Perceptually Motivated Blind Source Separation of Convolutive Audio Mixtures

Ram Mohana Reddy Guddeti

A thesis submitted for the degree of Doctor of Philosophy.
The University of Edinburgh.
September 2005

Abstract

Blind source separation aims to recover independent sources from their multiple observed mixtures using independent component analysis (ICA). However, when applying this technique to audio mixture problem such as a number of people talking in a room, the performance of the system is greatly reduced by the effect of room reflections and ambient noise.

In contrast, two human microphones (the ears) perform well in such a real *cock-tail party* environment. Over the last few years the application of psycho-acoustic principles (the human auditory perception) has led to the successful development of MPEG audio coding standard which is the basic technique behind MP3 players and music availability on the internet.

The first objective of this thesis is to apply psycho-acoustic principles to the spatial processing of speech signals in noisy and reverberant environment. The key assumption that will be adopted is that modern signal processing has failed to mimic the *cock-tail party effect* because there has been no attempt to adequately incorporate the psycho acoustical phenomenon of audio masking to aid source separation. A quasi linear mechanism for mimicking *simultaneous frequency masking* and *temporal masking* (post masking) techniques is developed. This framework is used to construct blind source separation algorithms that exploit audio masking prior to source separation (preprocessor) and after source separation (postprocessor).

The final objective of this thesis is to exploit the perceptual irrelevancy of some of the input speech spectrum using the perceptual masking techniques before utilising the subspace method as a preprocessor of the frequency-domain ICA (FDICA) which reduces the effect of room reflections in advance and the remaining direct sounds then being separated by ICA.

Incorporating the perceptual masking techniques prior to the application of FDICA with the subspace method as preprocessor not only reduces the computational complexity of similarity measure for solving the permutations but also avoids the so-called permutation problem by targeting a specific speech signal more intelligible than the available microphone signals. Experiments carried out in both synthetic and real room scenarios and the results shown good objective performance in terms of signal-to-interference ratio (SIR) and enhanced modified Bark spectral distortion (EMBSD) confirm the validity of the proposed solutions.

Dedicated To My Parents

Acknowledgements

Of the many people to whom I owe a great deal of heartfelt thanks for their invaluable assistance during the course of my PhD, the following deserve special mention:

- Professor Bernard Mulgrew, my supervisor, for his support, guidance, invaluable advice, encouragement and his willingness to share his deep insight during the various stages of this work. Also for reading and checking this thesis during time when his attention is greatly demanded by so many other people.
- Professor Steve McLaughlin, my 2nd supervisor, for his support and guidance.
- Association of Commonwealth Universities and British Council for providing funding to pursue my Doctoral studies in U.K. through Commonwealth Scholarship.
- My colleagues in the Institute for Digital Communications for their assistance in one way or another during the last three years. Special thanks to Dr. David Blanco, Moti Tabulo and Amit Mishra for providing valuable assistance.
- The staff of the Institute for Digital Communications, particularly Dr. Dave Laurenson and Dr. James Hopgood, who have at some stage or another provided valuable help.
- David Stewart, Michael Gordon, Chris Rudd and Bryan Tierney for their instantaneous computer support and tolerance towards my high computing usage.
- My parents for inspiration and brothers and brother-in-laws for lending me a wonderful family. My friend Kesava, for continuous help and encouragement from Hyd, India.
- Vijaya for giving me cheerful company through out the course of this work and bearing with all the difficulties of running the home smilingly.
- Special thanks to Prof. P. Narasimha Reddy, Prof. C. V. S. Rao, Prof. P. S. R. Murthy, Prof. Kumar Eswaran of SNIST and Prof. P. S. Moharir of NGRI, Hyderabad, for their encouragement to carryout research through Commonwealth Scholarship.
- All those who made our stay in Edinburgh an unforgettable and rewarding experience.

