1

MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS
STANDARD REFERENCE MATERIAL 1010a
(ANSI and ISO TEST CHART No. 2)

National Library
of Canada

Bibliothèque nationale
du Canada

Canadian Theses Service

Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

**MODIFICATION OF SHORT-TIME FOURIER TRANSFORMED
SPEECH MAGNITUDE USING IMAGE FILTER OPERATORS**

by

RUSSELL F. BYERS, B.ENG.

A thesis submitted to the
Faculty of Graduate Studies and Research
in partial fulfilment of the requirements
for the degree of
Master of Engineering

Ottawa-Carleton Institute for Electrical Engineering
Faculty of Engineering
Department of Systems and Computer Engineering
Carleton University

April, 1991

© copyright

1991, Russell F. Byers

National Library
of Canada

Bibliothèque nationale
du Canada

Canadian Theses Service Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-68890-4

Canada

The undersigned hereby recommends to the Faculty of Graduate Studies and Research
acceptance of the thesis,

**"MODIFICATION OF SHORT-TIME FOURIER TRANSFORMED SPEECH
MAGNITUDE USING IMAGE FILTER OPERATORS"**

submitted by **RUSSELL F. BYERS, B.ENG.**, in partial fulfilment of the
requirements for the degree of Master of Engineering

Thesis Co-supervisor

Thesis Co-supervisor

Chairman,
Department of Systems and
Computer Engineering

Department of Systems and Computer Engineering
Faculty of Engineering
Carleton University
April, 1991

ABSTRACT

The speech spectrogram is derived using short-time Fourier analysis and represented as an image, with time and frequency as Cartesian x and y coordinates, and image luminance a function of magnitude. The representation of the spectrogram as a graphic image provides a basis and motivation for examining the feasibility of modifying speech in the frequency domain using image enhancement filters. This thesis is a qualitative investigation into the visual and aural effects produced when spectrograms of speech signals are modified using two-dimensional, finite impulse response filters.

Spectrograms of digitized, single word speech samples were modified using a selection of 2D spatial filters. Fourier synthesis was then applied to recreate speech samples using the modified spectrograms and original phase information as inputs. Visual effects produced in modified spectrograms, and aural effects produced in resynthesized speech samples are discussed.

ACKNOWLEDGEMENTS

I would like to extend my sincerest thanks to my thesis co-supervisors, Brian W. Tansley, Associate Professor of Psychology at Carleton University, and L. Robert Morris, Professor of Systems Engineering at Carleton University. Their longstanding patience and support made it possible to complete this work on a part-time basis.

I would also like to thank my employer, Northern Telecom Ltd. for the use of computer equipment which proved to be invaluable for data analysis and thesis paper preparation.

TABLE OF CONTENTS

1	INTRODUCTION.....	1
1.1	THE SPEECH SPECTROGRAM.....	1
1.2	SPECTROGRAMS AND SCENE IMAGES	4
1.3	THE SHORT-TIME FOURIER TRANSFORM.....	6
1.4	DIGITAL IMAGE ENHANCEMENT METHODS	8
1.5	RESEARCH OBJECTIVES.....	9
1.6	CONTRIBUTIONS.....	10
1.7	OUTLINE	10
2	SIGNAL PROCESSING METHODOLOGIES	11
2.1	OVERVIEW.....	11
2.2	STFT ANALYSIS, SYNTHESIS, AND MODIFICATION.....	12
2.2.1	Short-Time Fourier Analysis Equations.....	12
2.2.2	Application of the Sampling Theorem.....	13
2.2.3	The Overlap-Add Method of STFT Synthesis.....	14
2.2.4	Effects of STFT Spectrum Modifications.....	16
2.3	DIGITAL IMAGE ENHANCEMENT OPERATORS	17
2.3.1	Point Operators.....	18
2.3.2	Neighborhood Image Operators.....	20
3	SYSTEM DEVELOPMENT	27
3.1	OVERVIEW.....	27
3.2	SYSTEM HARDWARE.....	28
3.3	SOFTWARE DEVELOPMENT	29
3.3.1	Audio Sampling.....	29
3.3.2	STFT Analysis	30
3.3.3	STFT Data Storage Format	31
3.3.4	STFT Synthesis.....	33
3.3.5	STFT Magnitude Modification Algorithms	34
3.3.6	Conventions and Unit Conversions	35
3.3.7	Two dimensional FIR Filters.....	37
3.3.8	Non-Linear High-Pass Filter Effects	39

3.3.9	Simulated Grey Level Spectrogram Display.....	41
3.3.10	Three Dimensional Graphic Display.....	43
3.3.11	Time Domain Signal Processing.....	43
4	SPECTROGRAM MODIFICATION EXPERIMENTS.....	45
4.1	OVERVIEW.....	45
4.1.1	Selection of Speech Samples.....	46
4.1.2	Selection of Spatial Filters.....	47
4.2	OBSERVATIONAL CRITERIA.....	49
4.2.1	Visual Effects in Spectrograms.....	50
4.2.2	Auditory Effects in Synthesized Signals.....	51
4.3	ON THE USE OF SPECTRAL PRE-EMPHASIS.....	51
4.4	EXPERIMENT A: NK X 1 HIGH-PASS FILTERS.....	52
4.5	EXPERIMENT B: 1 X NS HIGH-PASS FILTERS.....	58
4.5.1	Filters Applied to Original Spectrograms.....	58
4.5.2	Filters Applied to Spectrograms with Pre-Emphasis.....	64
4.6	EXPERIMENT C: NK X NS HIGH-PASS BOX FILTERS.....	69
4.6.1	Filters Applied to Original Spectrograms.....	69
4.6.2	Filters Applied to Spectrograms with Pre-Emphasis.....	71
4.7	EXPERIMENT D: NK X NS LOW-PASS FILTERS.....	84
4.8	SUMMARY.....	88
5	CONCLUSIONS.....	91
5.1	OVERVIEW.....	91
5.2	GENERAL COMMENTS.....	91
5.3	FURTHER APPLICATIONS OF THE METHODS.....	93
	REFERENCES.....	95

LIST OF FIGURES

Figure 1.1	Example of a Speech Spectrogram.....	2
Figure 2.1a	3x3 Low-Pass Box Filter Kernel.....	23
Figure 2.1b	3x1 Low-Pass Box Filter Frequency Response.....	23
Figure 2.2a	3x3 High-Pass Box Filter Kernel.....	25
Figure 2.2b	3x1 High-pass box Filter Frequency Response.....	25
Figure 3.1	System Function	28
Figure 3.2	Three Dimensional View of a Spectrogram.....	36
Figure 3.3	Spectrogram of the Word "intercept"	42
Figure 3.4	Time-Signal Normalization Process	44
Figure 4.1	Pre-emphasis Weighting Function	52
Figure 4.2a	"Broad.." Original Spectrogram.....	55
Figure 4.2b	"..cast" Original Spectrogram.....	55
Figure 4.3a	"Broad.." Filter I Spectrogram	56
Figure 4.3b	"..cast" Filter I Spectrogram.....	56
Figure 4.4a	"Broad.." Filter II Spectrogram	57
Figure 4.4b	"..cast" Filter II Spectrogram.....	57
Figure 4.5a	"Broad.." Filter III Spectrogram	61
Figure 4.5b	"..cast" Filter III Spectrogram.....	61
Figure 4.6a	"Broad.." Filter IV Spectrogram	62
Figure 4.6b	"..cast" Filter IV Spectrogram.....	62
Figure 4.7a	"Broad.." Filter V Spectrogram	63
Figure 4.7b	"..cast" Filter V Spectrogram.....	63
Figure 4.8a	"Broad.." Spectrogram with Pre-emphasis	66
Figure 4.8b	"..cast " Spectrogram with Pre-emphasis	66
Figure 4.9a	"Broad.." Pre-emphasis & Filter IV	67
Figure 4.9b	"..cast " Pre-emphasis & Filter IV	67
Figure 4.10a	"Broad.." Pre-emphasis & Filter V	68
Figure 4.10b	"..cast" Pre-emphasis & Filter V	68
Figure 4.11a	"Broad.." Filter VIa Spectrogram.....	74
Figure 4.11b	"..cast" Filter VIa Spectrogram.....	74

Figure 4.12a	"Broad.." Filter VIb Spectrogram.....	75
Figure 4.12b	"..cast" Filter VIb Spectrogram.....	75
Figure 4.13a	"Broad.." Filter VIIc Spectrogram.....	76
Figure 4.13b	"..cast" Filter VIIc Spectrogram.....	76
Figure 4.14a	"Broad.." Pre-emphasis & Filter VIa Spectrogram	77
Figure 4.14b	"..cast" Pre-emphasis & Filter VIa Spectrogram	77
Figure 4.15a	"Broad.." Pre-emphasis & Filter VIb Spectrogram	78
Figure 4.15b	"..cast" Pre-emphasis & Filter VIb Spectrogram.....	78
Figure 4.16a	"Broad.." Pre-emphasis & Filter VIIc Spectrogram	79
Figure 4.16b	"..cast" Pre-emphasis & Filter VIIc Spectrogram	79
Figure 4.17	"Flashing" Filter VIb Spectrogram.....	80
Figure 4.18	"Flashing" Filter VIIc Spectrogram.....	80
Figure 4.19	"Flashing" Pre-emphasis & Filter VIb Spectrogram.....	81
Figure 4.20	"Flashing" Pre-emphasis & Filter VIIc Spectrogram	81
Figure 4.21	"Intercept" Filter VIb Spectrogram.....	82
Figure 4.22	"Intercept" Filter VIIc Spectrogram.....	82
Figure 4.23	"Intercept" Pre-emphasis & Filter VIb Spectrogram.....	83
Figure 4.24	"Intercept" Pre-emphasis & Filter VIIc Spectrogram	83
Figure 4.25a	"Broad.." Filter VIII Spectrogram	86
Figure 4.25b	"..cast" Filter VIII Spectrogram	86
Figure 4.26	"..Broad" 3D View of Original Spectrogram	87
Figure 4.27	"Broad" 3D View of Filter IV Spectrogram.....	87
Figure 4.28	"Broad" 3D View of Filter VIII Spectrogram.....	88

LIST OF TABLES

Table 4.1	Phonemic Content of Speech Samples.....	47
Table 4.2	Summary of Spatial Filter Characteristics	49
Table 4.3	Aural Effects of Filtering Original Spectrograms	70
Table 4.4	Aural Effects of Filtering Pre-emphasized Spectrograms.....	72

CHAPTER 1

INTRODUCTION

1.1 THE SPEECH SPECTROGRAM

Over the course of decades of speech research, investigators have sought to better understand speech through the use of graphic representations of the speech signal. Such representations provide the investigator with insights into different aspects of the nature of speech. Time domain representations of speech are the basis of some types of speech analysis: Pitch extraction, energy profiling, and autocorrelation functions are some examples [1]. Although the time vs. amplitude graphic representation of a speech signal contains the whole of the information in the signal, it conveys only a very weak notion of how signal energy varies in time with respect to its spectral content.

The speech spectrogram is usually represented as a two dimensional graphic display of the time-varying spectral characteristics of a speech signal. A spectrogram can be thought of as being composed of the output of a bank of narrowband filters [1,12]. The filter bank's input is the time signal, and its output is a set of component narrowband signals. A *horizontal* cross section of a spectrogram corresponds to the time-varying magnitude of the output of one filter, having some bandwidth Δf . A spectrogram can also be described in terms of a series of *vertical* cross sections. Each vertical cross section corresponds to the magnitude component of the discrete Fourier transform (DFT) of a *short*

segment of the time signal [1]. Filter banks and short-time Fourier transforms are two different implementations of spectrogram generation.

Figure 1.1 is an example of a spectrogram. Signal frequency changes along the y axis, and time changes along the x-axis in the figure. The spectrogram shown in the figure was created by a sound spectrograph.

Figure 1.1 Example of a Speech Spectrogram¹

Spectrograms first appeared in the early 1940's, and until more recently, were produced using a device known as the *sound spectrograph* [2,3]. With the evolution of modern computers and digital signal processing methods, the short-time Fourier transform emerged as the primary means of generating spectrograms from time signals, and the cathode ray tube (CRT) terminal became the device of choice for displaying spectrograms [4,5]. Nowadays, spectrograms can be generated and displayed interactively, using a high performance microcomputer with a graphic display device[6].

¹ Figure 1.1 was taken from *Visible Speech*, reference [7].

