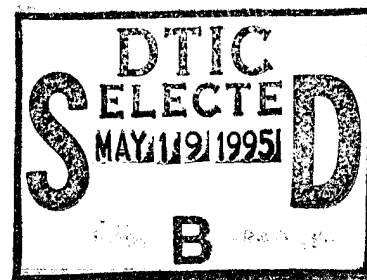


NAVAL POSTGRADUATE SCHOOL

Monterey, California



THESIS

SPEAKER IDENTIFICATION USING THE TWO-DIMENSIONAL CEPSTRUM TRANSFORM

by

Ioannis Lelakis

March, 1995

Thesis Advisor:
Co-Advisor:

Monique P. Fargues
Ralph Hippenstiel

Approved for public release; distribution is unlimited

19950518 024

DTIC QUALITY INSPECTED 5

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.			
1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE March 1995	3. REPORT TYPE AND DATES COVERED Master's Thesis	
4. TITLE AND SUBTITLE Speaker Identification Using the Two-Dimensional Cepstrum Transform		5. FUNDING NUMBERS	
6. AUTHOR(S) Ioannis Lelakis			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey CA 93943-5000		8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSORING/MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.			
12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited		12b. DISTRIBUTION CODE	
13. ABSTRACT (maximum 200 words) This thesis investigates the application of the two-dimensional cepstrum transform to a speaker identification system. Two distance measures are implemented for identification decision; the Euclidean distance and a weighted two-dimensional cepstral distance. The study considers three words to be tested under several noise levels. The effect of speaking rate during recordings is examined and is shown to be critical. Results show identification rates in the range of 95% to 98.5% for 50 dB signal to noise ratio and 57.65% to 80.7% for 0 dB signal to noise ratio.			
14. SUBJECT TERMS Speech Processing, Speaker Identification, Two-Dimensional Cepstrum.		15. NUMBER OF PAGES 96	
		16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18

Approved for public release; distribution is unlimited.

SPEAKER IDENTIFICATION USING THE TWO-DIMENSIONAL CEPSTRUM TRANSFORM

Ioannis Lelakis
LTJG, Hellenic Navy
Hellenic Naval Academy, 1987

Submitted in partial fulfillment
of the requirements for the degree of

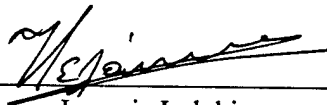
MASTER OF SCIENCE IN ELECTRICAL ENGINEERING

from the

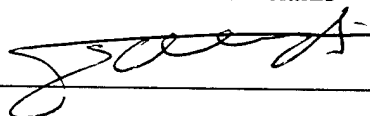
NAVAL POSTGRADUATE SCHOOL

March, 1995

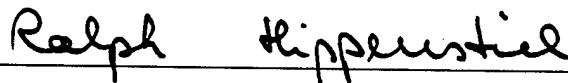
Author: _____


Ioannis Lelakis

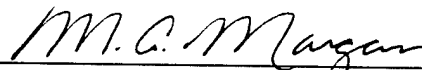
Approved by: _____



Monique P. Fargues, Advisor



Ralph Hippenstiel, Co-Advisor



Michael A. Morgan, Chairman
Department of Electrical and Computer Engineering

Accession For	
NTIS GRA&I	☒
DTIC TAB	☐
Unannounced	☐
Justification	
By _____	
Distribution/ _____	
Availability Codes	
Dist	Avail and/or Special
A-1	

ABSTRACT

This thesis investigates the application of the two-dimensional cepstrum transform to a speaker identification system. Two distance measures are implemented for the identification decision; the Euclidean distance and a weighted two-dimensional cepstral distance. The study considers three words to be tested under several noise levels. The effect of speaking rate during recordings is examined and is shown to be critical. Results show identification rates in the range of 95% to 98.5% for 50 dB signal to noise ratio and 57.65% to 80.7% for 0 dB signal to noise ratio.

TABLE OF CONTENTS

I. INTRODUCTION.....	1
II. SPEECH ANALYSIS	3
III. CEPSTRAL ANALYSIS	13
A. ONE-DIMENSIONAL CEPSTRUM	13
1. Introduction	13
2. Real Cepstrum	14
3. Complex Cepstrum	17
B. LIFTERING	19
C. APPLICATIONS OF THE REAL CEPSTRUM	22
D. TWO-DIMENSIONAL CEPSTRUM.....	25
1. Introduction	25
2. Examples	27
a. One Complex Exponential.	28
b. Two Complex Exponentials.	35
c. Phoneme /@/ From The Word "Man".	35
IV. SPEAKER IDENTIFICATION - PREPARATION	45
A. PROBLEM DEFINITION	45
B. DISTANCE MEASURES	46

C. DATA COLLECTION - PREPROCESSING	48
V. TESTS AND RESULTS	51
A. TESTS SET-UP	51
B. RESULTS	54
1. Speaker Recognition In Noisy Conditions.	54
2. Word Length Effects.	58
3. Robustness Of The Distance Measures.	58
4. Effects Of Frame Length And Overlap.	62
VI. CONCLUSIONS	73
APPENDIX A. MATLAB CODE FOR 2-D CEPSTRUM TRANSFORM.	75
APPENDIX B. MATLAB CODE FOR LIFTERING OPERATION	77
APPENDIX C. MATLAB CODE FOR 2-D CEPSTRAL DISTANCE.	79
LIST OF REFERENCES.	81
INITIAL DISTRIBUTION LIST.	83

LIST OF FIGURES

1.	Average formant locations for vowels in American English	7
2.	Time waveform and spectrum of the phoneme /@/ from the word 'man'	8
3.	Time waveform and spectrum of the phoneme /i/ from the word 'beat'	9
4.	Plot of second formant frequency versus first formant frequency for vowels by a wide range of speakers	11
5.	Speech production model	13
6.	Block-diagram for the computation of the Real Cepstrum.	16
7.	Block-diagram for the computation of the stRC using the FFT	16
8.	Computation of the complex cepstrum	18
9.	Block-diagram of liftering operation. The block stRC corresponds to the cepstrum computation of Figure 6	19
10.	Three types of low-time lifters	21
11.	Log spectra of individual frames for the word "man"	23
12.	Cepstral coefficients $c_s(n;m)$ obtained for the word "man", sampling frequency 8192 Hz, FFT size 512	24
13.	Block-diagram for "cepstral smoothing"	25
14.	Log spectrum S_{km} of $x(n)=\exp(j2\pi 0.3n)$; window length=512; 20% overlap	29
15.	S_{km} of $x(n)$ for $m=1$	30
16.	Magnitude of d_{qm} (1-D FFT of log spectrum S_{km} , FFT length=512, $f_s=1000$ Hz)	31
17.	Magnitude of 2-D cepstrum coefficients c_{qp} obtained for one complex exponential ($f_s=1000$ Hz, FFT length=64)	32
18.	Lower left quadrant of the cepstral matrix of Figure 17	33
19.	Unwrapped phase of c_{qp} of $x(n)$	34

20.	Log spectrum S_{km} of $y(n)=[\exp(j2\pi 0.3n), \exp(j2\pi 0.1n)]$; window length=512; 20% overlap	37
21.	S_{km} of $y(n)$ for $m=1$	38
22.	Magnitude of cepstrum coefficients c_{qp} for $0 < q < 256$ and $0 < p < 32$ of $y(n)$ ($f_s=1000$ Hz, FFT length along p-axis=64, FFT length along q-axis=512)	39
23.	Unwrapped phase of c_{qp} of two complex exponentials	40
24.	Magnitude of 2-D cepstrum coefficients for $1 \leq q \leq 256$ and $0 \leq p \leq 32$ for the phoneme /@/ ($f_s=8192$ Hz)	41
25.	Magnitude of the liftered c_{qp} coefficients (Lifter: raised sine, frame length= 32 msec, overlap= 75%, $f_s=8192$ Hz)	42
26.	Magnitude of the liftered c_{qp} coefficients (Lifter: raised sine, frame length= 32 msec, overlap=20%, $f_s=8192$ Hz)	43
27.	Magnitude of the liftered c_{qp} coefficients (Lifter: raised sine, frame length= 32 msec, overlap=20%, $f_s=8192$ Hz)	44
28.	General representation of the speaker recognition problem	46
29.	Block-diagram representing the preprocessing sequence for each word; f_s is the sampling frequency of the A/D conversion, and f_c is the cutoff frequency of the highpass filter	49
30.	Example of test set up for REF1 and test word spkr1(6)	52
31.	Word "beat" for two different speakers. In (a), the /t/ is clearly seen after the short period of silence following the phoneme /i/. In (b), the end of the word is not obvious since the /t/ is not clearly seen	57
32.	Identification performance for four combinations of overlap and frame length for the word "man", Euclidean distance	65
33.	Identification performance for four combinations of overlap and frame length for the word "beat", Euclidean distance	66
34.	Identification performance for four combinations of overlap and frame length for the word "indigestible", Euclidean distance	67

LIST OF TABLES

1.	Phonemes used in American English.	5
2.	Phonetic Alphabets.	6
3.	Identification rates for the word " <i>man</i> ", for SNR = 50, 20, 10, 5, 0, -5 dB for the Euclidean and 2-D cepstral distances; 2-D cepstral coefficients computed with 75% overlap and 256 time samples frame length.	55
4.	Identification rates for the word " <i>beat</i> ", for SNR = 50, 20, 10, 5, 0, -5 dB for the Euclidean and 2-D cepstral distances; 2-D cepstral coefficients computed with 75% overlap and 256 time samples frame length.	55
5.	Identification rates for the word " <i>indigestible</i> ", for SNR = 50, 20, 10, 5, 0, -5 dB for the Euclidean and 2-D cepstral distances; 2-D cepstral coefficients computed with 75% overlap and 256 time samples frame length.	56
6.	Identification rates of individual utterances for the word " <i>man</i> ", word length effect, Euclidian distance.	59
7.	Identification rates of individual utterances for the word " <i>beat</i> ", word length effect, Euclidean distance.	60
8.	Identification rates of individual utterances for the word " <i>indigestible</i> ", word length effect, Euclidean distance.	61
9.	Normalized distances and standard deviations for the case REF1 and TEST1 (Euclidean distance).	64
10.	Identification rates for the word " <i>man</i> ", for SNR = 50, 20, 10, 5, 0, -5 dB for the Euclidean distance, 2-D cepstral coefficients computed with 75% overlap and frame length 256 time samples.	68
11.	Identification rates for the word " <i>man</i> ", for SNR = 50, 20, 10, 5, 0, -5 dB for the Euclidean distance, 2-D cepstral coefficients computed with 75% overlap and frame length 512 time samples.	68
12.	Identification rates for the word " <i>man</i> ", for SNR = 50, 20, 10, 5, 0, -5 dB for the Euclidean distance, 2-D cepstral coefficients computed with 75% overlap and frame length 512 time samples.	69
13.	Identification rates for the word " <i>beat</i> ", for SNR = 50, 20, 10, 5, 0, -5 dB for the Euclidean distance, 2-D cepstral coefficients computed with 20% overlap and frame length 256 time samples.	69

14.	Identification rates for the word " <i>beat</i> ", for SNR = 50, 20, 10, 5, 0, -5 dB for the Euclidean distance, 2-D cepstral coefficients computed with 20% overlap and frame length 512 time samples.	70
15.	Identification rates for the word " <i>beat</i> ", for SNR = 50, 20, 10, 5, 0, -5 dB for the Euclidean distance, 2-D cepstral coefficients computed with 75% overlap and frame length 512 time samples.	70
16.	Identification rates for the word " <i>indigestible</i> ", for SNR = 50, 20, 10, 5, 0, -5 dB for the Euclidean distance, 2-D cepstral coefficients computed with 20% overlap and frame length 256 time samples.	71
17.	Identification rates for the word " <i>indigestible</i> ", for SNR = 50, 20, 10, 5, 0, -5 dB for the Euclidean distance, 2-D cepstral coefficients computed with 20% overlap and frame length 512 time samples.	71
18.	Identification rates for the word " <i>indigestible</i> ", for SNR = 50, 20, 10, 5, 0, -5 dB for the Euclidean distance, 2-D cepstral coefficients computed with 75% overlap and frame length 512 time samples.	72

I. INTRODUCTION

There have been several studies in the past dealing with the communication between humans and machines through speech. The problem of speaker identification is examined in this thesis, where speaker identification refers to the task of identifying a given speaker among a group of several known speakers using test utterances. The applications of speaker identification cover a wide area which may include efficient banking and business transactions, controlled access of a specific space or information to selected individuals for security purposes, etc.

The method of speaker identification focuses on transforming the speech signal into a set of parameters that will efficiently represent the individual speaker and then compare these parameters with a library of stored reference templates of parameters of a group of people. The task is then for the machine to find the closest match and make a decision. Several signal processing techniques have been applied in order to create these parameters. The vocal tract characteristics uniquely describe the voice of each individual and these characterizing parameters may be obtained by using one-dimensional analysis. The two-dimensional cepstrum is an extension of the one-dimensional in that it takes into account the variations of the cepstral coefficients. Therefore, in this thesis, the two-dimensional cepstrum transform is used to build word patterns to represent each speaker.

The use of three words is considered: two simple monosyllable and one longer, in order to examine the performance of the system in combinations of voiced and unvoiced speech. The robustness of the recognizer is also investigated in noisy conditions when background noise is added with a user-defined Signal to Noise Ratio.

Chapter II gives a brief analysis of speech and its characteristics. Chapter III presents the basics of cepstral analysis. We first introduce the one-dimensional cepstrum, and then present the two-dimensional cepstral extension. Chapter IV analyzes the data collection and preparation process. Chapter V describes the test set-up and the results

obtained. Finally, Chapter VI presents conclusions and recommendations for future research.

II. SPEECH ANALYSIS

Speech signals are used for communication and exchange of information between two or more people. Fundamentally, speech is made up of sound pressure waves that are produced by the mouth of a speaker when and traveling through a medium are perceived by a listener. These pressure waves are primarily produced in the lungs. The resulting flow of air passing through the trachea, glottis, larynx, pharynx, mouth and nose generates the various sounds. These sounds, depending on the position of the vocal tract articulators, namely the vocal cord, tongue, lips and velum, produce the phones, words and phrases that make up every spoken language.

Speech is divided into voiced and unvoiced sections depending on the means of excitation. Voiced speech is produced when air is blown through the glottis or between the vocal folds. The vocal cords then vibrate due to their tension in a quasi-periodic fashion. The sound produced in that way is called voice or phonation. Some examples of voiced sounds are the sounds /i/ in "eve", /E/ in "met" and /@/ in "at". Unvoiced speech is produced when there is a constriction at some point of the vocal tract. Examples of unvoiced sounds are /s/ in "see", /f/ in "for", /T/ in "thin" and /S/ in "she".

Speech can be represented as the concatenation of elements from a well defined set of symbols. The basic units or symbols from which each sound can be classified are called phonemes. When the phonemes are combined, they produce the words and phrases that are used for communication. The combinations of the phonemes follow certain rules, and the science studying them is called linguistics. Phonetics is the study and classification of speech sounds. Coarticulation is the term used to refer to the change in phoneme articulation and acoustics caused by the influence of another sound in the same utterance. Articulators are the finer anatomical features like the vocal cords, the velum, the tongue, the teeth and the lips that move to different positions to produce various speech sounds. These movements, that most of the time overlap in time, have an effect on the transitions from one sound to the next one, as well as on the duration of phonemes. For example, consider the duration of the utterance "an" spoken by itself and when spoken in the

sentence "an open door". The duration is significantly reduced when spoken in the phrase. The vocal and nasal tracts can be represented as tubes of nonuniform cross-sectional area. The resonant frequencies of the vocal tract tube, when sound propagates through it, are called formant frequencies which depend upon the shape and dimensions of the vocal tract.

Phonemes can be classified depending on properties related to the time waveform, characteristic frequencies, manner of articulation, place of articulation, type of excitation and stationarity of phoneme. In general, phonemes are divided into continuants and noncontinuants. Continuant is a phoneme whose sound is produced by a steady-state vocal tract configuration. Noncontinuant is a phoneme where there is a change in the vocal tract configuration. Table 1 presents a classification of phonemes of American English and Table 2 lists their respective phonetic representation, as well as, examples of how each phoneme is pronounced in the context of a word.

Phones are defined as the actual sounds produced by speakers, which lead to the understanding of the intended meaning of the sounds. Phoneme sounds, in normal speech, have transition periods between them, therefore, with each phoneme a group of transitional phone variations called allophones are associated [Ref. 1].

Vowels are among the phonemes with the largest amplitudes. They are produced by exciting a fixed vocal tract with quasi-periodic pulses of air generated by vibration of the vocal cords. They are distinguished by the frequency location of their first three formants, whose averages are shown in Figure 1 [Ref. 1]. The group of sounds consisting of /w/, /l/, /r/ and /y/ are called semivowels. They are classified as either liquids (/w/, /l/) or glides (/r/ and /y/). Note that they are called semivowels because they have similar spectral characteristics to vowels.

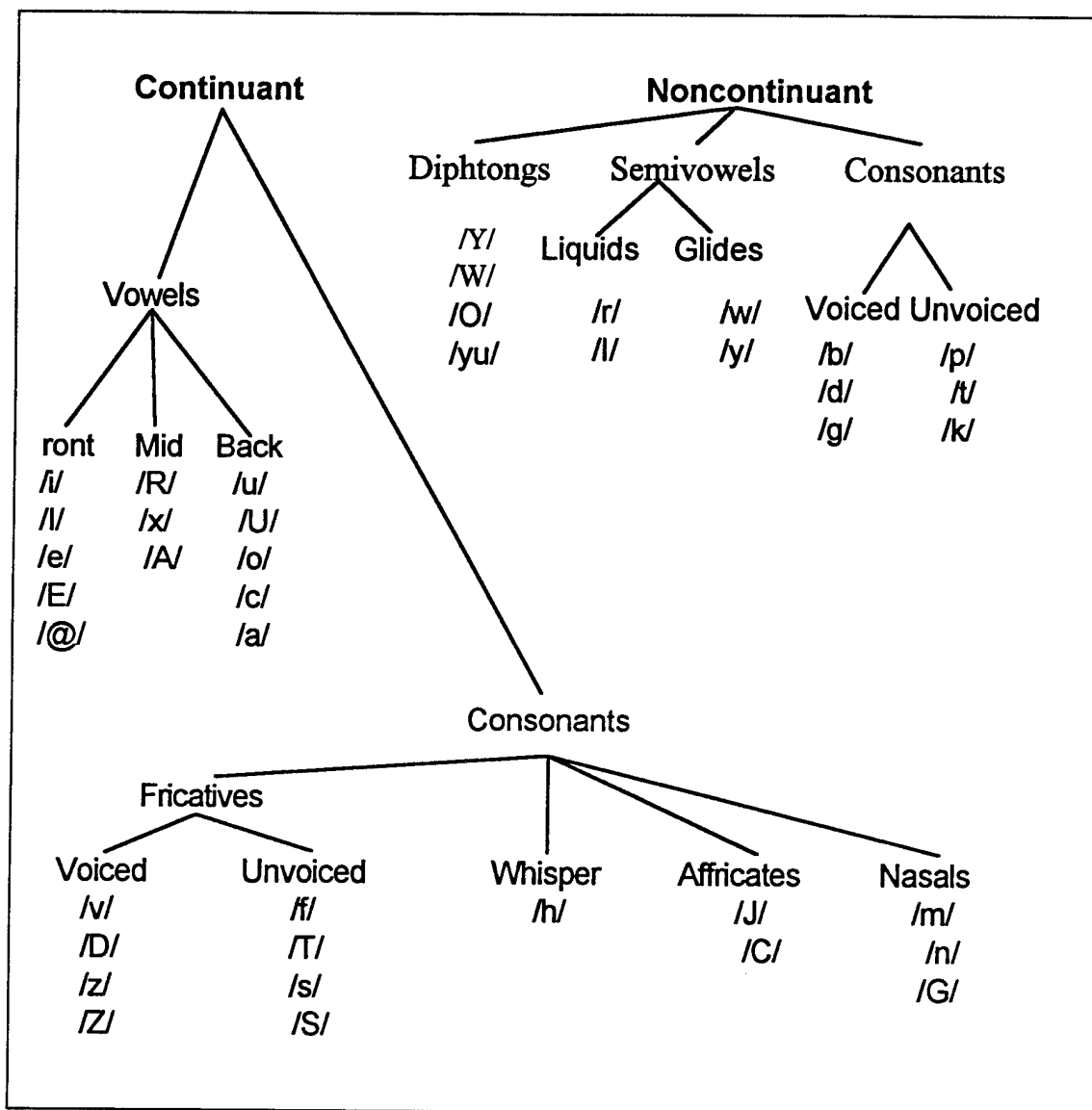


Table 1. Phonemes used in American English.

