Real-Time Implementation of a Dual Microphone Beamformer

A thesis presented in partial
fulfilment of the requirements for the degree of
Master of Engineering
in
Computer Systems

at Massey University,
Albany,
New Zealand.

Vaitheki Yoganathan

2005

To Whom It May Concern:

This thesis is submitted for the degree of Master of Engineering at Massey University and it is not previously submitted to this or any other institution for any degree, diploma or other qualification. It is the result of my independent work and the information obtained from the published or unpublished source of others have been acknowledged in the text and a list of references is given.

Vaitheki Yoganathan

Vaitheki Yoganathan

Date: 10-03-2005

Abstract

The main objective of this project is to develop a microphone array system, which captures the speech signal for a speech related application. This system should allow the user to move freely and acquire the speech from adverse acoustic environments. The most important problem when the distance between the speaker and the microphone increases is that often the quality of the speech signal is degraded by background noise and reverberation. As a result, the speech related applications fails to perform well under these circumstances. This unwanted noise components present in the acquired signal have to be removed in order to improve the performance of these applications.

This thesis contains the development of a dual microphone beamformer in a Digital Signal Processor (DSP). The development kit used in this project is the Texas Instruments TMS320C6711 DSP Starter Kit (DSK). The switched Griffiths-Jim beamformer was selected as the algorithm to be implemented in the DSK. Van Compernolle developed this algorithm in 1990 by modifying the Griffiths-Jim beamformer structure. This beamformer algorithm is used to improve the quality of the desired speech signal by reducing the background noise. This algorithm requires at-least two input channels to obtain the spatial characteristics of the acquired signal. Therefore, the PCM3003 audio daughter card is used to access the two microphone signals.

The software implementation of the switched Griffiths-Jim beamformer algorithm has two main stages. The first stage is to identify the presence of speech in the acquired signal. A simple Voice Activity Detector (VAD) based on the energy of the acquired

signal is used to distinguish between the wanted speech signal and the unwanted noise signals. The second stage is the adaptive beamformer, which uses the results obtained from the VAD algorithm to reduce the background noise.

The adaptive beamformer consists of two adaptive filters based on the Normalised Least Mean Squares (NLMS) algorithm. The first filter behaves like a beam-steering filter and it's only updated during the presence of speech and noise signal. The second filter behaves like an Adaptive Noise Canceller (ANC) and it is only updated when a noise alone period is present. The VAD algorithm controls the updating process of these NLMS filters and only one of these filters is updated at any given time.

This algorithm was successfully implemented in the chosen DSK using the Code Composer Studio (CCS) software. This implementation is tested in real-time using a speech recognition system. This system is programmed in Visual Basic software using the Microsoft Speech SDK components. This dual microphone system allows the user to move around freely and acquire the desired speech signal. The results show a reasonable amount of enhancement in the output signal, and a significant improvement in the ease of using the speech recognition system is achieved.

Acknowledgments

I would like to express my sincere thanks to my mentor Dr. Tom Moir for his guidance during this research work. Without his invaluable advice, help and suggestions this work would not have been possible. I would also like to thank the late Dr R. Chassaing for his assistance.

Finally I would also like to thank my family and friends for being there and supporting me throughout my studies.

Table of Contents

List of Figures

List of Abbreviations

ADC	Analog-to-Digital Converter
ANC	Adaptive Noise Canceller
C6711	TMS320C6711
CCS	Code Composer Studio
COFF	Common Object File Format
DAC	Digital-to-Analog Converter
DSK	Digital Signal Processor Starter Kit
DSP	Digital Signal Processor
GCC	Generalised Cross Correlation
GSC	Generalised Sidelobe Canceller
HOIT	Home Oriented Informatics and Telematics
LMS	Least Mean Squares
MFLOPS	Million of Floating Point Operations Per Second
ML	Maximum Likelihood
MSC	Magnitude Squared Coherence
NLMS	Normalized LMS
ROM	Read Only Memory
SDK	Software Development Kit
SNR	Signal-to-Noise Ratio
SDRAM	Synchronous Dynamic Random Access Memory
TASI	Time Assigned Speech Interpolation
TDOA	Time Difference Of Arrival
TI	Texas Instrument
VAD	Voice Activity Detector

1. Introduction

During the past 20 or more years, speech acquisition in adverse acoustic environments have received considerable attention due to the increased need for hands-free and voice controlled applications (Krasny & Oraintara, 2002; Martin, 1976). The main difficulty when acquiring speech from this type of environment is that often speech is corrupted by background noise and reverberation. Consequently, this corrupted speech complicates and degrades the performance of these speech related applications.

