted Speech) Average Power (dB)	Average Power at period of desired speech incoming to Microphone 1 (dB)	Average Power at period of only unwanted speech incoming to Microphone 1 (dB)
a	-49.46	-65.38	-21.56	-26.48
b	-59.32	-82.36	-21.56	-26.48
c	-41.19	-52.93	-21.56	-26.48
d	-45.36	-53.51	-21.56	-26.48
e	-57.69	-77.27	-21.56	-26.48
f	-55.34	-72.27	-21.56	-26.48
g	-53.39	-71.60	-21.56	-26.48
h	-58.04	-62.57	-21.56	-26.48
i	-63.15	-68.60	-21.56	-26.48
j	-52.28	-56.82	-21.56	-26.48

Table 5-14 Improved SNR with reference to Figure 5-49

Spectrograms	$SNR_{filtered}$ dB	$SNR_{pre-filtered}$ dB	$SNR_{improved}$ dB
a	15.92	3.92	12
b	23.04	3.92	19.12
c	11.74	3.92	7.82
d	8.15	3.92	4.23
e	19.58	3.92	15.66
f	16.93	3.92	13.01
g	18.21	3.92	14.29
h	4.53	3.92	0.61
i	5.45	3.92	1.53
j	4.54	3.92	0.62

As shown in Table 5-14, in order to simplify the design, we can only employ Microphone 1 and 2 for the Wiener filter inputs. Therefore, we have the results as in Table 5-14 a, c, e, g and i. The SNR is 12dB, 7.82dB, 15.66dB, 14.29dB or 1.53dB respectively. As all experiments in this paper use an English phase “Open the door” as desired speech and the unwanted speech is recorded from a Chinese radio station. The SNR is calculated by the period of this phase and the period of unwanted speech. The reason why c and i have low SNRs is that the unwanted speech is far away from Microphone 1, which acts as noise reference.

A hybrid real-time adaptive filter is built to operate within a geometrical zone defined around the head of the desired speaker in an automobile environment. Any sound outside of this zone is considered to be noise and suppressed. When speech alone is incoming from

outside of a desired zone, the technique uses three microphones to define a geometric based voice-activity detector (VAD) to mute the unwanted speech. When unwanted speech and desired speech are incoming simultaneously, the VAD falls to identify the unwanted speech or desired speech, so then an adaptive Wiener filter is switched on for noise reduction. The experiments show that a Wiener filter can be updated in the defined period, a frame, which is ready for speech recognition.

In the case of desired and unwanted speech incoming simultaneously, when the frame is 9.07 ms (M=100 samples at an11025 Hz sample rate), the SNR is improved by up to 18.9dB.

5.4 Summary

Experiments have been conducted in real-time on a combined three-microphone VAD and noise-cancelling system. The VAD assumes that the desired speech falls within a desired geometric zone (the active zone) which is most appropriate for an automobile environment. The noise-cancelling is only required when noise is present during desired speech as the VAD will mute any solo noise-source outside of the zone. Although the experiment only uses pre-recorded phrases, this work clearly demonstrates the ability of the algorithm to cancel speech outside of the active zone. In the Chapter 5.2.3, a desired voice incoming from a desired zone without any interference/noise from outside of the desired zone, and a NLMS ANC is applied, we recorded 6 dB SNR and 93% hit-rate at RS-07 speech recognition kit. When a desired voice incoming from a desired zone, and an interference/noise from outside of the desired zone, Figure 5-50 shows the experimental results. Both SNR and Signal quality are improved whilst adaptive Wiener filter is applied.

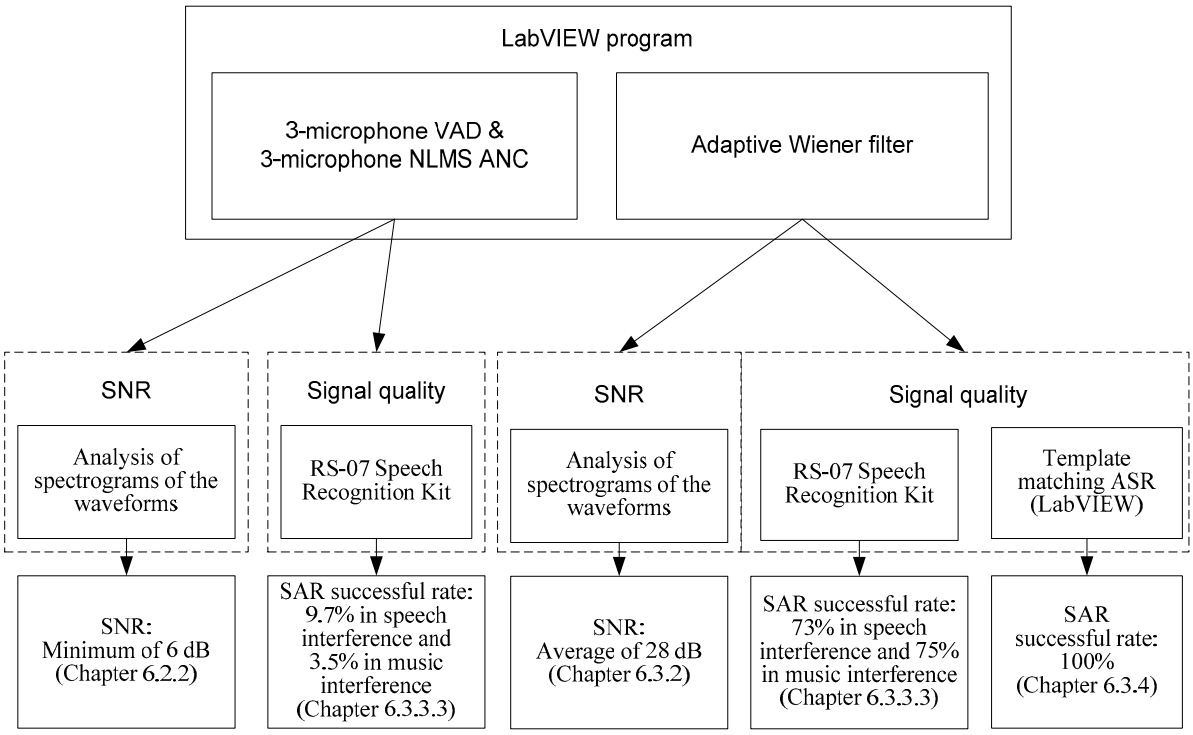

Figure 5-50 Summary of experiment

Previous research suggested that a Wiener filter W matrix can be updated by knowledge of noise periods and signal + noise periods. The W matrix always acts as a filter on the noisy speech command signal. Therefore a VAD is needed to distinguish between these noise periods or signal + noise periods. This thesis uses a method which eliminates entirely the need of a VAD switch using a continuously adapting Wiener filter. The experiments show

that a Wiener filter can be updated during noise periods and signal + noise periods, improving the SNR by slightly better than 28 dB. As an experiment, this thesis introduces a novel design in real-time environments for recognizing small vocabulary control commands. In order to receive the driver's voice whilst suppressing background noise e.g. voice from a radio, a simple and novel speech recognition system is designed using LabVIEW. Experiments show this voice recognizer has high accuracy and is capability of speaker-dependence in tests in a car when the car radio was on and the engine was idling. The filtered signals were clearly recognizable during informal listening tests and sounded vastly improved to the original. However, as with all of these kinds of automated 'smart-car' technologies, (in contrast with some other similar technologies) it is not necessary for the enhanced signal to sound better to the human ear, but only needs to be good enough to provide a Boolean on or off command.

In addition, chapter 5.2.4 "Comparison of NLMF and NLMS ANC" shows that NLMS has advantage of NLMF in noise cancellation as NLMS is much more stable and ability of SNR.

In chapter 5.3.5, an important experiment for the size of Wiener filter matrix in noise cancellation has an important engineering discover because we can decide to apply a 100x100 matrix to reduce the calculation time to enable the possibility of running in real-time.

In chapter 5.3.6, a result of experiment suggests an engineering consideration for the best location of microphones. It confirms the test plan as Figure 3-7 as a proposed engineering solution.

Finally, the experiment has also confirmed the adaptive Wiener filter has much more advantage of NLMS filter in 3-microphne noise cancellation as the SNR of adaptive Wiener filter has 28dB (the details can be found at chapter 5.3.2 Case study two: Analysis of average power in the spectrograms) and the NLMS filter only has 6 dB (the details can be found at chapter 5.2.1 Three-microphone VAD in a car).

6 Conclusions and Future Work

6.1 The improvements on the other researches

This thesis is based on the works as:

1. Chen and Moir's 3-microphone VAD (Chen and Moir, 1999).
2. Wiener-Hopf equation derived by Doclo & Moonen (Doclo & Moonen, 2002).

Chen and Moir derived their 3-microphone VAD and they built a test plan in lab using 3 microphones in 1999. The sampling frequency was 25.6 kHz and the separation of each pair of microphones is 20 cm. Some simulation experiments had been conducted to present the performance of a word boundary detection algorithm using three microphones. All the investigations showed the accuracy of the word boundary detecting percentage success rate to be more than 80%.

Using the 3-microphne VAD, this thesis contributes a method of smoothly updating the spectrum recursively at each FFT frame. The sampling frequency in this thesis is 11025Hz and the separation of each pair of microphones is 50 cm. This thesis also presents experiments in a car in a real-time environment. The 3-microphone VAD approach in this thesis shows that it is relatively simple and not too much of a computational overload for real-time system.

This thesis applied the Wiener-Hopf equation derived by Doclo & Moonen (Doclo & Moonen, 2002). Doclo & Moonen had two assumptions: short-term stationarity of the noise, and statistical independence of the speech and noise signals. Normally the NLMS approach is the technique used in updating this adaptive Wiener filter. Whereas the NLMS approach uses weight estimation to minimize a mean-square error, therefore, it is not suitable the applications of this solution in non-stationary noise environment. This thesis contributes a method of updating the **W** matrix in real-time. This alternative approach constructs the Wiener filter from estimation of covariance matrices.

6.2 Major findings and contributions

This thesis presents a hybrid system in a car which contains a novel 3-microphone beamformer with a 3-microphone VAD and a 3-microphone NLMS ANC, and a novel adaptive Wiener filter.