Single Symbol Version	Upper case Version	Examples	Single Symbol Version	Upper case Version	Examples	Single Symbol Version	Upper case Version	Examples
i	IY	heed	d	D	deep	t	T	tea
I	IH	hid	k	K	kick	F	DX	batter
e	EY	hayed	g	G	go	Q	Q	quit
E	EH	head	f	F	five	w	W	want
@	AE	had	v	V	vice	y	Y	yard
a	AA	hod	T	TH	thing	r	R	race
c	AO	hawed	D	DH	then	C	CH	church
o	OW	hoed	s	S	so	J	JH	just
U	UH	hood	z	Z	zebra	H	WH	when
u	UW	who'd	S	SH	show	b	B	bat
R	ER	heard	Z	ZH	measure	p	P	pea
x	AX	ago	h	HH	help	M	EM	some
A	AH	mud	m	M	mom	N	EN	son
Y	AY	hide	n	N	noon	X	IX	roses
W	AW	how'd	G	NX	sing	L	EL	cattle
O	OY	boy	l	L	love			

Table 2. Phonetic Alphabets.

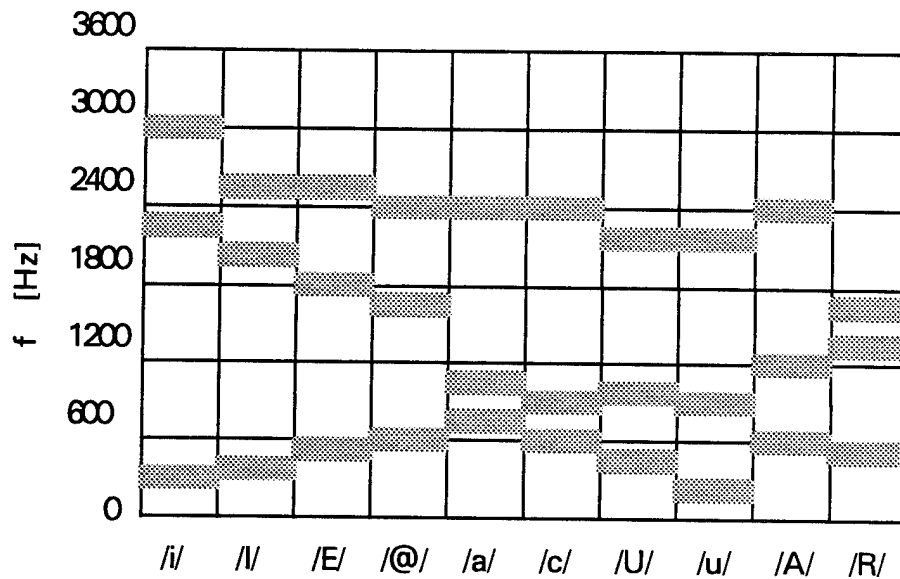


Figure 1. Average formant locations for vowels in American English, [Ref 1].

The quasi-periodic features of vowels are illustrated in Figure 2 and Figure 3. They respectively show the time domain and frequency domain representations for the vowels /@/ from the word "man" and /i/ from the word "beat".

Diphtongs are produced by the movement from one vowel toward another. This movement is done by varying in time the vocal tract during this transition. Semivowels are weaker than vowels and are classified as glides and liquids. They are transitional and vowel-like sounds and, hence, are similar in nature to the vowels and diphtongs. Nasals are voiced sounds produced by the glottal waveform exciting an open nasal cavity and closed oral cavity.

Fricatives are divided into voiced and unvoiced phonemes. The voiced fricatives are produced by exciting the vocal tract by a steady air flow and a region of the vocal tract is constricted. The location of the constriction determines the sound spoken. The voiced

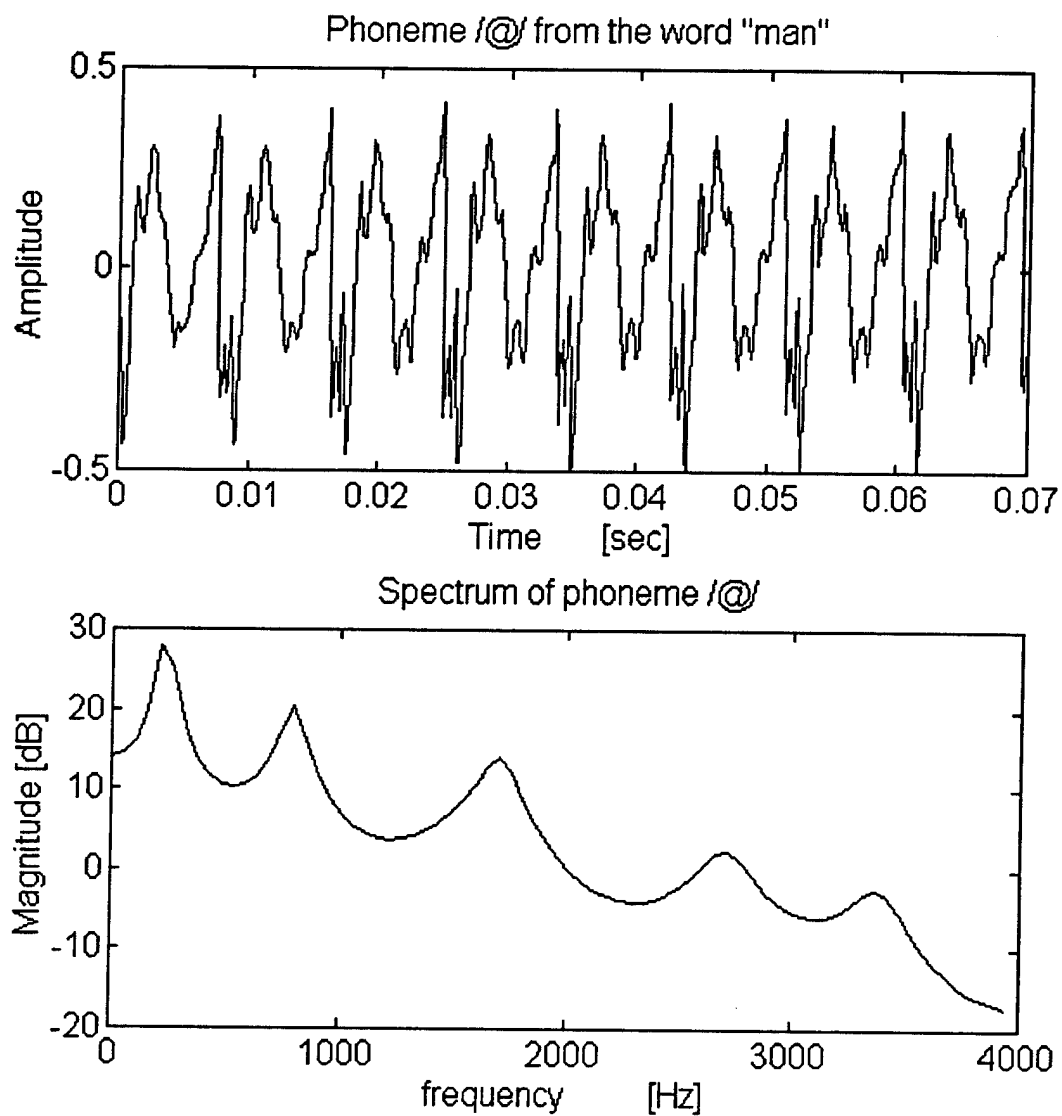


Figure 2. Time waveform and spectrum of the phoneme /@/ from the word 'man'.

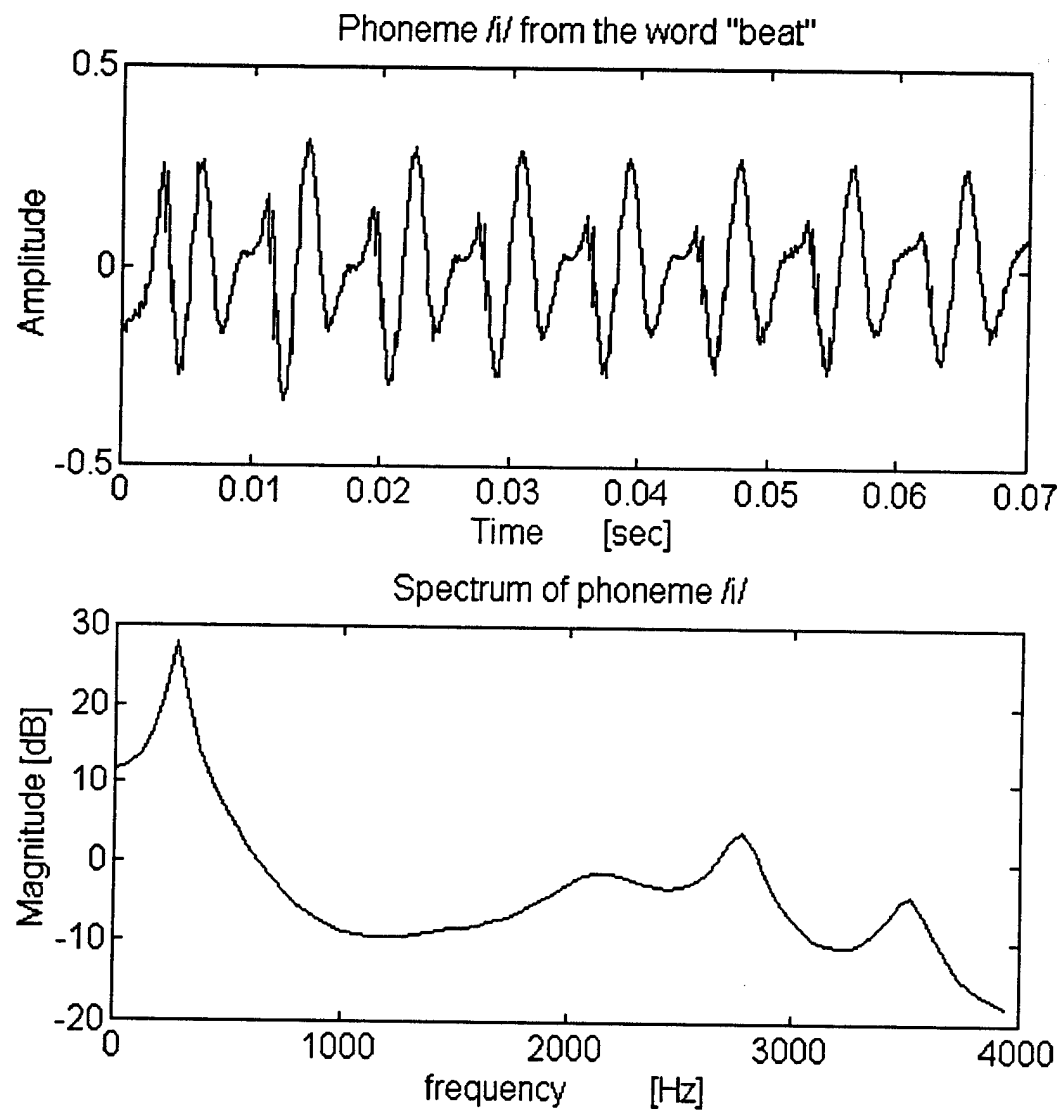


Figure 3. Time waveform and spectrum of the phoneme /i/ from the word 'beat'.

region of the vocal tract. There are also periodic glottal pulses exciting the vocal tract, which mixed with the excitation described before, produce the voiced fricatives. These two excitation sources cause two distinct spectral components. Stop consonants are transient, noncontinuant sounds that are produced by pressure built-up due to total constriction, followed by a sudden release of this pressure. The voiced stops differ from the unvoiced ones in that their production also includes vocal fold vibration. Affricates are sounds produced by transitions from a stop to a fricative.

Each specific individual has unique vocal tract characteristics which makes his/her own voice different from others. Each sound can be characterized by the vocal tract configuration that is used in its production. The changing resonant structure in the vocal tract is reflected as shifts in formant frequency locations. Figure 4 shows a plot of second formant frequency as a function of first formant frequency for several vowels spoken by a wide range of speakers [Ref. 2]. It can be easily seen that the broad ellipses drawn show the approximate range of variation in formant frequencies for each of these vowels. From this last figure, one can observe that not all speakers produce the same frequencies for each phoneme, although their locations vary within certain limits, but everyone has his/her unique characteristic way of speaking and producing sounds.

III. CEPSTRAL ANALYSIS

A. ONE-DIMENSIONAL CEPSTRUM

1. Introduction

As discussed in the previous chapter, speech is produced by a flow of air passing through the vocal tract. In more engineering terms, this can be represented by the filter of Figure 5, where the excitation sequence $e(n)$ is filtered with a time-varying linear filter to create the speech signal $s(n)$. The impulse response $\theta(n)$ of this filter corresponds to the vocal tract characteristics. The output speech signal $s(n)$ can be either voiced or unvoiced, depending on the type of the excitation sequence.

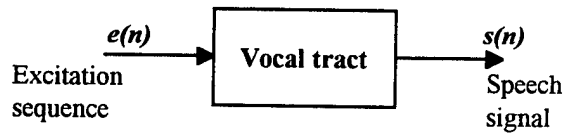


Figure 5. Speech production model.

The excitation sequence $e(n)$ is described by a set of periodic impulses when the resulting speech signal is voiced, and white noise when the speech signal is unvoiced [Ref. 1]. Thus, $e(n)$ is given by the following expression:

$$e(n) = \begin{cases} \sum_{q=-\infty}^{\infty} \delta(n - qT_p) & \text{voiced case} \\ \text{white; Gaussian; noise} & \text{unvoiced case} \end{cases}, \quad (1)$$

where n is the number of samples and T_p is the pitch period. Therefore, the speech signal can be expressed as:

$$s(n) = e(n) * \theta(n). \quad (2)$$

Since the excitation and impulse response of a linear time invariant system are combined in a convolutional manner, the problem of speech analysis can also be viewed as a problem in separating the components of the convolution to isolate the vocal tract characteristics represented by $\theta(n)$ [Ref. 2]. Such a separation can be obtained using cepstral analysis.

Historically, the cepstrum has its roots in the general problem of the deconvolution of two or more signals and was first proposed to decouple vocal tract characteristics from the excitation source by Bogert, Healy and Tukey (1963) [Ref. 3] and Noll (1967) [Ref. 4]. Two implementations of the cepstrum exist; the real cepstrum (RC) and the complex cepstrum (CC). Both implementations are discussed further and compared, although only the real cepstrum is used in the experiments.

2. Real Cepstrum

The one-dimensional real cepstrum of a speech signal $s(n)$ is given by:

$$c_s(n) = F^{-1} \{ \log |F\{s(n)\}| \} = \frac{1}{2\pi} \cdot \int_{-\pi}^{\pi} \log |S(\omega)| \cdot e^{j\omega n} d\omega, \quad (3)$$

where the operator $F\{\cdot\}$ denotes the DFT and ω the digital frequency.

Recall that a speech signal $s(n)$ is obtained from convolving the vocal system impulse response $\theta(n)$ with an excitation sequence $e(n)$ such that $s(n) = e(n) * \theta(n)$. Thus, Fourier transforming both sides of equation (3) leads to:

$$S(\omega) = E(\omega) \cdot \Theta(\omega). \quad (4)$$

Taking the logarithm of the magnitude of $S(\omega)$ we get:

$$C_s(\omega) = \log \{ |S(\omega)| \}$$

$$\begin{aligned}
&= \log \{ |E(\omega) \cdot \Theta(\omega)| \} \\
&= \log \{ |E(\omega)| \} + \log \{ |\Theta(\omega)| \} \\
C_s(\omega) &= C_e(\omega) + C_\theta(\omega).
\end{aligned} \tag{5}$$

Note that $C_s(\omega)$ is the linear combination of the two components $C_e(\omega)$ and $C_\theta(\omega)$, and also that it is real and even. The last step of the cepstrum operation is the inverse Fourier transform of $C_s(\omega)$ which leads to:

$$c_s(n) = \frac{1}{2\pi} \cdot \int_{-\pi}^{\pi} C_s(\omega) \cdot e^{j\omega n} d\omega. \tag{6}$$

However, due to the fact that $C_s(\omega)$ is real and even, we can substitute a straight Fourier transform for the inverse transform operation without changing the final expression for $c_s(n)$:

$$c_s(n) = c_e(n) + c_\theta(n). \tag{7}$$

The new domain introduced with the cepstrum transformation is called the queffrequency domain and has time dimensions. The coefficient $c_s(0)$ is related to the energy of the speech signal and is usually discarded, as discussed later. Experiments show that for speech signals the two components $c_e(n)$ and $c_\theta(n)$ occupy different parts of the queffrequency axis. The part $c_e(n)$ corresponds to the excitation source and is represented by a decaying train of periodic impulses which occupies the higher queffrequency portion. The period of the impulses represents the pitch period. The part $c_\theta(n)$ characterizes the vocal tract and occupies the low queffrequency portion up to the first impulse of $c_e(n)$.

Graphically, the real cepstrum transformation can be represented by the block-diagram given in Figure 6.

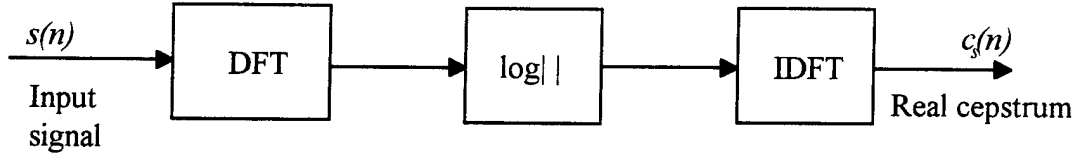


Figure 6. Block-diagram for the computation of the Real Cepstrum.

Note that in practical implementations we define the short-term real cepstrum (stRC) over finite-time windows. The only difference between the long-term cepstrum and the short-term is that for the short-term cepstral transformation the speech signal is first separated in frames of length N and each of these frames is processed individually. FFT's are used to compute cepstral coefficients, as it is more computationally efficient. In addition, zero-padding the frames of the speech signal is usually required to avoid aliasing [Ref. 1]. The block-diagram for the stRC computation is shown in Figure 7, where m denotes the time sample at which the N -length frame of speech ends, and $f(n; m) = s(n)w(m - n)$. The sequence $w(m - n)$ denotes the window used to separate the speech signal $s(n)$ into successive frames.

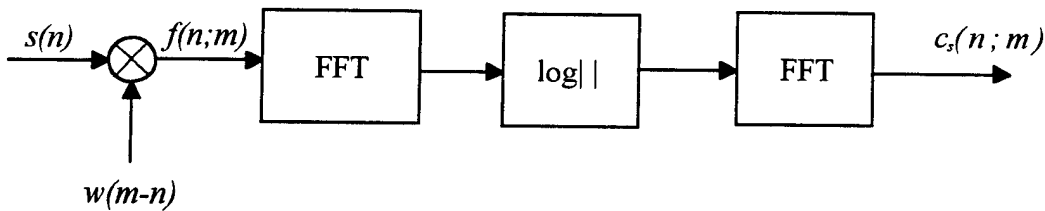


Figure 7. Block-diagram for the computation of the stRC using the FFT.

3. Complex Cepstrum

Backtransforming the real cepstrum information to recover the original signal is impossible, since the phase information is lost when the log spectrum is computed. This drawback is corrected with the complex cepstrum (CC) transformation by replacing the $\log|S(\omega)|$ operation with the complex logarithm of the DFT, which preserves the phase information. The complex cepstrum $\gamma_s(n)$ of the signal $s(n)$ is defined as:

$$\gamma_s(n) = F^{-1} \{ \log(F\{s(n)\}) \} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log S(\omega) e^{j\omega n} d\omega, \quad (8)$$

where ω is the digital frequency, and the logarithm is complex. Recall that the logarithm of a complex number is defined as:

$$\log z = \log |z| + j \arg \{z\}, \quad (9)$$

and consequently the complex logarithm of the spectrum $S(\omega)$ is equal to:

$$\log S(\omega) = \log |S(\omega)| + j \arg \{S(\omega)\}. \quad (10)$$

It is obvious that this transformation preserves the phase of the spectrum, thereby allowing us to return to the original time domain if so desired. The computation of the complex cepstrum is shown in the block-diagram of Figure 8, where the two branches for the computation of the complex logarithm are illustrated. The resulting cepstrum coefficients are given by:

$$\gamma_s(n) = \gamma_e(n) + \gamma_\theta(n), \quad (11)$$

where $\gamma_e(n)$ represent the coefficients corresponding to the excitation sequence and $\gamma_\theta(n)$ represent the vocal tract characteristics. Similarly to the real cepstrum, each component of $\gamma_s(n)$ occupies a different part of the queffreny axis. The lower portion of the axis is

occupied by $\gamma_e(n)$ and $\gamma_o(n)$ occupies the higher portion of the frequency axis. Now, either of the two components can be separated with appropriate windowing and back-transformed to the original time domain.