It is important to improve the quality of the acquired speech signal in order to improve the performance of these applications. This improvement is achieved by suppressing the unwanted noise components present in the acquired signal (without harming the speech signal). This background noise could consist of several components propagating from different sources such as computer fan, engine noise, air conditioner, audio equipments, or competing speech.

Noise reduction in the corrupted speech signal remains an important problem in many speech related applications. Some of these applications include:

- Videoconference and Teleconferencing (Elko, 1996)
- Hands-free telephony (Bouquin-Jeannès, Faucon, & Ayad, 1996; Campbell, 1999)
- Mobile telephony in moving vehicle environment (Cho & Krishnamurthy, 2003; Ezzaidi, Bourmeyster, & Rouat, 1997; Hussain, Campbell, & Moir, 1997; Lin, Lin, & Wu, 2002),

- Hearing aids (Greenberg, Desloge, & Zurek, 2003; Ventura, 1989; Wang et al., 1996; Widrow, 2001; Widrow & Luo, 2003; Wilson, 2003)
- Speech recognition (Chien & Lai, 2004; Moore & McCowan 2003; Van Compernolle, 1992a)
- Speech coding (Collura, 1999; Li & Hoffman, 1999)
- Robotics (Choi, Kong, Kim, & Bang, 2003; Mumolo, Nolich, & Vercelli, 2003; Seabra Lopes & Teixeira, 2000; Valin, Michaud, Rouat, & Letourneau, 2003)

Most of these applications can be categorised as either human-to-human communication (such as communication over the traditional telephone or data networks), or human-to-machine communication (such as communications with robots and computers). In human communication, the most natural and quickest form of high-level language is speech. Over the years, many people have been trying to extend this ability towards human-to-machine communication. However, this procedure has gained some progress only in the recently years. A more detailed discussion on using voice as input for human-to-machine communication can be found in the following literatures (Choen & Oviatt, 1995; Roe & Wilpon, 1994).

When a person's hands are busy and/or are unable to use them due to medical reasons, they could use speech to control the appliances. This is another attractive motivation to use speech communication as an interface to control appliances. For example, it is much safer for a driver to use his voice to dial the phone, rather than dial the phone by hand while driving. Recently, this application has received considerable attention due to the increased number of road accidents happened because the driver was preoccupied with the hand held devices. By using the hands-free technology, these incidents could

have been prevented. However, when acquiring speech from these adverse acoustic environments, the speech signal is more likely to be distorted by noise. As a result, under these circumstances the speech related applications fails to perform as expected.

One effective solution to improve the quality of the received speech signal is to use a microphone near the user, which requires the user to always wear or hold the microphone. However, using wearable microphones is impractical and is not desirable in many applications. Some examples of these situations include fast-food drive through outlets, un-manned service stations and voice pickup in large rooms (Flanagan, Johnston, Zahn, & Elko, 1985).

Directive microphones have been used to overcome this problem of wearable microphones. However, directive microphones capture the desirable speech as well as the background noise. As a result, the quality of the desired speech is degraded. Therefore, it does not achieve our objectives in adverse acoustic environments. To overcome these problems speech enhancement techniques can be used with the directive microphones to achieve a better performance.

Speech enhancement techniques can be divided into two main categories depending on the number of microphones used in the algorithm, they are single microphone system (Cole, Moody, & Sridharan, 1993; Scalart & Filho, 1996) and multi-microphone (microphone array) system (Cao & Sridharan, 1993; Van Compernolle & Van Gerven, 1995; Yan, Du, Wei, & Zeng, 2003). An overview of the available techniques for speech enhancement using single and multi microphone algorithms are given in the

following literatures (Lim, 1983; Ortega-Garcia & Gonzalez-Rodriguez, 1996; Van Compernolle, 1992b).

Single microphone systems have some limitations such as the interference has to be stationary and the input signal-to-noise ratio (SNR) over most of the frequency range has to be positive. (SNR is the ratio between the power of signal and the power of noise, and it is usually given in dB.) On the other hand, microphone array systems can handle non-stationary or very strong interference signals. By using more than one microphone, we can also obtain the spatial information such as the location of the acquired signal. Due to these reasons, this thesis will be focusing on the microphone array system to improve the acquired signal.