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

AAC	Advanced Audio Coding
ACR	Absolute Category Rating
ALS	Alternating Least Squares
AM	Auditory Masking
ATH	Absolute Threshold of Hearing
BM	Basilar Membrane
BSS	Blind Source Separation
CASA	Computational Auditory Scene Analysis
CB	Critical Bandwidth
CCR	Comparision Category Rating
CIFC	Combined Inter Frequency Correlation
DOA	Direction of Arrival
DMOS	Degradation Mean Opinion Score
EMBSD	Enhanced Modified Bark Spectral Distortion
ERB	Equivalent Rectangular Bandwidth
FDICA	Frequency Domain Independent Component Analysis
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
GMM	Gaussian Mixture Models
HOS	Higher Order Statistics
ICA	Independent Component Analysis
IFC	Inter Frequency Coherency
IFFT	Inverse Fast Fourier Transform
IFSEC	Inter Frequency Spectral Envelope Correlation
IHC	Inner Hair Cell
IIR	Infinite Impulse Response
ISO	International Standards Organization
ISTFFT	Inverse Short Time Fast Fourier Transform
JND	Just Noticeable Distortion

LPC	Linear Predictive Coding
MA	Moving Average
MBSD	Modified Bark Spectral Distortion
ML	Maximum Likelihood
MOS	Mean Opinion Score
MPEG	Moving Picture Experts Group
MP3	MPEG Audio Coding Standard Layer 3
MSFD	Multi Stage Frequency Domain
MSICA	Multi Stage Independent Component Analysis
NM	Noise Maskers
NMR	Noise-to-Mask Ratio
OHC	Outer Hair Cell
PCA	Principal Component Analysis
PEAQ	Perceptual Evaluation of Audio Quality
PN	Power Normalization
PSD	Power Spectral Density
SIR	Signal-to-Interference Ratio
SMR	Signal-to-Mask Ratio
SNR	Signal-to-Noise Ratio
SOC	Source Output Crosscorrelation
SOS	Second Order Statistics
SPL	Sound Pressure Level
SR	Strong Reflections
STFFT	Short Time Fast Fourier Transform
SVD	Singular Value Decomposition
TDDA	Time Delayed Decorrelation Algorithm
TDICA	Time Domain Independent Component Analysis
TM	Tonal Maskers
WR	Weak Reflections

Nomenclature

$\mathbf{A}$	Mixing Filter Matrix
$\mathbf{A}(\omega)$	Overall Mixing Filter Transfer Function
$\mathbf{A}'(\omega)$	Mixing Filter Transfer Function due to Strong Reflections
$\mathbf{A}_{er}(\omega)$	Mixing Filter Transfer Function due to Weak and Strong Reflections
$\bar{\mathbf{A}}(\omega)$	Permuted Mixing Filter Transfer Function
$\mathbf{A}_{m,n}(\omega)$	Mixing Filter Transfer Function from nth Source to mth Microphone
$\mathbf{A}^{er}_{m,n}(\omega)$	Reflected Part of $\mathbf{A}(\omega)$ from nth Source to mth Microphone
$\hat{\mathbf{A}}(\omega)$	Estimated Mixing Filter Transfer Function
$\mathbf{B}$	Unmixing Filter Matrix
$\mathbf{B}(\omega)$	Overall Unmixing Filter Transfer Function
$\mathbf{B}^{-1}(\omega)$	Inverse of Overall Unmixing Filter Transfer Function
$\mathbf{B}^{+}(\omega)$	Pseudoinverse of Overall Unmixing Filter Transfer Function
$\widetilde{\mathbf{B}}^{-1}_{\widetilde{m}}(\omega)$	Arbitrary Scaling Matrix due to PCA Method
$\widetilde{\mathbf{B}}^{+}_{\widetilde{m}}(\omega)$	Arbitrary Scaling Matrix due to Subspace Method
$BW_c(f)$	Critical Bandwidth in Hz
$\mathbf{C}$	Global separation Filter Matrix
$C(k)$	Confidence Measure for Computing Permutations
$C_d(j)$	Cognizable Distortion of jth Cognizable Segment
Corr	Correlation Between Adjacent Frequency Components
$\mathbf{D}$	Non-Singular Diagonal Filter Matrix
D	Number of Sound Sources
$D_{xy}(.)$	Loudness Difference Between Original and Separated Speech Vectors
$\mathbf{E}$	Eigen Vector Column Matrix
$E[.]$	Expectation
$F(\mathbf{P})$	Cost Function For Computing Permutations
$F(\mathbf{P}, \hat{k})$	Cost Function For Computing Correct Permutations at Reference Freq., $\hat{k}$
$F_{IFC}(\mathbf{P})$	Cost Function For Computing Permutations by IFC Method
$F_{IFSEC}(\mathbf{P})$	Cost Function For Computing Permutations by IFSEC Method
$\mathbf{F}(\omega)$	Frequency Domain Separated Filters after Solving Scaling and Permutation