When a desired voice incoming from a desired zone without any interference/noise from outside of the desired zone, the experiments in chapter 5 confirmed the capability of the 3-

microphone beamforming speech enhancement in a car: it has minimum of 6dB SNR and 93% ASR successful rate at RS-07 speech recognition kit.

When a desired voice incoming from a desired zone with an interference/noise from outside of the desired zone, the experiments verify that the adaptive Wiener filter in a car is not only improving SNR by 28dB, but also able to be recognized by an ASR in high hit-rate. Using RS-07 speech recognition kit with improving SNR by 28dB, the hit-rate can be 75%.

A LabVIEW program is designed using template matching ASR and an adaptive Wiener filter. In result, it improves the SNR by 28dB, the hit-rate can be 100% (with proper selection of fingerprint).

6.3 Suggestions for Future Work

As engineering applications are a recommendation of future work, further investigates are including

- As an ordinary Wiener filtering normally requires matrix inversion method, an innovative matrix inversion method is the highest priority to ensure a low computational workload in real-time applications. Moir (Moir, 2008) proposed a new method for finding the inverse of the covariance (or correlation) matrix. The technique is based on a previous method known as automatic variance control. The technique naturally gives rise to the inverse due to the properties of negative feedback itself rather than by traditional inversion methods. The method can be applied to the non-stationary problem and can form part of a recursive parameter estimation scheme.
- Since a Wiener filter can be updated by refreshing the auto-correlation matrix of the input signal during speech plus additive noise periods and refreshing an auto-correlation matrix of the input signal during noise-alone periods, alterative update method for Wiener filtering is applying a Multivariable Smoothing to update the auto-correlation matrix of the input signal then updating the Wiener filter.
- An open-source ASR is needed to ensure the finger-print training can be directly done using filtered signal from adaptive Wiener filter e.g. when the Wiener filter outputs a m-dimensional vector, such a m-dimensional vector can directly form the frame of a neural network in the ASR system. A FPGA can be a good solution to integrate an ASR and an adaptive Wiener filter.

Further to these exercise, there remain many applications to investigate including e.g.

- A voice controlled switch for Smart house project. A novel design of noise cancellation is needed in Smart-House interface which is using "Vanessa" (from Guile 3D)(Moir & Lindroth II, 2008). Vanessa is a Microsoft Agent which is programmed to switch lights on/off, TV on/off, Curtain open/close, fan on/off etc in an environment of a house. An adaptive Wiener filter based switch with 2 microphones can be employed to mute the TV to ensure speech recognition system work in clean environment to improve the successful rate. This design can use an individual built-in template matching ASR. A special command "TV mutes" can be recognized when the TV is on. When the special command is recognize, the system will send IR remote control code "mute" to the TV.
- A voice-controlled wheelchair. A wheelchair is normally controlled by hands. However, a voice controlled wheelchair has advantages. An US patent “Wheelchair mounted control apparatus” was logged in 1980: “A voice actuated wheelchair control apparatus is disclosed which allows a quadriplegic to control the speed and direction of travel of a motorized wheelchair by spoken commands. The apparatus includes means for executing spoken commands for initiating incremental changes in the chair's speed or direction of travel. The apparatus also includes means for executing spoken commands which modify or override the wheelchair motion initiated by previously spoken commands.”(Myron Youdin, Mario W. Clagnaz, Jr., & Louie, 1980) Since this invention was issued, many investigate and designs upon the voice controlled wheel chair had been done. However, few products had been used as they are not really user-friendly e.g. they require the users to wear a headset or speaking nearby a microphone. Otherwise there is a very low successful-rate at speech recognition since the wheelchairs normally are moving in very noisy environment. Therefore, a built-in Wiener filter with multi-microphone can be a considerable solution.

7 Reference

Agaiby, H., & Moir, T. J. (1997). *A robust word boundary detection algorithm with application to speech recognition.* Paper presented at the Digital Signal Processing Proceedings, 1997. DSP 97., 1997 13th International Conference on.

Alexandre, P., Boudy, J., & Lockwood, P. (1993). *Root homomorphic deconvolution schemes for speech processing in car noise environments.* Paper presented at the Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on.

Ban, Y., Banno, H., Takeda, K., & Itakura, F. A. I. F. (2002). *Synthesis of car noise based on a composition of engine noise and friction noise.* Paper presented at the Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on.

Barrault, G., Costa, M. H., Bermudez, J. C. M., & Lenzi, A. (2005). *A new analytical model for the NLMS algorithm.* Paper presented at the Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on.

Barrett, J. F., & Moir, T. J. (1987). A Unified Approach to Multivariable Discrete-Time Filtering Based on the Wiener-Theory. *Kybernetika 23*, 177-197.

Beamforming. (2008). Wikipedia, The Free Encyclopedia Retrieved 26 Feb, 2008, from http://en.wikipedia.org/w/index.php?title=Beamforming&oldid=191416312

Bellanger, M. (2001). *Adaptive digital filters*. New York: Marcel Dekker.

Brown, R. G., & Hwang, P. Y. C. (1996). *Introduction to Random Signals and Applied Kalman Filtering* (3 ed.). New York: John Wiley & Sons.

Cao, Y., Sridharan, S., & Moody, M. (1997). *Speech separation by simulating the cocktail party effect with a neural network controlled Wiener filter.* Paper presented at the Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on.

Capon, J. (1969). High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE, 57*(8), 1408-1418.

Carter, G., Knapp, C., & Nuttall, A. (1973). Estimation of the magnitude-squared coherence function via overlapped fast Fourier transform processing. *Audio and Electroacoustics, IEEE Transactions on, 21*(4), 337-344.

Chan, M. K., Zerguine, A., & Cowan, C. F. N. (2003). *An optimised normalised LMF algorithm for sub-Gaussian noise.* Paper presented at the Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on.

Chen, A., Vaseghi, S., & McCourt, P. (2000). *State based sub-band LP Wiener filters for speech enhancement in car environments.* Paper presented at the Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on.

Chen, W., N, & Moir, T. J. (1999). Active wood boundary detection using three microphones. *Signal Processing Systems, SiPS 99. 1999 IEEE Workshop*, Page(s):615 - 624.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of Acoustical Society of America, 25*(5), 975--979.

Cho, Y., & Ko, H. (2004). *Speech enhancement for robust speech recognition in car environments using Griffiths-Jim ANC based on two-paired microphones.* Paper presented at the Consumer Electronics, 2004 IEEE International Symposium on.

Cohen, I. (2003). Analysis of two-channel generalized sidelobe canceller (GSC) with post-filtering. *Speech and Audio Processing, IEEE Transactions on, 11*(6), 684-699.

Cohen, I., & Berdugo, B. (2003). *Two-channel signal detection and speech enhancement based on the transient beam-to-reference ratio.* Paper presented at the Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on.

Cornelius, P., Yermeche, Z., Grbic, N., & Claesson, I. A. C. I. (2004). *A spatially constrained subband beamforming algorithm for enhancement.* Paper presented at the Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2004.

Dabeer, O., & Masry, E. (2002). Analysis of mean-square error and transient speed of the LMS adaptive algorithm. *Information Theory, IEEE Transactions on, 48*(7), 1873-1894.

Dahl, M., Claesson, I., & Nordebo, S. (1997). *Simultaneous echo cancellation and car noise suppression employing a microphone array.* Paper presented at the Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on.

Dam, H. Q., Nordholm, S., Dam, H. H., & Low, S. Y. A. L. S. Y. (2006). *Post-Filtering Techniques For Directive Non-Stationary Source Combined With Stationary Noise Utilizing Spatial Spectral Processing.* Paper presented at the Circuits and Systems, 2006. APCCAS 2006. IEEE Asia Pacific Conference on.

Davis, A., Siow Yong, L., Nordholm, S., & Grbic, N. A. G. N. (2005). *A subband space constrained beamformer incorporating voice activity detection [speech enhancement applications]*. Paper presented at the Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on.

de Haan, J. M., Grbic, N., Claesson, L., & Nordholm, S. A. N. S. (2002). *Design and evaluation of nonuniform DFT filter banks in subband microphone arrays.* Paper presented at the Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on.

Ding, P., He, L., Yan, X., Zhao, R. A. Z. R., & Hao, J. A. H. J. (2006). *Robust Technologies towards Automatic Speech Recognition in Car Noise Environments.* Paper presented at the Signal Processing, The 8th International Conference on.

Diniz, P. S. R. (2002). *Digital signal processing: system analysis and design*: Cambridge University Press.

Dmochowski, J., & Goubran, R. (2004). *Noise cancellation using fixed beamforming.* Paper presented at the Haptic, Audio and Visual Environments and Their Applications, 2004. HAVE 2004. Proceedings. The 3rd IEEE International Workshop on.

Dmochowski, J., & Goubran, R. (2005). *Combined Beamforming and Noise Cancellation.* Paper presented at the Instrumentation and Measurement Technology Conference, 2005. IMTC 2005. Proceedings of the IEEE.

Dmochowski, J. P., & Goubran, R. A. (2007). Decoupled Beamforming and Noise Cancellation. *Instrumentation and Measurement, IEEE Transactions on, 56*(1), 80-88.

Doclo, S., & Moonen, M. (2002). GSVD-based optimal filtering for single and multimicrophone speech enhancement. *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on], 50*(9), 2230-2244.

Doclo, S., & Moonen, M. (2005). On the output SNR of the speech-distortion weighted multichannel Wiener filter. *Signal Processing Letters, IEEE, 12*(12), 809-811.

Erzin, E., Cetin, A. E., & Yardimci, Y. (1995). *Subband analysis for robust speech recognition in the presence of car noise.* Paper presented at the Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on.

Ferrara, E., Jr., & Widrow, B. (1981). Multichannel adaptive filtering for signal enhancement. *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on, 29*(3), 766-770.

Finite impulse response. (2008). Wikipedia, The Free Encyclopedia Retrieved 23 Feb, 2008, from http://en.wikipedia.org/w/index.php?title=Finite_impulse_response&oldid=190969975

Frost, O. L., III. (1972). An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE, 60*(8), 926-935.

Fuchs, M., Haulick, T., & Schmidt, G. (2004). *Noise suppression for automotive application based on directional information.* Paper presented at the icassp2004.

Grbic, N., Nordholm, S., & Johansson, A. (2001). *Speech enhancement for hands-free terminals.* Paper presented at the Image and Signal Processing and Analysis, 2001. ISPA 2001. Proceedings of the 2nd International Symposium on.