The spectrogram shown in figure 1.1 contains a number of the visual structures common to speech spectrograms in general. Spectral energy is proportional to darkness in the format used for this figure. Narrower bands of energy which run primarily in the direction of the horizontal axis correspond to *formants*, or resonances of the vocal tract in voiced speech sounds. Formant locations are, in fact, determined by the time-varying characteristics of the human vocal tract: Control of the vocal tract originates in the brain, and is effected by the muscles of the articulatory apparatus. Phonemes classified as *vowels*, *semi-vowels*, *nasals*, and *diphthongs* are typically distinguished by three or four formants occupying distinct frequency bands [1,7]. The different phonemes found in these classes are differentiated from one another by the frequencies and bandwidths of their constituent formants. When a formant is horizontal it conveys an unchanging vocal tract section, whereas a formant which is rising or falling on the frequency axis over time indicates a changing cross section. The latter is known as a formant transition. Diphthongs and semi-vowels are differentiated from simple vowels in that they characteristically contain formant transitions.

Phonemes classified as *fricatives* and *affricates* appear in a spectrogram as aperiodic, broadband patterns of energy. A voiced fricative contains a formant structure in addition to the wideband characteristic. The "s" sounds in figure 1.1 are examples of unvoiced fricatives. The class of phonemes known as *stops*, or *plosives* is more complex in nature. A stop sound is produced when air pressure is built up in the mouth, then suddenly released as a burst of energy. Stops correspond to consonants such as *b*, *d*, *g*, *k*, *p*, or *t*, and are usually preceded by or followed by a vowel sound. Stops generally consist of combinations of wideband structures and formants. In an unvoiced stop, the formants may appear before or after the energy burst. In a voiced stop, the formants appear more nearly simultaneous with the energy burst.

From the early days of the spectrogram, efforts have been made to teach humans to recognize speech by visual examination of the patterns of speech. The book *Visible Speech*, published in 1947, outlines the principles of spectrogram reading in detail [7]. Experiments carried out by Zue, Cole and others have verified that humans can learn to recognize speech by this method [8, 9]. In these experiments, observers were trained to make associations between the visual patterns found in spectrograms and corresponding samples of aural speech, which consisted of words or phrases. Following a period of training, the observers were shown unfamiliar spectrograms of speech, and asked to effectively "recognize" the speech content. The success rate of recognition by this method was largely dependant on the amount of effort spent on training. These studies suggest that if the learning effort invested in spectrogram reading was comparable to that of a child learning to understand speech aurally, that vocabularies of comparable size could be developed. The results of these experiments support the notion that there are unique mappings between the visually perceived morphological features of a speech spectrogram, and the aurally perceived patterns of speech.

1.2 SPECTROGRAMS AND SCENE IMAGES

As noted earlier, the speech spectrogram is displayed such that the frequency index is mapped to the y-axis, time index is mapped to the x-axis, and image luminance is a function of magnitude in a two dimensional Cartesian coordinate system. In this respect, a spectrogram has much in common with other types of images. All images contain patterns which vary with respect to shape, size, and contrast. The coarser aspects of the feature morphology found in speech spectrograms can be suggestive of patterns found in some types of scenes. For example, the "ridges" which comprise formant tracks in speech spectrograms resemble ranges of mountains as might be found in an aerial view of terrain.

Speech spectrograms generally consist of segments of formant tracks interspersed with other patterns. In the science of digital image processing, filters are used to enhance the perceptually recognizeable features found in images by manipulating the amplitude of their constituent picture elements, or pixels. The results produced by image enhancement techniques are often dramatic. For example, obscure features buried within an image can be made to stand out in stark contrast to their surroundings when an appropriate image filter is applied. Another example of image enhancement is the systematic removal of Gaussian noise using two dimensional smoothing filters.

Some examples of digital image processing methods applied to speech spectrograms can be found in the literature. Zue [10] used a Gaussian classifier to pre-process spectrograms which were then fed into an automated spectrogram-based speech recognition algorithm. Several other more pertinent examples of image processing on spectrograms are discussed in subsequent sections.

Thus far, the discussion has been focused on the speech spectrogram and its characteristics as an image. The motivation for this thesis is based on some questions which are related to both speech processing and image processing. These questions are the following:

- (i) If we take a spectrogram, modify it using digital image processing techniques, then synthesize a time domain signal using the modified spectrogram, what will be the aural effects observed?
- (ii) How do visually perceived changes in spectrograms relate to aurally perceived changes in speech synthesized from them?

(iii) Is this a viable means of speech enhancement?

Prerequisite to an investigation of these questions is the availability of a system capable of (1) generating spectrograms from sampled speech, (2) applying image processing methods to spectrogram images and (3) recreating speech signals from modified spectrograms.

1.3 THE SHORT-TIME FOURIER TRANSFORM

The short-time Fourier transform (STFT) is an identity transform [1,13] in that, if a time signal is transformed into a series of STFT's by analysis, that same time signal can be recreated exactly using STFT synthesis. The identity property of the STFT provides some advantages. Specifically, when an STFT is modified and synthesized to recreate a time signal, the changes in the resulting time signal are due to the modifications performed on the STFT, and are not artifacts of the analysis-synthesis process. Details of the theory and methods used in STFT analysis and synthesis are given in more detail in chapter 2.

The development of signal analysis and synthesis methodologies based on the discrete Fourier transform (DFT) took place largely in the 1960's and 1970's, in step with the development of the modern digital computer, and more recently the DSP microprocessor. In 1973, Schafer and Rabiner [11] published a paper describing a speech analysis-synthesis system based on the STFT, which was evolved from earlier filter bank models used to describe the process. Other publications which followed produced further refinements on STFT based analysis-synthesis. In 1975, Portnoff [12] described an identity analysis/synthesis system based on the STFT. This system was implemented using the fast Fourier transform (FFT) to improve computational efficiency. Allen [13]

outlined the prerequisites for spectral modification of STFT's in 1977, and in 1980, Crochiere [14] developed a technique for STFT analysis-synthesis known as the weighted overlap-add method, which further improved computational efficiency. Griffin and Lim showed, in 1983, that STFT synthesis could be achieved in the absence of the phase component of the signal by means of an iterative signal estimation algorithm [15].

One of the proposed applications for the STFT was to perform spectral modification as a means of manipulating the time-domain signal indirectly, as indicated in the writings of Allen [13] and others. Following the development of STFT analysis-synthesis methodologies, works on the modification of STFT magnitude have been produced by a number of individuals. Portnoff [16] described how the time scale of speech could be altered using STFT analysis and synthesis. A method for frequency scale compression by STFT modification was investigated in a thesis project by Green [17] at Carleton University.

A number of published papers outline methods for removing noise from speech by modifying the STFT magnitude. Spectral subtraction [18] is one of the better known methods found in the literature. Another is spectral amplitude compression, in which the STFT magnitude of each datum point is modified according to a non-linear transfer characteristic [19,20]. The general objective is to suppress low amplitude spectra using the transfer characteristic, removing noise from areas of the spectrogram where there is no appreciable speech energy. Both of these methods have had limited success at removing noise from the signal, and tend to break down when the noise level is high, producing artifactual effects known as *tonals*, which result from partial removal of the noise energy.

A more elaborate technique of enhancing noisy speech by modification of the STFT magnitude was published by Ariki, Kajimoto, and Sakai in 1986 [21]. An interesting aspect of this work is that it makes use of a two dimensional convolutional smoothing filter

to modify spectral magnitude data. This spatial filter and its application as outlined in the paper are markedly similar to smoothing filters used to remove high frequency noise from digital images of scenes. The paper by Ariki et al. predates this thesis.

1.4 DIGITAL IMAGE ENHANCEMENT METHODS

Digital image processing science makes use of an assortment of operators to enhance images. These generally fall into two classes, point operators and neighborhood operators.

Point operators are functions in which each image pixel is transformed according to a function which uses the pixel's original value as the input variable. A mapping function specifies the relationship between the input and output values over the dynamic range of the pixels. A *contrast stretch* is a commonly used point operator. Its effect is to spread out the intensity distribution of concentrations of pixels on the intensity histogram so that a greater part of the available dynamic range is exploited. The overall result is greater intensity contrast, and better definition of features in the image.

Neighborhood operators are also known as two dimensional finite impulse response (FIR) filters. In the literature, two of the configurations most frequently mentioned are the low and high-pass box filters [22,23]. In a low-pass box filter, all of the coefficient values in the impulse response are unity. Hence, the transformation is simply an average of pixels in a neighborhood about the central pixel. When convolved with an image, a low-pass filter can remove high frequency noise components such as "snow"².

2. The term "snow" refers to the additive Gaussian noise component in video reception. It generally arises from a low signal to noise ratio in the received signal.

The high-pass box filter is also referred to as an edge enhancement filter. As the name suggests, it provides greater definition to high frequency image features such as edges.

Some of the works on STFT magnitude modification cited thus far employ techniques which are markedly similar to those used to modify digital images. Spectral amplitude compression systems as proposed by Munday [19], and Eger et al. [20] make use of techniques very similar to the point operator. The noise reduction system described by Ariki et al. [21] uses a filter similar to the low-pass box filter to perform smoothing of spectral magnitude variations.

1.5 RESEARCH OBJECTIVES

The primary objective of this thesis is to investigate the effects produced by application of digital image processing operators to speech spectrograms generated using the STFT. Consideration is to be given to the effects produced in visual representations of the STFT spectrogram, and to effects in the speech signal produced by Fourier synthesis of the modified STFT spectrogram. Correlation, and comparison of the visually perceived effects of spectrogram modification with the aural effects produced in the synthesized signal are of substantial importance to the work. The entire focus of the investigation is on modification of the STFT magnitude, and its effects. The phase component of the signal is preserved in every case examined.

The speech material used for the investigation consists of limited selection of single word utterances. This constraint was necessary so that processing times would be reasonable for the compute-intensive nature of the numeric processing required. Overall, the investigation is qualitative in nature. It does not attempt to quantify the results via

formal psychophysical testing. The aim is to obtain some preliminary indications of the potential usefulness of modifying speech using image enhancement methods.

Algorithms developed to carry out the objectives are implemented so that they will run on an IBM PC compatible computer with a standard math coprocessor. No consideration was given to the feasibility of performing the types of speech modification outlined in real time.

1.6 CONTRIBUTIONS

The primary contribution of this work to speech processing is that it provides some insight into the feasibility of enhancing speech by the direct application of image enhancement methods to spectrograms. It also gives indications of the types of results one can expect as spatial filter parameters are varied. A discussion of the problems associated with the application two-dimensional spatial filters, as they relate to spectrograms, is also an original contribution.

1.7 OUTLINE

Chapter 2 outlines the signal processing methodologies which form the basis of this work. Chapter 3 describes in detail the system development and implementation of the signal processing methodologies. The experimental results obtained through application of the algorithms to samples of speech utterances are outlined in chapter 4. A discussion of the significance of the experimental results obtained and the conclusions comprise chapter 5.

CHAPTER 2

SIGNAL PROCESSING METHODOLOGIES

2.1 OVERVIEW

In this chapter, the signal processing theory and methodologies which form the basis for a speech processing system are outlined. Topics discussed are drawn from two separate areas of signal processing. These are digital image processing, and short-time Fourier analysis-synthesis of speech. The discussion converges on the speech spectrogram, which has the dual characteristic of being simultaneously the output of STFT analysis, and a digital image.

The first part of the discussion describes a set of STFT analysis and synthesis methods in mathematical terms, and identifies the requirements for sampling in frequency and time.

The second part of the discussion deals with image processing, with a focus on point and neighborhood operators. It is shown that in certain other related investigations of speech processing, these types of image processing methods have been applied to speech spectrograms, probably without the respective investigator's awareness of the commonalities between speech and image processing which were exploited.

2.2 STFT ANALYSIS, SYNTHESIS, AND MODIFICATION

In this section, the short-time Fourier transform analysis/synthesis method is outlined in brief. A detailed explanation of the theory can be found in Rabiner and Schafer [1], as well as in previously cited papers on STFT analysis-synthesis, and a description of an implementation of Crochiere's weighted overlap-add method is given in Green [17].

2.2.1 Short-Time Fourier Analysis Equations

The discrete short-time Fourier transform X_k of a signal is defined in the following equation.

$$X_k[n] = \sum_m w[n-m] x[m] e^{-j2\pi km/M}, \quad k = 0 \dots M-1 \quad (2.1)$$

In equation 2.1, $x[m]$ is the discrete time domain signal, $w[n-m]$ is the analysis window, n is the discrete time index, k is the frequency index, and M is the number of frequency points in the transform. The transformation of a time domain signal into a series of STFT frames is achieved by sliding the analysis window along the signal by some interval, and generating the transform at each window position. This can be expressed as,

$$X_k[sR] = \sum_m w[sR-m] x[m] e^{-j2\pi km/M} \quad (2.2)$$

R is the step interval for the analysis window expressed as a number of discrete points, and s is the frame index such that $s = 0, 1, 2, \dots$

2.2.2 Application of the Sampling Theorem

It can be shown that the original signal $x[n]$ can be recovered from $X_k[sR]$ in the case where $s=R$ (zero overlap condition), provided no modifications are made to X_k . However, if a time domain signal $y[n]$ is to be synthesized using a modified version of $X_k[sR]$ as input, then it is necessary to satisfy the requirements of the sampling theorem such that no aliasing occurs in $y[n]$.