G	Gain Constant for the Nonlinear Score Function
$.^H$	Hermitian Transpose
$H_{m,n}(\omega)$	Magnitude of $\mathbf{A}(\omega)$ from nth Source to mth Microphone
$\mathbf{I}$	Identity Matrix
$\Im(.)$	Imaginary Part of the Complex Data
K	Reference Range for Computing Permutation Matrix
L	A Positive Constant Used in Spectral Envelope Estimation
L_T	Number of Tonal Maskers
$L_x(.)$	Normalized Loudness Vectors of Original Speech
$L_y(.)$	Normalized Loudness Vectors of Separated Speech
$\mathbf{M}_i$	Multiple Time Delayed Decorrelation Matrices
M	Number of Microphones (Sensors)
$M_d(.)$	Perceptual Distortion Indicator
M_N	Number of Noise Maskers
N	FFT Length
N_c	Number of Critical Bands
N_f	Number of Cognizable Segments
$NMTb$	Noise-Masking-Tone for Each of the Partition Index, b
$NMTh$	Noise Masking Threshold Without the Spreading Function, $SF(i,j)$
$\mathbf{P}$	Permutation Filter Matrix
$P(k)$	Power Spectral Density Estimate for a Particular Frequency Bin, k
$P_d(j)$	Perceptual Distortion of the jth Cognizable Segment
$P_{TM,NM}(k)$	Power Spectral Density of Tonal (Noise) Maskers
$Q_d(j)$	Postmasking Distortion of the jth Cognizable Segment
$\Re(.)$	Real Part of the Complex Data
$\mathbf{R}(\omega,\tau)$	Time-Lagged Spatial Correlation Matrix
$\mathbf{R}(\omega)$	Spatial Correlation Matrix at $\tau = 0$
$SF(i,j)$	Spreading Function of Masking from Masker bin j to Maskee bin i
S_T	Tonal Set
SNR_b	Signal-to-Noise Ratio in Each Partition Index, b
$\mathbf{T}_c(.)$	Complex Rotational Filter Matrix
T_g	Global Masking Threshold of Audibility Using Model 1
Th_w	Total Energy Threshold of Audibility Using Model 2

TMN_b	Tone-Masking-Noise for Each of the Partition index, b
T_{NM}	Noise Masking Thresholds
T_q	Threshold in Quiet
T_{TM}	Tonal Masker Thresholds
$.^T$	Transpose of a Matrix
$\mathbf{U}(\omega)$	ICA Filter Matrix
$\mathbf{W}(\omega)$	Subspace (PCA) Filtering Matrix
$Z_c(f)$	Critical Band Number in Barks
$\mathbf{\Phi}$	Perceptual Gain
$\mathbf{\Phi}_{th}$	Perceptual Masking Threshold Matrix
$\mathbf{\Lambda}$	Diagonal Matrix of Eigenvalues
$\hat{\mathbf{P}}$	Estimated Permutation Matrix
λ	Eigenvalues
d	Delay Associated with Mixing Filter Matrix, $\mathbf{A}(\omega)$
$\mathbf{e}$	Eigenvectors
r	Number of Matrices to be Simultaneously Diagonalized
span$(\mathbf{A})$	Subspace Spanned by the Columns of $\mathbf{A}$
τ	Time Delay (Shift) Parameter
$\tau_{m,n}$	Propagation Time from nth Source to the mth Microphone
$\widetilde{m}$	An Arbitrary Microphone Number
$*$	Linear Convolution Operator
ε	Moving Average Operator
$\neq$	Not Equal
$\simeq$	Almostly Equal
$\forall$	For All Values
$\gg$	Much Greater
$\in$	Belongs to
ρ	Correlation Coefficient of Spectral Envelopes
sim(.)	Similarity Measure for Computing Permutations
diag(.)	Diagonal of Matrix Values
$\varphi(.)$	Nonlinear Score (Activation) Function
η	Learning Parameter
γ	Postmasking Factor of Model 2