Griffiths, L. (1978). *An adaptive lattice structure for noise-cancelling applications.* Paper presented at the Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '78.

Hai Huyen, D., Nordhohm, S. E., Dam, H. H., & Siow Yong Low, A. S. Y. L. (2004). *Speech enhancement using the microphone array and short-term spectral amplitude estimator.* Paper presented at the Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on.

Hai Quang, D., Nordholm, S., Hai Huyen, D., & Siow Yong Low, A. S. Y. L. (2005). *Adaptive beamformer for hands-free communication system in noisy environments.* Paper presented at the Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on.

Hai Quang, D., Siow Yong, L., Hai Huyen, D., & Nordholm, S. A. N. S. (2004). *Space constrained beamforming with source PSD updates.* Paper presented at the Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on.

Haykin, S. (1996). *Adaptive filter theory* (Third ed.). Upper Saddle River, NJ.: Prentice-Hall, Inc.

Haykin, S. (2002). *Adaptive Filter Theory* (4 ed.): Prentice Hall.

Haykin, S., & Chen, Z. (2006). The machine cocktail party problem. In S. Haykin (Ed.), *New Directions in Statistical Signal Processing: From Systems to Brains (Neural Information Processing)*. Cambridge, MA: The MIT Press.

Hodgkiss, W. (1980). *Dynamic beamforming of a random acoustic array.* Paper presented at the Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '80.

Ifeachor, E. C. (1993). *Digital signal processing A practical approach*: Addison-Wesley.

Infinite impulse response. (2008). Wikipedia, The Free Encyclopedia Retrieved 23 Feb, 2008, from

http://en.wikipedia.org/w/index.php?title=Infinite_impulse_response&oldid=189899892

Jabloun, F., Cetin, A. E., & Erzin, E. (1999). Teager energy based feature parameters for speech recognition in car noise. *Signal Processing Letters, IEEE, 6*(10), 259-261.

Jabloun, F., & Enis Cetin, A. (1999). *The Teager energy based feature parameters for robust speech recognition in car noise.* Paper presented at the Acoustics, Speech, and Signal Processing, 1999. ICASSP '99. Proceedings., 1999 IEEE International Conference on.

Jeong, S., & Hahn, M. (2001). Speech quality and recognition rate improvement in car noise environments. *Electronics Letters, 37*(12), 800-802.

Johnson, D. *Introduction to Estimation Theory*. Retrieved 28 March, 2005, from http://cnx.rice.edu/content/m11263/latest/

Johnson, D. (2003). *Introduction to Estimation Theory*. Retrieved 23 Feb, 2008, from http://cnx.org/content/m11263/latest/

Kaiser, T. (2005). *Smart antennas : state of the art*. New York, NY, USA: Hindawi Pub. Corp.

Kim, L., Hasegawa-Johnson, M., & Koeng-Mo, S. (2006). *Generalized Optimal Multi-Microphone Speech Enhancement Using Sequential Minimum Variance Distortionless Response(MVDR) Beamforming and Postfiltering.* Paper presented at the Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on.

Kolmogorov, A. N. (1941). Interpolation and extrapolation of stationary random sequences. *Mathematics series, 5*(314).

Krolik, J., Eizenman, M., & Pasupathy, S. (1986). *Time delay estimation via generalized correlation with adaptive spatial prefiltering.* Paper presented at the Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86.

Least mean squares filter. (2008). Wikipedia, The Free Encyclopedia Retrieved 29 Feb, 2008, from http://en.wikipedia.org/w/index.php?title=Least_mean_squares_filter&oldid=193011334

Lecomte, I., Lever, M., Boudy, J., & Tassy, A. A. T. A. (1989). *Car noise processing for speech input.* Paper presented at the Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on.

Li, J., & Stoica, P. (Eds.). (2005). *Robust Adaptive Beamforming*: Wiley-Interscience.

Li, W., Takeda, K., & Itakura, F. (2005). Adaptive log-spectral regression for in-car speech recognition using multiple distributed microphones. *Signal Processing Letters, IEEE, 12*(4), 340-343.

Liang, H., Rosca, J., & Balan, R. (2003). *Independent component analysis based single channel speech enhancement.* Paper presented at the Signal Processing and Information Technology, 2003. ISSPIT 2003. Proceedings of the 3rd IEEE International Symposium on.

Lim, T. J., and Macleod, M. D.,. (1994). Adaptive allpass filtering for nonminimum phase identification. *IEE Proceedings Vis. Image Processing, 141*, pp 373-379.

Lockwood, P., Boudy, J., & Blanchet, M. (1992). *Non-linear spectral subtraction (NSS) and hidden Markov models for robust speech recognition in car noise environments.* Paper presented at the Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on.

Lyons, R. G. (2004). *Understanding digital signal processing*: Prentice Hall PIR.

Ma, L., Shangguan, W., & Zang, Y. (2007). *Design of Speech Control System IN Car Noise Environments.* Paper presented at the Mechatronics and Automation, 2007. ICMA 2007. International Conference on.

Mailloux, R. J. (1994). *Phased Array Antenna Handbook*. Boston: Artech House, Boston.

Meyer, J., & Simmer, K. U. (1997). *Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction.* Paper presented at the Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on.

Moinuddin, M., Zerguine, A., & Sheikh, A. U. (2005). *Tracking analysis of the NLMF algorithm in the presence of both random and cyclic nonstationarities.* Paper presented at the Signal Processing and Its Applications, 2005. Proceedings of the Eighth International Symposium on.

Moir, T. J. (2008). Control systems approach to the sample inverse covariance matrix. *J. Franklin Inst.*

Moir, T. J., & Lindroth II, G. (2008). *From science fiction to science fact: A Smart-House interface using speech technology and a photo-realistic avatar.* Paper presented at the M2VIP 2008.

Myers, L. J., Erim, Z., & Lowery, M. M. (2004). Time and frequency domain methods for quantifying common modulation of motor unit firing patterns. *Journal of Neuroeng Rehabil*.

Myron Youdin, Mario W. Clagnaz, Jr., & Louie, H. (1980). USA Patent No. 4207959.

Nascimento, V. H., & Bermudez, J. C. M. (2005). *When is the least-mean fourth algorithm mean-square stable?* Paper presented at the Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on.

Nordholm, S., Claesson, I., & Dahl, M. (1999). Adaptive microphone array employing calibration signals: an analytical evaluation. *Speech and Audio Processing, IEEE Transactions on, 7*(3), 241-252.

Oh, S., Viswanathan, V., & Papamichalis, P. (1992). *Hands-free voice communication in an automobile with a microphone array.* Paper presented at the Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on.

Ortega, A., Lleida, E., Masgrau, E., & Gallego, F. A. G. F. (2002). *Cabin car communication system to improve communications inside a car.* Paper presented at the Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on.

Patzewitsch, J., Srinath, M., & Black, C. (1979). Near-field performance of passive coherence processing sonars. *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on, 27*(6), 573-582.

Plackett, R. L. (1950). Some theorems in least squares. *Biometrika, 37*(1-2), pp.149-157.

Puder, H., & Dreiseitel, P. (2000). *Implementation of a hands-free car phone with echo cancellation and noise-dependent loss control.* Paper presented at the Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on.

Qi, Z., & Moir, T. (2007a). *An Adaptive Wiener Filter for an Automotive Application with Non-Stationary Noise.* Paper presented at the 2nd International Conference on Sensing Technology, Palmerston North.

Qi, Z., & Moir, T. (2007b). *A Design of Automotive Voice Recognizer Using LabVIEW*. Paper presented at the Proceedings of the Fourteenth Electronics New Zealand Conference, Wellington.

Qi, Z., & Moir, T. (2008). An Adaptive Wiener Filter for Automatic Speech Recognition in a Car Environment with Non-Stationary Noise. In S. Mukhopadhyay & G. S. Gupta (Eds.), *Smart Sensors and Sensing Technology*. Berlin, Heidelberg, New York: Springer Verlag.

Qi, Z., & Moir, T. J. (2005). *An Automotive three-microphone Voice Activity Detector and noise canceller.* Paper presented at the 2005 International Conference on Intelligent Sensors, Sensor Networks and Information, Melbourne.

Qi, Z., & Moir, T. J. (2006). Automotive 3-microphone Noise Canceller in a Frequently Moving Noise Source Environment. *INTERNATIONAL JOURNAL OF SIGNAL PROCESSING, 3*(4), 298-304. http://www.enformatika.org/ijsp/

Resende, L. S., Romano, J. M. T., & Bellanger, M. G. (2004). Split Wiener filtering with application in adaptive systems. *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on], 52*(3), 636-644.

Rorabaugh, C. B. (1999). *DSP primer*: McGraw Hill.

Rulph, C. (2002). *DSP applications using C and the TMS320C6x DSK*: J. Wiley.

Ryan, J. G., & Goubran, R. A. (1997). *Near-field beamforming for microphone arrays.* Paper presented at the Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on.

Shozakai, M., Nakamura, S., & Shikano, K. (1998). *Robust speech recognition in car environments.* Paper presented at the Acoustics, Speech, and Signal Processing, 1998. ICASSP '98. Proceedings of the 1998 IEEE International Conference on.

Skidmore, I. D., & Proudler, I. K. (2001). KAGE: a new fast RLS algorithm. *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*.

Smith, S. W. (1999). *Digital Signal Processing*. San Diego: California Technical Publishing.

Sullivan, T. M. (1996). *Multi-Microphone Correlation-Based Processing for Robust Automatic Speech Recognition.* Carnegie Mellon University, Pittsburgh.

Swarts, F. (1999). *CDMA techniques for third generation mobile systems*. Boston: Kluwer Academic Publishers.

Van Compernolle, D. (1990). *Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings.* Paper presented at the Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on.

Walach, E., & Widrow, B. (1984). The least mean fourth (LMF) adaptive algorithm and its family. *Information Theory, IEEE Transactions on, 30*(2), 275-283.

Wang, J., Yang, C., & Chang, K. (2004). *Subspace tracking for speech enhancement in car noise environments.* Paper presented at the Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on.

Wang, S., & Tang, M. (2004). Exact confidence interval for magnitude-squared coherence estimates. *Signal Processing Letters, IEEE, 11*(3), 326-329.