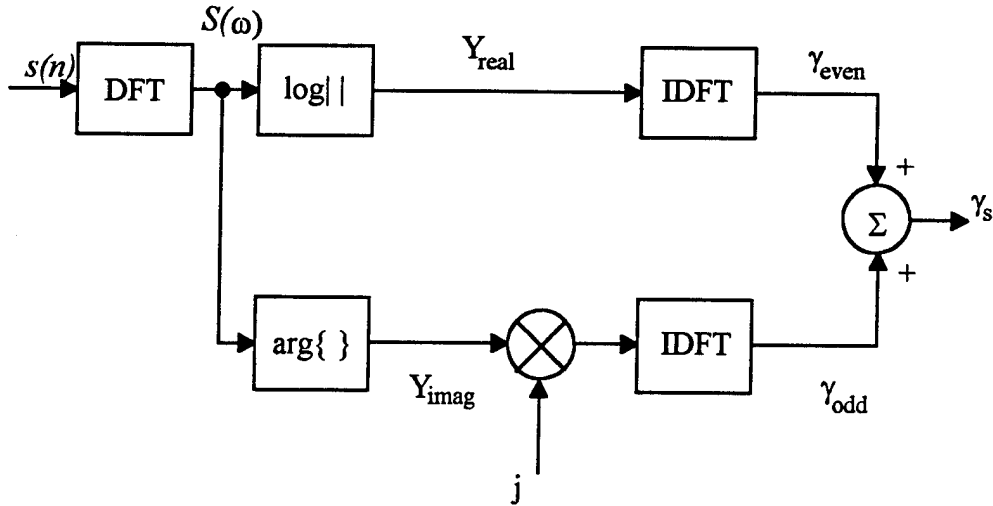


Figure 8. Computation of the complex cepstrum.

Note that the complex logarithm is multivalued, thus, discontinuities in the complex cepstrum phase may appear when the imaginary part of the logarithm is computed modulo 2π . Such a discontinuity is not allowed from the definition of the complex cepstrum, which requires that the imaginary part of $\log S(\omega)$ be a continuous and periodic function of ω . This problem can be avoided by unwrapping the phase of $\log S(\omega)$ which changes phase jumps greater than π to their 2π complement, thereby eliminating discontinuities in the phase curve. In addition, the complex cepstrum can also be implemented in the short-term sense, as we did with the real cepstrum.

The real cepstrum is equivalent to the even part of the complex cepstrum. If the application under study does not require back-transformation to the original time domain, then it is more preferable to use the real cepstrum due to its simplicity.

B. LIFTERING

It is possible to separate the excitation from the vocal tract characteristics using cepstral analysis, as described in section A. In order to eliminate one of the two components, we apply a linear filter, which is called a 'lifter' in the cepstrum domain. The procedure of the liftering operation is shown in Figure 9.

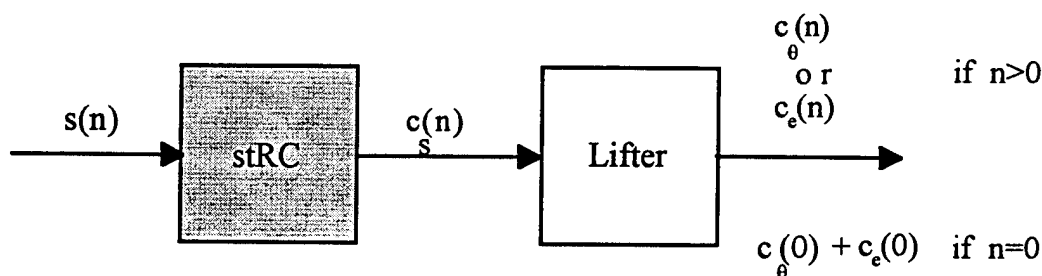


Figure 9. Block-diagram of liftering operation. The block stRC corresponds to the cepstrum computation of Figure 6.

The most popular lifters used in speech applications are low-time lifters, which eliminate the higher part of the cepstrum corresponding to the excitation sequence. The specific size and shape vary according to the type of application considered. Some of the more frequently used low-pass lifters include the rectangular lifter, the triangular lifter and the raised sine lifter, which are shown in Figure 10. The time-domain expressions for these three lifters are given by:

rectangular lifter,

$$w_1(k) = \begin{cases} 1 \\ 0 \end{cases}, \quad \begin{matrix} k = 1, \dots, L \\ \text{otherwise} \end{matrix},$$

triangular lifter,

$$w_2(k) = \begin{cases} 1 + \frac{L}{2} \cdot (k-1)/(L-1) \\ 0 \end{cases}, \quad \begin{matrix} k = 1, \dots, L \\ \text{otherwise} \end{matrix}, \quad (12)$$

raised sine lifter,

$$w_3(k) = \begin{cases} 1 + \frac{L}{2} \cdot \sin\left(\frac{k\pi}{L}\right) \\ 0 \end{cases}, \quad \begin{matrix} k = 1, \dots, L \\ \text{otherwise} \end{matrix},$$

where L is the length of the window and is usually chosen to be less than one pitch period.

Note that the raised sine lifter, initially proposed by Juang et al. [Ref. 4], is mostly used for cepstral smoothing, as it reduces the variation of the cepstrum coefficients between different speakers. This lifter allows the user to reduce the effects due to low and high-order coefficients which have higher variance. Variability in lower order coefficients is a result of the variations in transmission and speaker characteristics, and thus, it is desired to reduce them. Variability in higher quefrency coefficients is an artifact of the procedure of the cepstral transformation, especially when the cepstral coefficients are derived using LPC analysis. Such liftering can be very useful, especially in speech recognition, where it was shown to increase the performance of the recognizer.

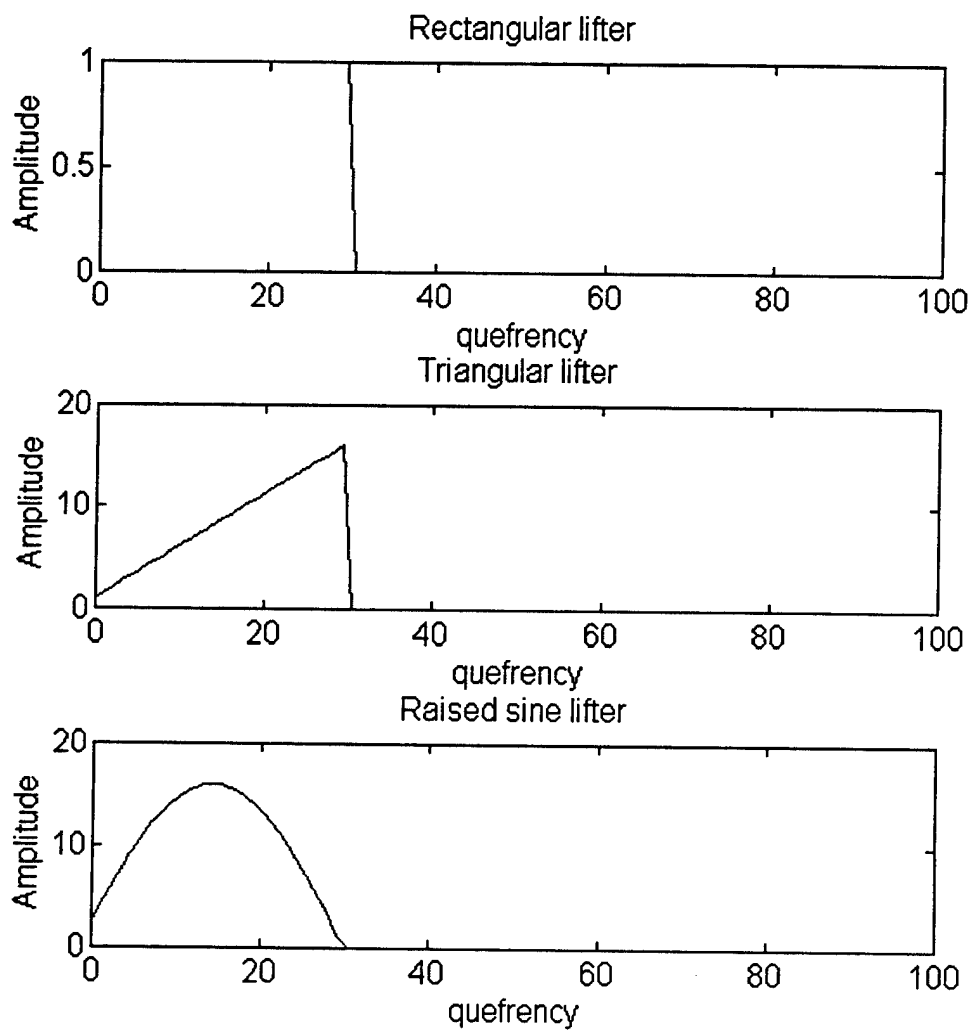


Figure 10. Three types of low-time lifters.

C. APPLICATIONS OF THE REAL CEPSTRUM

Cepstral analysis is used in speech processing to separate the excitation source of speech from the vocal tract characteristics. The real cepstrum is mostly used in pitch detection, formant estimation and speech recognition. All three of these applications are briefly discussed in what follows .

The pitch period is defined only for voiced speech signals and such information is included in the excitation source used for their production. In [Ref. 1], it is shown that the higher cepstrum coefficients correspond to the excitation source and are approximated by a periodic impulse train with period equal to the pitch period. The part $c_s(n)$ that corresponds to the vocal tract characteristics usually decays rapidly with respect to the pitch period. Therefore, the peaks are easily distinguished from the rest of the cepstrum, and since the quefrency axis has time dimensions, pitch period can be estimated, as shown in the following example.

The word "*man*" was recorded and digitized with a sampling frequency of 8192 Hz. First, the word sequence was divided into frames of length $N=256$, which corresponds to time duration of 32 ms, with a 50% overlap. Figure 11 plots the respective log spectra, zero padded to 512, obtained for each frame. Next, we compute the FFT of the log spectra to obtain the one-dimensional cepstrum coefficients, as shown in Figure 12. The peaks that appear at approximately 75 and 150 time samples (i.e. 9.15 msec and 18.3 msec), on the quefrency axis correspond to the pitch period and twice the pitch period of the speech signal. The absence of peaks in the first and last frames indicates unvoiced speech or silence. Note that the window length must be long enough to cover at least two periods of the voiced portion of the speech signal. Otherwise, the resulting cepstrum no longer consists of an impulse train, and the pitch cannot be detected. The coefficient $c_s(0)$ represents the energy of the signal and is not shown in the plots, since it has been observed that absolute power measures of the signal are unreliable and the use of $c_s(0)$ has been de-emphasized in the literature [Ref. 6]. Therefore, from this point on, $c_s(0)$ is not included in the set of coefficients used in our study.

Log spectra of individual frames for the word "man"

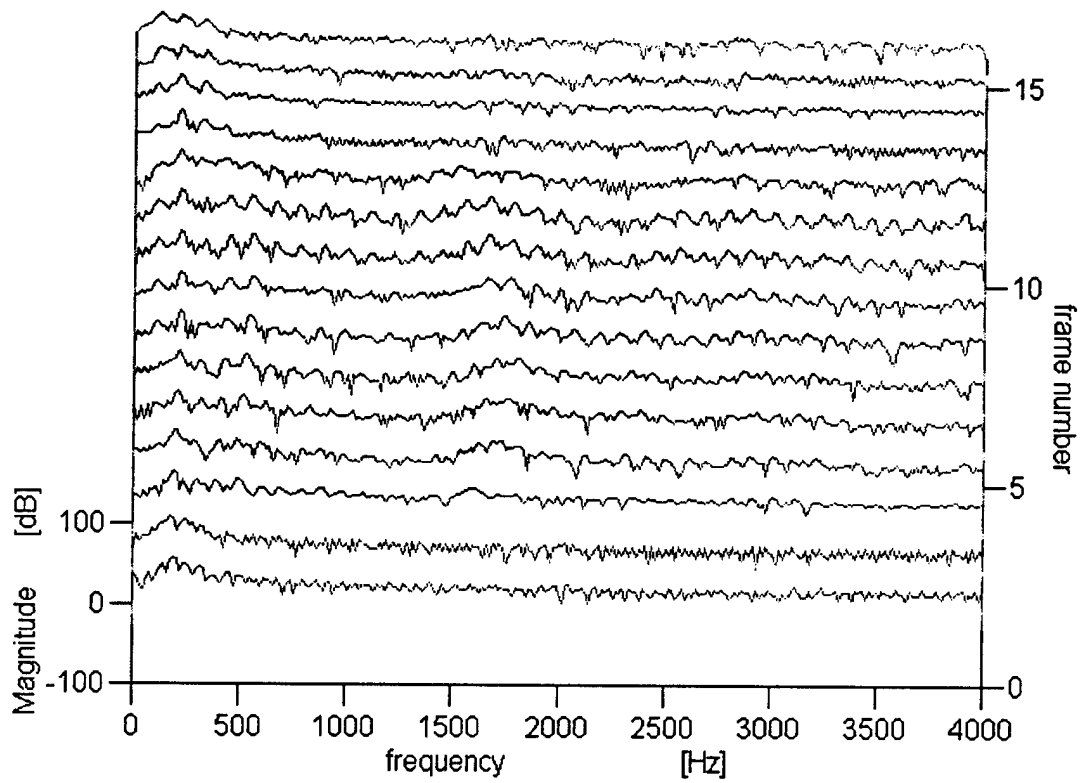


Figure 11. Log spectra of individual frames for the word "*man*".

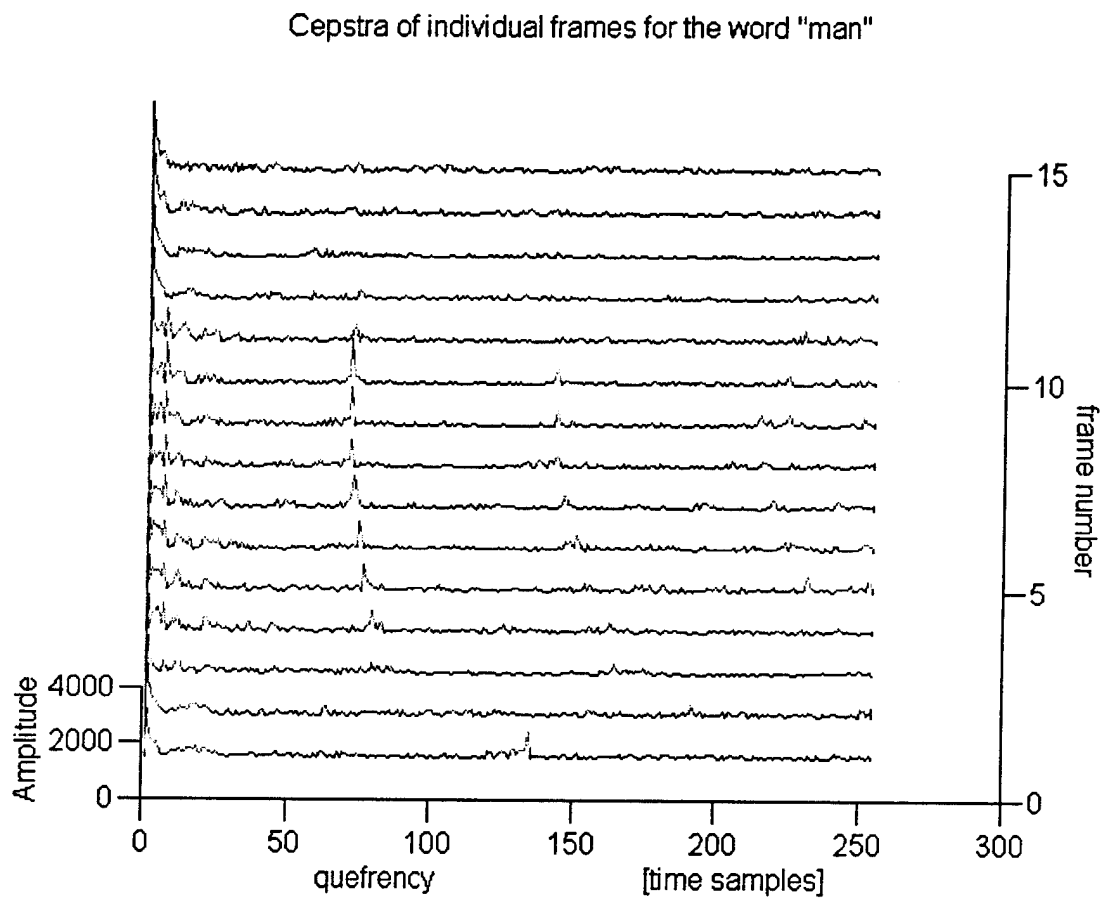


Figure 12. Cepstral coefficients $c_s(n,m)$ obtained for the word "man", sampling frequency 8192 Hz, FFT size 512.

A second application of the cepstral analysis is in formant estimation, which is done by "cepstral smoothing" to produce an estimate of $\Theta(\omega, m)$ or $\log|\Theta(\omega, m)|$. The procedure to generate the set of characteristic parameters is shown in Figure 13.

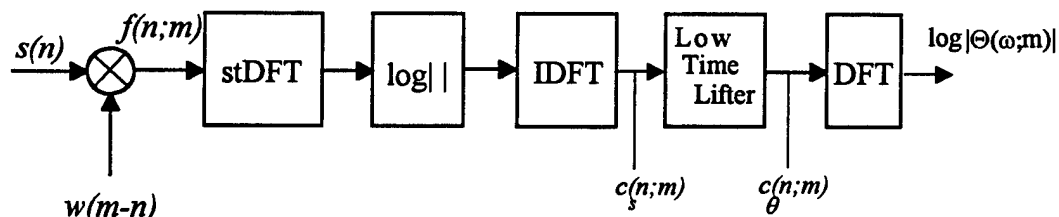


Figure 13. Block-diagram for "cepstral smoothing".

The coefficients $c_\theta(n; m)$ can be isolated from $c_s(n; m)$ by applying a low-time lifter. The estimate of $\log|\Theta(\omega, m)|$ can be obtained by computing the DFT of $c_\theta(n; m)$ [Ref. 1].

Speech recognition is an area where cepstral analysis is mostly applied. Another way of computing the cepstral coefficients is recursively from the LPC parameters associated with the speech signal. Such resulting cepstral coefficients produce a smoothed version of the cepstrum coefficients derived using FFT's as previously discussed. Hence, they provided a far superior performance when used in speech recognition since the differences between various speakers are reduced.

D. TWO-DIMENSIONAL CEPSTRUM

1. Introduction

The two-dimensional cepstrum is the extension of the one-dimensional cepstrum, as described earlier. The two-dimensional cepstrum represents both static and dynamic features of speech, as well as, frequency and time variations of speech at the same time [Ref. 7].

The generation of the two-dimensional cepstral matrix is similar to that of the one-dimensional cepstrum. First, the speech signal is divided into frames of fixed length long enough to include at least two periods of the voiced part of the speech signal, usually 32 to 64 ms. Then, the log spectrum S_{km} of each frame is computed by:

$$S_{km} = 10 \log \left| \sum_{n=0}^{N-1} s_{nm} W_1^{-nk} \right|^2, \quad (13)$$

where s_{nm} represents the n th point of the m th frame, N is the length of each frame, M is the number of frames, k is the frequency index, m denotes the frame number and,

$$W_1 = \exp(j2\pi/N), \quad 0 \leq k \leq N-1 \text{ and } 0 \leq m \leq M-1. \quad (14)$$

The two-dimensional cepstral coefficient c_{qp} is obtained by applying a two-dimensional FFT to S_{km} , which leads to:

$$C_{qp} = \frac{1}{NM} \sum_{k=0}^{N-1} \sum_{m=0}^{M-1} S_{km} W_1^{-kq} W_2^{-mp}, \quad (15)$$

where $W_2 = \exp(j2\pi/M)$, with $0 \leq q \leq N-1$ and $0 \leq p \leq M-1$. Due to symmetry properties of the two-dimensional FFT, only one quarter of the matrix needs to be used. The q -axis is called quefrency and has time dimension. Each row along the q -axis represents the one-dimensional cepstrum of each frame. The p -axis is called time-frequency and has frequency dimension. This axis indicates the variation of given cepstral coefficients along the frames.