The microphone array techniques are a potential replacement for the wearable and the directive microphones to use in the speech related applications. This idea of using the microphone array system to improve the desired speech signal isn't new to this field. However, more affordable solutions became available only after the availability of the inexpensive digital signal processors (DSPs) during the 1980's. In the past decade, many literatures have been published on microphone array techniques and their applications by many authors, such as (Affes & Grenier, 1997; Brandstein & Ward, 2001; Campbell, 1999; Farrell, Mammone, & Flanagan, 1992; Fischer & Simmer, 1996; Kaneda & Ohga, 1986; McCowan, 2001a, 2001b; Silverman, 1987; Van Compernolle, 1990).

The most important objective of a microphone array is to provide a high quality version of the desired speech signal in an adverse acoustic environment. The microphone array

achieves this via beamforming techniques, which are designed to reduce the level of background noise signals, while minimising distortion to speech. A beamformer does spatial filtering by separating the desired signal and the interference signals that originate from different directions but have the same temporal frequency band.

Finding an optimal beamformer algorithm for a particular application depends highly on the available hardware and the computational resources. The Texas Instruments (TIs) TMS320C6711 DSP Starter Kit (DSK) is the hardware platform available for this project. Since the DSK supports only one input channel, the PCM3003 daughter card is used to obtain the two input signals received at the microphones. When choosing an optimal algorithm, more priority is given to less computational cost algorithms; this is due to the limited computational resources in the DSK.

Characteristics of the target application environment such as the type and the level of noise and reverberation are also equally important when choosing an algorithm for a particular application. At present, there is a broad range of speech related applications that require enhancement of speech. The interferences that degrade the quality of the speech signal for each of these applications may differ from application to application. For example, indoor applications may have interference coming from audio equipment, computer fan, etc and outside applications may have interference coming from vehicle engine, birds, etc. Therefore, it is impractical to suggest one beamformer algorithm that is generally applicable to all speech related applications.

This thesis will focus on a specific application in order to choose an optimal algorithm. However, with simple necessary modifications this implementation can work for any

speech related applications. One possible target application for this microphone array beamformer is the Massey University Smart House (Diegel et al., 2005). This smart house is designed for the disable or elderly people to give them independence, quality of life, and the safety they require.

One of the smart technologies in this house is the smart management system. This is a computer based software program called Jeeves, which is programmed to works as a virtual butler for the house. The basic idea behind this system is to interact with the occupants of the smart house, in order to receive commands to control the household appliances. A more detailed discussion about the functions of the smart house can be found in the following literature (Diegel et al., 2005), which is to be presented at the Home Oriented Informatics and Telematics (HOIT) conference in 2005 (copy of this paper is given in the Appendix D).

In order to make it easy as possible for the occupants of the smart house to interact with Jeeves, the medium of communication must be as natural as possible. A natural interface should allow the user to interact with Jeeves directly using their voice to request commands to perform some action. Therefore, a speech recognition system is required in order to analyse the speech and extract the necessary commands to control the household appliances. Commercially available speech recognition software (Visual Basic with Microsoft Speech SDK) is used in this project to interpret the human speech.

For a speech recognition system to work efficiently, it typically requires a SNR of greater than 20dB. In an ideal house, there will be a considerable amount of background noise propagating from different sources such as computer fan, radio, TV,

and other talkers. The speech signal acquired in this environment will be distorted by these background noises. This could lead to poor performance of the speech recognition system.

A simple solution is to use a wearable microphone to acquire the speech from the user. Since the smart house is designed for the disable and elderly people, it will not be convenient for the occupants of the house to use a headset every time they want to issue a command. Moreover, majority of the people wouldn't be fond of wearing a microphone in their house at all times. Therefore, a microphone array system (Chien, Lai, & Lai, 2001) that allows the users to move around freely and interact easily with Jeeves by using their voice is required. This thesis discusses the real-time implementation of such a microphone array beamformer on a DSP.

This thesis is organised as follows: *Chapter 2* discusses the historical context behind the noise reduction techniques and the speech detection algorithms. It gives an overview of the previous approaches done on this area and explains in-depth about the chosen algorithm for this project. The switched Griffiths-Jim beamformer algorithm was selected as the algorithm to be implemented on the TMS320C6711 DSK. Hardware and software implementation of the chosen algorithm in real-time is described in *Chapter 3*. This completed system is tested, and the results can be found in *Chapter 4*. The conclusions and future work are given in *Chapter 5*.