Wiener filter. (2008). Wikipedia, The Free Encyclopedia Retrieved 23 Feb, 2008, from http://en.wikipedia.org/w/index.php?title=Wiener_filter&oldid=190480874

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

Yang, R., Yang, R., & Haavisto, P. (1995). *Noise compensation for speech recognition in car noise environments*

Noise compensation for speech recognition in car noise environments. Paper presented at the Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on.

Yermeche, Z., Cornelius, P., Grbic, N., & Claesson, I. A. C. I. (2004). *Spatial filter bank design for speech enhancement beamforming applications.* Paper presented at the Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2004.

Yermeche, Z., Garcia, P. M., Grbic, N., & Claesson, I. A. C. I. (2002). *A calibrated subband beamforming algorithm for speech enhancement.* Paper presented at the Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2002.

Yoon, B.-J., Tashev, I., & Acero, A. (2007). *Robust Adaptive Beamforming Algorithm using Instantaneous Direction of Arrival with Enhanced Noise Suppression Capability.* Paper presented at the Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on.

Yu, K., Boling, X., Mingyang, D., & Chongzhi Yu, A. C. Y. (2000). *Suppressing cocktail party noise for speech acquisition.* Paper presented at the Signal Processing Proceedings, 2000. WCCC-ICSP 2000. 5th International Conference on.

Zerguine, A. (2000). *Convergence behavior of the normalized least mean fourth algorithm.* Paper presented at the Signals, Systems and Computers, 2000. Conference Record of the Thirty-Fourth Asilomar Conference on.

Zhang, X., & Hansen, J. H. L. (2003). *CFA-BF: A Novel Combined Fixed/Adaptive Beamforming for Robust Speech Recognition in Real Car Environments.* Paper presented at the Eurospeech-2003, Switzerland.

Zhang, X., & Hansen, J. H. L. (2003). *CSA-BF: novel constrained switched adaptive beamforming for speech enhancement & recognition in real car environments.* Paper presented at the Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on.

Zhang, Z., Liu, Z., Mike, S., Alex, A., Li, D., Jasha, D., et al. (2004). *Multi-sensory microphones for robust speech detection, enhancement and recognition.* Paper presented at the ICASSP 2004.

Appendix

A.1 Tools for acoustic digital signal processing

Microphone technology

The first microphone was invented by Emile Berliner but the first practical carbon microphone was commercialized by Thomas Edison in 1876. Its main element is a thin and flexible diaphragm as in Figure A-1. When this thin piece of material struck by sound waves, it vibrates. The vibration is converted into electrical signals.

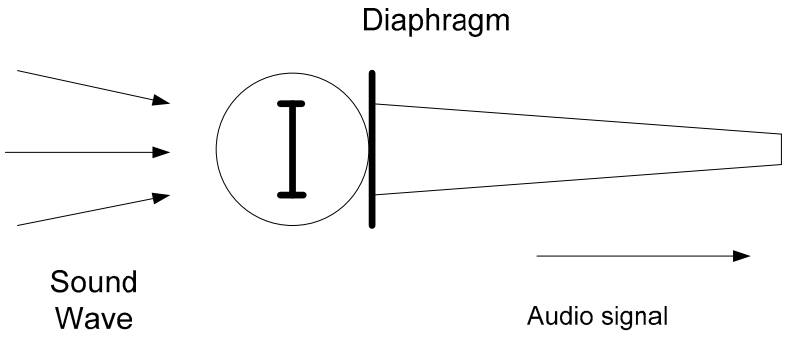

Figure A-1 Structure of microphone

A microphone is normally classified in two ways: sensitivity polar patterns and physical materials.

Classification by sensitivity polar pattern

As Classification by sensitivity polar pattern, microphone types are classified as non-directional (omni-directional) and directional (bi-directional and unidirectional) according to sensitivity patterns as showed in Figure A-2. (*Live Sound Microphone: Buying Guide*, 2008)

Omni-directional microphone

The diaphragm in omni-directional microphone is designed to be sensitive to signals incoming from any direction as in Figure A-2(A). It cannot detect a direction of arrival (DOA) or a distant sound source. A single omni-directional microphone will measures the acoustic signal directly and output proportionally. (*Microphone*, 2008)

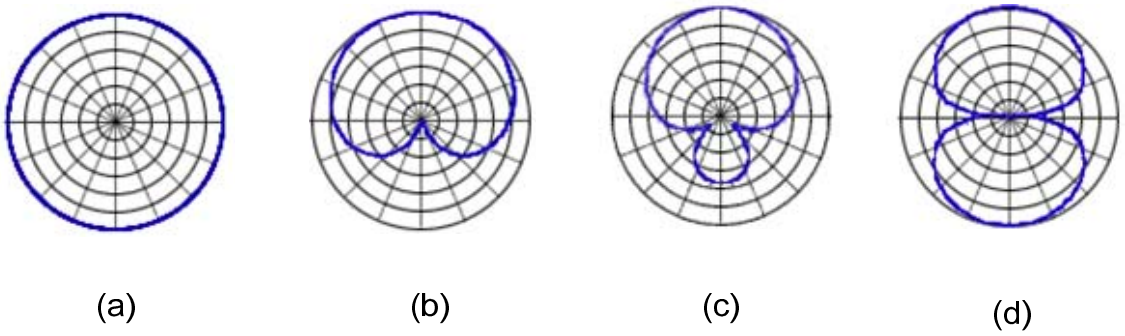

Figure A-2 Sensitivity polar patterns: (A) omnidirectional (B) unidirectional (cardioid) (C) unidirectional (hypercardioid) (D) bidirectional

Unidirectional

A unidirectional microphone uses an internal acoustic time delay to have the directional response of the microphone towards a desired direction. The unidirectional microphones are more sensitive to acoustic waves arriving from a direction but are less sensitive to waves incoming from other directions. This microphone normally appears as cardioid as Figure A-2(B) or hypercardioid directivity patterns as Figure A-2(C).

Bi-directional

Bi-directional microphone is a pressure gradient microphone and is also a noise-cancelling microphone. It uses properties of a gradient microphone to build a noise cancellation.

As showed in Figure A-2 (D), any noise from the side is cancelled and the sound pressure never arrives at the front and the back of the microphone at the same time. (*Microphone*, 2008)

Classification by physical material

On the type of physical materials, a microphone can be built as dynamic (electromagnetic) or variable condenser (electrostatic) models.

Dynamic microphone

A dynamic microphone consists of a coil, a diaphragm and a magnetic. A coil of fine wire couples to a flexibly-mounted diaphragm. As the coil is mounted in the air gap of a magnet, when sound reaches the diaphragm surface causing vibrates. These vibrations couples to the coil, which moves in the magnet field. Therefore, an electrical current is created in the wire

of the coils as it cuts through lines of magnetic force. So the changes of current reflect on the sound wave incoming to the diaphragm. (*Microphone*, 2008)

Condenser microphone

Condenser microphone is the most common microphone. It consists of a capacitor with a pair of metal plates separated by a dielectric. One of these plates responses to the motion of the sound pressure. The polarizing voltage in the microphone and the distance of these plates builds the sensitivity. In comparison of the dynamic microphone, condenser microphone needs an external power supply. (*Microphone*, 2008)

Electret condenser microphone

Electret condenser microphone is a type of the condenser microphone however it does not need power to charge the diaphragm; instead it requires a power supply for the built-in pre-amplifier. This is a small in size, low cost, good performance at high frequencies. The telephone handset normally uses Electret condenser microphone. (*Microphone*, 2008)

Signal sampling and Anti-alias filter

In signal processing, when a signal is sampled, it is reconstructed as an alias of the original signal. It is called signal alias. Whilst sampling a human speech, for instance, if we sample it with a frequency that is too low and reconstruct the speech using a digital to analog converter, then low-frequency aliases of the under-sampled high frequencies are created. As an example, Figure A-3 shows a low frequency sample at a high frequency signal, an alias waveform is created. Therefore, before the sampling, the best solution is to remove the high frequencies using a low-pass filter. Such a low-pass filter is called Anti-alias filter(*Anti-aliasing*, 2008).

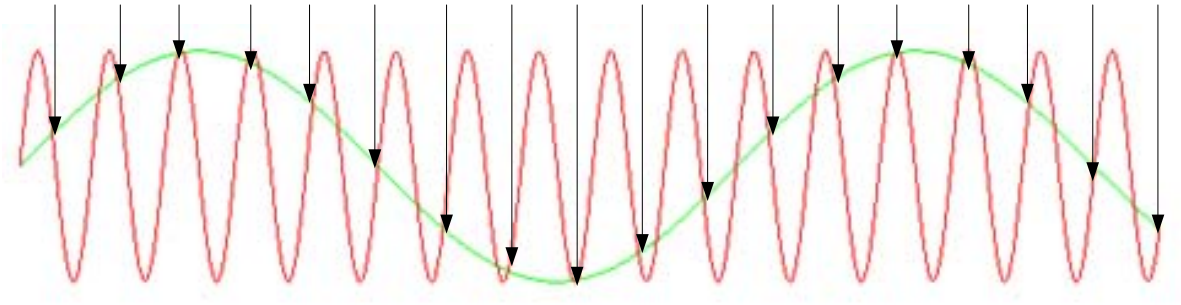

Figure A-2 Low sample rate causes alias

Digital Signal Processing hardware and software review

Texas Instrument DSP TMS320C6711 DSK provides a good development environment to build a real time test e.g. LMS filter. However, there are not much supports of multi-channel signal processing. Therefore, a Data Acquisition System (DAQ) PMD-1208FS was built to apply Universal Library for LabVIEW to test algorithms to produce a novel innovative solution to the problem in speech recognition in automotive.

The PMD-1208FS (*Universal Library for LabVIEW User's Guide*, 2004) features eight analog inputs, two 12-bit analog outputs, 16 digital I/O connections, and one 32-bit external event counter. The PMD-1208FS is powered by the +5 volt USB supply from your computer. No external power is required.

The PMD-1208FS analog inputs are software configurable for either eight 11-bit single-ended inputs, or four 12-bit differential inputs. Sixteen digital I/O lines are independently selectable as input or output in two 8-bit ports. A 32-bit counter can count TTL pulses. A

SYNC (synchronization) input / output line allows you to pace the analog input acquisition of one PMD module from the clock output of another. The PMD-1208FS is shown in Figure A-5. I/O connections are made to the screw terminals located along each side of the PMD-1208FS.