The higher components c_{qp} on the q axis correspond to the fine structure of the spectrum. The lower components correspond to the spectral envelope. The fine structure corresponds to the excitation source, and the spectral envelope corresponds to the vocal

tract characteristics. Higher components on the p -axis correspond to local time variation and lower components correspond to global time variation. [Ref. 7]

In order to get a better understanding of the two-dimensional cepstrum, we replace the two-dimensional operation in (15) with two successive one-dimensional FFTs in order to study the effects of each FFT operation onto S_{km} . Thus, applying a first FFT to S_{km} along the k axis leads to:

$$d_{qm} = \frac{1}{N} \sum_{k=0}^{N-1} S_{km} W_1^{-kq}. \quad (16)$$

Next, applying the second one-dimensional FFT to d_{qm} along the m axis leads to the resulting cepstral coefficient c_{qp} :

$$c_{qp} = \frac{1}{M} \sum_{m=0}^{M-1} d_{qm} W_2^{-mp}. \quad (17)$$

Note that the coefficients d_{qm} represent the cepstral coefficients obtained at a given frame m . Next, the coefficients c_{qp} represent the variation of the q^{th} spectral coefficient in the p -frequency domain.

2. Examples

In this section, we apply the two-dimensional cepstral transformation to a few signals to investigate the transform properties. The signals considered are:

- i. One complex exponential,
- ii. Two complex exponentials,
- iii. Phoneme /@/.

a. One Complex Exponential

The complex exponential signal $x(n)$ considered is given by:

$$x(n) = \exp(j2\pi(0.3)n), \quad (18)$$

where 0.3 represents the normalized frequency of the signal and the sampling frequency is $f_s=1000$ Hz. The data length is equal to 2000 points (2 sec). A window with length equal to 512 points (512 msec) with a 20% overlap is used in the study. Thus 20% of the length of the window corresponds to 102 points (102 msec). Hence, the first data frame begins at point 1 and ends at point 512. The second frame begins at point 411 and ends at point 922. Finally, we separate the data into four frames of length equal to 512 points and ending at points 512, 922, 1332 and 1742, respectively. Figure 14 represents the expression S_{km} obtained by applying (13) to $x(n)$. Note that S_{km} shows a peak at 0.3, as illustrated in Figure 15, which plots a cross-section of S_{km} for m fixed to 1. Figure 16 plots the magnitude of the coefficients d_{qm} obtained by applying a one-dimensional FFT of size 512 along the columns of the log spectrum quantity S_{km} . Figure 16 indicates the presence of a dc value for all the frames m , as expected. Next, Figure 17 plots the magnitude of the two-dimensional cepstral coefficients c_{qp} obtained by applying a one-dimensional FFT of size 64 along the rows of the coefficients d_{qm} . Note that the coefficients c_{qp} exhibit symmetry at $p=32$ and $q=256$, thus all the information contained in c_{qp} is present for $1 \leq p \leq 32$ and $1 \leq q \leq 256$, i.e., in the lower left quadrant of the set of coefficients c_{qp} , as shown in Figure 18.

Further, note that the frequency information of $x(n)$ is contained in the phase of c_{qp} . The phase of the coefficients c_{qp} is unwrapped in order to replace jumps greater than π with their 2π complement. The phase can be represented by a plane whose cross-section is shown in Figure 19. The frequency information can be extracted from its slope as follows; Recall that the phase of the cepstral coefficients is first calculated modulo 2π and is unwrapped next. For this specific case the slope is computed to be equal to -1.885. The actual normalized frequency is obtained by dividing the slope 1.885 by 2π .

Log spectrum of one complex exponential

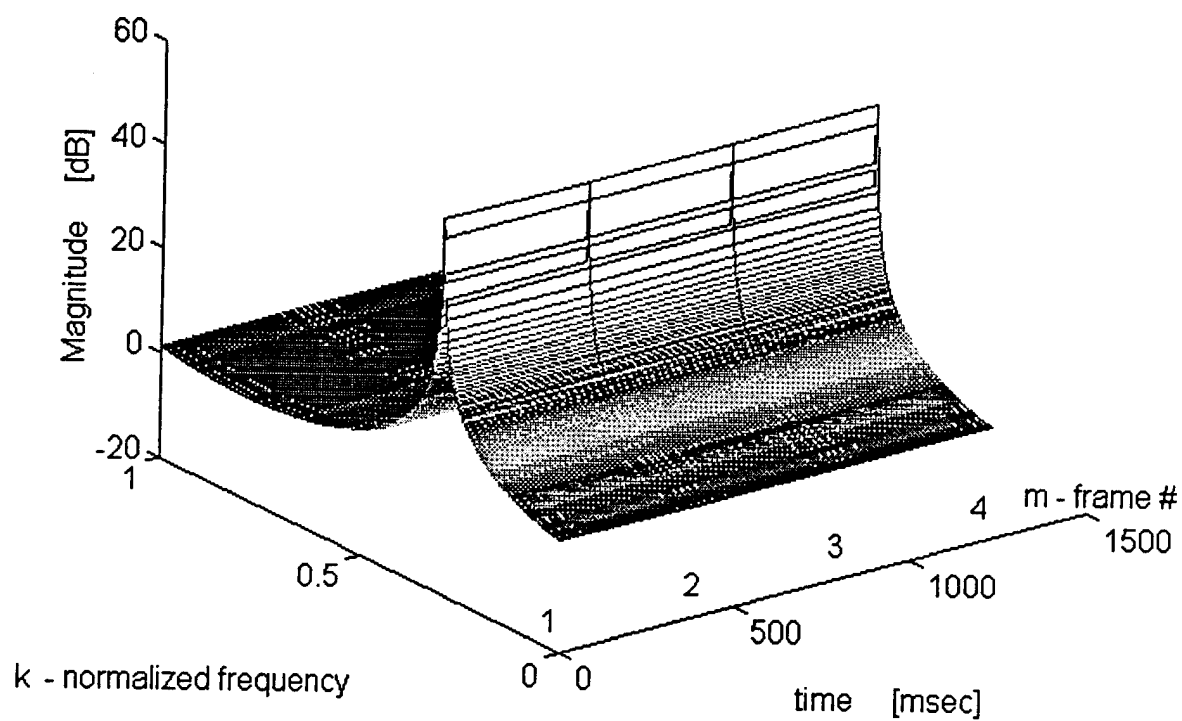


Figure 14. Log spectrum S_{km} of $x(n) = \exp(j2\pi 0.3n)$; window length = 512; 20% overlap.

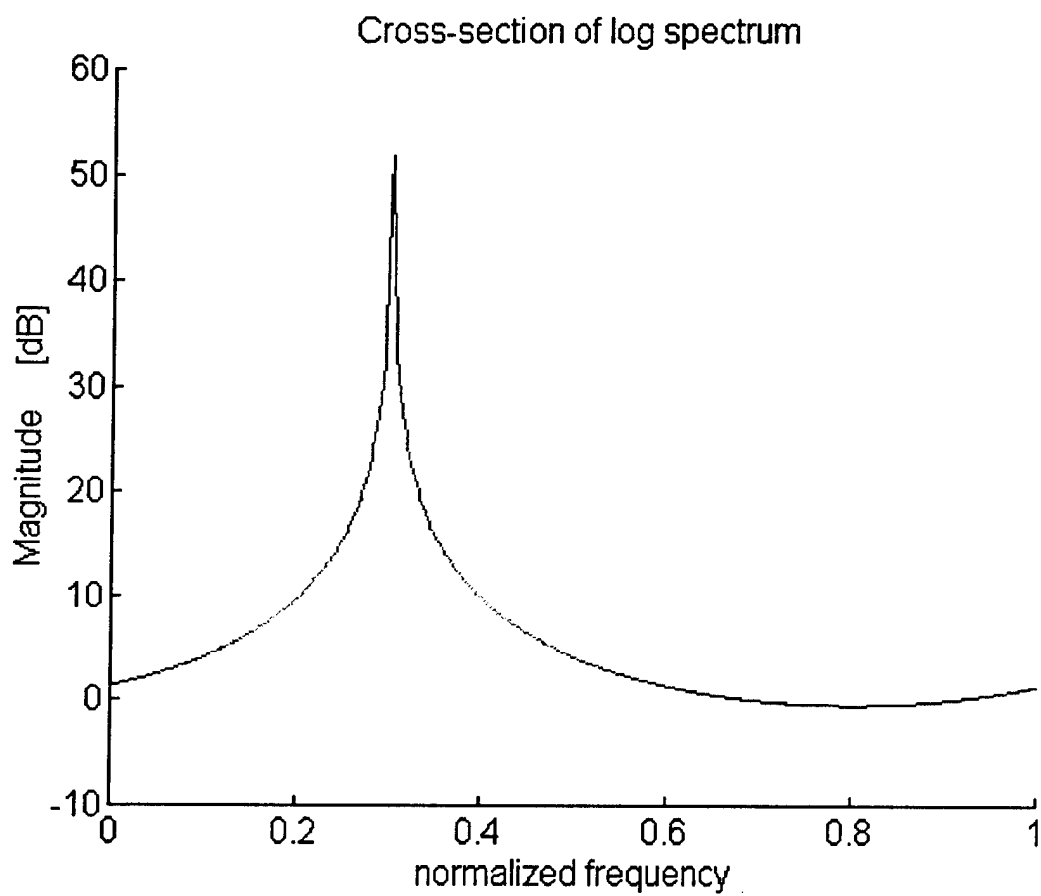


Figure 15. S_{km} of $x(n)$ for $m = 1$.

1-D FFT of Log spectrum

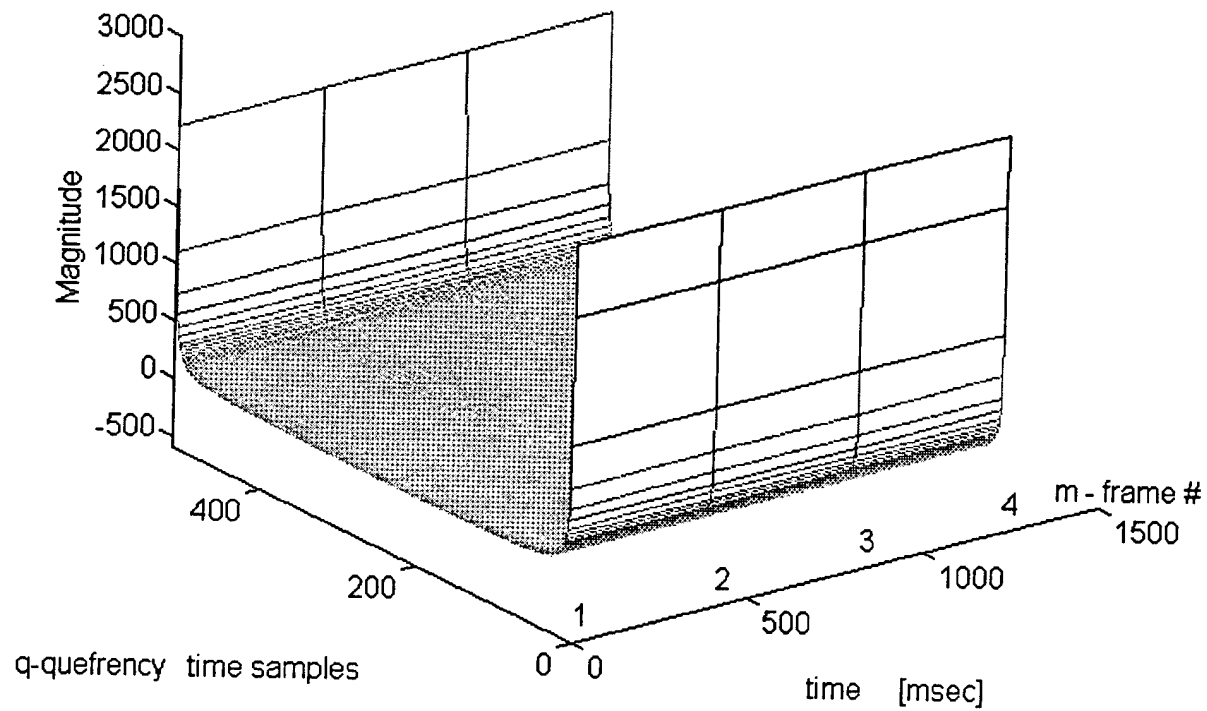


Figure 16. Magnitude of d_{qm} (1-D FFT of log spectrum S_{km} , FFT length = 512, $f_s = 1000$ Hz).

Magnitude of 2-D cepstrum coefficients of one complex exponential

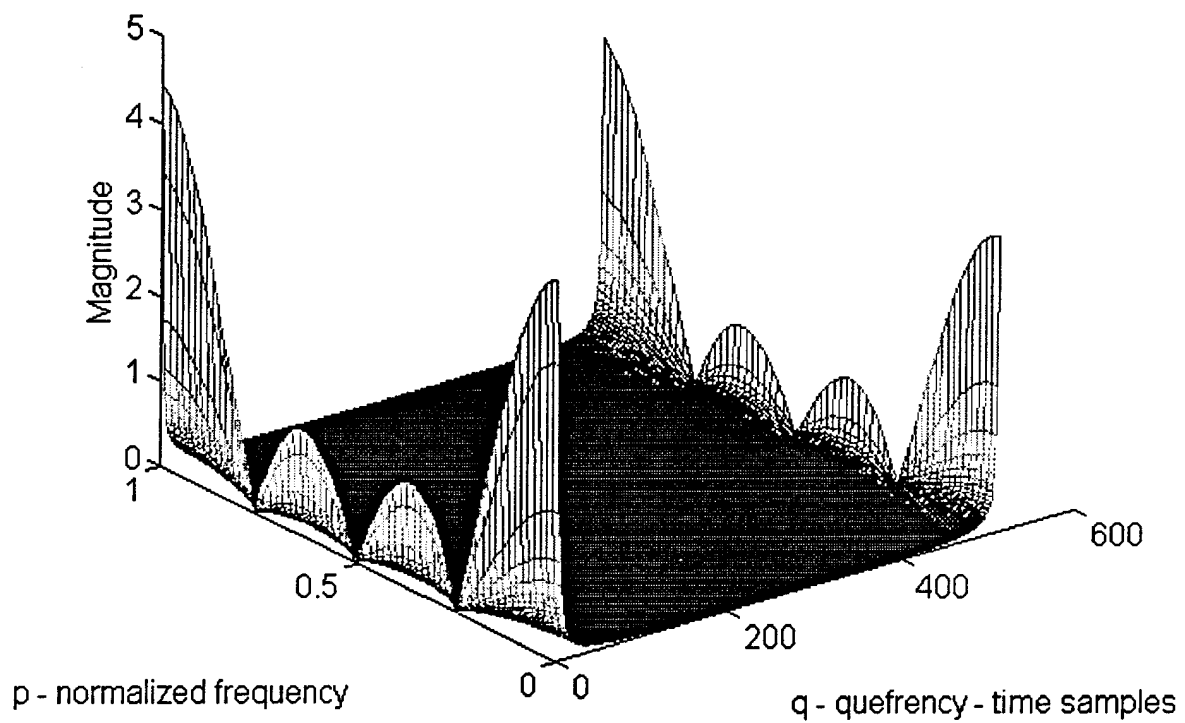


Figure 17. Magnitude of 2-D cepstrum coefficients c_{qp} obtained for one complex exponential ($f_s = 1000$ Hz, FFT length = 64).

Magnitude of 2-D cepstrum coefficients of one complex exponential

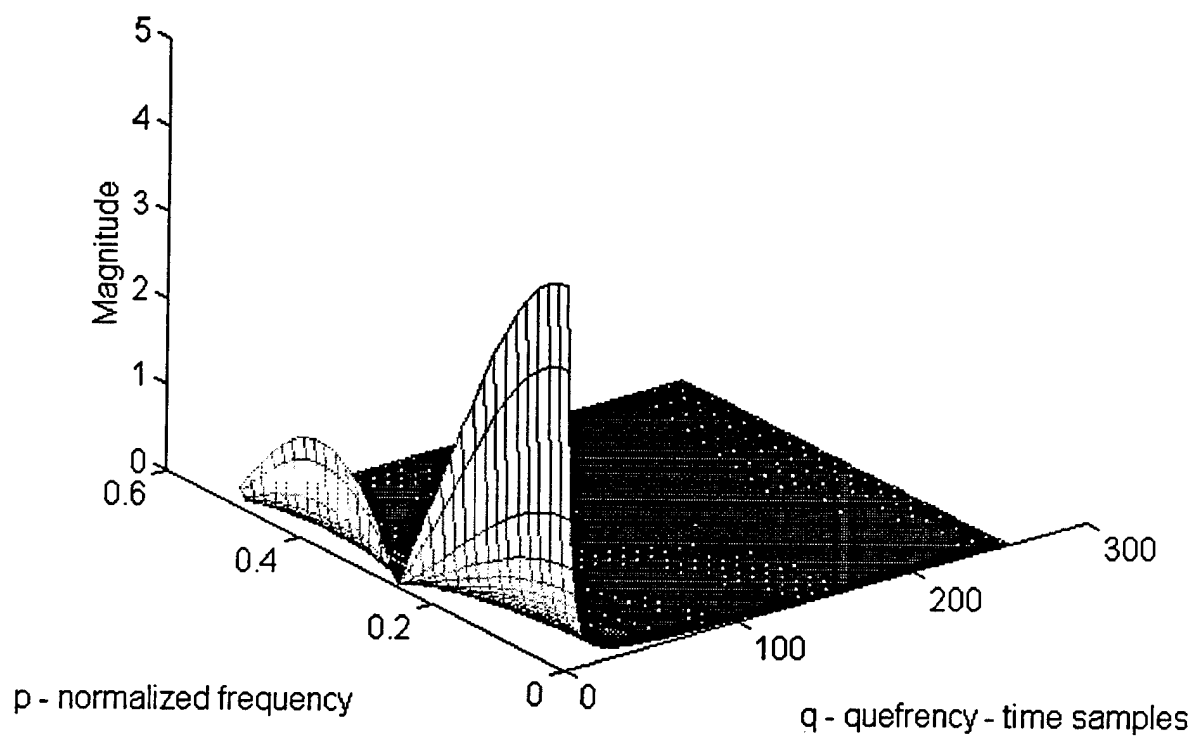


Figure 18. Lower left quadrant of the cepstral matrix of Figure 17.