$argmax_{\mathbf{P}}[F(\mathbf{P})]$	The Argument $\mathbf{P}$ that maximises $F(\mathbf{P})$
$max_{\mathbf{P}}[F(\mathbf{P})]$	Maximise $F(\mathbf{P})$ in terms of $\mathbf{P}$
bb	Index for Convolved Energy and Unpredictability Measures of Model 2
bc_b	Power Ratio of Model 2
$bmax$	Largest Value of Calculation Partition Index, b
bvb	Median Bark Value of the Partition
c_b	Weighted Unpredictability of Model 2
c_w	Unpredictability Measure of Model 2
ct_b	Convolved Weighted Unpredictability of Model 2
e_b	Energy in Each Partition of Model 2
ec_{bb}	Convolved Partioned Energy of Model 2
en_b	Normalized Energy of model 2
f	Frequency in Hz
$\mathbf{f}(i)$	Time Domain Separated Filters after Solving Scaling and Permutation
f_0	Fundamental Frequency Corresponding to Voice Pitch
f_w	Phase of the Complex Spectra of Model 2
$\hat{F}_w$	Predicted Phase of the Complex Spectra of Model 2
$iblen$	Shift Length Parameter of Model 2
j	Complex Phasor Notation Equal to $\sqrt{-1}$
$\bar{k}$	Geometrical Mean of Spectral Lines
l	Lower Spectral Line Boundary of Critical Band
mvb	Lower Limit for SNR in the Partition (Stereo Unmasking Effects)
nb	Number of Bits of Digital Audio Samples of Model 1
nb_b	Actual Energy Threshold for Each Partition of Model 2
nb_w	Threshold Energy Spread Over FFT Spectral Lines of Model 2
$p(.)$	Probability Density Function
rn_b	Normalization Coefficient of Model 2
r_w	Magnitude of the Complex Spectra of Model 2
$\hat{r}_w$	Predicted Magnitude of the Complex Spectra of Model 2
$\tanh(.)$	Sigmoid Tanh Nonlinearity Function
t_{bb}	Tonality Index of Model 2
u	Upper Spectral Line Boundary of Critical Band
v	Cognizable Unit of Model 2

$a_i(\omega, t)$	Magnitude of the Spectral Envelope for Each Output
$\phi_i(\omega, t)$	Phase of the Spectral Envelope for Each Output
$\mathbf{n}(t)$	Noise Vector of General ICA/BSS Framework
$\mathbf{n}(\omega, t)$	Noise Spectra Vector of FDICA Framework
$\mathbf{s}(t)$	Source Vector of General ICA/BSS Framework
$\mathbf{s}(\omega, t)$	Source Spectra Vector of FDICA Framework
$w(i)$	Windowing Function
wlb	Lowest Frequency Line in the Partition of Model 2
whb	Highest Frequency Line in the Partition of Model 2
$\mathbf{x}(t)$	Sensor Vector of General ICA/BSS Framework
$\mathbf{x}(\omega, t)$	Sensor Spectra Vector of FDICA Framework
$\mathbf{x}_f(\omega, t)$	Sensor Spectra Vector of Perceptual FDICA Framework
$\mathbf{x}(n)$	Normalized Digital Audio Samples of Sensor Input
$\widetilde{\mathbf{x}}(n)$	Digital Audio Samples Without Normalization
$\mathbf{x}_i(n)$	Reconstructed 1024 Speech Data of Model 2
$\mathbf{y}(t)$	Output Vector of General ICA/BSS Framework
$\mathbf{y}_{PCA}(t)$	Output of PCA Filter Network
$\mathbf{y}(\omega, t)$	Output Spectra Vector of TDDA Framework
$\mathbf{y}_f(\omega, t)$	Output Spectra Vector of Perceptual TDDA Framework
$\hat{\mathbf{z}}(\omega)$	Estimated Spectral Envelope of the Separated Output
$\mathbf{z}(\omega, t)$	Output Spectra Vector of Complex Infomax Framework
$\mathbf{z}_f(\omega, t)$	Output Spectra Vector of Perceptual Complex Infomax Framework