Requirements For Sampling in Time

Using the filter bank interpretation of the Fourier Transform for a fixed frequency ω , $X_n[e^{j\omega}]$, the DFT of $x[m] \cdot w[n-m]$ can be shown to be the output of a filter with impulse response $w[n]$, whose Fourier transform is $W[e^{j\omega}]$. For most windows used, $W[e^{j\omega}]$ is a low-pass filter. If we let B represent the effective bandwidth of the window, the Nyquist criterion requires that $X_n[e^{j\omega}]$ be sampled at a rate of at least $2B$ to avoid aliasing. This is the requirement for sampling in time, as it determines the frame rate to be used in the discrete domain.

Requirements for Frequency Sampling

The inverse of $X_n[e^{j\omega}]$ is $x[m] \cdot w[n-m]$. If $w[n-m]$ is of length L samples, then so must be $x[n]$, given that the window is non-zero over $n=0..L-1$. Hence, $X_n[e^{j\omega}]$ must consist of at least L samples or discrete values of ω over the range 0 to 2π .

$$\omega_k = 2\pi k/L, \quad k = 0, 1, 2, \dots, L-1 \quad (2.3)$$

Sampling With the Hamming Window

The discrete-time response of the Hamming window function [25] is defined as,

$$\begin{aligned} w[n] &= 0.54 - 0.46 \cos(2\pi n/(L-1)), \quad 0 \leq n \leq L-1; \\ &= 0, \quad \text{otherwise.} \end{aligned} \quad (2.4)$$

The window has a 40 dB bandwidth of approximately,

$$B_{40\text{dB}} = 2F_s/L \quad (2.5)$$

F_s is the sampling frequency of the time signal, and L is the window length. To satisfy the requirements of time sampling, $X_n[e^{j\omega}]$ must be sampled at a rate of $2B=4F_s/L$. In other words, $X_n[e^{j\omega}]$ must be sampled at least four times over the length of the window. Expressed in terms of the frame rate R of the STFT, $R = L/4$ is the critical frame step size. Finally, to satisfy the requirements of sampling in frequency, the value of M (number of frequency points) in the STFT must be greater than or equal to the length of the Hamming window, L .

2.2.3 The Overlap-Add Method of STFT Synthesis

To reconstruct a signal from a series of short-time Fourier transform frames by the method of overlap-addition, the following transformation is applied.

$$y[n] = \sum_s \{1/N\} \left\{ \sum_k X_k[sR] e^{j2\pi kn/N} \right\} \quad (2.6)$$

The reconstructed time signal $y[n]$ is a sum of overlapping inverse short-time Fourier transforms $y_r[n]$. In equation 2.7, the equation for $y[n]$ has been rewritten in terms of the original, windowed time signal.

$$y[n] = \sum_s y_s[n] = \sum_s x[n] w[sR-n] = x[n] \sum_s w[sR-n] \quad (2.7)$$

it can also be shown that when $X_k[sR]$ is properly sampled in time,

$$\sum_s w[sR-n] = W[e^{j0}]/R \quad (2.8)$$

A detailed derivation of this result can be found in [1]. Combining equations 2.7 and 2.8, it becomes clear that $y[n]$ is identical to $x[n]$ by a constant multiple as shown in equation 2.9.

$$y[n] = x[n] W[e^{j0}]/R \quad (2.9)$$

The summation indices used to derive $y[n]$ in equation 2.6 can be exploited by computer algorithms to increase computational efficiency. An efficient algorithm for implementation of overlap-add synthesis has been described by Crochiere[14].

2.2.4 Effects of STFT Spectrum Modifications

It has now been demonstrated that analysis and synthesis based on the STFT is an identity process, provided it meets the sampling criteria set out above. The next step is to examine the effect of spectral modifications on the output of synthesis in mathematical terms.

Fixed in Time Modifications

Equation 2.10 shows the effect of a spectral modification which varies in frequency, but is constant in time. In practice, this is a weighting function which is applied to each STFT frame in the entire sequence.

$$y'[n] = \sum_s \{ 1/N \} \left\{ \sum_k X_k[sR] P_k e^{j2\pi kn/N} \right\} \quad (2.10)$$

In equation 2.10, P_k is the spectral weighting function and $y'[n]$ is the time signal computed using the modified spectrum as input. Manipulation of equation 2.10 can be used to show that $y'[n]$ is a convolution of $x[n]$ with the time domain response $p[n]$ as shown in equation 2.11

$$y'[n] = W\{e^{j0}\}/R \cdot \{ x[n] * p[n] \} \quad (2.11)$$

A convolution of two sequences of length N produces a sequence of maximum length $2N-1$. Hence, when spectral modifications of this type are performed in the frequency domain, provisions must be made to prevent aliasing. Anti-aliasing can be

realized by padding the window function with zeros to reduce its characteristic length in relation to the length of the STFT [13].

Varying in Time Modifications

The effect of a spectral modification $P_k[sR]$, which varies in time is shown in equation 2.12.

$$y'[n] = \sum_s \{1/N\} \left\{ \sum_k X_k[sR] P_k[sR] e^{j2\pi kn/N} \right\} \quad (2.12)$$

$$y'[n] = \sum_l x[l] \sum_s w[sR-l] p_s[n-l] \quad (2.13)$$

A substitution of variables in equation 2.12 produces the result shown in equation 2.13. A significant characteristic of the convolution of the window function with a time varying modification function is that the modification function is bandlimited by the window, but is not time-limited by it.

2.3 DIGITAL IMAGE ENHANCEMENT OPERATORS

This section contains a discussion of some of the more common operators used to enhance digital images. These are divided into two categories, point operators, and neighborhood operators or two dimensional FIR filters. It will be shown that some recent works on STFT-based speech processing employ methods which are very similar to some

of these image operators. This overlap of methods between diverse areas of signal processing, and their mutual application forms part of the basis of this work.

2.3.1 Point Operators

A point operator is a function which transforms each image pixel magnitude value as a function of its original value as shown in equation 2.14.

$$x'[i,j] = f \{ x[i,j] \} \quad (2.14)$$

The point transformation function $f\{x\}$ may be a linear or nonlinear function. A frequently used point operator is the *contrast stretch*, in which the pixel intensity values are remapped over the available range of discrete magnitude to enhance low contrast features. A contrast stretch function may be applied globally to the entire range of pixel intensities found in the image, or locally to a subrange of intensities. The enhancement effects produced by the contrast stretching images are often dramatic, bringing out features which were previously not discernable. Various methods exist for determining the mapping function of the contrast stretch. Some are based on the histogram the of discrete intensity distribution, maximizing contrast over input intensity subranges occupied by the greatest numbers of pixels, and reducing contrast over intensity subranges which are sparsely populated.

De facto examples of point operators applied to speech spectrograms can be found in Murday [19] and Eger et al. [20]. In each of these papers, a system for modifying the STFT magnitude component using a point transformation function is described. Equation 2.15 shows the function employed by Eger et al.

$$x'_k = \{ (x_k/\alpha)^N / (1 + (x_k/\alpha)^N) \} \cdot x_k \quad (2.15)$$

In equation 2.15, x'_k is the output spectral magnitude, x_k is input spectral magnitude, k is the DFT frequency index, α is the threshold magnitude, and N is the power index.

The effect of this function is to suppress spectra (pixels) having magnitudes less than α , while passing those greater than α without change. The value for N determines the shape of the transition region near α . Munday's amplitude transformation is somewhat similar, but more sophisticated in that it varies the mapping function over subranges of the DFT spectrum to optimize the output with respect to the location of formants in speech. The purpose of the transformation is to remove noise from the speech signal component over the entire spectrum. This is obviously somewhat different from a contrast stretch operation in its intent, but the mechanics are identical.

The works of Munday and Eger et al. are highly relevant to this investigation, because their approaches to speech enhancement are quite similar in many respects to the speech modification method to be described in the following chapters. In each case, windowed STFT analysis was used to generate the frequency domain image, and overlap-add synthesis was employed to synthesize modified speech. Only STFT magnitude is modified, and the original phase information is reused in synthesis.

Spectral subtraction, as described in reference [18], is another means of STFT amplitude modification in which a point operator is applied. The objective of spectral subtraction is also to remove noise from speech. The method only really differs from those described in Munday and Eger et al. in that it is a different type of point operator function. The algorithm is essentially the subtraction of a constant value from each pixel, with

truncation of negative values to zero. In terms of image processing, spectral subtraction is approximately equivalent to making the entire image slightly darker. Truncation of negative-going pixels to zero is a non-linear transformation and is similar to introducing a veiling luminance to the image.

2.3.2 Neighborhood Image Operators

Neighborhood operators are also known as two dimensional FIR filters, and as point spread functions. Equation 2.16 is the general form of a two dimensional FIR. Each pixel value $I'[i,j]$ is a weighted sum of itself and its neighboring pixels as defined by the filter impulse response. The filter impulse response, or *kernel* is convolved with the image.

$$I'[i,j] = \{G / (m \cdot n)\} \left\{ \sum_{k=1}^{k=m} \sum_{l=1}^{l=n} Q[k,l] \cdot I[[i-m/2], [j-n/2]] \right\} + C \quad (2.16)$$

In equation 2.16, I is an array of pixel values in the neighborhood of the central pixel $I[i,j]$, m and n are the dimensions of the spatial filter kernel in the respective Cartesian coordinates i and j , Q is an array of filter kernel coefficients, G is a gain constant, and C is an additive offset constant.

In the following sections, two configurations of the two dimensional FIR are described. One is a high-pass, or edge enhancement filter. The other is a low-pass, or averaging filter. The design of these filters is optimized only with respect to computational efficiency. In each configuration, the filter impulse response consists primarily of positive

or negative unity values. Unity-valued impulse response filters are widely used in digital image processing because they are computationally efficient [22,23]. They are not optimized with respect to stopband rolloff slope, passband ripple, or sidelobe attenuation in their frequency response characteristics. In this investigation, most of the spatial filters applied are based on unity-valued impulse response kernels because they were found to be satisfactory for the purposes of a qualitative analysis. Some experimentation made use of filters based on other impulse response derivations, such as the window design method [25].

In the foregoing discussion about point operators, a number of examples were given to demonstrate that these operators were applicable to STFT spectrograms of speech. With regard to neighborhood operators, such examples are not so easily found. One notable example in the literature is described in Ariki et al. in reference [21]. The paper outlines a method named TDSS (two dimensional spectral smoothing). The TDSS function is a two dimensional low-pass FIR filter. Its purpose is to enhance the formant structure of speech corrupted by additive white noise.

The Low-pass Box Filter

Figure 2.1a. shows the impulse response of a typical two dimensional low-pass box filter kernel, and figure 2.1b. shows its frequency response characteristic along either Cartesian coordinate i or j . The mathematical expression for convolution of an image with this filter is given in equation 2.17. Its impulse response is simply an average of all points about $I(i,j)$ in a neighborhood defined by the filter dimensions m and n .

$$I[i,j] = \{G / (m \cdot n)\} \left\{ \sum_{k=1}^{k=m} \sum_{l=1}^{l=n} I[|i-m/2|, |j-n/2|] \right\} + C \quad (2.17)$$

The effect of the low-pass box filter is to remove high spatial frequency information such as Gaussian noise (speckle) from images. The filter bandwidth can be specified by the two dimension parameters n and m . The dimension parameters are always odd integers so that the filter impulse response will be symmetrical about its central point $[(m/2+1), (n/2+1)]$. In the general case, where n does not equal m , the bandwidth varies with respect to direction about the center of the filter kernel.

1	1	1
1	1	1
1	1	1

Figure 2.1a 3x3 Low-Pass Box Filter Kernel

Figure 2.1b 3x1 Low-Pass Box Filter Frequency Response

The High-Pass Box Filter

The high-pass box box filter is a simple form of high-pass FIR, in which all points in the impulse response are equal to -1 with the exception of the central point, as shown in equation 2.18. In an undiluted high-pass box filter, the central point is given a value such that the sum of all elements in the impulse response is 0.

$$I[i,j] = \{G/(m \cdot n)\} \cdot \left\{ (W+1) \cdot I[i,j] - \sum_{k=1}^{k=m} \sum_{l=1}^{l=n} I[|i-m/2|, |j-n/2|] \right\} + C \quad (2.18)$$

The symbol W is a weighting parameter used to determine the strength of the filter. If W is equal to the sum of all negative coefficient values in the impulse response, then the filter is strictly a high-pass filter. If the value of W is greater than the sum of all negative weights, then the high-pass effect is reduced, and the filter is no longer high-pass in a strict sense, but is a high-boost filter. The filter strength is sometimes expressed as a parameter K where,

$$K = (W + 1) / (m \cdot n) \quad (2.19)$$

Figure 2.2a shows the impulse response of a two dimensional high-pass box filter, and figure 2.2b shows its frequency response characteristic.