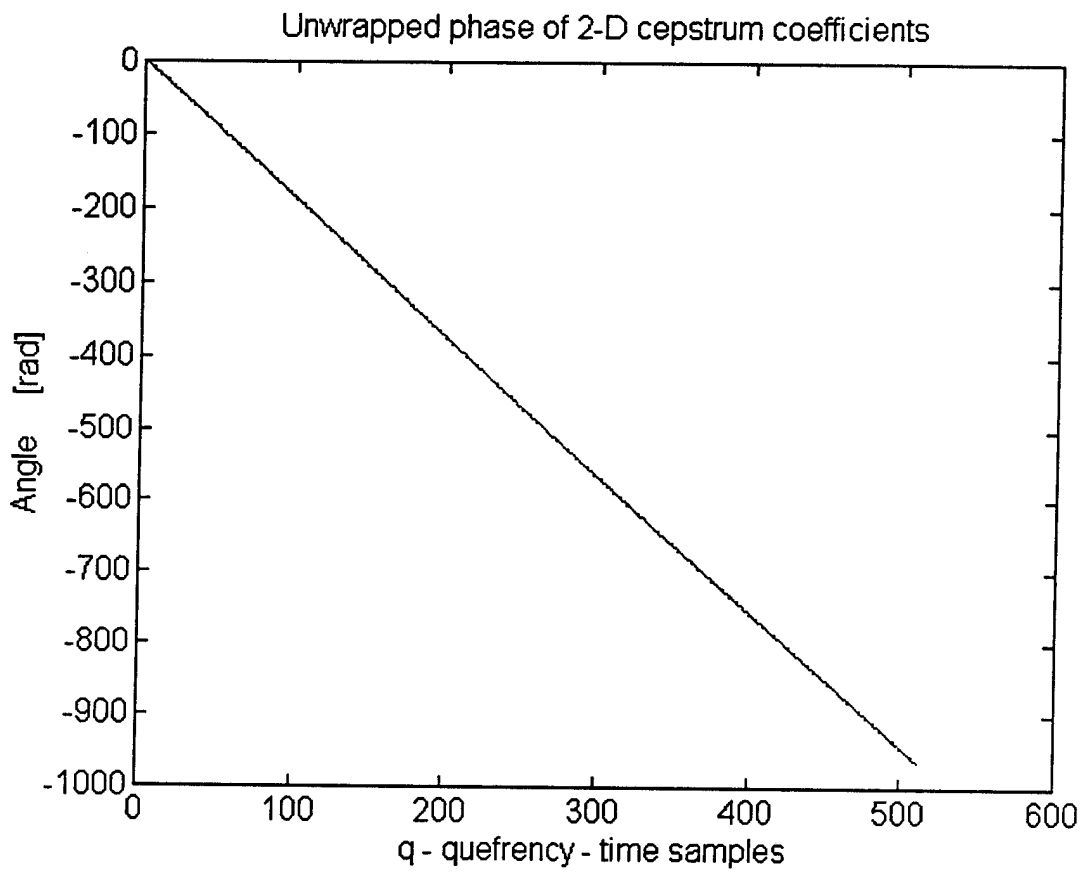


Figure 19. Unwrapped phase of c_{qp} of $x(n)$.

b. Two Complex Exponentials

Consider now the signal $y(n)$, which consists of the concatenation of two single complex exponentials of different frequencies. The signal $y(n)$ is:

$$y(n) = [x_1(n), x_2(n)], \quad (19)$$

where $x_1(n) = \exp(j2\pi(0.1)n) \quad 0 \leq n \leq 2000, \quad (20)$

$$x_2(n) = \exp(j2\pi(0.3)n) \quad 2201 \leq n \leq 6000, \quad (21)$$

and $y(n) = 0$ for $2001 \leq n \leq 2200$. The time duration of the signal is 6 seconds, given that the sampling frequency is $f_s = 1000$ Hz. Following the same procedure as before, we obtain the log spectrum S_{km} shown in Figure 20. The two peaks at frequencies 0.1 and 0.3 are illustrated in the cross-section of S_{km} for $m=1$ in Figure 21. Next, the cepstral coefficients c_{qp} are computed, and only the absolute value of the coefficients for $1 \leq p \leq 32$ and $1 \leq q \leq 256$ are plotted in Figure 22. Note that in the case of two complex exponentials the phase plot has again the shape of a plane, but the frequency information contained in its slope is close to the average of the two frequencies in the signal. Similarly to the one complex exponential case, the slope is found to be -1.5 and divided by 2π gives a result of 0.23, which is approximately the average the frequencies 0.1 and 0.3 contained in the signal. The cross-section of the unwrapped phase of the coefficients c_{qp} is shown in Figure 23.

c. Phoneme /@/ From The Word "Man"

For the third example we consider the phoneme /@/ from the word "man". The length of the phoneme is 120 msec and is separated into frame lengths of 32 msec with 75% overlap, in order to apply the two-dimensional cepstrum transformation. The absolute value of the coefficients c_{qp} is plotted in Figure 24. Since the cepstral matrix is symmetric around $p=256$ and $q=32$, when the two-dimensional FFT is of size (512,64), only one fourth of the two-dimensional FFT transform is shown. The first column, for

$p=1$, represents the energy of the speech signal and is discarded from the plot, since it is not used in our computations, as mentioned in the previous section. Peaks shown around $q=70, 140, 210$ in Figure 24, represent multiples of the pitch period. Next we apply a low-time lifter in order to isolate the part of the two-dimensional cepstrum that corresponds to the vocal tract. Experimentally, in [Ref. 7] it is shown that the range of values $1 \leq q \leq 15$ and $0 \leq p \leq 4$ is sufficient and contains sufficient information about the speech signal concerning the vocal tract characteristics. We use a raised sine lifter to deemphasize the effects due to low-order and high-order cepstral coefficients which have higher variance. Figure 25 shows the magnitude of the resulting liftered coefficients c_{qp} for $1 \leq q \leq 15$ and $0 \leq p \leq 4$, where the frequency axis is no longer normalized. The choice of the frame length and of the amount of overlap used for the computation of the two-dimensional cepstrum coefficients is not unique, and varying this combination changes the shape of the two-dimensional cepstral surface. For example, note that reducing the amount of overlap from 75% to 20% and keeping the frame length fixed to 32 msec, leads to fewer frames obtained from the signal. Figure 26 shows the magnitude of the resulting liftered coefficients c_{qp} for the range (q, p) as defined above. Note that if we extend the range of the frequency p from 5 to 15 we obtain the plot of Figure 27. Comparing Figures 25 and 27, we observe that the magnitudes of the coefficients have similar shapes for the different ranges of the frequency axis. Thus, reducing the amount of overlap "stretches" the two-dimensional cepstral plot, which may result in the loss of some information if we keep the range of the frequency axis fixed when changing the amount of overlap. Therefore, the choice of the range of the frequency axis p , may need to be modified when changing the amount of window overlap in order to keep the amount of information contained in the set of c_{qp} coefficients the same. This simple example illustrates the fact that the range of p over which the cepstral coefficients are to be used for analysis is dependant upon the choice of window length and overlap.

Log spectrum of two complex exponentials

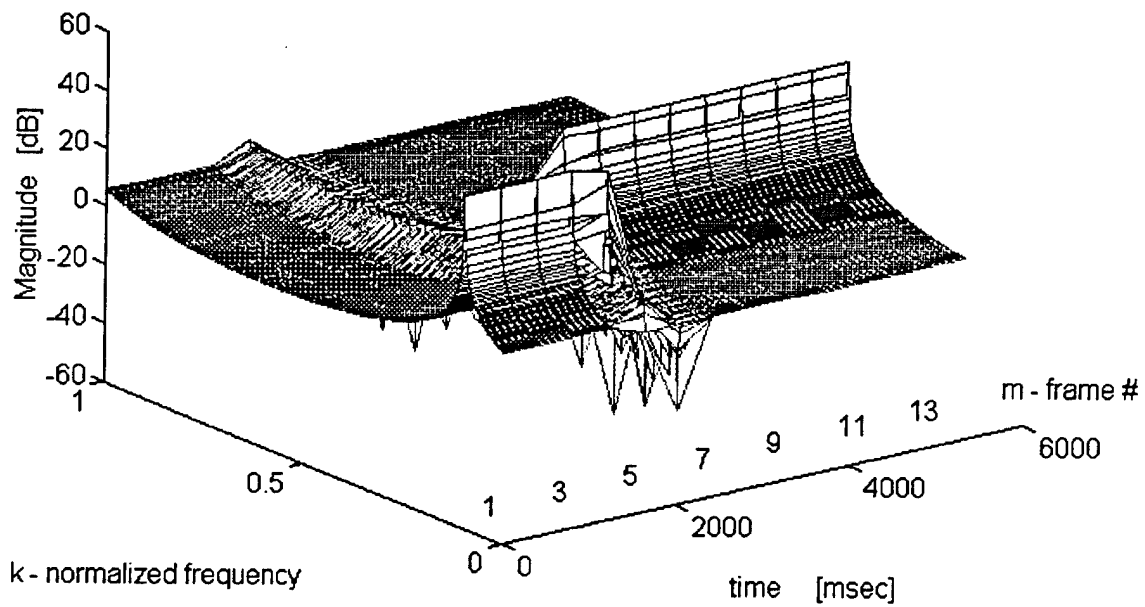


Figure 20. Log spectrum S_{km} of $y(n)=[\exp(j2\pi 0.1n), \exp(j2\pi 0.3n)]$; window length = 512; 20% overlap.

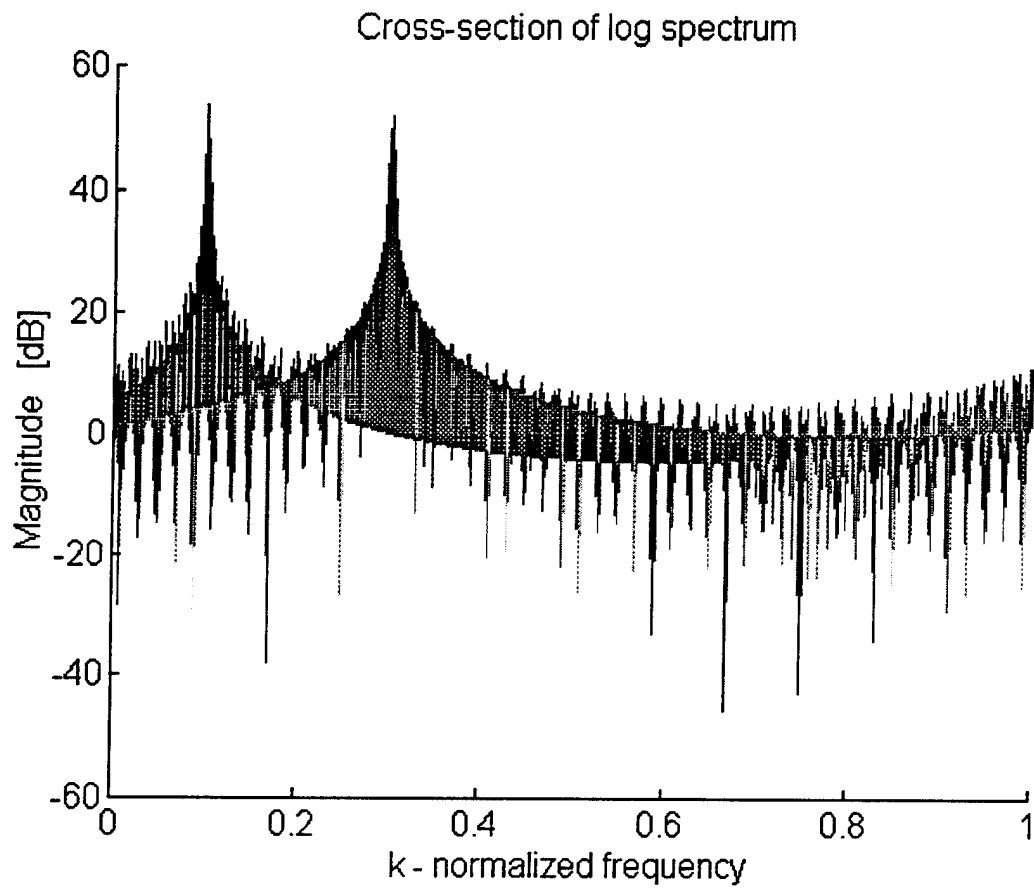


Figure 21. S_{km} of $y(n)$ for $m = 1$.

Magnitude of 2-D cepstrum coefficients of two complex exponentials

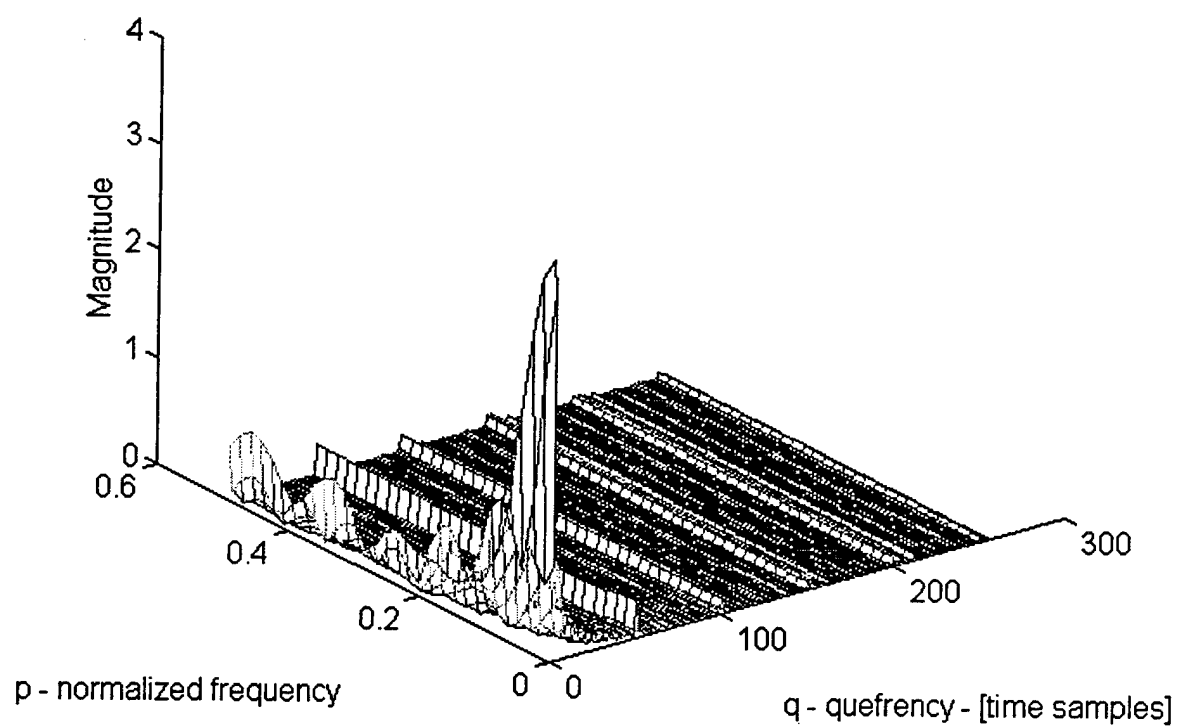


Figure 22. Magnitude of cepstrum coefficients c_{qp} for $0 < q < 256$ and $0 < p < 32$ of $y(n)$ ($f_s = 1000$ Hz, FFT length along p -axis = 64, FFT length along q -axis = 512).

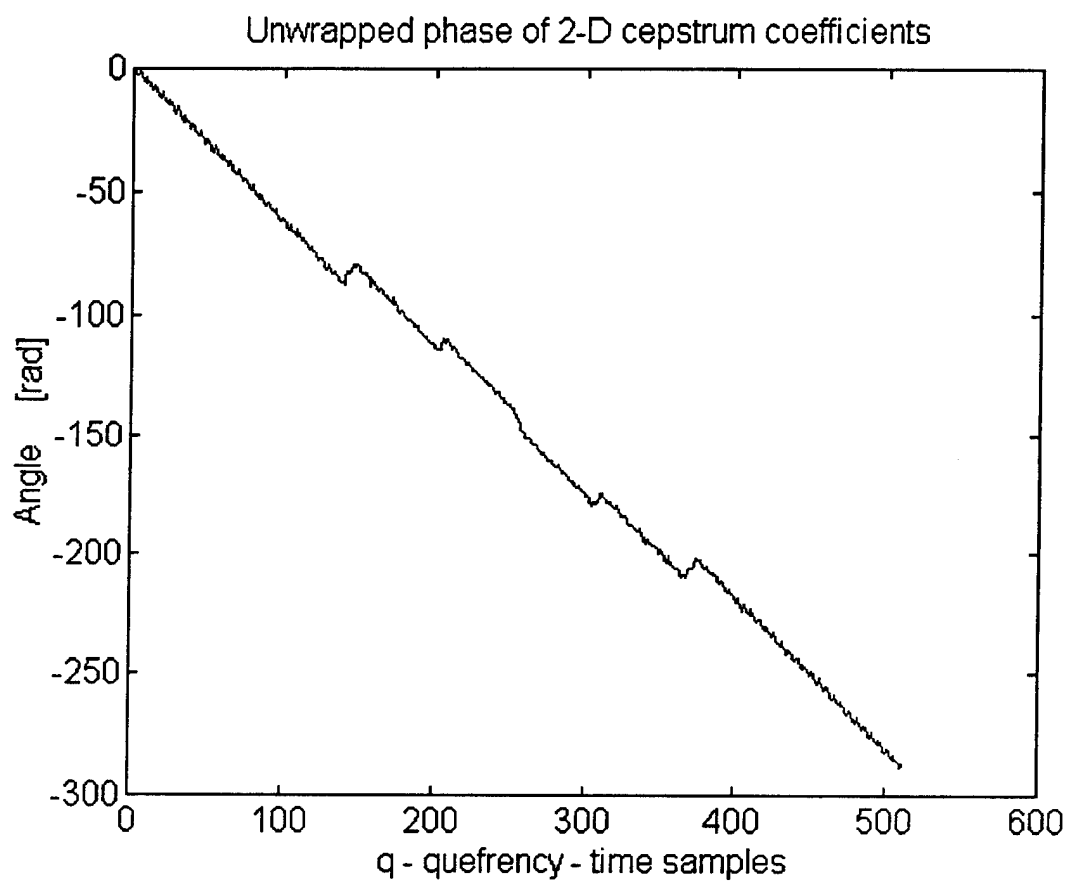


Figure 23. Unwrapped phase of c_{qp} of two complex exponentials.

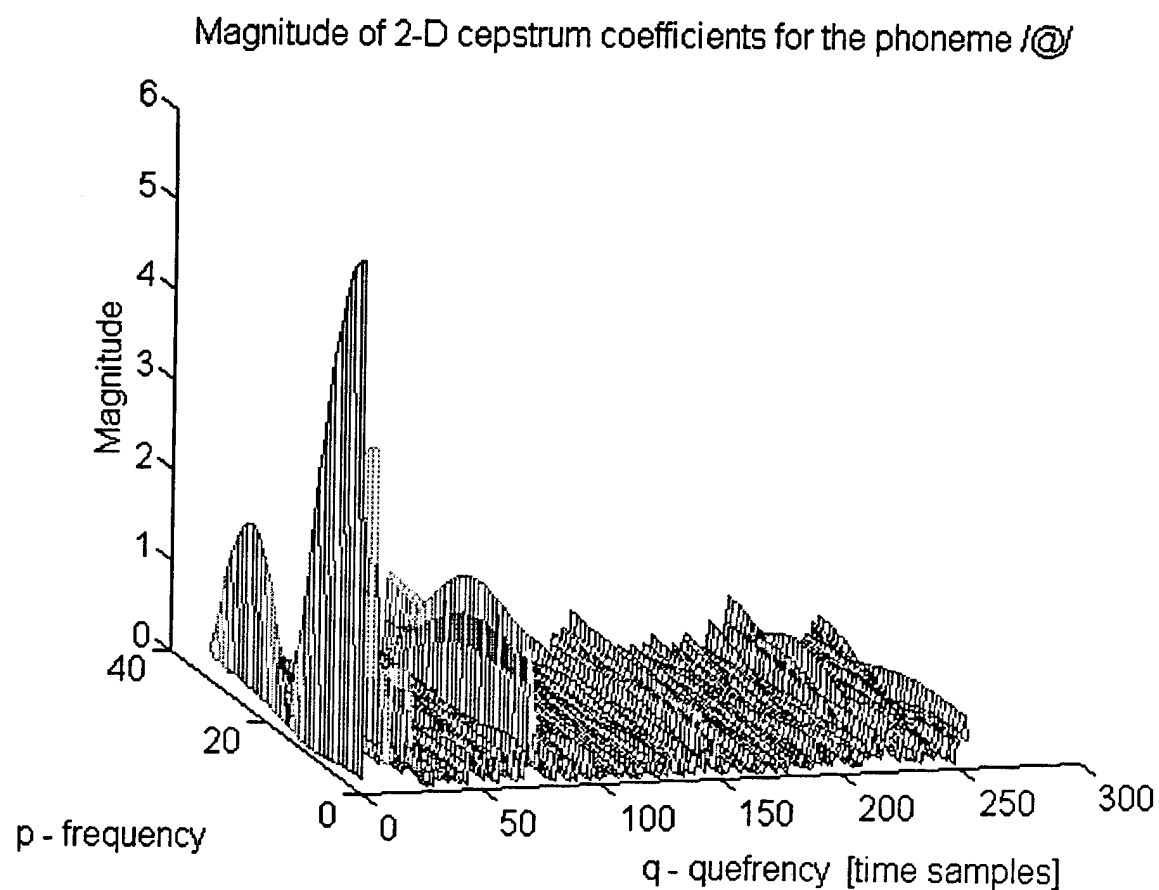


Figure 24. Magnitude of 2-D cepstrum coefficients c_{φ} for $1 \leq q \leq 256$ and $0 \leq p \leq 32$ for the phoneme /@/ ($f_s=8192$ Hz).