Chapter 1
Introduction

This chapter gives a general introduction to the work presented in this thesis, providing a brief overview of the research field, indicating current areas of interest and identifying the focus of the investigation. The aims of this work are then addressed, followed by an outline of the content of other chapters presented in the thesis.

1.1 Audio Source Separation

Humans exhibit a remarkable ability to extract a sound source of interest from an auditory scene captured by the brain. The human brain can perform this everyday task in real time using only the information acquired from a pair of microphones (sensors), i.e. two ears. Imagine the situation of attending a *cocktail party* function busy with a lot of activities. Our ears capture a huge variety of sound sources: music, other people speaking, mobile phones ringing, glasses tinkling etc. However, we can concentrate on a specific source that is of more interest at that point of time. For example, we may listen to what our friend is saying. Getting bored, we can overhear somebody else's conversation, pay attention to an annoying mobile ringtone or even listen to the music played by the sound system, only to understand it is a popular song.

Thus, the human brain can automatically focus on and separate a specific sound source of our interest. In general, source separation is the process aiming to separate a finite number of source signals from a finite set of recorded (observation) signals. Audio source separation can be defined as the problem of decomposing a real world sound mixture (auditory scene) into independent audio objects. A perceptually motivated analysis using a computer (machine) that exploits the irrelevancy of captured auditory scene through a number of sensors in a noisy and reverberant environment is the main objective of this thesis. Although this is a relatively simple task for the human auditory system, a perceptually motivated audio source separation can be considered one of the most challenging topics in the current research.

Different approaches were proposed to solve this audio source separation problem (*cocktail party effect*) are reviewed in the following subsections.

1.1.1 Computational Auditory Scene Analysis

A possible approach to address the problem will be to analyse and finally emulate the way humans perform audio source separation using a computer. Psychoacoustics is a special area of research studying how people perceive, process and deduce information from sounds. Such studies construct experimental stimuli consisting of a few simple sounds such as sine tones or noise bursts, and then record human subjects perception of these test sounds [3]. Audio source separation may be regarded as one aspect of a more general process of auditory organization of these simple structures, which is able to untangle an acoustic mixture in order to retrieve a perceptual description of each constituent sound source [4].

Computational Auditory Scene Analysis (CASA) was one of the first methods that tried to decrypt the human auditory system in order to perform an automatic audio source separation system [4–6]. Conceptually, CASA may be divided into two stages. In the first stage, the acoustic mixture is decomposed into sensory elements (segments). CASA employs either complete ear models (outer and middle ear, cochlear filtering etc) or computer vision techniques in order to segment the auditory scene into several audio elements. The second stage (grouping) then combines segments that are likely to have originated from the same sound source [4]. Psychological and psychoacoustic research of this kind has uncovered a number of cues or grouping rules which may describe how to group different parts of an audio signal into a single source, such as i) common spatial origin, ii) common onset characteristics, i.e., energy appearing at different frequencies at the same time, iii) amplitude or frequency modulations in the harmonics of a musical tone, iv) harmonicity or periodicity, v) proximity in time and frequency, vi) continuity (i.e. temporal coherence). Usually, CASA employs one or two sensor signals, as the main goal is to emulate humans way of performing auditory scene analysis [6].