A block diagram of PMD-1208FS is shown as Figure A-4.

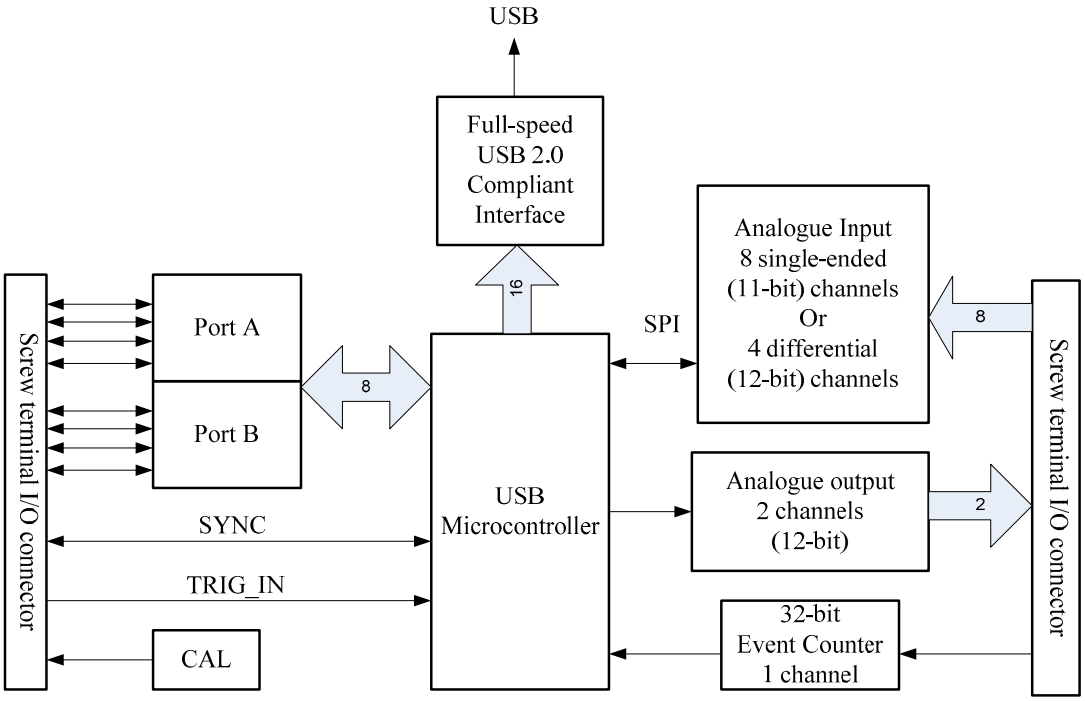

Figure A-3 A block diagram of PMD-1208FS

Research in this thesis used an upgraded version PMD-1608FS, which is an USB-based DAQ module with 8 channels of 16-bit analog input, 8 DIO bits. Sample rate is 200 kS/s max total throughput (max 50 kS/s for any channel).

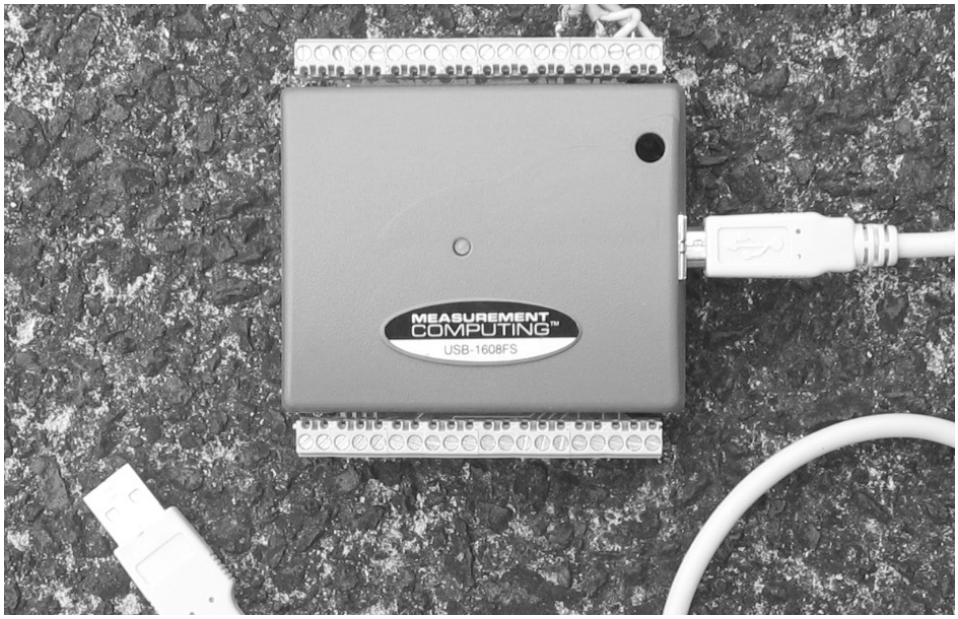

Figure A-4 DAQ PMD-1608FS

With the Universal Library for LabVIEW software, Universal Library Extension VIs or traditional DAQ VIs with Measurement Computing data acquisition and control boards is available for LabVIEW programming.

Universal Library Extension VIs and traditional DAQ VIs can be used in the same application. An existing LabVIEW program can be installed with both MCC and NI hardware.

The Universal Library for LabVIEW (*PMD-1208FS Personal Measurement Device brand USB-based Analog and Digital I/O Module User's Guide*, 2004) supports LabVIEW version 6 or greater. LabVIEW 7 Express DAQmx VIs is not supported in this release.

The Universal Library for LabVIEW includes a set of LabVIEW virtual instruments (VIs). Each low-level VI corresponds to one Universal Library function.

Software tools for Digital Signal Processing

LabVIEW

LabVIEW (short for Laboratory Virtual Instrumentation Engineering Workbench) is a platform and development environment for a visual programming language from National Instruments.

LabVIEW program mainly consist of Front Control Panel (as Figure A-6) and Block Diagram (as Figure A-7).

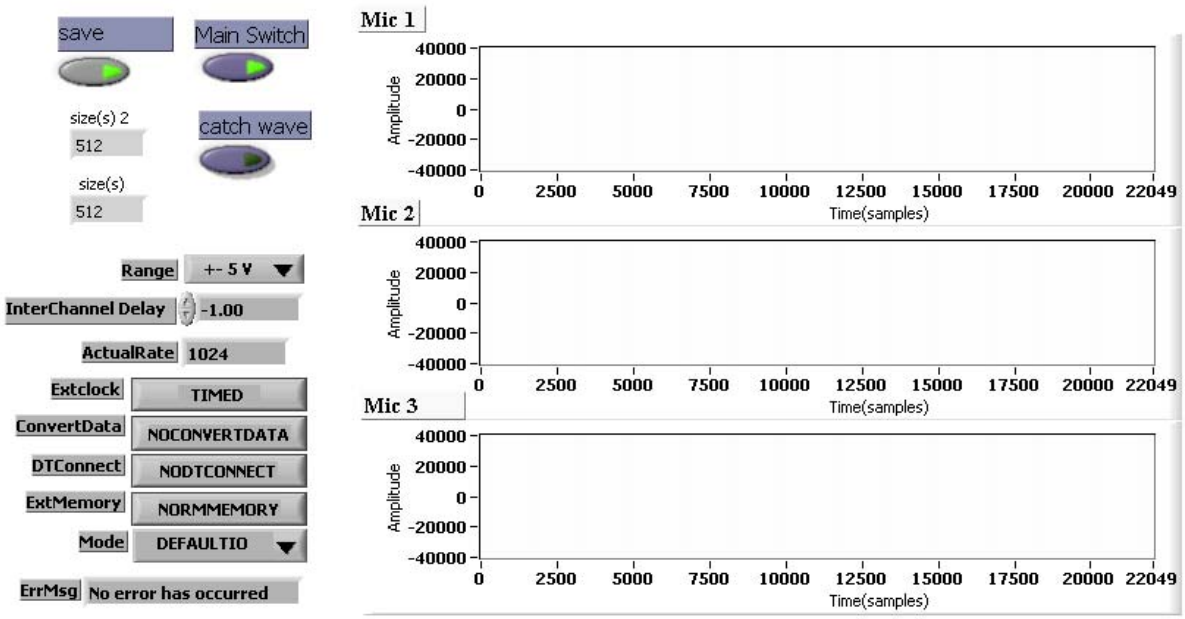

Figure A-5 An example of LabVIEW Control panel

As an example, Figure A-6 shows a control panel for 3-microphone ADC sampling via ADC hardware. Figure A-7 shows part of Block Diagram on Figure A-6.