Magnitude of liftered 2-D cepstrum coefficients

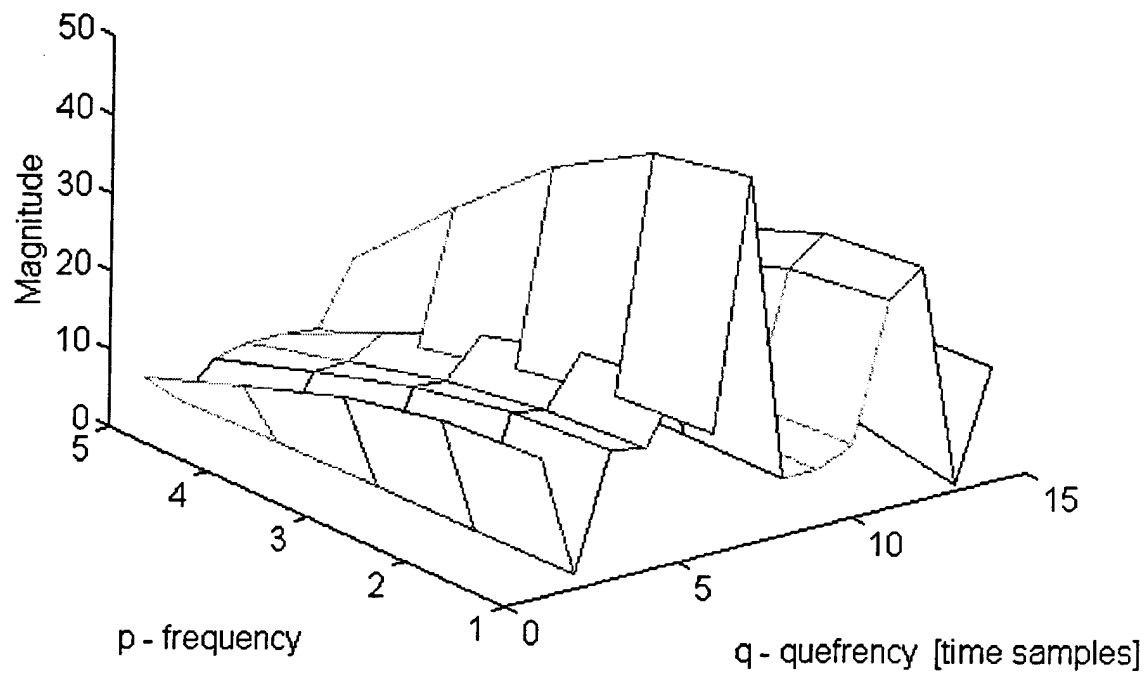


Figure 25. Magnitude of the liftered c_{qp} coefficients. (Lifter: raised sine, frame length= 32 msec, overlap: 75%, $f_s = 8192$ Hz).

Magnitude of liftered 2-D cepstrum coefficients

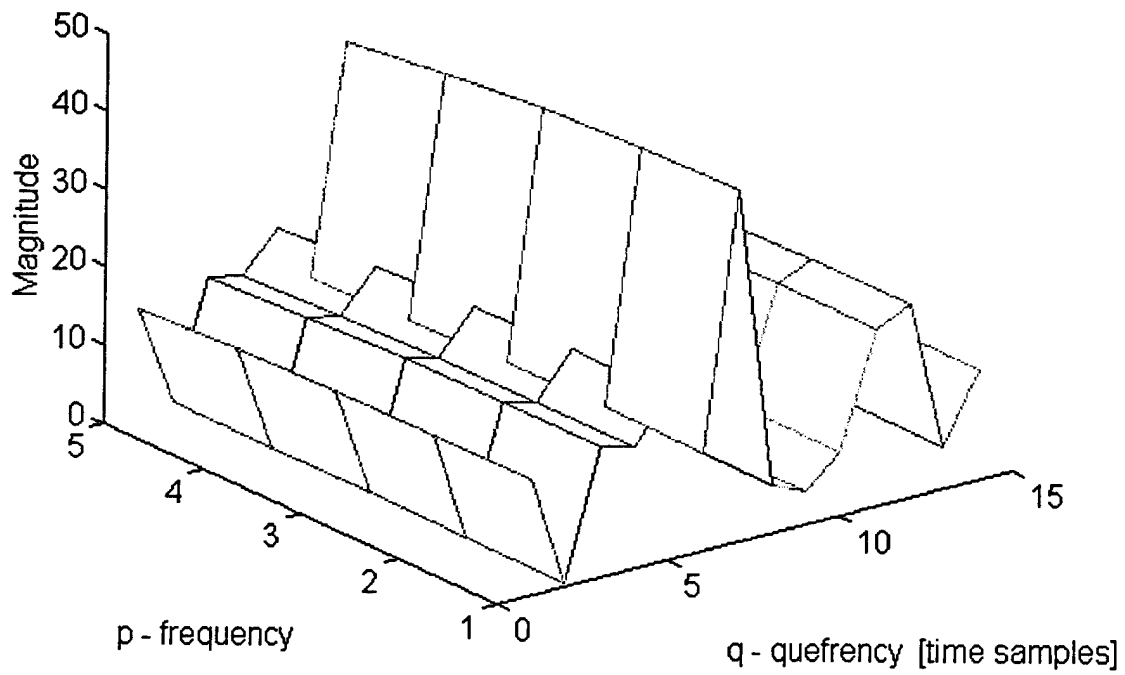


Figure 26. Magnitude of the liftered c_{qp} coefficients. (Lifter: raised sine, frame length= 32 msec, overlap=20%, $f_s = 8192$ Hz)

Magnitude of liftered 2-D cepstrum coefficients

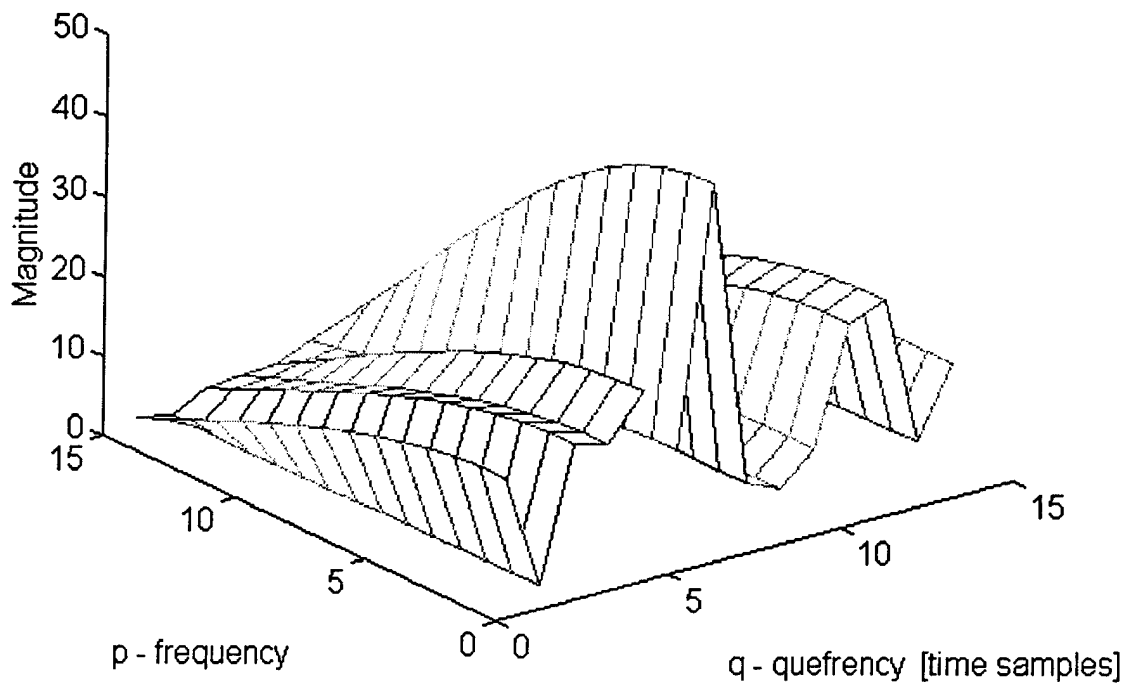


Figure 27. Magnitude of the liftered c_{qp} coefficients. (Lifter: raised sine, frame length = 32 msec, overlap=20%, $f_s = 8192$ Hz)

IV. SPEAKER IDENTIFICATION - PREPARATION

A. PROBLEM DEFINITION

Every individual has a different voice and the ability to recognize a person from his/her voice only is called speaker recognition. Recall that variations between speakers are mainly caused by the anatomical differences related to the shape and size of the vocal tract. Other variations can result from the different ways people have learned to produce speech. Aside from the variations between different speakers, there are also variations within the same speaker. Such variations are caused by different factors in speaking rates, emotional state, health, etc. In speaker recognition applications, it is desirable to have low variations within the same speaker but high variations between different speakers [Ref. 8].

Speaker recognition consists of two distinct parts; speaker identification and speaker verification. Speaker identification deals with the task of identifying a given speaker among a group of several known speakers using test utterances. Speaker verification deals with the task of verifying a speaker's identity. Note that the speaker verification task is much simpler than the identification problem, since it only requires a binary decision; to accept or reject the claimed speaker. On the other hand, speaker identification requires comparison with reference utterances from all speakers in the group, and this problem becomes increasingly difficult when additional speakers are added to the group.

A general representation of the speaker recognition problem is shown in the block-diagram of Figure 28. Several utterances of a specific speech signal are recorded and then processed in order to extract some average information required to accurately represent each person. This information is used to form a reference pattern or template. A test speech signal $s(n)$ is then processed identically to signals used to create the reference templates. Next, reference and test templates are compared for identification or verification purposes. Note that speaker verification and speaker identification problems are similar in terms of the signal processing parts and the main differences occur in the decision logic. In the present study, only the speaker identification problem is investigated.

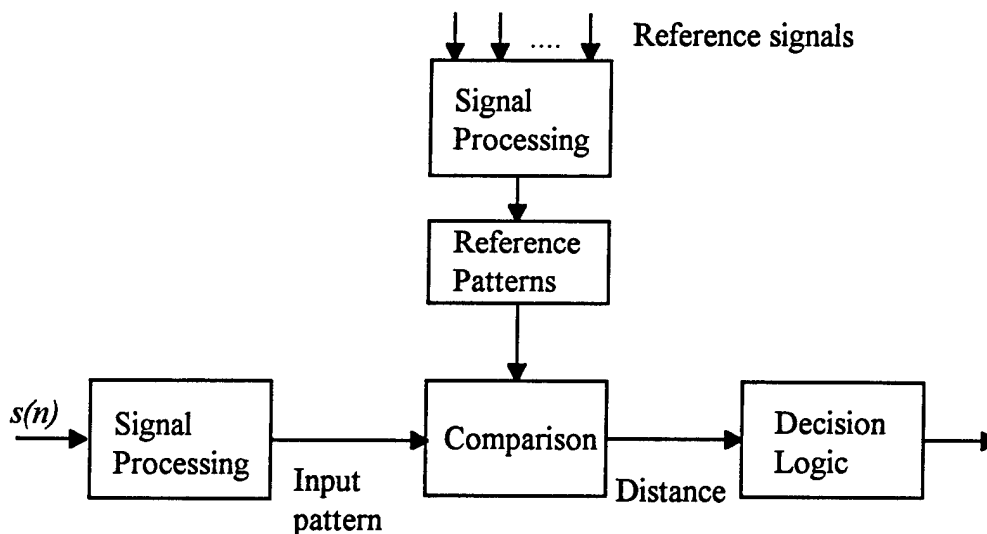


Figure 28. General representation of the speaker recognition problem.

The technique applied in the signal processing block of Figure 28 depends on the desired information to be extracted from the speech signal. Such information can be the pitch period, the formant frequencies, the LPC parameters or the cepstrum coefficients. The technique applied throughout this work is the two-dimensional cepstrum transformation, which was described in Chapter III. The coefficients derived from this transformation are used to form a reference pattern to represent each speaker.

B. DISTANCE MEASURES

The purpose of the distance calculation is to provide a measure of similarity between the reference pattern and the input test pattern. For the case of the

two-dimensional cepstrum coefficients used in our study, c_r will denote the reference pattern and c_t will denote the test pattern. There are a number of distances introduced in the literature but only two of these are examined here; the Euclidean distance measure and the weighted two-dimensional cepstral distance. Recall that a distance measure must have the following three properties:

$$\text{Nonnegativity:} \quad \begin{aligned} D(c_r, c_t) &> 0, & c_r &\neq c_t \\ D(c_r, c_t) &= 0, & c_r &= c_t \end{aligned}$$

$$\text{Symmetry:} \quad D(c_r, c_t) = D(c_t, c_r). \quad (22)$$

$$\text{Triangle inequality:} \quad D(c_r, c_{t1}) \leq D(c_r, c_{t2}) + D(c_{t1}, c_{t2}).$$

The Euclidean distance measure is defined as $D(c_r, c_t) = \|c_r - c_t\|^2$ and is perhaps the most popular distance measure that satisfies these relations [Ref. 6].

The second distance measure used in this study is a weighted two-dimensional cepstral distance. This distance measure takes into account the variations of the cepstral coefficients along the frames by computing their derivatives, with respect to the frame number. The distance measure is given by:

$$D(c_r, c_t) = \sum_{q=1}^{N-1} \sum_{p=0}^{M-1} (\beta p^2 + q^2 + 1) \cdot |c_r(p, q) - c_t(p, q)|^2, \quad (23)$$

where β is a combinational factor defined to introduce the effects due to the derivative terms into the distance measure, N is the range of the queffreny axis q , and M is the range of the frequency axis p . Pai and Wang showed experimentally that the optimum values of the factor β are in the range between one and three [Ref. 9]. Experiments have also shown that this two-dimensional distance measure is very promising, especially for speech recognition applications [Ref. 9,10].

C. DATA COLLECTION - PREPROCESSING

The words selected to be used in this study are two simple monosyllable words containing three phonemes each: "man", "beat" and one more complicated word: "indigestible" which contains several voiced and unvoiced phones. Although cepstral analysis is mostly designed for voiced speech, the third word was also chosen to examine the behavior of the method in unvoiced speech. Fourteen speakers were used for the experiments, thirteen male and one female. Both U.S. and foreign speakers were used, with an average age of thirty. Three groups of speakers were formed, and one group was used for each word. Each group contained ten speakers. All speakers were recorded on the same machine, a Sun Sparc-1, under the same conditions and with the same microphone. Each speaker recorded the same word ten times and each repetition was saved in an individual file. For some of the recorded repetitions, speakers were asked to distort their voice by closing their nose during the recording, or speaking at faster or slower rate than normal. As discussed later, this was done in order to examine the performance of the recognizer using slightly distorted words or words spoken at rates different than normal. The other speakers were directed to speak at their usual rate and using their normal voice.

All words were digitized as recorded using a sampling frequency of 8192 Hz. Silence or noise present before and after the digitized word was removed visually. The microphone used in the digitization process usually introduces undesired side effects, such as line frequency noise, at a frequency around 50 or 60 Hz. Therefore, a highpass FIR filter, of order 24, with a cutoff frequency of 100 Hz was applied to each speech signal to remove any unwanted noise caused by the equipment.

Note that each individual spoke with slightly different levels of loudness. Therefore, each word was energy normalized after having its mean removed. The normalization used is the following:

$$s_N(n) = \frac{s_F(n)}{\sqrt{s_F(n) \cdot s_F(n)^T}},$$

where $s_N(n)$ is the normalized sequence and $s_F(n)$ is the filtered speech signal. As a result, all words have the same energy level, equal to one, regardless of the speaker or the word spoken. The block-diagram of Figure 29 represents the preprocessing sequence followed for the preparation of each word.

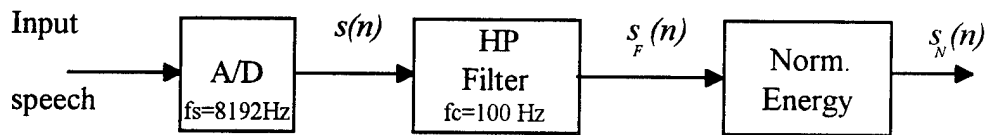


Figure 29. Block-diagram representing the preprocessing sequence for each word; f_s is the sampling frequency of the A/D conversion, and f_c is the cutoff frequency of the highpass filter.

V. TESTS AND RESULTS

A. TESTS SET UP

The two-dimensional cepstral coefficients, computed after the processing procedure described in the previous chapter, are used to form reference and test templates. Each word is repeated ten times by each speaker. Five of the repetitions are used to create a given speaker's reference template by averaging the two-dimensional cepstral coefficients obtained with each repetition. The other five repetitions are tested against the reference templates of all speakers. Four sets of reference and test templates are formed, so that each word repetition obtained from each speaker is included twice in each of the reference and test groups. The four sets formed for each word and each speaker were the following:

- i. REF1=[1,2,3,4,5], TEST1=[6,7,8,9,10],
- ii. REF2=[2,4,6,8,10], TEST2=[1,3,5,7,9],
- iii. REF3=[6,7,8,9,10], TEST3=[1,2,3,4,5] and
- iv. REF4=[1,3,5,7,9], TEST4=[2,4,6,8,10],

where [1,2,3...,10] represent the respective repetitions of each word. Several averaging procedures, which are described next, are conducted during the experiments in order to make the results more statistically relevant. Figure 30 illustrates the set-up used for the experiments, where the combination of reference and test groups is composed of the groups denoted REF1 and TEST1, as described above. Note that in Figure 30, the term $\text{spkrX}(i)$ denotes the i^{th} repetition of a given word spoken by the X^{th} speaker. Recall that ten repetitions of each of the three words are available for a given speaker. Five of the repetitions are used to build the average reference information obtained for that given word and speaker. The other five repetitions are used in the testing phase of the identification procedure. Figure 30 illustrates the test procedure conducted on the first repetition of a given word spoken by Speaker1. The set of cepstral coefficients obtained

from this specific repetition is compared to the average sets of cepstral coefficients obtained for each of the speakers using the statistical distances introduced earlier. These ten successive comparisons lead to ten different distance values. Note that a correct decision would be obtained when the minimum value from this set of distances is the one which results from comparing the test template from Speaker1 to the reference template from Speaker1. The resulting identification decision for this given repetition of the word for a given speaker is recorded as "correct" or "incorrect". This procedure described in Figure 30 is repeated for each of the five repetitions contained in the test templates, leading to the estimation of the recognition rate for a given combination of reference and

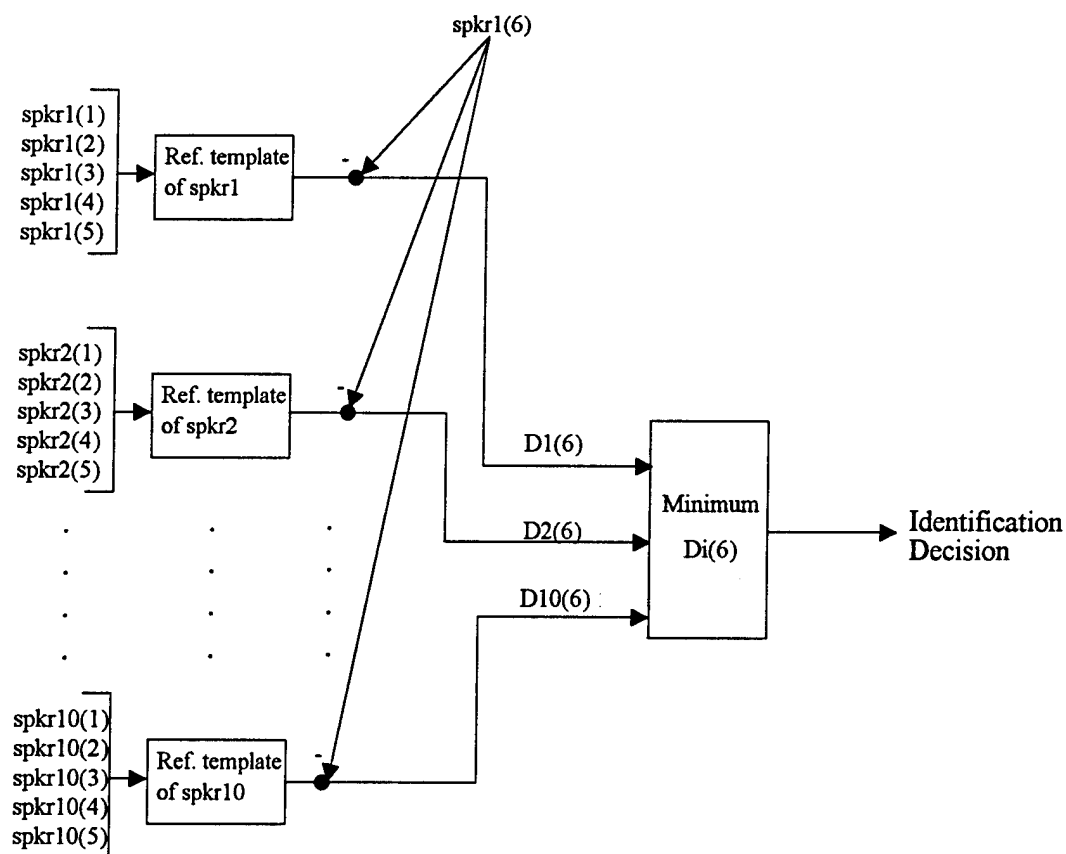


Figure 30. Example of test set up for REF1 and test word spkr1(6).

test groups. Next, the above process is repeated again for each of the four combinations of reference/test (REFi, TESTi) groups defined above to limit potential bias in the results. The overall recognition rate is obtained from averaging the recognition rates obtained in the four combinations of reference /test groups.