-1	-1	-1
-1	8	-1
-1	-1	-1

Figure 2.2a 3x3 High-Pass Box Filter Kernel

Figure 2.2b 3x1 High-Pass Box Filter Frequency Response

Boundary Conditions for Neighborhood Operators

In real images, a boundary condition must be applied at all edges, because the filter kernel effectively "hangs over the edge" when it is applied to pixels near the edge of the image. This occurs whenever the distance between image pixel $I[i,j]$ and the edge of the image is less than $n/2$ or $m/2$, where m, n are the filter kernel dimensions. In practice, a

border around the edge of the image is usually left unfiltered. For a modest sized filter kernel, this amounts to a border of only a few pixels, which does not have a great overall effect.

CHAPTER 3

SYSTEM DEVELOPMENT

3.1 OVERVIEW

This chapter describes a system that was developed for the purpose of performing STFT analysis, spectral modification, and synthesis. This system consists primarily of a set of programs composed in a high level programming language, and which run on an IBM PC-compatible computer. These programs are implementations of the signal processing methodologies described in the previous chapter. A functional diagram of the system is shown in figure 3.1. The following sections describe the functional blocks in detail, and discuss the selection of parameters used for speech modification experiments.

Figure 3.1 System Function

3.2 SYSTEM HARDWARE

The development system consists of an IBM-XT compatible configured with a Hercules monochrome graphics card, 640 kilobytes of memory, and a 20 megabyte hard disk. To improve data throughput for numeric processing, an 8087 coprocessor was installed, although the software developed was capable of running with or without it. For

audio sampling, a Data Translation DT2823 analog I/O card was installed. This peripheral provides 16-bit resolution in the input and output analog/digital conversion channels.

To fulfill the requirements for anti-alias filtering of the analog to digital (A/D) input channel and sampling clock removal from the D/A output channel, analog filters having a 40 dB stop band cutoff frequency of 4000 Hz. and a 3 dB passband cutoff frequency of 3400 Hz, and a unity response over the range 0 to 3400 Hz. were connected to the Data Translation peripheral. This configuration made it possible to sample analog signals at a rate of 8000 samples per second.

3.3 SOFTWARE DEVELOPMENT

The overall strategy for software development was to implement each of the functional blocks required to create the system shown in figure 3.1 as a collection of independent programs. Audio data in the form of time samples or Fourier transformed frequency samples were stored in binary formatted files. Each program required one or more input files of specific data types and produced output in the form of files or graphically displayed information. A 100 byte file header contained all information each program needed in order to be able to determine the type, format and length of data contained in the file.

3.3.1 Audio Sampling

Given the limited data storage and processing rates of the PC, it was necessary to minimize the speech sample rate in order to maintain reasonable processing capacity. A sampling rate of 8000 samples per second was chosen because it met the minimal sampling

requirements for digitization of speech. Taking into account the analog system components, the effective sampled frequency range was d.c. to 3400 Hz.

3.3.2 STFT Analysis

The STFT analysis algorithm used in this work is a slightly modified version of a Fortran implementation listed in Appendix A of reference [17]. The original implementation runs on a PDP-11. All necessary changes were made so that the program would compile in Microsoft Fortran 3.3 and run on a PC-compatible. A Hamming window option was also installed in place of the Kaiser window in the original program.

Four parameter values were required to specify the STFT analysis characteristics. These are,

- (i) time sampling window type
- (ii) time sampling window length
- (iii) DFT length
- (iv) DFT sample interval.

A low-pass Hamming window was selected for STFT analysis because it was considered to be a good compromise in terms of its frequency resolution and sidelobe attenuation characteristics [25].

Following some experimentation, a DFT length of 256 points was chosen, because it provided a sufficiently fine resolution in frequency, without requiring excessive computing resources, as would have been required to generate higher resolution DFT's. A window length of 256 points sampled at 8000 samples per second has a time length of 32

milliseconds. Because changes in the phonemic content speech occur with a minimum length of 10 to 20 milliseconds [1], a shorter analysis window length was required. The chosen analysis window was of length 128 points, which is equivalent to 16 milliseconds. The remaining 128 points of the analysis input were set to zero. Zero-padding of the analysis input increases the frequency sample density of the DFT, and decouples the analysis window length from the DFT length. This is equivalent to over-sampling in the frequency domain. As discussed in chapter 2, over-sampling in frequency is a means of preventing aliasing effects which can result from spectral modifications.

In chapter 2, it was stated that the sample interval R , required to satisfy the sampling-in-time requirement of the STFT is $L/4$, where L is the analysis window length of a Hamming window. Application of this formula to the 128 point window gives a value of $R=32$ samples, which is equivalent to 4 milliseconds.

3.3.3 STFT Data Storage Format

To reduce the data storage and handling requirements of the STFT generated by analysis, two methods were applied. One reduced the amount of data to be stored on disk, and the other separated the STFT data into magnitude and phase components.

Data reduction was achieved by exploiting conjugate properties of DFT's derived from real input $x[n]$. Each point X_k on the upper half of the complex Z-plane for $k=1$ to $N/2 - 1$, has a corresponding complex conjugate X_{N-k} on the lower half of the plane. The relationship used to extract the real and imaginary parts of points on the lower half of the Z-plane from points on the upper half is shown in equation 3.1. For complex valued

variables, the symbol *Re* denotes the real part, and the symbol *Im* denotes the imaginary part.

$$\text{Re}(X_{N-k}) = \text{Re}(X_k), \quad k=1, 2, \dots, N/2 - 1 \quad (3.1a)$$

$$\text{Im}(X_{N-k}) = -\text{Im}(X_k), \quad k=1, 2, \dots, N/2 - 1 \quad (3.1b)$$

Hence, it is not necessary to store any of the points on the lower half of the *Z*-plane. These data points can always be recovered by taking the conjugate of points X_k , with k between 1 and $N/2-1$. Only the first $N/2+1$ points of each STFT frame are stored in the output file.

Spectral modifications in this work are applied only to the magnitude component of the STFT spectra. This permitted a further step to be taken to optimize STFT data storage and handling. The magnitude and phase components of each analysis output datum point were separated and written to different output files. Each output file consisted of a header followed by an array of data points in binary 16-bit integer format. Equations 3.2a and 3.2b were used to extract magnitude and phase components from complex X_k values, which make up the DFT.

In the case of phase, taking the arctangent of X_k yielded a phase value between $-\pi/2$ and $\pi/2$ radians. The signs of the real and imaginary parts of X_k were then used to determine the quadrant of origin on the complex plane. The value π was then added or subtracted as required to map the phase angle onto the correct quadrant in the range $-\pi$ to π radians. Finally, the phase value was multiplied by a scaling factor such that the range $-\pi$ to π occupied the full 16 bit integer range of -32768 to 32767.

$$\text{Magnitude}(X_k) = \{(\text{Re}(X_k))^2 + (\text{Im}(X_k))^2\}^{1/2} \quad (3.2a)$$

$$\begin{aligned} \text{Phase}(X_k) &= \arctan\{\text{Im}(X_k) / \text{Re}(X_k)\} + \pi, \text{ if } \text{Im}(X_k) > 0 \text{ and } \text{Re}(X_k) < 0; \\ &= \arctan\{\text{Im}(X_k) / \text{Re}(X_k)\} - \pi, \text{ if } \text{Im}(X_k) < 0 \text{ and } \text{Re}(X_k) < 0; \\ &= \arctan\{\text{Im}(X_k) / \text{Re}(X_k)\}, \text{ otherwise} \end{aligned} \quad (3.2b)$$

3.3.4 STFT Synthesis

As was the case for STFT analysis, the STFT synthesis program was borrowed from Appendix A of reference [17] and modified for use with a PC compatible. The synthesis program is an implementation of overlap-add synthesis as described in chapter two.

Modifications to the synthesis program were required so that it would accept a STFT phase data file and a STFT magnitude data file as input, and read both files simultaneously, combining the phase and magnitude to form the complex valued DFT points X_k for $k=0$ to $k=N/2+1$, then deriving the conjugate points X_k for $k= N/2+2$ to $N-1$. This is simply the reverse of the steps taken in STFT analysis to generate the STFT magnitude and phase data files.

The output of STFT synthesis is a time domain audio file in the same file format as the input audio file. If synthesis is performed on the unmodified STFT, then the resulting audio is identical within a scale factor, provided a sufficient pad of silence exists at the beginning of the input time domain file to accommodate startup transients characteristic of the synthesis algorithm. To synthesize audio using a modified STFT magnitude data file as

input, the data file containing the modified STFT magnitude and the original STFT phase data file are specified as inputs to the synthesis program.

3.3.5 STFT Magnitude Modification Algorithms

This section provides details of the implementation of algorithms used to modify STFT data files. In terms of the functional block diagram shown in figure 3.1, these algorithms are represented by the processing block entitled "STFT modification" which loops back to itself, to indicate that one or perhaps several STFT modifications may be performed in sequence to generate a final result for synthesis.

Spectral Weighting Function

Modification of the STFT using a spectral weighting function provided a convenient and direct means of applying a digital filter of arbitrary frequency response. This method was used to apply pre-emphasis to the spectrum, so that the energy level of high frequency spectral content would be comparable to that of the lowest frequencies. Energy at frequencies below 80 hz. was also removed using this weighting function. This function is a variation of the *point operator* as it is known in image processing, because its input variable is the pixel value $X_M[k,s]$. It differs from other point operators in that the function also uses the k index as an input variable. The spectral weighting program is a direct implementation of equation 3.3.

$$X_M'[k,s] = X_M[k,s] P[k], \quad \text{for } k = 0,1,\dots, N/2+1; \quad s = 0,1,2,\dots \quad (3.3)$$

In equation 3.3, k is the STFT frequency index, s is the frame index, $P[k]$ is the spectral weighting array as a function of frequency, $X_M[k,s]$ is the input STFT magnitude, and $X_M'[k,s]$ is the reweighted output. The frequency index k ranges from 0 to $N/2+1$ due to the fact that the conjugates of X_M have been suppressed. In the implementation of this equation, the values of $P[k]$ are derived from a generalized frequency response profile using piecewise linear interpolation. The frequency response profile consists of a set of two or more data pairs, each of which contains a frequency value and a corresponding magnitude value. It is stored as an ASCII file.

3.3.6 Conventions and Unit Conversions

The customary convention for mapping the STFT magnitude to the two dimensional Cartesian coordinate system for spectrogram display is to depict time as the x-axis, frequency as the y-axis, and the magnitude as darkness, or grey level. This convention is followed throughout this work. The STFT indices for frequency k and time s , are used interchangeably with x and y image coordinates respectively.

Wherever the spectrogram is depicted as a three dimensional figure (waterfall plot), using the xyz Cartesian coordinate system, the z-axis is used to represent magnitude. A spectrogram displayed as waterfall plot depicts magnitude as changes in height along a horizontal frequency contour, as shown in figure 3.2. Each horizontal frequency contour corresponds to a single DFT frame, or a vertical slice of a 2D spectrogram.

Figure 3.2 Three Dimensional View of a Spectrogram

Spatial frequency, as used in the context of spectrogram images, is the two-dimensional Fourier transform of pixel value configured in spatial coordinates (x,y) , and is thereby directly related to the indices (k,s) of the STFT. Hence, indirect relationships between audio units and spatial units can be derived. So as not to confuse spatial frequency, which is used in the context of image representations, with audio frequency, any mention of spatial frequency will be explicitly identified as such. The units of spatial frequency used are k^{-1} or s^{-1} .

Unity-coefficient spatial filters are often identified by their dimensions in x and y coordinates. The notations used to identify spatial filters in this way are n_k and n_s . This notation is equivalent to the symbols m and n used in previous discussions on spatial filters to denote their dimensions.

3.3.7 Two dimensional FIR Filters

The capability of applying two dimensional FIR filters to STFT magnitude data files was a core requirement for carrying out the objectives of this work. A program was created to provide the following functionalities:

- (i) Define 2D FIR filters as two dimensional arrays of weighting values, store them in ASCII files, load filters into memory from file storage, and display their contents.
- (ii) Apply 2D FIR filters as defined in (i) to magnitude spectrogram files and generate filtered output files.