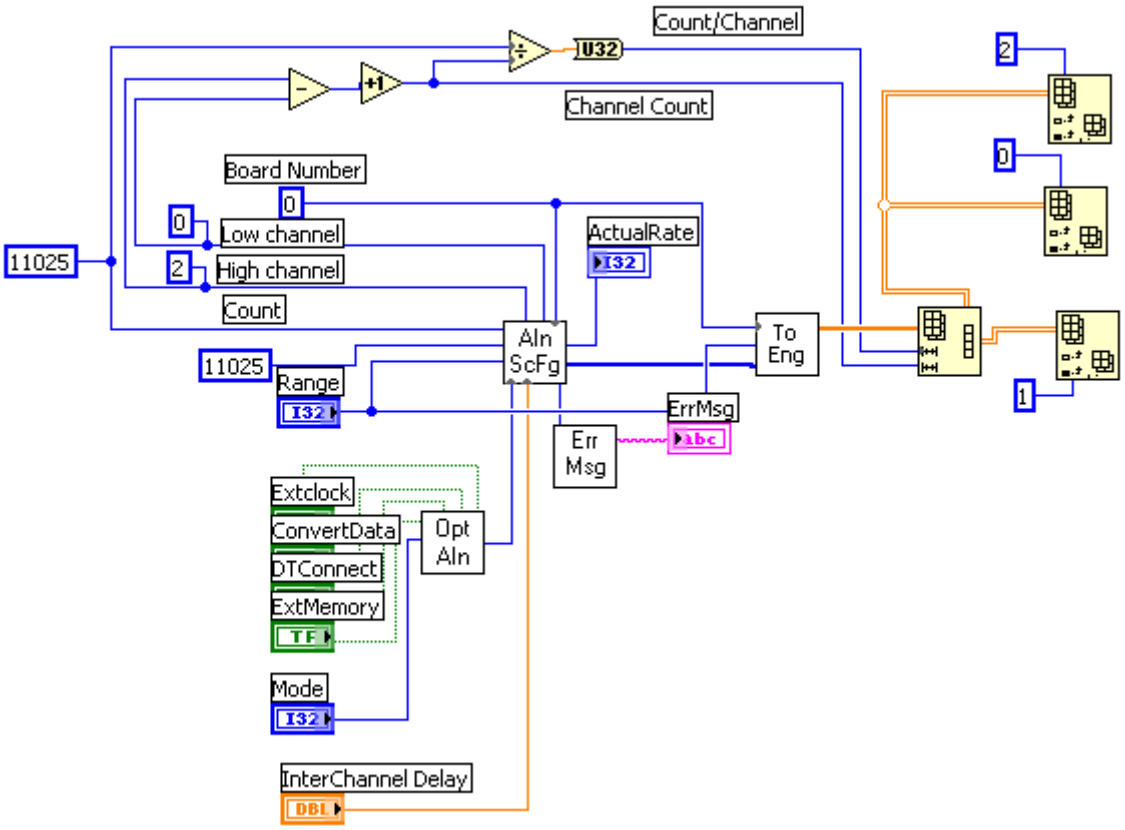

Figure A-6 An example of LabVIEW Block diagram

CoolEdit

Adobe Audition (formerly Cool Edit Pro) is a digital audio editor computer program from Adobe Systems featuring both a multitrack, non-destructive mix/edit environment and a destructive-approach waveform editing view that has been referred to as the "Swiss army knife" of digital audio. Originally, Cool Edit was a shareware program with some crippleware features.

CoolEdit has Features: create multitrack mixes with unlimited stereo tracks, multichannel encoder to produce high-quality 5.1 surround sound, remove audio flaws with easy-to-use effects, accepts third-party DX and VST plugins, and supported audio CD burning. CoolEdit is property of Adobe, Inc.

In this thesis, CoolEdit is used to analyse waveforms for experiment as Figure A-8 and Figure A-9.

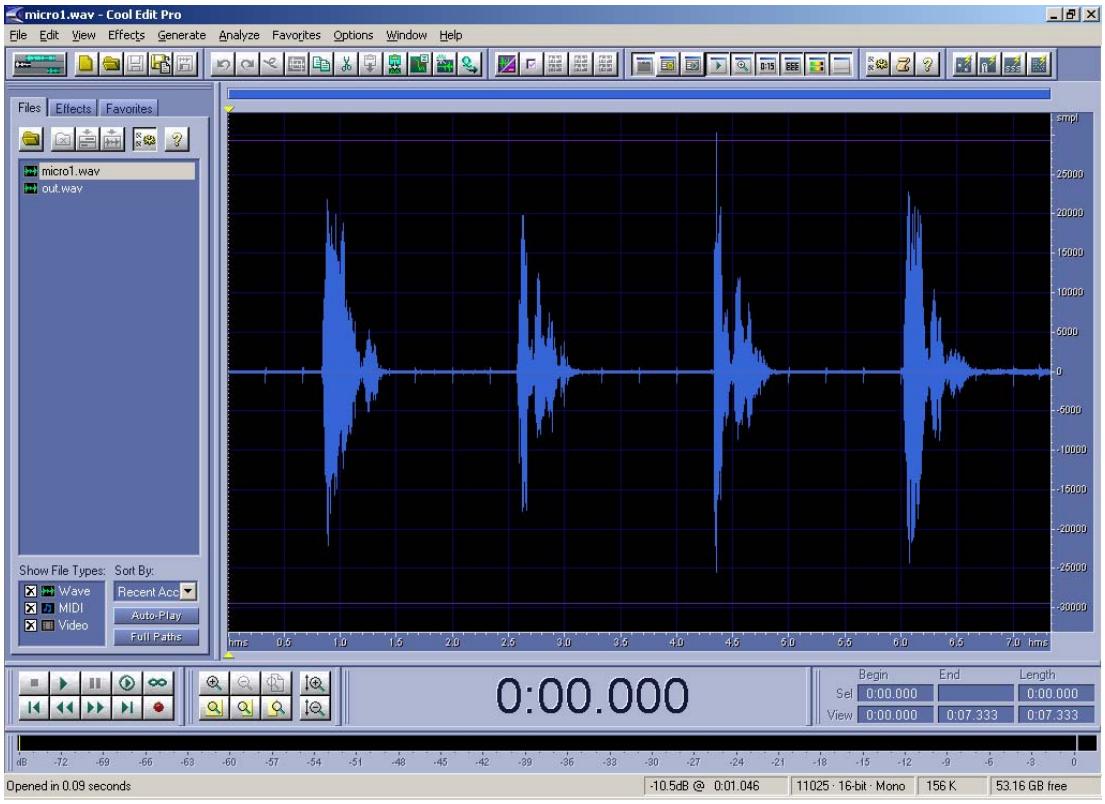

Figure A-7 A demo of CoolEdit program

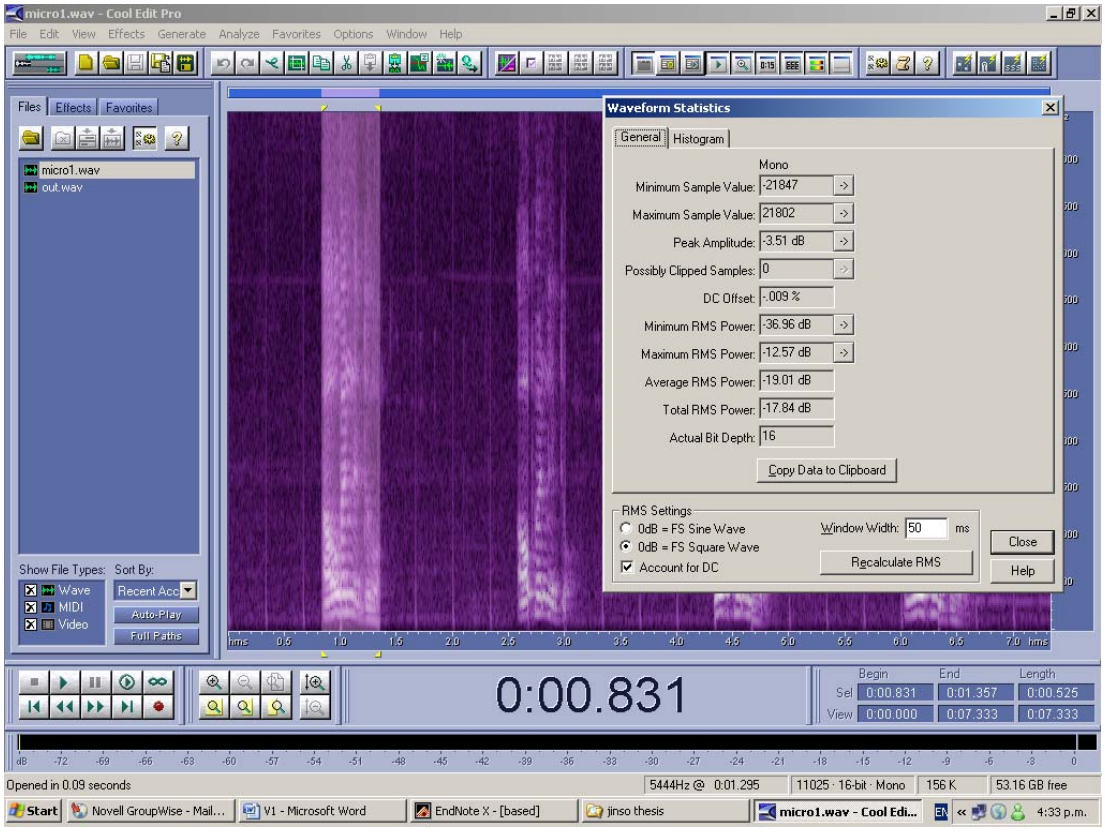

Figure A-8 A demo on Average RMS Power calculation using Waveform Statistic in CoolEdit

Maple 7

Maple 7 is a comprehensive computer system for advanced mathematics. It includes facilities for interactive algebra, calculus, discrete mathematics, graphics, numerical computation and many other areas of mathematics. It also provides a unique environment for rapid development of mathematical programs using its vast library of built-in functions and operations.

In this thesis, Maple 7 program is used for creating 2D or 3D graphic demos for the multi-microphone desired geometrical zone.

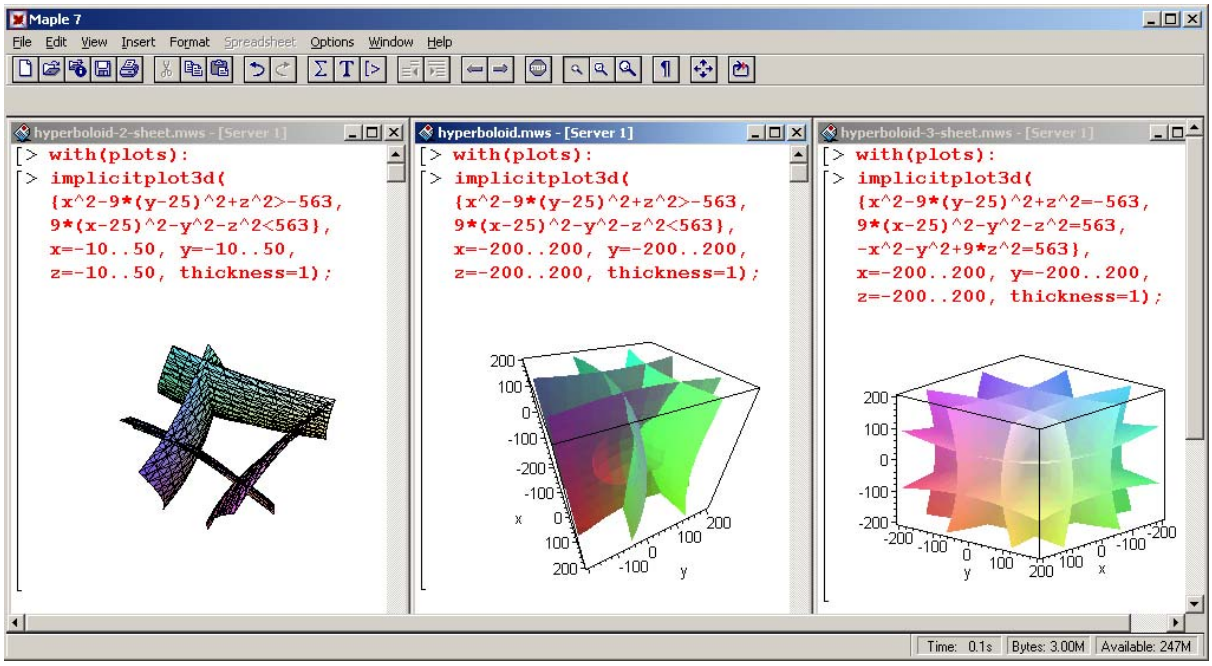

Figure A-9 Demos in Maple 7 program

As an example, in Figure A-10, the left and the middle graphs indicate 3-microphone desired geometrical zone and the graph on the right shows a 4- microphone desired geometrical zone.