Finally, some of the experiments investigated the robustness of the identification procedure in the presence of additive white Gaussian noise. For these experiments, the entire procedure described above was repeated ten times for a given SNR level to obtain more statistically relevant results. In these experiments, the overall recognition rate is the average of those obtained in each of the ten individual trials conducted at a given SNR level for each of the reference/test groups combinations, i.e. the average value obtained from 40 recognition rates. Finally recall that both the Euclidean and two-dimensional cepstral distance are used in some of the experiments to compare the identification performance obtained with each. Note that the cepstral coefficients were previously lifted with a raised sine lifter when using the Euclidean distance. However, no lifter was used for the case of the two-dimensional cepstral distance.

In order to examine the robustness of the method in additive noise, a white Gaussian sequence with a user-defined Signal to Noise Ratio (SNR) was added to the original time signals before the computation of the two-dimensional coefficients. SNR's of 50, 20, 10, 5, 0, -5 dB were used, assuming that the original recording was noise free. Background noise can take the form of speech from other speakers, equipment sounds or even noise produced from the speaker as in breath noises, lip smacks, etc. Other parameters that were varied throughout the tests were the amount of overlap and frame length, as used in the computation of the two-dimensional cepstral coefficients. Four cases were examined:

- i. overlap=75%, frame length=256,
- ii. overlap=20%, frame length=256,
- iii. overlap=20%, frame length=512 and

iv. overlap=75%, frame length=512,

where 256 and 512 time samples correspond to 32 and 64 msec, respectively.

B. RESULTS

1. Speaker Recognition In Noisy Conditions

The robustness of the method in background noise is shown in Tables 1, 2 and 3 for the three words tested. It can easily be seen that the performance degrades as the Signal to Noise Ratio decreases. Results using the two different distance measures considered are presented in the same Tables. Note that the two-dimensional cepstral distance did not show any significant improvement over the Euclidean distance, but instead degraded more readily with the decrease of the SNR level.

The words "*man*" and "*indigestible*" gave similar results in performance. The word "*beat*" had a slightly worse performance. This is mainly caused by the existence of the stop consonant /t/ at the end of the word. As seen in Figure 30(a), there is a period of silence before the burst of energy which produces the /t/. It was noticed in the experiments that some speakers pronounced the /t/ clearly, but some others did not, as seen in Figure 30(b). For the speakers who did not utter the /t/ clearly, it was more difficult to detect the end of the word, since there was no obvious point to differentiate from the speech signal and silence. This caused some unwanted variations of the cepstral coefficients, resulting in decreased performance. Note also, that the period of silence before the /t/ is not negligible compared to the duration of the phoneme /i/. Therefore, the effect of the noise in this section of the word is more noticeable, which also leads to degraded performance.

The results shown in Tables 3, 4 and 5 are obtained using cepstral coefficients computed with 75% overlap and a frame length of 256 time samples (32 msec).

MAN overlap=75% frame length=256				
Correct Identification rate				
	Euclidean Distance		2-D Cepstral Distance	
SNR (dB)	Mean	Standard Deviation	Mean	Standard Deviation
50	97.7%	0.7232	94.5%	1.797
20	96.45%	1.6	96.05%	1.894
10	95.7%	1.951	93.7%	5.214
5	90.7%	2.919	89.45%	3.762
0	80.7%	5.115	73.15%	6.108
-5	59.52%	5.25	54%	5.026

Table 3. Identification rates for the word "man", for SNR = 50, 20, 10, 5, 0, -5 dB for the Euclidean and 2-D cepstral distances; 2-D cepstral coefficients computed with 75% overlap and 256 time samples frame length.

BEAT overlap=75% frame length=256				
Correct Identification rate				
	Euclidean Distance		2-D Cepstral Distance	
SNR	Mean	Standard Deviation	Mean	Standard Deviation
50	95.9%	2.216	94.6%	2.134
20	91.5%	4.619	90.65%	3.932
10	80.4%	5.047	79.15%	4.682
5	71.4%	5.715	69.85%	5.289
0	57.65%	5.061	56.2%	4.898
-5	46.9%	5.771	46.1%	7.016

Table 4. Identification rates for the word "bear", for SNR = 50, 20, 10, 5, 0, -5 dB for the Euclidean and 2-D cepstral distances; 2-D cepstral coefficients computed with 75% overlap and 256 time samples frame length.

INDIGESTIBLE overlap=75% frame length=256				
Correct Identification rate				
	Euclidean Distance		2-D Cepstral Distance	
SNR	Mean	Standard Deviation	Mean	Standard Deviation
50	95%	2.935	94.5%	2.592
20	92.7%	4.826	88.7%	4.381
10	88.65%	5.072	85.9%	6.008
5	85.7%	5.388	82.3%	5.88
0	79.7%	5.845	74.25%	6.23
-5	62.35%	6.208	57.65%	5.201

Table 5. Identification rates for the word "*indigestible*", for SNR = 50, 20, 10, 5, 0, -5 dB for the Euclidean and 2-D cepstral distances; 2-D cepstral coefficients computed with 75% overlap and 256 time samples frame length.

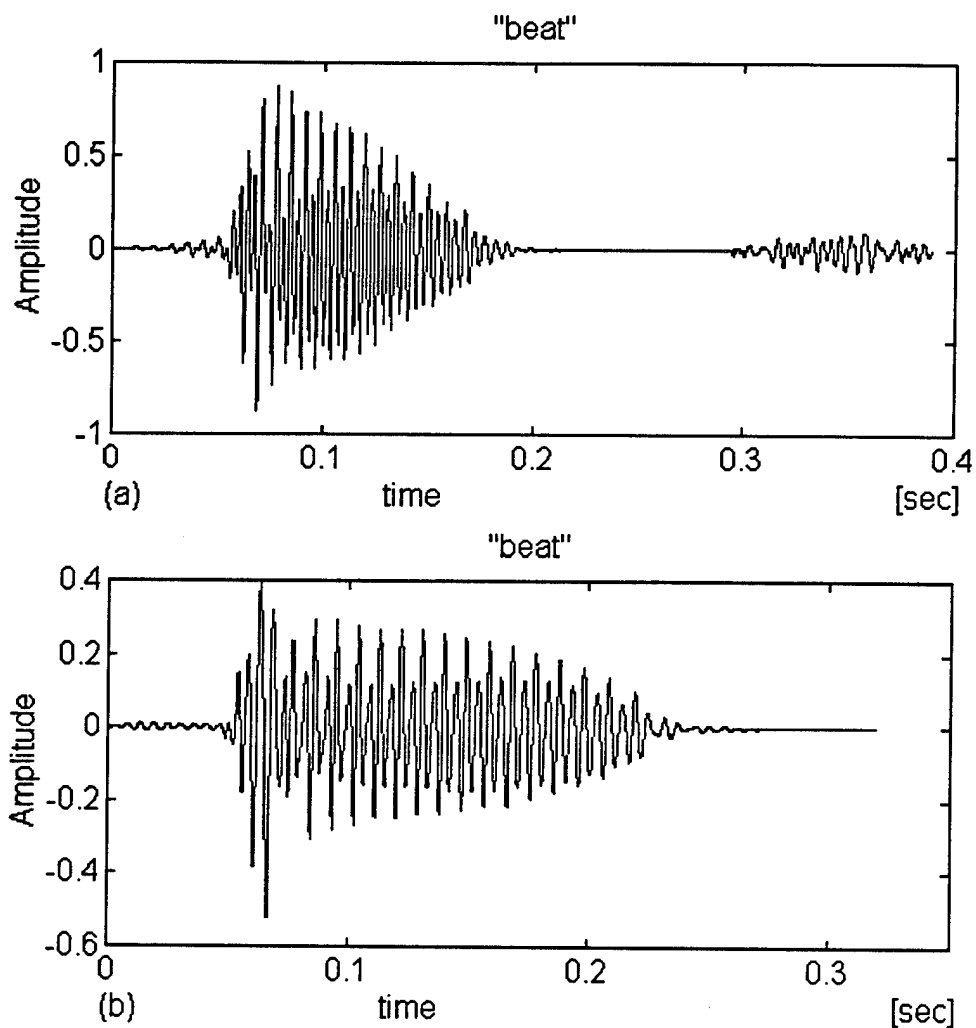


Figure 31. Word "beat" for two different speakers. In (a), the /t/ is clearly seen after the short period of silence following the phoneme /i/. In (b), the end of the word is not obvious since the /t/ is not clearly seen.

2. Word Length Effects

Next we consider the effect of word length on the performance of the recognizer, to investigate the effect due to different speaking rates. Note that no time alignment between different repetitions of a given word was applied. Thus, repetitions had different lengths according to the speaking rate of the speaker. In addition, some speakers were asked to distort their original voice, for some of their recordings, by speaking at rates different than normal to emphasize such a difference. When the repetitions of these speakers were tested against the speaker's own reference template, increased error rates were obtained. We noted that the words that failed the speaker identification test were those with average length much different from the average for high SNR levels. Tables 6, 7 and 8 show the identification rates of specific utterances for certain speakers and for various SNR's for the three words tested. The second column of each table shows the deviation of the length of the word tested from the speaker's average word length. It was noted that repetitions that were spoken at rates different than normal intentionally almost always failed. On the other hand, we noted that repetitions where the difference in average length resulted from normal speaking, only failed in low SNR's, as shown in Tables 6 to 8. The lengths that were made intentionally different from the average can be identified in the tables from their values of deviation higher than 10%. It was interesting to note that speakers who were recorded mainly at a constant speaking rate were 100% correctly identified, even with SNR's as low as -5 dB.

3. Robustness Of The Distance Measures

The Euclidean distance computed in the tests is normalized by the number of cepstral coefficients included in the reference or test templates. The difference between the distance obtained from comparing a speaker's test utterance tested against its own reference template and that obtained when comparing to a different speaker can give us indications regarding the robustness of the procedure. Misidentification could more easily occur when the gap between these two distances is small.

<u>MAN</u>							
	Deviation from avg. length	50 dB	20 dB	10 dB	5 dB	0 dB	-5 dB
spkr4(1)	+11%	0%	0%	0%	0%	0%	0%
spkr4(6)	+13%	60%	45%	45%	45%	0%	0%
spkr4(8)	-17%	100%	45%	25%	25%	20%	15%
spkr4(10)	+14%	100%	60%	50%	25%	15%	0%
spkr2(1)	+17%	100%	55%	20%	20%	20%	20%
spkr3(1)	+22%	100%	65%	60%	0%	0%	0%
spkr6(9)	-5.5%	100%	100%	100%	100%	95%	70%
spkr7(3)	+3%	100%	100%	100%	100%	100%	100%
spkr1(5)	+6%	100%	100%	100%	100%	80%	75%
spkr5(8)	-0.4%	100%	100%	100%	100%	100%	60%

Table 6. Identification rates of individual utterances for the word "*man*", word length effect, Euclidean distance.

<u>BEAT</u>							
	Deviation from avg. length	50dB	20 dB	10 dB	5 dB	0 dB	-5 dB
spkr3(4)	-19%	0%	0%	0%	0%	0%	0%
spkr1(5)	+10%	0%	0%	0%	0%	0%	0%
spkr1(7)	+30%	30%	30%	25%	5%	0%	0%
spkr7(6)	-17%	65%	5%	0%	0%	0%	0%
spkr10(2)	+19%	100%	50%	45%	30%	25%	0%
spkr3(7)	+11%	100%	30%	0%	0%	0%	0%
spkr4(5)	-0.8%	100%	100%	100%	100%	100%	60%
spkr7(8)	+1.8%	100%	100%	100%	90%	60%	60%
spkr9(9)	-1.7%	100%	100%	100%	100%	80%	65%
spkr8(7)	+4%	100%	100%	100%	100%	90%	75%

Table 7. Identification rates of individual utterances for the word "*beat*", word length effect, Euclidean distance.

<u>INDIGESTIBLE</u>							
	Deviation from avg. length	50 dB	20 dB	10 dB	5 dB	0 dB	-5 dB
spkr1(2)	+32%	0%	0%	0%	0%	0%	0%
spkr8(2)	+31%	0%	0%	0%	0%	0%	0%
spkr7(4)	+6%	50%	50%	50%	50%	50%	50%
spkr9(9)	+6%	50%	50%	50%	50%	50%	50%
spkr1(9)	-6%	50%	20%	0%	0%	0%	0%
spkr1(7)	+6%	50%	50%	50%	50%	20%	20%
spkr4(9)	-0.3%	100%	100%	100%	100%	100%	100%
spkr6(8)	-3%	100%	100%	100%	100%	100%	80%
spkr5(4)	+0.8%	100%	100%	100%	100%	100%	60%
spkr10(6)	+1%	100%	100%	100%	80%	65%	50%

Table 8. Identification rates of individual utterances for the word "*indigestible*", word length effect, Euclidean distance.

Table 9 summarizes the distances for the test case where REF1 was used as reference utterances and TEST1(1) was used as test utterance for the word "*man*". Respective distances for the other words were found to have similar behavior. The table shows the distances obtained between all the test speakers tested against all the speakers' reference templates. The values shown are the average distances (with their respective standard deviations -std- below) of ten iterations of the same test at a given SNR of 50 dB. Note that bold characters denote the minimum distances and correct identification is obtained when the minimum distance is found on the main diagonal. Note that for only one speaker (speaker 4) the minimum distance is off the main diagonal, which means that this specific speaker was misidentified. The specific utterance of this speaker is 13% longer from the average length of the ten repetitions of the given word, as shown in Table 7.

It is obvious from Table 9 that there is a significant gap between the minimum distances and the rest. Experiments showed that the gap decreases, as the SNR level decreases, thereby increasing the error rate.

4. Effects Of Frame Length And Overlap

As mentioned earlier in the chapter, the tests were repeated for different amounts of overlap and frame lengths, as applied in the computation of the two-dimensional cepstral coefficients. Figures 31 through 33, and Tables 10 to 18 show the relative performance of the four cases for the three words tested. Recall that the four cases examined are i) overlap = 75% and frame length = 256 (32 msec), ii) overlap = 20% and frame length = 256, iii) overlap = 20% and frame length = 512 (64 msec) and iv) overlap = 75% and frame length = 512.

It can be observed that for the words "*man*" and "*beat*", cases i, ii and iv behave very similarly. Case iii exhibits a degraded performance, especially when noise is added. Note that fewer frames of the speech signal are obtained for this combination of overlap and frame length, as discussed in the examples of Chapter II. This leads to less information available for the computation of the two-dimensional cepstral coefficients, and since the

range of the frequency axis p and the quefrency axis q remain fixed, the identification procedure is subject to more errors.

The word "*indigestible*", due to its complexity and increased time duration relative to the other two words, leads to slightly degraded performance. The differences between the individual test cases are observed in the region of low SNR's, since for noise-free conditions all cases indicate similar results. The higher identification rates are obtained when the shorter frame length of 256 time samples (32 msec) is used. The lower identification rate is obtained when the overlap is 20% and the frame length is 512 time samples, as for the other two words. Since in that word several combinations of phonemes exist, such a long frame length results in obtaining frames that include both voiced and unvoiced portions of the speech signal. Thus, the identification performance is reduced, especially when noise is added, as there is not enough information to accurately represent the vocal tract characteristics.

TEST ----- REF	spkr1 av. dist (std)	spkr2 av. dist (std)	spkr3 av. dist (std)	spkr4 av. dist (std)	spkr5 av. dist (std)	spkr6 av. dist (std)	spkr7 av. dist (std)	spkr8 av. dist (std)	spkr9 av. dist (std)	spkr10 av. dist (std)
spkr1	0.1131 (0.0119)	0.2979 (0.0364)	0.5188 (0.0850)	0.4804 (0.0841)	0.6281 (0.1226)	0.5527 (0.1175)	0.6111 (0.1178)	0.5028 (0.0789)	0.5198 (0.0997)	0.4796 (0.0817)
spkr2	0.3784 (0.1537)	0.1624 (0.0192)	0.5146 (0.0660)	0.5141 (0.0954)	0.6917 (0.1275)	0.5467 (0.1158)	0.6319 (0.1192)	0.5427 (0.0812)	0.5293 (0.0968)	0.5415 (0.0958)
spkr3	0.4857 (0.0869)	0.4156 (0.0698)	0.1974 (0.0105)	0.5622 (0.1006)	0.5762 (0.1064)	0.5305 (0.1075)	0.6159 (0.1154)	0.5248 (0.0839)	0.5325 (0.0992)	0.5195 (0.0843)
spkr4	0.3701 (0.0628)	0.3731 (0.0316)	0.4672 (0.0662)	0.2899 (0.0467)	0.4326 (0.0794)	0.3731 (0.0739)	0.5948 (0.1011)	0.2759 (0.0450)	0.2827 (0.0560)	0.2458 (0.0366)
spkr5	0.6307 (0.1335)	0.6001 (0.1110)	0.5108 (0.1084)	0.4281 (0.0886)	0.1915 (0.0457)	0.4241 (0.0906)	0.6462 (0.1234)	0.3854 (0.0908)	0.3029 (0.0647)	0.3585 (0.0720)
spkr6	0.5203 (0.1022)	0.4769 (0.0789)	0.5738 (0.1127)	0.3006 (0.0482)	0.4281 (0.0815)	0.2814 (0.0695)	0.6166 (0.1157)	0.3521 (0.0673)	0.2825 (0.0510)	0.348 (0.0707)
spkr7	0.5648 (0.0949)	0.5915 (0.0867)	0.6454 (0.1130)	0.4651 (0.0646)	0.5683 (0.0856)	0.4641 (0.0827)	0.1789 (0.0220)	0.5835 (0.0944)	0.5906 (0.1015)	0.5684 (0.0898)
spkr8	0.5439 (0.1116)	0.5083 (0.0852)	0.5568 (0.1167)	0.3321 (0.0758)	0.3733 (0.0832)	0.4033 (0.0801)	0.6836 (0.1330)	0.1472 (0.0442)	0.2253 (0.0515)	0.2429 (0.0547)
spkr9	0.5526 (0.1181)	0.5146 (0.0918)	0.5724 (0.1194)	0.3203 (0.0739)	0.3205 (0.0726)	0.3296 (0.0625)	0.6723 (0.1296)	0.2675 (0.0674)	0.0931 (0.0250)	0.2048 (0.0491)
spkr10	0.5168 (0.1146)	0.5059 (0.0835)	0.5605 (0.1103)	0.2668 (0.0535)	0.3687 (0.0740)	0.3251 (0.0605)	0.6418 (0.1204)	0.2711 (0.0668)	0.2296 (0.0448)	0.1151 (0.0257)

Table 9. Normalized distances and standard deviations for the case REF1 and TEST1 (Euclidean distance).