Item (ii) is a direct application of equation 2.16, the general equation for filtering using two dimensional FIR's. In the implementation, two additional non-linear parameters were used to define the filter: a low value exclusion threshold, and a output saturation threshold. Input data points having a value less than the low value exclusion threshold are passed through without filtering, the purpose being to instruct the filter to ignore insignificant data values. The output saturation threshold is compared to each output data point before it is converted from floating point to integer format for storage. If its value exceeds the threshold, then it is clipped. A tally of clipping events was recorded and displayed the CRT display upon completion of a filtering operation. A large number of clipping events associated with a given filtering operation was indicative of inappropriate scaling values.

Filter Boundary Condition Applied to Spectrograms

For points $X[k,s]$ of the STFT spectrogram which are close to the edges of the image, the convolution of X with a spatial filter kernel Q having dimensions n_k and n_s cannot be computed for all points the kernel, because, as noted in chapter 2, Q effectively hangs over the edge of the spectrogram. The boundary condition in the time index s was handled by not filtering at the boundaries. A pad of zeros was inserted at the beginning and end of each audio sample used to generate spectrograms. The length of zero padding of the time sample was sufficiently long that a significant number of DFT's at the beginning and end of the spectrogram were all zeros.

A different boundary condition was applied in the k direction, to filter spectrogram pixels representing the lowest and highest frequencies. Equation 3.4a shows the equation used to compute output values of spectrogram points $X_M'[k,s]$ near the boundaries in k . This equation is a variation on the computation used to apply a filter to points not near a boundary. Equations 3.4b and 3.4c show the expansions of index variables p and q . X_M' is the convolution of X_M with filter Q in the neighborhood of $X_M[k,s]$, for all points where Q overlays the image.

$$X_M'[k,s] = \left\{ (A \cdot (b-a+1)/(n_k \cdot n_s)) \cdot \sum_{i=a}^{i=b} \sum_{j=1}^{j=n_s} X_M[p,q] Q[i,j] \right\} + B, \quad \text{for } k=1,2,\dots, N/2+1; \quad s=(n_s/2+1),2,\dots, (S-1-n_s/2) \quad (3.4a)$$

$$\text{where: } p = k+i-1 - (n_k/2) \quad (3.4b)$$

$$q = s+j-1 - (n_s/2) \quad (3.4c)$$

In equation 3.4, A is the scale factor, B is an offset value, $Q [p,q]$ is the spatial filter kernel, N is the number of points in the DFT, S is the total number of STFT frames, n_k is the filter kernel size in the k (frequency) dimension, n_s is the filter kernel size in the s (time) dimension, a is the lower limit of the filter kernel frequency index, and b is the upper limit of the filter kernel frequency index.

The upper and lower limits of the frequency index are clipped at the edges of the STFT, and are computed by applying the following logical conditions near the boundaries.

$$\begin{aligned}
 &\text{if } (k \leq n_k/2) \\
 &\quad \text{then } a = n_k/2 + 2 - k; \quad b = n_k \\
 &\quad \text{else if } ((N/2 + 1) - k < n_k/2) \\
 &\quad \quad \text{then } a = k - n_k/2; \quad b = (N/2 + 1) - k + 2 \\
 &\quad \text{else} \\
 &\quad \quad a = 1; \quad b = n_k
 \end{aligned}
 \tag{3.5}$$

This implementation of the general filter equation handles the k -axis boundary condition by computing the output only for the points in the filter kernel which overlay the image, and rescaling based on the number of input points actually used.

3.3.8 Non-Linear High-Pass Filter Effects

By definition, a spectrogram is composed only of positive sample points. Therefore, the average or mean value component of all sample points is a positive value. In the spatial frequency domain, this corresponds to a spectral component at zero frequency.

When a high-pass filter is convolved with a spectrogram, some negative-going sample points are produced because the zero-frequency component of the signal is removed

by the filter. However, the resulting spectrogram must contain only positive valued sample points. Hence, negative output values must be somehow rectified. This can be achieved using one of two methods.

The first method is to add a constant, positive offset to the filtered sample points, such that all become greater than or equal to zero. In equation 3.4, the constant term "B" incorporates this offset into the general filter algorithm. Adding an offset as such allows the filter to retain a linear transfer characteristic. In the spatial frequency domain, the additive offset reintroduces a zero frequency component.

The second method is to simply truncate all negative output sample points to zero. Clearly, this approach introduces a non-linear characteristic into the filter transformation. In the implementation of the general box filter algorithm, both types of negative output rectification were implemented. In fact, all negative output sample points are truncated to zero, if a constant "B" term does not provide the required compensation.

In practise, it was found that in the application of high-pass filters to spectrograms, it was more appropriate to use truncation rather than add an offset to the output data. Unlike most other images, the spectrogram is created by splitting a complex-valued STFT into its magnitude and phase components. Although the phase component is absent in a spectrogram, it must still be taken into consideration when a spectrogram undergoes modification, if the result is to be used in resynthesis. In general, it would be undesirable to boost the magnitude of sample points in regions of the spectrogram where there is not a significant amount of speech energy, because the phase and magnitude components represent background noise.

Applying a high-pass filter to a neighborhood of low energy sample points will produce small-valued positive and negative output samples about a mean value of zero. If a constant offset value is added to rectify the output, the magnitude of low level samples is now raised to form a "floor" about the offset level. In other words, formerly insignificant samples representing background noise acquire substantial magnitude. If the filtered spectrogram is now recombined with its matching phase in STFT synthesis, the result is a noisy signal.

Truncation of negative values, on the other hand, introduces non-linear distortion into the spectrogram. However, the effect was minimal because truncation of filtered samples most often occurs in areas of minimum overall signal energy. Some examples of the zero truncation effect can be observed in spectrograms shown in chapter 4. Energy troughs between harmonics typically contain significant evidence of truncation, as do areas preceding plosives.

3.3.9 Simulated Grey Level Spectrogram Display

A Hercules monochrome Graphics peripheral with a resolution of 720 columns by 350 rows was used for spectrogram display. In order to simulate the effect of discrete intensity grey levels, a dithering algorithm was devised.

The dithering kernel for a 256 point DFT reduced to $N/2+1=129$ points was two pixels in the vertical direction (frequency). In the horizontal dimension (time), the resolution was variable, depending on the total number of STFT frames displayed at one time. The total number of columns available for STFT display after labeling was 600. Between each frame a space of one column of dark pixels was inserted such that a

maximum of 300 frames could be shown at time, given a minimal sized dithering kernel of 2x1 pixels. Reducing the number of frames displayed to 200 allowed the dithering kernel size to be increased to 2x2 pixels. Reducing the display size to 150 frames allowed the dithering kernel size to be set to a maximum of 2x3 pixels. Figure 3.3 shows a spectrogram of the word "intercept" produced using a 2x2 dithering kernel.

Figure 3.3 Spectrogram of the Word "intercept"

Magnitude Mapping Function

For each kernel, the number of pixels set to the "on" state was proportional to the magnitude in decibels of the related data point of the STFT, over a selected range of magnitude values. Data points having magnitude greater than the selected range were assigned the maximum number of "on" pixels, while those having magnitude less than the

selected range were assigned zero "on" pixels. Parameter inputs to the spectrogram display program allowed the range of magnitude values, start STFT frame, and end STFT frame to be specified.

3.3.10 Three Dimensional Graphic Display

In the three dimensional Cartesian coordinate system, DFT frequency was mapped onto the x-axis, STFT frame index was mapped onto the y-axis, and magnitude in decibels was mapped onto the z-axis, with scaling and labelling. The three dimensional display algorithms were provided by the Grafmatic graphics subroutine software package [24].

3.3.11 Time Domain Signal Processing

In order to be able to make aural comparisons between different speech samples presented using the digital to analog conversion and playback equipment, it was necessary to normalize the signal energy in the samples being compared to a common reference level, so that they would have equal loudness. To meet this requirement, an algorithm for measuring the average speech energy of an audio sample, and applying a scale factor to the sample was developed. Figure 3.4 shows this process in the form of a set of functional blocks with data inputs and outputs.

Figure 3.4 Time-Signal Normalization Process

Initially, a sliding window is applied to the time domain audio file samples, and the average RMS value is computed for each window position. A voice/silence threshold parameter is then input by the user or computed based on the minimum RMS value in the sample. This threshold is applied to the RMS signal to determine a set of ordered pairs of pointers which indicate segments of the time-signal containing voice. Estimates of the voice power and the DC bias are computed as an average of all the time samples inside these segments. The extracted power value, in combination with an parameter input specifying the desired power level, is used to generate a normalization scale factor. In the final step of normalization, the DC bias is subtracted, and the scale factor is applied to each time signal datum point to generate a normalized version of the speech file as the output.

CHAPTER 4

SPECTROGRAM MODIFICATION EXPERIMENTS

4.1 OVERVIEW

Having put in place a system capable of performing STFT analysis and synthesis, the next step was to apply a series of filter operators to some speech samples and examine the results. Initially, it was necessary to try out an assortment of filters to get a general feel for the effects produced in both the displayed spectrogram image and in the audio signal produced by synthesis of the modified spectrogram. Initial efforts at filtering, some of which proved to be quite destructive to speech, led to refinements which produced more promising results.

Experimentation with speech samples was quite constrained by the limited processing power of the PC-based system. A strategy for experimentation was adopted to work within these constraints and produce some observational results. To begin with, only a few single word utterances were selected. Processing entire phrases, and/or large numbers of word utterances would have demanded unreasonably long periods of time for analysis-synthesis and filtering. The first constraint relates also to the second part of the experimentation strategy, observation by critical listening. It would have been unrealistic to generate large sample sets of utterances as required for conducting formal psychophysical measurement of observer responses as a means of determining whether or not aural enhancements had been produced. Observations regarding the effect of various filters on

speech samples were formed by means of extensive critical listening on the part of the author.

A description of the speech modification experiments is organized in terms of general filter types applied to the speech samples. In each experiment, a set of test cases consisting of the original spectrograms generated by STFT analysis were processed and observations were made by comparing the filtered result to the original. In some of the experiments, spectral pre-emphasis was applied to the spectrogram prior to processing by FIR filters to flatten the spectral slope across the range of frequencies in the DFT. Wherever applicable, observations are presented in tabular form. All experiments outlined are based on the application of some variation of the two-dimensional FIR filter to the STFT magnitude component of the aforementioned word utterances.

4.1.1 Selection of Speech Samples

For the purposes of generating modified speech, three samples were selected from a list of single word utterances recorded by a professional female speaker. The primary selection criterion was that each word contain a representative mixture of basic phoneme types found in speech. These word utterances were "broadcast", "flashing", and "intercept". Table 4.1 lists the constituent phonemes of these chosen words.

Sample Word	Characters	Phoneme Symbol	Phoneme Type
broadcast	/b/	B	voiced stop
	/r/	ER	mid vowel
	/oa/	AO	mid vowel
	/d/	D	voiced stop
	/c/	K	unvoiced stop
	/a/	AE	front vowel
	/s/	S	unvoiced fricative
	/t/	T	unvoiced stop
flashing	/f/	F	unvoiced fricative
	/l/	EL	liquid semivowel
	/a/	AE	front vowel
	/sh/	SH	unvoiced fricative
	/i/	IY	front vowel
	/ng/	NX	nasal
intercept	/i/	IH	front vowel
	/n/	N	nasal
	/t/	T	unvoiced stop
	/cr/	AXR	mid vowel
	/c/	S	unvoiced fricative
	/e/	EH	front vowel
	/p/	P	unvoiced stop
	/u/	T	unvoiced stop

Table 4.1 Phonemic Content of Speech Samples³

4.1.2 Selection of Spatial Filters

Table 4.2 is a list of the spatial filters which are referred to in subsequent sections. Each filter is referenced with an Roman numeral symbol as shown in the first column of the table. Filter kernel dimensions are given in the 2nd and 3rd columns. The symbol n_k is the

3. The phoneme symbols (column 3) used in Table 4.1 were taken from Cole, Rudnick, Zue & Reddy, reference [9], page 16. A table of phoneme categories from Rabiner & Schafer, reference [17], page 43 was also used as an aid to classification.

filter size in units of the spectrogram frequency index k , and the symbol n_s is the filter size in units of the spectrogram time index s .

Spatial frequency pass band characteristics of the filters are summarized in the 4th through 6th columns. Only low-pass and high-pass filters were used, hence in each case the pass band is defined by one parameter in each spatial dimension. Spatial frequency parameter values in k^{-1} units were derived by taking the discrete Fourier transform of each filter impulse response with respect to its k dimension only, and establishing the point at which frequency response was -3dB of unity. Similarly, s^{-1} spatial frequency values were derived by taking the discrete Fourier transform of each filter impulse response with respect to its s dimension only. Hence, spatial frequency response parameters listed are orthogonal in k and s , and the spatial frequency characteristics in directions having both a k and s component are not shown. The frequency response in a direction having a k and s component is approximately equal to that of the response along the axis which is closer with respect to angle about the central point (origin).