A.2 Automatic speech recognition

Automatic Speech Recognition (ASR) is to identify human speech using computer and computer program. Starting in the 1950s, rule-based ASR method had very little success. An

alternative method is to use pattern-matching techniques. But major improvement did not occur until 1970s (Holmes, 2001).

Template matching based speech recognition

Template matching based speech recognition is to collect templates of speech and then compare the incoming speech to compute a possibility of matching between the incoming speech and the templates. A template is a collection of a particular pronunciation, which is normally a word. These words are typically called frames. During ASR is running, each incoming new frame is compared with all of the reference frames (templates). Once the incoming frame associates with the template with the smallest distance, we decide it matched. (Gaudard, Aradilla, & Bourlard, 2007)

In 1970's, Campbell and Saridis (Campbell & Saridis, 1979) described a system developed on a PDP 11/45 computer which recognizes spoken commands and controls a Scheinmann MIT mechanical arm to serve the needs of an immobilized patient. The speech recognition is discrete word with template matching, using as features zero crossing rate, average absolute magnitude, and frequency and normalized error as derived from the two pole linear predictive analysis. Commands are made up of words from a 16-word vocabulary obeying a specified, syntax. Upon recognition of a command the arm performs the corresponding pre-programmed task. A recognition time of two seconds is attained (including the one second sampling interval), plus an additional one second to verify the word. With the command syntax a recognition rate of 88% is attained, as compared with only 56% when each word is matched against every word in the vocabulary.

Although Template Matching has been applied in speech recognition for long time, many researches using Template Matching has been applied in speech recognition for long time have been reported recently. Shi et al. (Shi, Liu, & Liu, 2001) described a single-chip speech recognition system. It contains the speech functions of prompt, playback, speaker-dependent speech recognition, suitable for the voice activated systems in toys, games, consumer electronics, office devices, etc. The chip is designed based on the SOC (system on chip) philosophy and an 8-bit MCU, RAM, ROM, ADC/DAC, PWM, I/O ports and other peripheral circuits are all embedded in it. Software modules including control/communication, speech coding and speech recognition algorithms are implemented in an 8051 compatible microcontroller core, resulting in the extremely low cost of the chip. The speech recognition adopts the template matching technique. It recognizes up to 20 phrases

with an average length of 1 second and the recognition accuracy reaches more than 95% with the background SNR above 10 dB. Speech coding uses continuous variable slope delta modulation (CVSD) algorithm. The bit rate is 16 kbits/s.

Guoliang et al. (Guoliang, Hui, Fang, & Wenhu Wu, 2004) presented an extended template matching strategy, which imports the filler models of a keyword spotting strategy into the template matching strategy. Because this recognition strategy not only makes use of the context information and the background knowledge by grammar template, but also adopts filler models to match extraneous speech and non-speech signals, it achieves high recognition accuracy and good robustness.

De Wachter (De Wachter et al., 2007) attempted to overcome these problems by relying on straightforward template matching. The basis for the recognizer is the well-known Dynamic Time Warping algorithm. However, classical Dynamic Time Warping continuous speech recognition results in an explosion of the search space. The traditional top-down search is therefore complemented with a data-driven selection of candidates for Dynamic Time Warping alignment.

Hidden Markov model based speech recognition

Recently most speech recognition systems are based on Hidden Markov model (HMM). In template matching we compare the incoming speech to compute a possibility of matching between the incoming speech and the templates. When HMM is used for this possibility calculation, we call it Hidden Markov model based speech recognition(Holmes, 2001).

Hidden Markov model based speech recognition has advantage of the time warping ability.

Neural Network based speech recognition

Neural network based speech recognition is a computational model based on neural networks to identify human speech. Since Neural Network can model the complex relationships between inputs and outputs to find patterns in data, it is a useful tool for speech recognition. Neural network based speech recognition has advantage of the pattern recognition capability(Gemello, Gemello, Albesano, & Mana, 2000).

Hybrid Hidden Markov model and Neural Network based speech recognition

When we integrate the Hidden Markov model based speech recognition with its advantage of the time warping ability and the neural network based speech recognition with its advantage of the pattern recognition capability, experiments show the Hybrid models have a mature technology highly competitive with Hidden Markov model based speech recognition (Gemello et al., 2000).

Speech Recognition using microcontroller in a car

For real-time environments, a small vocabulary speech recognition system is very often required but has limited resources e.g. CPU speed and memory. Previous research has been done by employing micro-controller to build a speech recognizer. Bernal-Ruiz [1] presented a compact system for vowels and small vocabulary recognition with a standard microcontroller device, with the typical requirements of a low range embedded application where memory and computer power are very limited. Fezari (Fezari, Bousbia-Salah, & Bedda, 2006) introduced an Electric wheelchair using a small vocabulary word recognition system implemented in a RISC architecture microcontroller adapted to a speech recognition development kit 'Voice Direct 364' (VD364). Hwang (Hwang & Kintigh, 1994) applied a speech recognition IC VCP200 on an intelligent Roving Robot. Hale (Hale & Nguyen, 1995) used fuzzy logic on an advanced microcontroller with integrated DSP hardware for voice command recognition. Gonzalez-Concejero (Gonzalez-Concejero, Rodellar, Alvarez-Marquina, Martinez de Icaya, & Gomez-Vilda, 2006) suggested that any front-end processing is done on a short-time frame by frame basis. The speech signal is divided into non-overlapped fixed-length blocks, each of size 128 samples. Two successive blocks form a frame, having a length of 2*128 = 256 samples, so the speech signal is divided into overlapping frames. From each frame, a set of frequency-domain or cepstral-domain parameters are derived to form the so-called feature vector. The processes involved are described as: Windowing, Fast Fourier Transforms, Mel-frequency filter banks, Inverse Discrete Cosine Transforms and Vector quantization. Microchip (Microchip, 2004) created the sdPICC30 Speech Recognition Library which operates in both clean and noisy conditions. The total signal-to-noise ratio (SNR) should be no less than 15 dB. However, A signal level above the noise threshold is presumed to be an incoming word.

Application to speech recognition in an automotive Background

Recently more people are concerned about car safety. As many new technologies have been used in cars, drivers pay more attentions on these new devices in cars such as navigation systems - electronic Global Positioning System (GPS) maps but not on the road. Many countries announced policies to minimize driver distraction, such as banning handheld cell phone use in the car. Many commercial companies e.g. Ford and General Motors, have espoused an "eyes-on-the-road, hands-on-the wheel" philosophy as they work to incorporate new electronic capabilities into automobiles. (Murray, 2000)

Microsoft Corp. is working with automotive manufacturers and suppliers to support the safe use of in-car technology with the recent release of the newly re-engineered speech system in Windows® CE for Automotive 3.5, its in-vehicle telematics software platform. (Redmond, 2002)

Microsoft Windows CE for Automotive 3.5 contains two unique speech components that make it easier than ever for drivers to interact in a natural voice with their in-car communication device: the Speech User Interface (SUI) and Speech Application Programming Interface (SAPI) 5.0. Microsoft's breakthrough SUI system enables easy implementation of voice recognition and human-like text-to-speech technologies. This allows speech applications to easily integrate with telematics devices in a safe and effective manner, thus helping to ensure that the driver can maintain visual contact with the traffic around the vehicle. The SUI works in tandem with SAPI 5.0, a global standard interface for speech applications, to provide a complete system capable of understanding natural speech and handling multiple interactions in a simple and familiar programming model (Redmond, 2002). However, automotive manufactures say they cannot bring eyes-on-the-road, hands-on-the-wheel techniques to vehicles unless they have good speech recognition systems. And a good speech recognition system requires a good filter to improve the Signal to Noise ratio (SNR). (Murray, 2000)

Applications of speech recognition in automotive have serious problems as noise is stronger in cars than in offices. So that a good filter become more and more important.

A normal two-microphone adaptive LMS filter is good at cancelling noise from desired directions. However, this filter cannot reduce the noise where the desired signal is from. Theoretically, three-microphone beamforming can focus on a desired area rather than a desired direction.

Automatic speech recognition hardware

Commercialized Automatic speech recognition (ASR) hardware is broadly using in toys. Most popular Integrate Circuit (IC) are HM2007 and Voice Direct ICs.

The HM2007 is a single-chip voice-recognition system previously manufactured by Hualon. It requires an external array of SRAM, but otherwise provides all necessary components for a voice-recognition system including the analog front-end, voice analysis, recognition processing, and system control. The HM2007 can recognize either 40 0.9-second long words or 20 1.92-second long words or phrases. The recognition technology is of the speaker-dependent discrete type meaning that it can only recognize words spoken in isolation and the speaker must "train" the system with each word to be recognized. The HM2007 is available in a 48-pin PDIP or 52-pin PLCC package. This report only addresses the PDIP package. (*HM 2007 Voice Recognition IC*, 2006)

Voice Direct™, from the Interactive Speech™ family of products, is a speaker-dependent speech recognition IC designed for use in cost-sensitive electronic products. In addition to performing speech recognition, Voice Direct plays speech prompts, performs system control functions, provides status outputs and interfaces to external ROM and Serial EEPROM. Voice Direct can be controlled by an external host processor or it can operate in a pin configurable stand-alone mode. Voice Direct employs a sophisticated neural network to recognize trained words or phrases with greater than 99% accuracy. The highly-integrated nature of the chip reduces external parts count. A complete recognition system can be built with few additional parts other than a battery, speaker, external memory, microphone, crystal, and audio input circuitry. Voice Direct is available as an IC or in a complete system module that includes a PCB and all external components(*Sensory*, 2008).