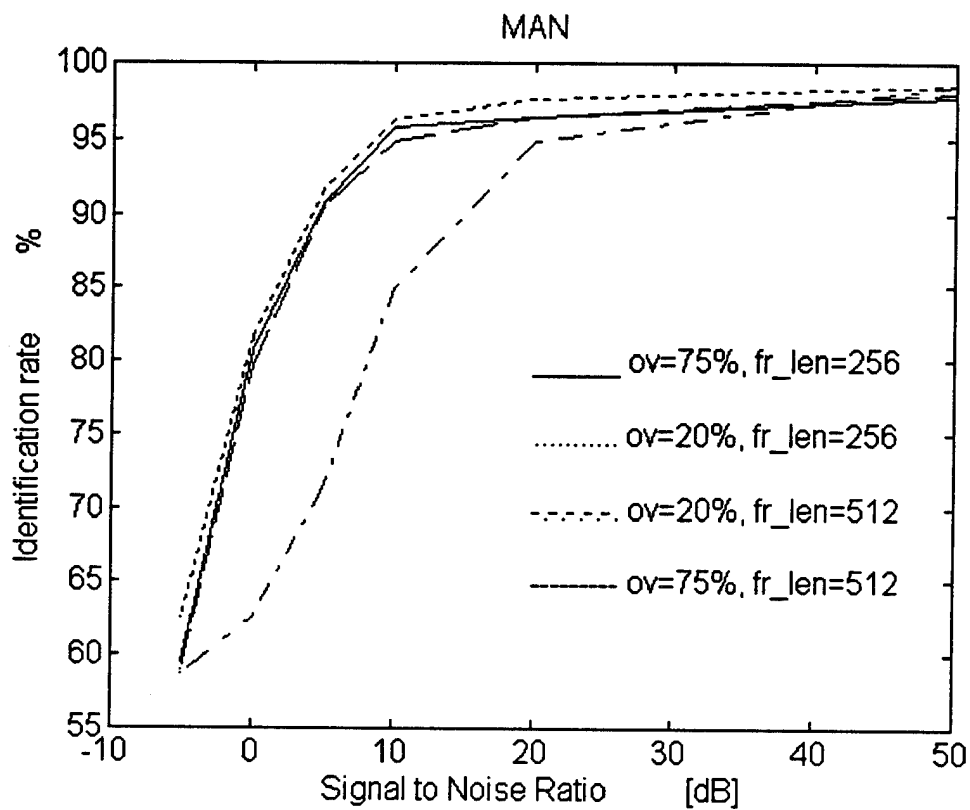


Figure 32. Identification performance for four combinations of overlap and frame length for the word "*man*", Euclidean distance.

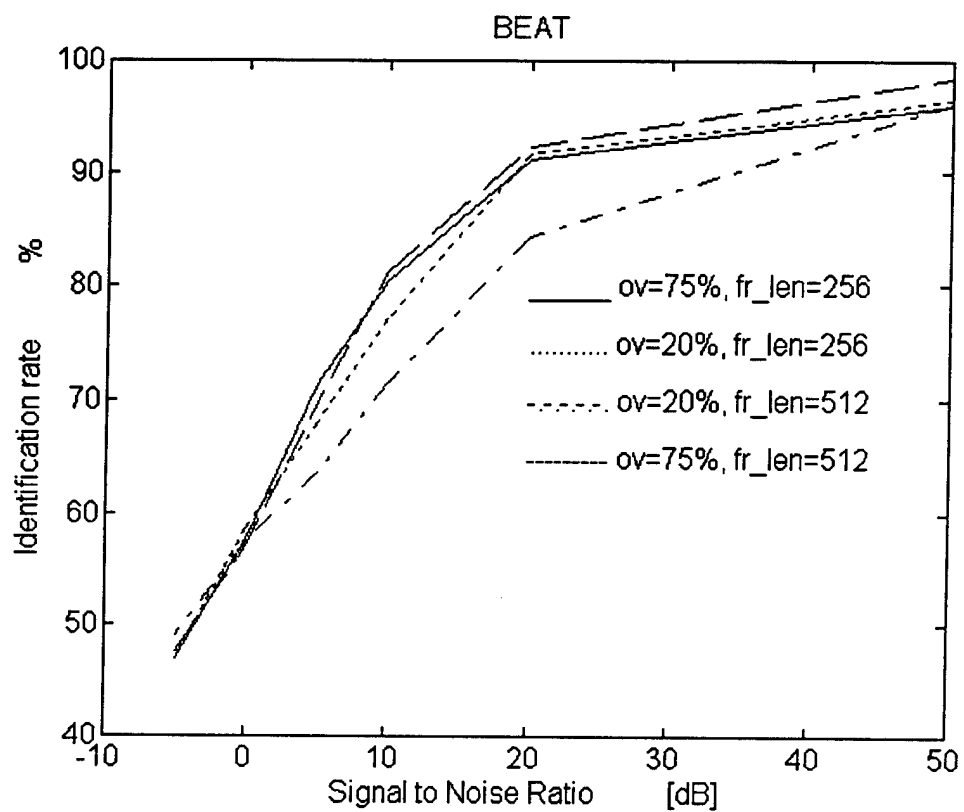


Figure 33. Identification performance for four combinations of overlap and frame length for the word "beat", Euclidean distance.

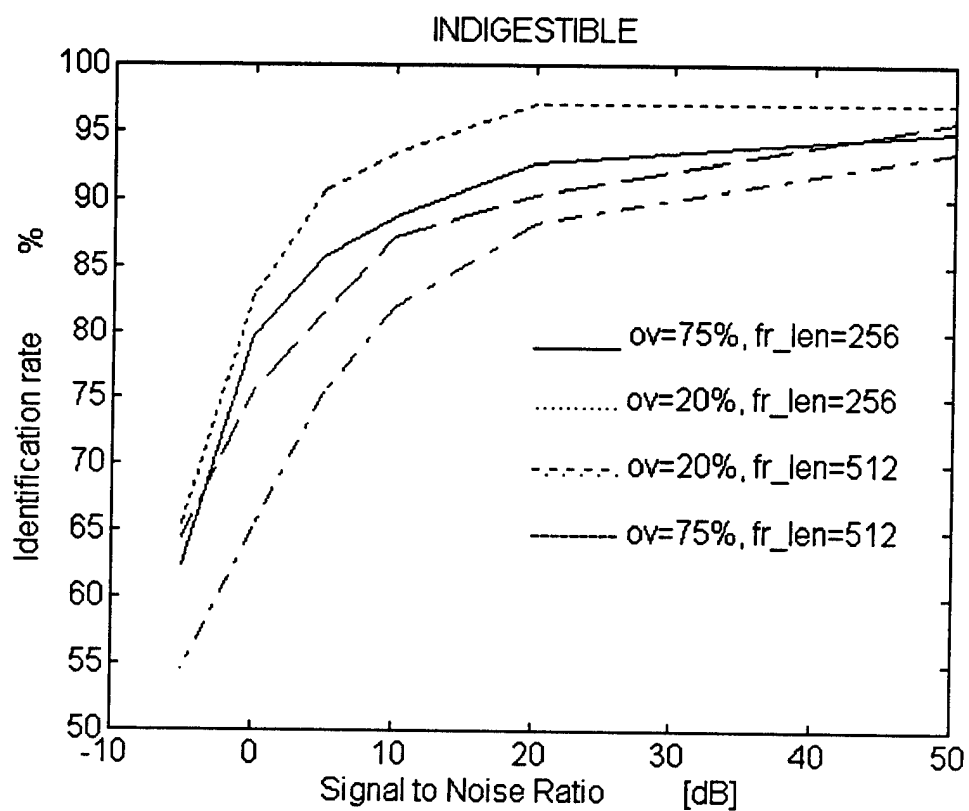


Figure 34. Identification performance for four combinations of overlap and frame length for the word "*indigestible*", Euclidean distance.

MAN overlap =20% frame length=256		
Euclidean Distance		
	Correct Identification rates	
SNR (dB)	Mean	Standard Deviation
50	98.5%	0.8771
20	97.65%	1.6259
10	96.35%	2.1668
5	91.65%	3.8401
0	81.65%	6.0576
-5	62.6%	7.0776

Table 10. Identification rates for the word "man", for SNR=50, 20, 10, 5, 0, -5 dB for the Euclidean distance. 2-D cepstral coefficients computed with 75% overlap and frame length 256 time samples.

MAN overlap=20% frame length=512		
Euclidean Distance		
	Correct Identification rates	
SNR (dB)	Mean	Standard Deviation
50	98.5%	1.6794
20	94.8%	2.388
10	84.85%	3.8133
5	71.25%	5.5412
0	62.6%	6.8717
-5	58.7%	6.1193

Table 11. Identification rates for the word "man", for SNR=50, 20, 10, 5, 0, -5 dB for the Euclidean distance. 2-D cepstral coefficients computed with 75% overlap and frame length 512 time samples.

MAN overlap=75% frame length=512		
Euclidean Distance		
	Correct Identification rates	
SNR (dB)	Mean	Standard Deviation
50	98.05%	1.395
20	96.4%	1.6455
10	94.8%	1.8564
5	90.5%	3.8364
0	79.6%	4.3958
-5	59%	5.8177

Table 12. Identification rates for the word "*man*", for SNR=50, 20, 10, 5, 0, -5 dB for the Euclidean distance. 2-D cepstral coefficients computed with 75% overlap and frame length 512 time samples.

BEAT overlap=20% frame length=256		
Euclidean Distance		
	Correct Identification rates	
SNR (dB)	Mean	Standard Deviation
50	96.4%	1.6455
20	91.6%	3.6289
10	77.2%	4.9157
5	68.1%	6.4839
0	58.7%	4.9313
-5	46.95%	6.729

Table 13. Identification rates for the word "*beat*", for SNR=50, 20, 10, 5, 0, -5 dB for the Euclidean distance. 2-D cepstral coefficients computed with 20% overlap and frame length 256 time samples.

BEAT overlap=20% frame length=512		
Euclidean Distance		
	Correct Identification rates	
SNR (dB)	Mean	Standard Deviation
50	96.15%	5.2112
20	84.3%	6.6956
10	71.5%	4.8092
5	63.75%	5.6783
0	57.45%	5.9912
-5	49.15%	6.2739

Table 14. Identification rates for the word "*beat*", for SNR=50, 20, 10, 5, 0, -5 dB for the Euclidean distance. 2-D cepstral coefficients computed with 20% overlap and frame length 512 time samples.

BEAT overlap=75% frame length=512		
Euclidean Distance		
	Correct Identification rates	
SNR (dB)	Mean	Standard Deviation
50	98.35%	1.6878
20	92.15%	2.8061
10	81.15%	4.0227
5	69.8%	5.7788
0	57%	6.0085
-5	47.55%	6.0847

Table 15. Identification rates for the word "*beat*", for SNR=50, 20, 10, 5, 0, -5 dB for the Euclidean distance. 2-D cepstral coefficients computed with 75% overlap and frame length 512 time samples.

INDIGESTIBLE overlap=20% frame length=256		
Euclidean Distance		
	Correct Identification rates	
SNR (dB)	Mean	Standard Deviation
50	97.13%	3.2051
20	97.1%	2.1219
10	93.35%	1.9942
5	90.6%	2.725
0	82.75%	3.9141
-5	65.5%	6.1227

Table 16. Identification rates for the word "*indigestible*", for SNR=50, 20, 10, 5, 0, -5 dB for the Euclidean distance. 2-D cepstral coefficients computed with 20% overlap and frame length 256 time samples.

INDIGESTIBLE overlap=20% frame length=512		
Euclidean Distance		
	Correct Identification rates	
SNR (dB)	Mean	Standard Deviation
50	93.6%	2.6096
20	88.25%	4.689
10	81.95%	2.8819
5	75.35%	3.6624
0	65.15%	6.5459
-5	54.6%	4.5336

Table 17. Identification rates for the word "*indigestible*", for SNR=50, 20, 10, 5, 0, -5 dB for the Euclidean distance. 2-D cepstral coefficients computed with 20% overlap and frame length 512 time samples.

INDIGESTIBLE overlap=75% frame length=512		
Euclidean Distance		
	Correct identification rates	
SNR (dB)	Mean	Standard Deviation
50	95.8%	1.9638
20	90.35%	3.0004
10	87.25%	3.0947
5	81.45%	2.6013
0	75.4%	2.9071
-5	64.45%	6.3648

Table 18. Identification rates for the word "*indigestible*", for SNR=50, 20, 10, 5, 0, -5 dB for the Euclidean distance. 2-D cepstral coefficients computed with 75% overlap and frame length 512 time samples.

VI. CONCLUSIONS

We investigated the application of the two-dimensional cepstral transform to the speaker identification problem. For practical implementations, the whole process from data recordings to identification decision can be automated. The two-dimensional cepstrum transform was shown to be efficient in decoupling the vocal tract characteristics from the excitation source. Thus, the two-dimensional cepstral coefficients as generated form an accurate representation of each speaker. The memory requirements of the process are significantly reduced as only a small part of the entire cepstral matrix is needed.

Three words were selected in the experiments: two simple monosyllables and one longer word. Four different reference and test groups were formed from a total of fourteen speakers used in the tests. Two different distance measures were implemented for the identification decision; the Euclidean distance and a weighted two-dimensional cepstral distance.

Results show identification rates from 95% to 98.5% for a 50 dB signal to noise ratio and from 57.65% to 80.7% for 0 dB signal to noise ratio. The high identification rates of the results, even under noisy conditions, seem promising enough. In addition, results show that the two-dimensional cepstral distance doesn't lead to significant improvements in the identification rates over those obtained using the basic Euclidean distance. Results also show that the choice of frame length and overlap must be dependent upon the range of the two-dimensional cepstral coefficients used in the identification process. Specifically, we have shown that when the frame is relatively long (512 time samples, i.e. 64 msec) and the overlap is small (20%), fewer frames of the speech signal are computed, hence less information is available. This, in turns, leads to degraded performance, especially at lower signal to noise ratios.

We noted that the most critical parameter that caused errors was the word length variation that resulted from differences in speaking rate during the recordings. Experiments showed that a higher rate of misidentification was obtained with test words which differed in length by 10% or more from the average word length obtained for a

given word and speaker. However, note that a 10% or more variation was obtained only when the speakers were specifically asked to speak at a rate different than normal. For the speakers that were directed to speak at a normal rate, although differences in word lengths existed, fewer errors occurred. Furthermore, error rate increased when some speakers intentionally distorted their voice. This leads to the conclusion that different repetitions of a given word need to be aligned in time (i.e. linearly matched) before being processed, in order to improve the results significantly.

Finally, we found out that the choice of the specific words to be tested also had an effect on the performance. Each word must be spoken clearly, and the beginning and end of each utterance must be easily identified, as illustrated with the word "*beat*" in our experiments.

Overall, the identification rates obtained from the tests are very promising especially under high SNR conditions. Future study of the specific subject should focus on the time alignment of the speech signals after their recording, in order to decrease large variations in word lengths.

APPENDIX A. MATLAB CODE FOR 2-D CEPSTRUM TRANSFORM

```
function [Cl,cres,SNR] = cepstrum(word,overlap,p_axis,frame_length,flag,SNR)
% Ver. 1.1 Modified by Ioannis Lelakis. 11/10/94
%
% This function takes a word of any length and returns
% a 5 x 14 cepstral coefficient matrix.
% If flag=1 the cepstral matrix is liftered by a raised sine lifter.
% If flag=0 the cepstral matrix is not liftered.
% The first column of the matrix is also removed, reducing the influence
% of speech energy.
%
% The user defines the amount of overlap and the frame length.
% White Gaussian noise is added to the speech signal as desired by SNR.
%
%usage:[Cl,cres,SNR]=cepstrum(word,overlap,p_axis,frame_length)

word_length=length(word);
n=frame_length;
i=1;
a=word_length-n;
overlaperc=round(frame_length*overlap/100);
word=reshape(word,1,length(word));
% ADD NOISE TO THE SPEECH SIGNAL
noise=randn(1,length(word));
sigma=std(word)^2/(std(noise)^2*10^(SNR/10));
word=word+sqrt(sigma)*noise;

% Preprocessing section
load hi100
word=filter(b,1,word);           % HighPass filtering Cutoff f=100Hz
word=normaliz(word);            % Remove the mean
word=word./norm(word);          % Normalize the power

    while a>(frame_length-overlaperc)
        n=i*frame_length-(i-1)*overlaperc;
        x(i,:)=word(n-frame_length+1:n);
        i=i+1;
        a=word_length-n;
    end
end
x=x.';                           %Orient frames columnwise for 1-D FFT
X=fft(x,512);                     %Step 1: fft of frames (zero padded)
```

```

Skm=20.*log10(abs(X)+eps);      %Step 2: log spectrum of frames
N=frame_length;
M=i;                            %Number of frames
C=(1/(N*M)) .* fft2(Skm,512,64); %Step 3: 2D FFT

C=C.';                          %Reorient to row orientation of frames
C=C(p_axis,2:15);              %Take first column out and reduce to
                                % 5 x 14 matrix
Cl=C;

if flag==1
    Cl=lifter(C);
end
[r,c]=size(Cl);
cres=reshape(Cl,1,r*c);        % Return 2-D cepstral coefficients in vector form.

```

APPENDIX B. MATLAB CODE FOR LIFTERING OPERATION

```
function c_lifword=lifter(cword)
% This function lifters the cepstral coefficients by weighting
% each frame of the cepstral matrix by the equation
%  $l(k)=1+(L/2)\sin(\pi*k/L)$ .
%
% Reference: Equation 6.53; Discrete-Time Processing of Speech Signals
% Deller et al.
%
% usage:      c_lifword=lifter(cword)
%

[m,L]=size(cword);

                                % Construct the lifter
for k=1:L
    l(k)=1+((L-1)/2)*sin(pi*(k-1)/(L-1));
end

                                % Lifter each frame of the cepstral matrix
for k=1:m
    c_lifword(k,:)=l.*cword(k,:);
end
```


APPENDIX C. MATLAB CODE FOR 2-D CEPSTRAL DISTANCE

```
function dis=papdist(average,testcep,p_axis)
% This function computes the 2-D cepstral distance of two signals
% given the cepstral coefficients according to [Ref. 10].
%
%
average=reshape(average,length(p_axis),14);
testcep=reshape(testcep,length(p_axis),14);
num_coef=length(p_axis)*14;

for u=1:length(p_axis)
    for v=1:14
        disp(u,v)=(v^2+u^2+1)*abs(average(u,v)-testcep(u,v))^2;
    end
end

dis=(1/num_coef)*sum(sum(disp));
```


LIST OF REFERENCES

- [1] J. R. Deller, Jr., J. G. Proakis, J. H. L. Hansen, *Discrete-time Processing of Speech Signals*, Macmillan Publishing Company, New York, New York, 1993.
- [2] L. R. Rabiner, R. W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall Inc, Englewood Cliffs, New Jersey, 1978.
- [3] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The quefrency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-spectrum and saphe cracking," *Proceedings of the Symposium on Time Series Analysis*, John Wiley and Sons, New York, New York, pp. 209-243, 1963.
- [4] A. M. Noll, "Cepstrum pitch determination," *Journal of the Acoustical Society of America*, Vol. 41, pp. 293-309, February 1967.
- [5] B. H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," *IEEE Transactions on ASSP*, Vol. ASSP-35, No. 7, July 1987.
- [6] J. W. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE*, Vol. 81, No. 9, pp. 1215-1247, September 1993.
- [7] Y. Ariki, S. Mizuta, T. Sakai, "Spoken-word recognition using dynamic features analysed by two-dimensional cepstrum," *IEE Proceedings*, Vol. 136, Pt. 1, No. 2, April 1989.
- [8] B. S. Atal, "Automatic recognition of speakers from their voices," *Proceedings of the IEEE*, Vol. 64, pp. 460-475, April 1976.
- [9] H. F. Pai, H. C. Wang, "Two-dimensional cepstral distance measure for speech recognition," *ICASSP-93*, Vol. II, pp. II.672-II.675, 1993.
- [10] Y. Tohkura, "A weighted cepstral distance measure for speech recognition," *IEEE Transactions on ASSP*, Vol. ASSP-35, No. 10, pp. 1414-1422, October 1987.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center. 2
Cameron Station
Alexandria, Virginia 22304-6145
2. Library, Code 52. 2
Naval Postgraduate School
Monterey, California 93943-5101
3. Chairman, Code EC. 1
Department of Electrical and Computer Engineering
Naval Postgraduate School
833 Dyer Road, Room 437
Monterey, California 93943-5121
4. Professor M. P. Fargues, Code EC/Fa. 4
Department of Electrical and Computer Engineering
Naval Postgraduate School
833 Dyer Road, Room 437
Monterey, California 93943-5121
5. Professor R. Hippenstiel, Code EC/Hi. 1
Department of Electrical and Computer Engineering
Naval Postgraduate School
833 Dyer Road, Room 437
Monterey, California 93943-5121
6. LT Ioannis Lelakis. 1
Pipinou 72-74 Str.
112 51 Athens, Greece