Filter strength is expressed as a percentage of maximum. This represents the ratio of filtered image and original image which are additively combined to create the resulting output image. The final column describes the design of each specific filter. For the most part, the high-pass box filter was used, because it was found that the different frequency response characteristics which could be obtained using various window designs [25] had little effect on the results obtained. High-pass filters created using the window design method were found to produce results similar to the basic high-pass box filters, when the filter bandwidths were approximately equal.

Filter Name Code	Dimensions		Bandpass Cutoff Frequency (3 dB)		Filter Type	Filter Strength	Design Description
	n_k	n_s	$1/k$	$1/s$			
I	19	1	0.10	-	high-pass	100%	Window design method, Hamming high-pass window.
II	5	1	0.15	-	high-pass	100%	High-Pass Box Filter. All coefficients = -1, except central coefficient.
III	1	19	-	0.10	high-pass	100%	Window design method, Rectangular high-pass window.
IV	1	31	-	0.02	high-pass	100%	Window design method, Rectangular high-pass window.
V	1	31	-	0.02	high-pass	67%	Window design method, Rectangular high-pass window.
VIa	3	15	0.19	0.05	high-pass	100%	High-Pass Box Filter. All coefficients = -1, except central coefficient.
VIb	3	31	0.19	0.025	high-pass	100%	High-Pass Box Filter. All coefficients = -1, except central coefficient.
VIc	3	59	0.19	0.012	high-pass	100%	High-Pass Box Filter. All coefficients = -1, except central coefficient.
VIIa	7	15	0.10	0.05	high-pass	100%	High-Pass Box Filter. All coefficients = -1, except central coefficient.
VIIb	7	31	0.10	0.025	high-pass	100%	High-Pass Box Filter. All coefficients = -1, except central coefficient.
VIIc	7	59	0.10	0.012	high-pass	100%	High-Pass Box Filter. All coefficients = -1, except central coefficient.
VIII	7	3	0.06	0.16	low-pass	100%	Low-Pass Box Filter. All coefficients = 1.

Table 4.2 Summary of Spatial Filter Characteristics⁴

4.2 OBSERVATIONAL CRITERIA

As stated previously, it was of interest to examine the visual effects of filtering on the two dimensional spectrogram image and on the three dimensional view of the spectrogram, and to compare them with the aural effects produced in the time-domain

4. The 'window design method' in this table refers to FIR filter design using time window functions. This design method is described in detail in chapter 8 of Stearns & David, reference [25].

signal synthesized from the modified spectrogram. The basic observable feature is the phoneme, which is associated with a specific aural signature in the utterance and with a corresponding combination of visual structures in the image spectrogram.

4.2.1 Visual Effects in Spectrograms

In the sections which follow, the two dimensional grey level representation of the spectrogram is used as the basis for most visual observations on the effects of spatial filtering. In general, the underlying structures of speech were more observable in grey level spectrogram images than in three dimensional images. The presence of high spatial frequencies in the three dimensional image was problematic, because it masked out information contained in middle and lower spatial frequencies. High spatial frequencies in the grey level representation of the spectrogram were attenuated by the limited quantization range of the monochrome display, and did not dominate the image. Another problematic characteristic of three dimensional representation was that features were hidden behind other features due to the nature of perspective.

Three dimensional views were found to be somewhat impractical for the general purpose of comparing visual and aural features of speech in most cases, because they did not benefit from high-pass filtering. High-pass filtering increased the dominance of the highest spatial frequencies in three dimensional views, rendering them quite useless for observation purposes. Low-pass filtering had the opposite effect, producing some improvement by removing high frequencies, thereby *enhancing* the appearance by allowing the lower spatial frequency content to stand out.

4.2.2 Auditory Effects in Synthesized Signals

The audio samples used to create test cases were relatively noise-free, making it possible to observe the effect of filtering on the signal itself, rather than on the additive noise component of the signal. In each case, the original sample was clearly intelligible in the 0-3400 hertz audio band. Observations were based on comparisons between the original signal, and a spatially filtered, synthesized version of it. In each comparison, an effort was made to discern the filter's effect on each individual phoneme, and on the transition areas between phonemes. An assessment was also made of the overall effect of the filter on the sample, in terms of its perceptual impact. For example, a word could be made to sound more emphatic than the original using certain high-pass spatial filters.

4.3 ON THE USE OF SPECTRAL PRE-EMPHASIS

In some of the filtering experiments to be described, a spectral weighting function was applied to the STFT magnitude data as a preprocessing step. The profile of the weighting function used is shown in figure 4.1. This type of frequency profile is a form of pre-emphasis filter. Pre-emphasis filters are commonly used to boost the weaker frequency components of signals before transmission through a channel. Telephone transmitters are one example of the application of pre-emphasis. Speech characteristically has lower energy in the higher frequencies, hence the filter boosts the high end of the signal in relation to the low frequencies, creating a balanced spectrum for analysis purposes. At frequencies below 80 hz, there is no significant speech energy, so that part of the spectrum is filtered out. Zeroing the lowest frequency spectra also helps reduce effects produced by the boundary conditions imposed on the use of two-dimensional FIR filters, as described in the previous

chapter. Overall, frequency equalization was expected to enable the filter process to produce a more balanced output than the frequency weighting of the original spectrogram.

Figure 4.1 Pre-emphasis Weighting Function

4.4 EXPERIMENT A: $N_K \times 1$ HIGH-PASS FILTERS

Filtering each of the STFT frames independently of their neighbors (i.e. filtering the spectrogram in vertical strips) was achieved by constraining the time dimension of the filter to be $s=1$. The word "broadcast" was processed using a series of high-pass filters, with cutoff frequency ranging from 0.01 to 0.2 units of spatial frequency (hertz-seconds). Figures 4.2a and 4.2b show the spectrograms of the original signal, and figures 4.3 through 4.4 show the spectrograms after filtering for two of the filters tried. Filter I has a 3 dB high-pass cutoff of 0.1 spatial frequency units with a steep rolloff characteristic, while Filter II has a 3 dB high-pass cutoff of 0.16 spatial frequency units and a gentler rolloff

characteristic. The word "broadcast" was broken into "broad" and "cast" for the purposes of spectrogram display to permit a higher resolution of intensity per pixel.

Observations

1. The harmonic components of vowels were emphasized by filtering. This is evident in the spectrograms shown in figures 4.3 and 4.4, in which the vowel harmonics dominate the images. Vowels in the synthesized audio were sonorous, or nasal sounding, due in part to the emphasis of the upper harmonics. In the spectrograms of figures 4.3 and 4.4, the domination of narrowband features in the horizontal dimension, made the formants appear less distinct in comparison to the unfiltered versions of figure 4.2. However, the aural difference between different vowels was still quite distinct. The optimal filter band for placing emphasis on vowel harmonics was found to be in the range of 0.05 to 0.15 spatial frequency units. Lower frequencies produced little change between the input and output.

2. Figures 4.3 and 4.4 show a marked destruction of the stops and fricative components of the signal. This is in greater evidence in the latter. This characteristic is highly evident in the audio, especially in the case of stops, which are compromised to the point of becoming indistinct with respect to similar sounding phonemes ("d" and "t" sounded alike). Overall, the effect of $n_k \times 1$ filters was a degradation in the signal, due mainly to the destruction of stops.

3. Non-linear effects produced by truncation of negative output values by the filter algorithm show up in the spectrograms as white space between vowel harmonics.

4. Three dimensional views of the spectrogram became *noisier* in appearance when high-pass filtered with $n_k \times 1$ filters. Isolated peaks tended to block out coarser features one looks for in the 3-D view. (An example of a 3-D view of a high-pass filtered spectrogram is shown in figure 4.27.)

Figure 4.2a "Broad.." Original Spectrogram

Figure 4.2b "..cast" Original Spectrogram

Figure 4.3a "Broad.." Filter I Spectrogram

Figure4.3b "...cast" Filter I Spectrogram

Figure 4.4a "Broad.." Filter II Spectrogram

Figure 4.4b "..cast" Filter II Spectrogram

4.5 EXPERIMENT B: $1 \times N_S$ HIGH-PASS FILTERS

In this experiment, the filter dimension corresponding to audio frequency was constrained to be $k=1$. Using this constraint, strips of the spectrogram along the time axis were effectively filtered independently of one another. Hence, the modified value of each pixel was a function of its own value and the values of its neighbors along the same spectral component. High-pass filtering in the time direction produced an emphasis on pixels in regions of changing magnitude with respect to time, while de-emphasizing the pixels in regions of unvarying magnitude.

A substantial number of test cases were generated for the application of this type of filter. The results produced using this type of filter proved to be of more value than those of experiment A, in which the stop sounds were audibly and visibly decimated. Figures 4.5 through 4.10 are spectrograms generated for the word "broadcast", with and without pre-emphasis.

4.5.1 Filters Applied to Original Spectrograms

Figures 4.5 through 4.7 are a selection of spectrograms of the results obtained by application of $1 \times n_s$ high-pass filters to the original spectrogram of the word "broadcast". Test cases were created for series of filters with cutoff frequencies ranging from 0.005 to 0.2 spatial frequency units. In some cases several different results were obtained using a single filter by varying the filter strength. When the filter strength is reduced, some of the original spectrogram is additively combined with the filtered output. Filters III, IV and V from table 4.2 are a representative subset of the $1 \times n_s$ filter configurations experimented with.

Observations

1. Filters having dimension $1 \times n_s$, and cutoff frequencies ranging from 0.01 to 0.05 spatial frequency units were found to be capable of producing visual emphasis in parts of the spectrogram where transitions between phonemes occur. This corresponded to perceivable aural emphasis in the resynthesized speech in the matching phonemes. The spectrogram shown in figure 4.6 compares favorably to the original spectrogram in figure 4.2 in most respects. In some cases, stops were audibly emphasized by filtering. This was the case with Filters IV and V, which had a cutoff frequency of 0.02 spatial units, but differed in strength, the relative strengths being 100% and 67% respectively. A particularly noticeable effect, although not necessarily a desirable one, was produced in the /b/r/ transition of the word "broadcast". After filtering, the "br" took on a rolling sound as if spoken with a strong accent. The other stops in the word were audibly emphasized. An optimal 3 dB pass band cutoff was found to be in the range of 0.01 to 0.025 spatial frequency units. For filters with a cutoff of greater than spatial frequency of 0.05, the speech began to take on a rough sound quality. This effect was quite pronounced for results produced by filter III, which had a cutoff frequency of 0.10 spatial frequency units. Figure 4.5 shows the spectrogram produced by filter III. The emphasis of stops diminished with filters having passband cutoff of less than 0.01 spatial units.

2. Filters which were determined to be effective at emphasizing stop sounds and transitions were found to compromise vowel sounds and fricatives in the word samples used. In the case of vowels, the regions having horizontal and parallel formant tracks (figure 4.7) were weakened significantly by filter IV. Vowel sounds in the synthesized audio were audibly compromised in that they were less distinct and rough sounding, although still clearly intelligible. An entirely different problem arose with the fricative sounds, particularly in the word "flashing". In the filtered spectrogram, there are

significant signs of improvement over the original spectrogram. However, the synthesized audio contains audible evidence of the destruction of these same fricatives, which no longer sound like spectrally weighted white noise. Unnatural sounding *tonals* as described in Munday [19] were evident in areas containing a Gaussian noise component. The presence of tonals can be explained by high-pass filtering of white noise producing non-linear truncation to zero of the lower magnitude pixels, effectively altering the Gaussian noise characteristic. A better sounding result was obtained using filter V, at 67% strength. Mixing some of the original spectrogram with the filtered version retained some of the filter's emphasis on stops and transitions, while allowing the vowels and fricatives to sound more natural.

3. Non-linear effects produced by truncation of negative output values by the filter algorithm show up in the spectrograms as white space between vowel harmonics, and in areas having low overall speech energy.

4. Three dimensional views of the spectrogram showed that $1 \times n_s$ high-pass filters produced dramatic degradations in feature definition. For this reason, extensive comparisons were not made using the 3-D views.