Voice Direct Speech Recognition IC

Voice Direct performs high quality speaker dependent speech recognition. The chip utilizes its neural network recognizer to recognize discrete words or short phrases. The chip performs three basic functions(*Sensory*, 2008):

Train - Users train the chip to identify a specific word by saying each word twice. After training, the two patterns are averaged and a template is stored.

Recognize - The user speaks a word and the chip compares the new pattern with the previously trained templates to identify which word was spoken. The chip then outputs the result of its analysis. (*Sensory*, 2008)

Erase - Users can delete previously trained words from the set of recognition templates.

In each of these functions, Voice Direct features integrated speech prompting providing a complete interactive user interface. (*Sensory*, 2008)

RS-07 Speech Recognition Kit

The SR-07 Speech Recognition Kit is a standalone circuit that can recognize up to 40 words (user selected words) lasting one second each or 20 words (user selected words or phrases) lasting 2 seconds each. The speech recognition circuit is multi-lingual, words to be trained for recognition may be in any language. The heart of the circuit is the HM2007 speech recognition integrated circuit. (*Speech Recognition Circuit Assembled SR-07*, 2008) The chip provides the options of recognizing either forty .96 second words or twenty 1.92 second words. This circuit allows the user to choose either the .96 second word length (40 word vocabulary) or the 1.92 second word length (20 word vocabulary). For memory the circuit uses an 8K X 8 static RAM. (*Build a Speech Recognition Circuit*, 2008) Figure A-11 showed a SR-07 Speech Recognition Kit.

This chip has two operational modes; manual mode and CPU mode. The CPU mode is designed to allow the chip to work under a host computer. This is an attractive approach to speech recognition for computers because the speech recognition chip operates as a co-processor to the main CPU. The job of listening and recognition doesn’t occupying any of the computer's CPU time. When the HM2007 recognizes a command it can signal an interrupt to the host CPU and then relay the command code. The HM2007 chip can be cascaded to provide a larger word recognition library(*Build a Speech Recognition Circuit*, 2008).

The SR-07 circuit are operating in the manual mode. The manual mode allows one to build a standalone speech recognition board that doesn't require a host computer and may be integrated into other devices to utilize speech control(*Build a Speech Recognition Circuit*, 2008).

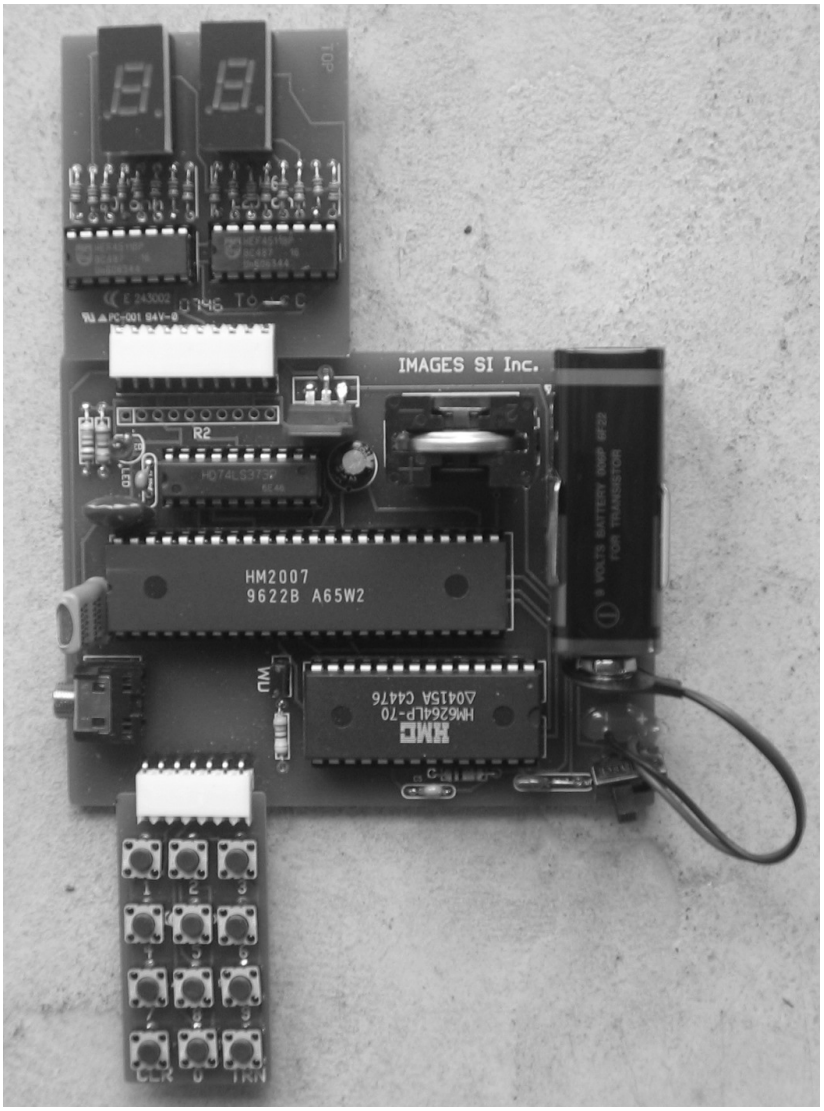

Figure A-10 SR-07 Speech Recognition Kit

In speaker dependent mode, systems are trained by the individual who will be using the system. These systems are capable of achieving a high command count and better than 95% accuracy for word recognition. The drawback to this approach is that the system only responds accurately only to the individual who trained the system. This is the most common approach employed in software for personal computers. (*Speaker Dependent / Speaker Independent*, 2008)

Reference

Anti-aliasing. (2008). Wikipedia, The Free Encyclopedia Retrieved 23 Feb, 2008, from http://en.wikipedia.org/w/index.php?title=Anti-aliasing&oldid=193170782

Build a Speech Recognition Circuit. (2008). Speech Recognition Circuits Retrieved 2 March, 2008, from http://www.imagesco.com/articles/hm2007/SpeechRecognitionTutorial01.html

Campbell, J., & Saridis, G. (1979). *A voice-controlled mechanical arm for immobilized patients.* Paper presented at the Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79.

De Wachter, M., Matton, M., Demuynck, K., Wambacq, P. A. W. P., Cools, R. A. C. R., & Van Compernolle, D. A. V. C. D. (2007). Template-Based Continuous Speech Recognition. *Audio, Speech, and Language Processing, IEEE Transactions on [see also Speech and Audio Processing, IEEE Transactions on], 15*(4), 1377-1390.

Fezari, M., Bousbia-Salah, M., & Bedda, M. (2006). *Voice and Sensor for More Security on an Electric Wheelchair*.

Gaudard, C. e., Aradilla, G., & Bourlard, H. e. (2007). *Speech Recognition based on Template Matching and Phone Posterior Probabilities*. IDIAP Research Institute Retrieved 12 March, 2008, from http://www.idiap.ch/ftp/reports/2007/gaudard-idiap-com-07-02.pdf

Gemello, R., Gemello, R., Albesano, D., & Mana, F. (2000). *CSELT hybrid HMM/neural networks technology for continuous speech recognition*

CSELT hybrid HMM/neural networks technology for continuous speech recognition. Paper presented at the Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on.

Gonzalez-Concejero, C., Rodellar, V., Alvarez-Marquina, A., Martinez de Icaya, E., & Gomez-Vilda, P. (2006). *Designing an Independent Speaker Isolated Speech Recognition System on an FPGA*.

Guoliang, Z., Hui, S., Fang, Z., & Wenhu Wu, A. W. W. (2004). *Robust speech recognition directed by extended template matching in dialogue system.* Paper presented at the Intelligent Control and Automation, 2004. WCICA 2004. Fifth World Congress on.

Hale, C., & Nguyen, C. (1995). *Voice command recognition using fuzzy logic*.

HM 2007 Voice Recognition IC. (2006). ITP Sensor Workshop Retrieved 20 March, 2008, from http://itp.nyu.edu/physcomp/sensors/Reports/HM2007VoiceRecognitionIC

Holmes, W. (2001). *Speech Synthesis and Recognition* (2 ed.): CRC Press.

Hwang, S., & Kintigh, B. P. (1994). *Implementation of an intelligent roving robot using multiple sensors*.

Live Sound Microphone: Buying Guide. (2008). Sweetwater Sound Inc. Retrieved 23 Feb, 2008, from http://www.sweetwater.com/shop/live-sound/microphones/buying-guide.php

Microchip. (2004). dsPIC30F Speech Recognition Library User's Guide.

Microphone. (2008). Wikipedia, The Free Encyclopedia Retrieved 23 Feb, 2008, from http://en.wikipedia.org/w/index.php?title=Microphone&oldid=192615484

Murray, C. J. (2000, 12/03/2000). Automakers struggle with speech recognition technology.

PMD-1208FS Personal Measurement Device brand USB-based Analog and Digital I/O Module User's Guide. (2004). Measurement Computing Corporation.

Redmond, W. (2002). *Microsoft's Breakthrough In-Car Speech System Promotes Safer, Enhanced Driving Experience*. http://www.microsoft.com/presspass/press/2002/apr02/04-22InCarSpeec

Sensory. (2008). Digi-Key Corporation Retrieved 20 March, 2008, from http://suppliers.digikey.com/Manufacturer.aspx?vendor=340&site=US&lang=en

Shi, Y., Liu, J., & Liu, R. (2001). Single-chip speech recognition system based on 8051 microcontroller core. *Consumer Electronics, IEEE Transactions on, 47*(1), 149-153.

Speaker Dependent / Speaker Independent. (2008). Speech Recognition Circuits Retrieved 2 march, 2008, from http://www.imagesco.com/articles/hm2007/SpeechRecognitionTutorial02.html

Speech Recognition Circuit Assembled SR-07. (2008). Speech Recognition Circuits Retrieved 2 March, 2008, from http://www.imagesco.com/kits/speech-recognition-kit.html

Universal Library for LabVIEW User's Guide. (2004). Measurement Computing Corporation.