1.1.2 Beamforming

Array signal processing is a research topic that developed during the late 1970s and 1980s mainly for telecommunications, radar, sonar and seismic applications. The basic array processing problem consists of obtaining and processing the information about a signal environment from the waveforms received at the sensor array (a known constellation of microphones). Generally, the signal environment consists of a number of emitting sources plus noise. Exploiting time difference information from the observed signals, one can estimate the number of sources present in the environment using direction of arrival (DOA) towards the array sensor [7].

The use of an array allows for a directional beam pattern. The beam pattern can be adapted to null out signals arriving from directions other than the specified look direction. This technique is known as spatial filtering or adaptive beamforming [8]. The reception of sound in large rooms, such as conference rooms and auditoria, is typically contaminated by interfering noise sources and reverberation. One can set up an array of microphones and apply the techniques of adaptive beamforming in the multiuser communication environment to perform several audio processing tasks. We can enhance the received amplitude of a desired sound source, while reducing the effects of the interfering signals and reverberation.

Moreover, we can estimate the direction or even the position of the sound sources in the near field present in the room (source localisation). Most importantly, if the auditory scene contains more than one source, we can isolate one source of interest, whilst suppressing the others, i.e. perform source separation. Beamforming assumes some prior knowledge on the geometry of the array, i.e. the distance between the sensors and the way they are distributed in the auditory scene. Generally, linear arrays are used to reduce the computational complexity of the source separation system. In addition, optimally the array should contain more sensors than the sources in the auditory scene. Exploiting the information of the extra sensors using subspace methods, we can localise and separate the audio sources [9, 10].

1.1.3 Blind Source Separation

In contrast to CASA and beamforming, blind source separation (BSS) is a statistical technique that draws inspiration neither from the mechanisms of auditory function nor from the geometry of the auditory scene. BSS systems can identify sound objects from the observed mixtures of original sources. Blind means that we hardly know anything about the original sources. By definition, in blind separation there is no available a priori knowledge concerning the exact statistical distributions of the source signals; no available information about the nature of the process by which the source signals were combined (mixing process) [11–13].

In reality, some assumptions must be made regarding the source signal distributions and a model of the mixing process must be adopted. However, these assumptions remain fairly general without undermining the strength of the method. BSS aims to separate the source signals from their multiple observed mixtures using Independent Component Analysis (ICA), a blind estimation framework that assumes that the sources are statistically independent [14–16]. This assumption together with a source prior can perform audio source separation task [17–19].

To model the mixing procedure, usually finite impulse response (FIR) filters are employed to describe the room's transfer function between the sources and the sensors [19]. In the thesis, we are mainly going to focus on this analysis method and more specifically on frequency-domain ICA (FDICA). However, as all the aforementioned approaches try to solve essentially the same problem, it might be beneficial to find some links between these methods in order to produce a more complete audio source separation system. In the thesis, we also explore whether blind source separation can incorporate some important features from beamforming (using subspace method) and the human auditory system (using psychoacoustic models).

1.1.4 Applications of Audio Source Separation

There are many applications where an audio source separation system can be useful:

- *Noise Suppression for mobile phones/hearing aids.* Having unmixed the sources that exist in an auditory scene, one can remove the unwanted noise sources in a multiple source environment. This process can serve as a denoising utility for mobile phones, hearing aids or any other recording facility.

- *Music transcription.* Unmixing a recording to the actual instruments that are playing in the recording is an extremely useful tool for all music transcribers. Listening to an instrument playing solo rather the actual recording facilitates the transcription process. This applies to all automated polyphonic transcription algorithms that have appeared in research. Combining an audio source separation algorithm with a polyphonic transcriber will lead to a very powerful musical analysis tool.

- *Efficient coding of music.* Each instrument has different pitch, attack, timbre characteristics, requiring different bandwidth for transmission. Decomposing a musical signal into sound objects (instruments) will enable different encoding and compression levels for each instrument resulting in a more efficient, high quality audio codec. This will be more in line with the general framework of MPEG-4 for video and audio.