Figure 4.5a "Broad.." Filter III Spectrogram

Figure 4.5b "..cast" Filter III Spectrogram

Figure 4.6a "Broad.." Filter IV Spectrogram

Figure 4.6b "...cast" Filter IV Spectrogram

Figure 4.7a "Broad.." Filter V Spectrogram

Figure 4.7b ".cast" Filter V Spectrogram

4.5.2 Filters Applied to Spectrograms with Pre-Emphasis

The pre-emphasis weighting function shown in figure 4.1 was applied to the original spectrograms of the word utterances, to equalize the energy of the upper and lower ranges of the frequency spectra. Pre-emphasis alone produced a significant change in the synthesized speech, boosting the energy of fricative sounds and stops as would be expected. It also introduced an audible *hiss*, caused by boosting the energy of the background noise in the upper range of frequencies. The effect of pre-emphasis is evident in the spectrogram image of the word "broadcast", as shown in figure 4.8, but the feature definition did not show a significant improvement.

The same $1 \times n_s$ filters as were applied to the original spectrogram were applied to the pre-emphasized spectrograms. Pre-emphasized, filtered results were compared to the pre-emphasized input to determine the filter effects. Spectrograms showing the effects of filtering with pre-emphasis are represented in figures 4.9-4.10.

Observations

1. The results of filtering were similar to those produced using the original STT-TM as input. Spectrograms showed increased resolution of phonemic features produced by filters having a passband cutoff of 0.01 to 0.05 spatial frequency units, particularly in stops and transition regions. Upper harmonics of vowels were also well defined in the image spectrograms. In the aural domain, the stress on stops produced by filtering pre-emphasized spectrograms was more dramatic than that produced by filtering the original versions. In most respects, the results were similar to those obtained by filtering the original version of the spectrogram. Filter IV proved to be destructive to vowel sounds.

Filter V, a weaker version of filter IV, improved the vowel sounds, while adding emphasis to stops.

2. Tonals in fricatives were significantly less noticeable in the audio synthesized from pre-emphasized, filtered spectrograms. This can be explained by the fact that the overall Gaussian noise energy of fricatives was increased by pre-emphasis. The increase in overall pixel magnitude in Gaussian noise regions resulted in a lower number of non-linear pixel transformations due to zero truncations. High-pass filtering of regions consisting of higher energy Gaussian noise reduced the presence of tonals to some extent.

Figure 4.8a "Broad.." Spectrogram with Pre-emphasis

Figure 4.8b ".cast " Spectrogram with Pre-emphasis

Figure 4.9a "Broad.." Pre-emphasis & Filter IV

Figure 4.9b ".cast" Pre-emphasis & Filter IV

Figure 4.10a "Broad.." Pre-emphasis & Filter V

Figure 4.10b "...cast" Pre-emphasis & Filter V

4.6 EXPERIMENT C: $N_K \times N_S$ HIGH-PASS BOX FILTERS

In experiments A and B, one-dimensional high-pass filters were applied to STFTM's, and an optimal range of passbands was determined for the $n_k \times 1$ and $1 \times n_s$ filters respectively. This experiment combines the characteristics of $n_k \times 1$ and $1 \times n_s$ filters in two dimensional $n_k \times n_s$ filters. Hence the new value of each point in the spectrogram is a function of its previous value and the values of its neighbors in both time and frequency dimensions. Two dimensional filters used in this experiment are filters VIa through VIIc as described in table 4.2. In each case the filter strength is set to 100%. One of the primary aims was to produce emphasis on stops and other transitions as in experiment B without compromising the vowel harmonics.

The approach taken is similar to that of experiment B. A series of filters were applied to the original spectrogram and a subset of the same series of filters was applied to the pre-emphasized spectrogram. Observations obtained by listening to the samples produced by synthesis of the filtered results are summarized in tables 4.2 and 4.3. Figures 4.11 to 4.24 are a selection of spectrogram images produced with $n_k \times n_s$ filters.

4.6.1 Filters Applied to Original Spectrograms

Filters having a 3 dB cutoff spatial frequency between 0.1 and 0.2 units on the frequency (k) axis, and 0.01 to 0.05 units on the time (s) axis were applied to the original STFTM's of three utterances "broadcast", "flashing", and "intercept". Table 4.3 summarizes the aural effects observed in the results of filtering.

FILTER ID	OBSERVATIONS (differences between filtered samples and original samples)			
	Nasals, Vowels, Glides	Voiced Stops, Unvoiced Stops	Fricatives	Other Notes
Vla (3x15)	Rough sound, distorted	Strong emphasis on transitions	Presence of tonals (whistling noises)	Greatly increased stress on transition areas, overall effect is rippled sounding, distorted
VIb (3x31)	Some distortion, reverberation, esp. in nasal sound "n"	Emphasis noticeable on some stops	Presence of tonals noticeable, especially on "sh" in "flashing"	Transition areas rippled sounding, improvement over filter Vla
VIc (3x59)	Lower distortion, some reverberation in nasals	Emphasis slightly noticeable	Weaker "f" in "flashing"	Overall effect is increase in overall emphasis in all three words.
VIIa (7x15)	Significant distortion	Noticeable increase in emphasis on stops	Tonal distortion in most fricative sounds	Vowels more sonorous, rippled transitions, distortion, better than filter Vla
VIIb (7x31)	Vowels more sonorous than with filter VIb	Increase in emphasis on some stops	Tonal distortion in some fricatives. "f" has shorter induration	Words sound very close to original with noted exceptions
VIIc (7x59)	Little noticeable change	Emphasis on some, others weakened	Slight de-emphasis of "s", "f" sounds	Overall, little noticeable change, except fricative sounds in "flashing"

Table 4.3 Aural Effects of Filtering Original Spectrograms

Observations

1. Filters VIa, VIb, and VIIa produced a significant amount of distortion in the audio. The distortion effects included strong tonals in fricatives, and a rippling amplitude effect on some phoneme transitions. In general, these filters produced unnatural sounding results, and significant emphasis on stop sounds.

2. Filters VIc, VIIb, and VIIc produced significantly more natural sounding results in that less distortion was present in the audio. Filter VIIb produced some tonal effects,

particularly on the "sh" of "flashing", but was an improvement over filter VIb. No single filter was clearly better than the others in this group, although overall, filters having $n_s=59$ produced less undesirable aural effects. Filters with $n_k=7$ resulted in cleaner vowel sounds and greater stress on fricatives.

3. Each of the filters listed in table 4.3 improved the appearance of spectrogram images relative to the original. Some of the aural effects described in the previous two paragraphs also have visual counterparts in the spectrogram images, shown in figures 4.11 to 4.13, 4.17, 4.18, 4.21, and 4.22. For spectrograms created using filters with $n_k=7$, the harmonic tracks of vowels are wider, and fricative regions occupy greater area than those created using filters with $n_k=3$.

4.6.2 Filters Applied to Spectrograms with Pre-Emphasis

In this section, the same filters as in the previous section were applied to the same three utterances with pre-emphasis. Overall, filtering the pre-emphasized version of the spectrogram produced results which were as good or better than results produced by filtering the original spectrogram. Table 4.4 provides details of the aural effects observed in samples synthesized from pre-emphasized, filtered spectrograms.

FILTER ID	OBSERVATIONS (differences between filtered sample and original with pre-emphasis)			
	Nasals, Vowels, Glides	Voiced Stops, Unvoiced Stops	Fricatives	Other Notes
Vla (3x15)	Rough sound, distorted	Strong emphasis on stops, transitions	Tonal distortion noticeable in most fricatives	All samples very distorted overall
Vlb (3x31)	Slightly noticeable distortion	Emphasis significant on most stops	Emphasis noticeable, but also significant tonal noise	"Broadcast", "flashing": slight distortion, sharper sound. "Flashing": emphasis in transitions.
Vlc (3x59)	Harmonics effect more balanced than original	Emphasis noticeable on some, others weakened	Slightly noticeable tonals on "f", "sh"	"Broadcast", "intercept": cleaner sound "Flashing": more emphasis on transitions, fricatives
Vlla (7x15)	Significant distortion, less than on original w. pre-emphasis	Emphasis significant on most stops	Tonal distortion in most fricative sounds	Quite distorted overall, better than filter Vla
Vllb (7x31)	Sounds "cleaner", de-emphasis of high frequency noise	Increase in emphasis on most stops	Tonal distortion in "sh". Shorter induration of "c" in "intercept"	"Broadcast": sharper, cleaner "Flashing": more emphasis on transitions, expressive. "Intercept": little difference
Vllc (7x59)	Sounds "cleaner", de-emphasis of high frequency noise	Emphasis noticeable in some stops	Slight de-emphasis of fricatives (shorter duration), cleaner sounding	"Broadcast", "Flashing": sharper, cleaner sound. "Intercept": little difference

Table 4.4 Aural Effects of Filtering Pre-emphasized Spectrograms

Observations

1. Filters Vla, Vlb, and Vlla added aural stress to some parts of the speech but produced significant distortions similar to those produced by filtering spectrograms without added pre-emphasis. The most significant difference was that tonal distortion was more evident when pre-emphasis was applied to a spectrogram before filtering.

2. Filters VIc, VIIb, and VIIc produced cleaner sounding results, but tonals were still significant in some fricatives in audio samples produced by filters VIIb and VIc. Overall, the filter VIIc produced the best results in terms of aural quality. Distortion was minimal, and there was noticeable positive stress on some phoneme transitions.

3. Following filtering, pre-emphasized spectrograms were almost identical in appearance to those which did not have pre-emphasis. In some cases, the upper spectra show some evidence of greater emphasis, but the difference is not as noticeable as it is in the aural domain. Figures 4.14 to 4.16, 4.19, 4.20, 4.23, and 4.24 are spectrograms of samples which underwent pre-emphasis prior to filtering.

Figure 4.11a "Broad.." Filter Via Spectrogram

Figure 4.11b "...cast" Filter Via Spectrogram

Figure 4.12a "Broad.." Filter Vlb Spectrogram

Figure 4.12b "..cast" Filter Vlb Spectrogram

Figure 4.13a "Broad.." Filter Vllc Spectrogram

Figure 4.13b ".cast" Filter Vllc Spectrogram

Figure 4.14a "Broad.." Pre-emphasis & Filter Via Spectrogram

Figure 4.14b "...cast" Pre-emphasis & Filter Via Spectrogram

Figure 4.15a "Broad.." Pre-emphasis & Filter Vlb Spectrogram

Figure 4.15b "..cast" Pre-emphasis & Filter Vlb Spectrogram

Figure 4.16a "Broad.." Pre-emphasis & Filter VIIc Spectrogram

Figure 4.16b "...cast" Pre-emphasis & Filter VIIc Spectrogram

Figure 4.17 "Flashing" Filter VIb Spectrogram

Figure 4.18 "Flashing" Filter VIIc Spectrogram

Figure 4.19 "Flashing" Pre-emphasis & Filter Vlb Spectrogram

Figure 4.20 "Flashing" Pre-emphasis & Filter VIIc Spectrogram

Figure 4.21 "Intercept" Filter VIb Spectrogram

Figure 4.22 "Intercept" Filter VIIc Spectrogram

Figure 4.23 "Intercept" Pre-emphasis & Filter Vlb Spectrogram

Figure 4.24 "Intercept" Pre-emphasis & Filter Vllc Spectrogram

4.7 EXPERIMENT D: $N_K \times N_S$ LOW-PASS FILTERS

A series of two dimensional low-pass filters were applied to STI-TM's to determine the overall effect. It was determined with a limited amount of experimentation that these filters were of little value for directly enhancing the 2D spectrogram or audio synthesized from the filtered spectrogram. Only the three dimensional view of the spectrogram benefitted from low-pass filtering. Filter VIII from table 4.2 is one of the low-pass filters which produced enhancement effects in 3D views. An example of the effect of a low-pass filter on a 2D spectrogram image is shown in figure 4.25. The effect of a low-pass filter on a 3D spectrogram representation is shown in figure 4.28.

Observations

1. In voiced sounds, such as vowels and nasals, a low-pass filter with sufficiently low cutoff frequency causes audible degradation in the harmonic structure of the phoneme. The pixels representing harmonic energy are averaged with those in areas lying between harmonics. The result is *dirty* sounding vowels. A filter spatial cutoff frequency of 0.1 units along the k (frequency) axis was capable of corrupting speech in this way. Stop sounds were also weakened by low-pass filtering. It was found that filters having a spatial cutoff frequency of 0.1 units along the s (time) axis produced significant audible degradation in stops.
2. The 2D spectrogram showed deterioration when filtered by any low-pass filter. The smaller the filter passband in the direction of either axis, the worse the result.
3. Three dimensional view was greatly improved by low-pass filtering. The 7×3 filter was found to be quite effective in evening out small scale features, making the large

scale ones, including formants, more discernable. Smaller filters, such as the 3x3, produced some degree of smoothing. Filters larger than 7x3 started to show signs of *washing out* features too much.

Figure 4.25a "Broad.." Filter VII: Spectrogram

Figure 4.25b "...cast" Filter VIII Spectrogram

2 of/de 2

MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS
STANDARD REFERENCE MATERIAL 1010a
(ANSI and ISO TEST CHART No. 2)

Figure 4.26 ".Broad" 3D View of Original Spectrogram

Figure 4.27 "Broad" 3D View of Filter IV Spectrogram

Figure 4.28 "Broad" 3D View of Filter VIII Spectrogram

4.8 SUMMARY

Experiments A through D examined the effects produced by four different spatial filter configurations on the visual characteristics of spectrograms, and on the aural characteristics in samples synthesized from those spectrograms. Experiment A established a range of filter bandwidth parameter values which could be used for high-pass filtering in the frequency (k) dimension of a spectrogram. Experiment B established a range of filter bandwidth parameter values which could be used for high-pass filtering in the time (s) dimension. The findings of experiments A and B were incorporated into experiment C, which applied two dimensional filters to spectrograms in an attempt to create a spatial filter which showed a capability of enhancing speech. Experiment D looked specifically at the effects produced by low-pass box filters in spectrograms and in synthesized speech.

A number of general statements can be made to summarize the results of experiments A through D. These are outlined in the following:

1. Two dimensional discrete grey level representations of spectrograms clearly benefitted from high-pass filtering by simple high-pass box filters. Filters having n_s set such that the 3 dB spatial cutoff frequency is between 0.01 and 0.05 units of s^{-1} were found to be the most effective. An unity valued filter with dimension parameter n_s ranging from 15 to 59 produced consistent improvements to the original spectrogram and to the pre-emphasized spectrogram. Filtering along the k (spectral frequency) dimension brought emphasis to harmonic structures, but compromised the distinctiveness of other types of features, particularly those corresponding to unvoiced structures. The optimal spatial frequency for filtering in the k dimension was found to be in the range of 0.1 to 0.2 units of k^{-1} .

2. High-pass filters having a 3 dB spatial cutoff frequency from 0.01 to 0.025 units of s^{-1} , and a spatial cutoff frequency from 0.1 to 0.5 units of k^{-1} are capable of producing aural enhancement effects in speech synthesized from the modified spectrogram. When a pre-emphasis weighting function is applied to the spectrogram to equalize the energy across the range of spectra, similar results are obtained. The type of enhancement effects produced include increased stress on stop sounds and emphasis on the transition regions associated with phonemes. There appears to be a tradeoff between stop, vowel, and fricative quality which depends on the filter characteristics. Filters which produce the most dramatic emphasis on stops tend to weaken vowels and introduce tonal coloring into fricatives.

3. There is a significant correlation between the grey level representation of the spectrogram and the aural result of Fourier synthesis based on the spectrogram, with

respect to effects produced by filtering. High-pass filtering the spectrogram usually produces an improvement over the original version. This is sometimes true in terms of the aural effects produced. Degradation in the audio, particularly in fricatives, does not show up to nearly the same degree in the spectrogram. When a low-pass filter is applied to the spectrogram, the visual representation and the aural audio quality degenerate in a proportion which is related to the size of the filter.

4. Some of the visual effects produced in high-pass filtered spectrograms can be attributed to zero truncation of negative-going output samples, a non-linear filter characteristic. It is unclear whether or not non-linear effects contributed significantly to the aural effects.

5. The units used in the above statements apply only to spectrograms created using the sampling parameter values outlined in chapter 3. Should these be changed, appropriate scaling would need to be applied. For example, units of k in this work correspond to audio frequency bands of 31.25 hertz. Hence, units of k^{-1} are equivalent to hertz^{-1} multiplied by a scaling factor of 31.25. Similarly, units of s correspond to time increments of 0.004 seconds. Therefore units of s^{-1} are equivalent to units of seconds^{-1} multiplied by 0.004. If the sampling parameters were to be changed in STFT analysis, the time and frequency units could be used determine new spatial units from which equivalent spatial filters would be specified.

CHAPTER 5

CONCLUSIONS

5.1 OVERVIEW

The stated primary objective of this work was to investigate the effects of spatial image filters on speech when applied to speech spectrograms derived using the short time Fourier transform. In order to meet this objective, a system for processing speech was developed and implemented in software on a PC compatible computer. Signal processing algorithms for STFT analysis-synthesis, spatial filtering were borrowed and modified as required in order to satisfy the objectives of the investigation. Having developed the necessary speech analysis, synthesis and modification tools, a series of pilot experiments were devised and executed. In each of these experiments, a series of convolutional image filters was applied to a selection of speech samples. Observations were taken with regard to the visual changes produced in the spectrogram, and with regard to the aural changes produced in speech samples synthesized from modified spectrograms and original STFT phase.

5.2 GENERAL COMMENTS

In some respects, the speech processing methods applied in this thesis resemble ones found in the literature. Eger et al. [20] describes a similar STFT analysis-modification-synthesis process. The paper by Ariki et al. [21] is, in part, an example of the application

of 2D spatial filtering to speech spectrograms. The algorithms used for STFT analysis and synthesis were borrowed in part, from Green [17] and others. However, there was not in existence any work so similar in nature that it could serve as a model for this investigation in a very broad sense. At the outset of this investigation, it could only be speculated upon as to what would be found by manipulating speech using the methods described. Therefore, much of the early work consisted of a trial and error approach to filtering spectrograms in order to see what would result. Some of the paths taken are not described in detail, as they did not lead anywhere. The four experiments described focus on the test cases that produced noteworthy results, in the visual domain or in the aural domain.

It was found to be possible to manipulate stops, vowels, fricatives and the other phoneme types in a systematic fashion, by varying the bandwidth of spatial filters with respect to the STFT time and frequency dimensions. It was also found that tradeoffs exist such that a filter capable of putting much emphasis on one type of phoneme, would produce distortions in or weaken other phoneme types. Stops and vowels were affected in this way. Fricatives tended to be problematic, as they have been found to be in other types of speech filtering based on the STFT, namely spectral subtraction and nonlinear spectral magnitude compression.

Regarding the visual enhancement of spectrograms, the simulated grey level resolution obtained by dithering on the Hercules monochrome monitor was found to be a serious limitation. Given a display with 8-bit or better discrete grey level resolution, it is expected that a more detailed study of visual spectrogram enhancement would be useful.

5.3 FURTHER APPLICATIONS OF THE METHODS

The results of this investigation suggest that further experimentation with application of image processing filters to speech is worthwhile. A number of alternative approaches could be taken, and some other objectives could be pursued. For example, a more direct approach to spatial filtering would be to take the two-dimensional Fourier transform of the spectrogram, and selectively modify it to attenuate or boost specific spatial frequencies. This approach would require an even greater amount of computing resources. Spatial filtering need not be limited to speech imagery in the form of spectrograms. It is possible that useful results could be obtained by applying spatial filters to the short-time cepstrum of speech.

The STFT is sometimes described in terms of a filter bank, with each spectral component corresponding to the output of a single filter [1,12]. In this filter bank model, the bandwidth of each filter is equal to that of every other. An alternative to this type of filter bank model is one based on the physiology of the human ear [26,27]. Filter bank models of the cochlea are composed of filters having equal bandwidths on a logarithmic frequency scale. Typically, the filters each have a bandwidth of 1/6 octave or 1/3 octave. It is possible that the application of filters to images of speech derived from perception-based models would yield an entirely different set of findings.

For any further work, it would be of great value to implement the processing algorithms on a state-of-the-art, high speed computer system. Ideally, the system would have some specialized hardware and software to perform numeric processing. It would also support a high resolution graphic display with variable luminance of at least 8 bit resolution. Such a system would benefit the investigator by providing results quickly for review, enabling him to maintain a train of thought while interacting with the system and

improvising on his methods. An apparatus for psychophysical testing and verification would be a probable requirement for quantitative work.

REFERENCES

1. L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, Prentice-Hall Inc., Englewood Cliffs, N. J. (1978).
2. W. Koenig, H. K. Dunn, and L. Y. Lacy, "The Sound Spectrograph", Journal of the Acoustical Society of America, Volume 17, (July 1946), pp. 19-49.
3. J. L. Flanagan, Speech analysis, Synthesis and Perception, 2nd Edition, Springer Verlag, New York (1972).
4. Alan V. Oppenheim, "Speech Spectrograms Using the Fast Fourier Transform", I.E.E.E. Spectrum, Volume 7, (August 1970), pp. 57-62.
5. Paul Mermelstein, "Computer-generated Spectrogram Displays for On-line Speech Research", I.E.E.E. Transactions on Audio and Electroacoustics, Vol AU-19, No. 3 (March 1971), pp. 44-47.
6. L. R. Morris, "Fast Speech Spectrogram Reduction and Display on Minicomputer/Graphics Processors", I.E.E.E. Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-23, No. 3 (June 1975), pp. 297-300.
7. R. K. Potter, G. A. Kopp, and H. G. Kopp, Visible Speech, D. Van Nostrand Co., New York, 1947. Republished by Dover Publications Inc. 1966.
8. Victor W. Zue and Ronald A. Cole, "Experiments on Spectrogram Reading", I.E.E.E. Conference Proceedings ICASSP, Washington, D.C. (1976), pp. 116-119.

9. Ronald A. Cole, Alexander I. Rudnick, Victor W. Zue, and D. Raj Reddy, "Speech as Patterns on Paper", Perception and Production of Fluent Speech, Ronald A. Cole ed., Lawrence Earlbaum Associates, Hillsdale, N.J. (1980), pp. 3-50.
10. Hong C. Leung and Victor W. Zue, "Visual Characterization of Speech Spectrograms", I.E.E.E. Conference Proceedings ICASSP, Tokyo, Japan (1986), Paper 51.1.
11. Ronald W. Schafer and Lawrence R. Rabiner, "Design and Simulation of a Speech Analysis-Synthesis System Based on Short-Time Fourier Analysis", I.E.E.E. Transactions on Audio and Electroacoustics, Vol AU-21, No. 3 (June 1973), pp. 165-173.
12. M. R. Portnoff, "Implementation of the digital Phase Vocoder Using the Fast Fourier Transform", I.E.E.E. Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-24, No. 3 (June 1976), pp. 243-248.
13. J. B. Allen, "Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform", I.E.E.E. Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-25, No. 3 (June 1977), 235-238.
14. R. E. Crochiere, "A Weighted Overlap-Add Method of Short-Time Fourier Analysis/Synthesis", I.E.E.E. Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-28, No. 1 (February 1980), pp. 99 - 102.
15. D. W. Griffin and J. S. Lim, "Signal Estimation from Modified Short-Time Fourier Transform", I.E.E.E. Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-32 (April 1984), 236-242.

16. M. R. Portnoff, "Time-Scale Modification of Speech Based on Short-Time Fourier Analysis", I.E.E.E. Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-29 (June 1981), pp. 374-390.
17. Timothy C. Green, "A Design Study for a DFT-Based Speech Modification System on a DSP Microchip", Master of Engineering Thesis Paper, Carleton University (1984).
18. S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", I.E.E.E. Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-27, No. 2 (April 1979), pp. 113-120.
19. E. Munday, "Noise Reduction using Frequency-domain Non-Linear Processing for the enhancement of Speech", British Telecom Technology Journal, Vol. 6, No. 2 (April 1988), pp. 71 - 83.
20. T. E. Eger, J. C. Su, and L. W. Varner, "A Nonlinear Spectrum Processing Technique for Speech Enhancement", I.E.E.E. Conference Proceedings ICASSP, Tokyo (1986), 18A.1.1 - 18A.1.4
21. Y. Arik, K. Kajimoto, and T. Sakai, "Acoustic Noise Reduction by Two Dimensional Spectral Smoothing and Spectral Amplitude Transformation", I.E.E.E. Conference Proceedings ICASSP, Tokyo (1986), Paper 3.5.
22. Rafael C. Gonzales and Paul Wintz, Digital Image Processing, Addison Wesley, Reading, Massachusetts (1977).
23. Robert A. Schowengerdt, Techniques for Image Processing and Classification in Remote Sensing, Academic Press, New York (1983), chapters 1-2.

24. Grafmatic Scientific/Engineering Graphics, user manual, (1984), Microcompatibles Inc., 301 Prelude Drive, Silver Spring M.D. 20901.
25. Samuel D. Stearns, Ruth A. David, Signal Processing Algorithms, Prentice-Hall Inc., Englewood Cliffs, New Jersey (1988).
26. J. B. Allen, "Cochlear Modelling", I.E.E.E. ASSP Magazine, Volume 2, No.1, (January 1985), pp. 3-29.
27. James C. Anderson and Campbell J. Searle, "Speech Analysis/Synthesis Based on Perception", I.E.E.E. Conference Proceedings ICASSP, Boston (1983), 28.10.

END

03·03·92

FIN