- *Medical applications.* Audio source separation algorithm might be useful in applications such as the separation of foetus's heartbeat from the mother's in the womb.

- *Surveillance applications.* The ability of discriminating between the audio objects of an auditory scene will enhance the performance of surveillance applications.

- *Remixing of studio recordings.* In future audio applications, with all the powerful tools that can search for songs similar to the ones we like or that sound like the artist we want, a personal remixing of a studio recording will be possible with audio source separation.

- *Post-processing of film recordings.* Source separation tools will be very useful for editing and special effects in the film industry. An audio source separation algorithm will help post-adjust actors' voice levels in a film take. Dubbing in different languages and any kind of post-processing will also be facilitated.

1.2 Aims of This Work

The research in this thesis focuses on the linear convolutive mixing of speech signals in a noisy and highly reverberant environment. The real *cocktail party* problem cannot be directly solved by general ICA framework because of two reasons: first, there is a reverberation effect due to actual observation of sound sources in a room that is no longer modeled by a linear instantaneous mixing process, but this can be modeled by a convolutive mixing process; second, in practice there are fewer microphones (sensors) than unknown sound sources. However, humans deal with this problem very effectively and easily by using 2 dynamic sensors (ears).

Perceptual audio coders use human hearing models to determine perceptual relevance and then eliminate redundancy with the minimal degradation of relevant information. When the sources are stationary, perceptual audio coders are frame or packet based because masking thresholds are computed for finite input blocks (1024 or 512 samples). This framework is known as block based perceptual masking approach. Similarly, when the sound sources are dynamic then we need to consider the sequential or adaptive perceptual masking approach.

The primary objective of this thesis is to reduce the computational complexity of solving the permutation problem of FDICA using both perceptual masking approaches. In the block based approach, simultaneous masking is applied between adjacent frequencies of the input speech block at the same time, to remove the irrelevant frequency components.

On the other hand, sequential based perceptual approach deals with non-simultaneous masking (also known as temporal masking). This framework will be used to construct BSS algorithms that exploit audio masking prior to the source separation (preprocessor) and after the source separation (postprocessor) in the frequency-domain.

The final objective of this thesis will be to develop an intelligent FDICA system that extracts a single source of interest from the mixtures. This can be achieved by exploiting the irrelevancy of some of the input speech spectrum using perceptual masking techniques before utilizing the subspace filtering method as a preprocessor of FDICA which reduces the effect of room reflections in advance and the remaining direct sounds then being separated by ICA.

1.3 Assumptions

In order to obtain a simplified understanding of the source separation problem, as well as to develop and test the proposed perceptually motivated method, the speech signal and masking models are simplified as much as possible. The following assumptions are made:

- Speech signals have a temporal structure that it is stationary for a period shorter than 50~60 ms but non-stationary if it is longer than 50~60 ms [20]. We used this time structure to construct BSS algorithms. The definition of stationarity adopted in this thesis is that described by Papoulis [21] for wide-sense stationarity: a signal whose mean and variance are constant. The description non-stationary may therefore be applied to signals that do not exhibit these properties.
- As backward masking tends to last only for a very short duration of 5 ms, we did not consider this category of temporal masking while estimating the masking threshold using ISO/MPEG-1 psychoacoustic model 2. Furthermore, we have considered the values of Calculation Partition Table and Absolute Threshold Table at 32 kHz sampling rate due to the constraints of the proposed perceptual FDICA algorithm.

1.4 Thesis Overview

This thesis is focused on the audio source separation problem in a noisy and highly reverberant room mixing environment. In this investigation, we address a couple of specific open problems in the field, as it is explained further on. Our solutions are based on the combined approach of psychoacoustic filtering (perceptual masking), subspace filtering method (beamforming) and blind source separation (using ICA) method aiming to decompose linear convolutive audio mixtures of statistical independent speech signals while reducing the computational complexity of solving the permutation ambiguity problem of FDICA.

- Chapter 2 introduces the key aspects of relevant background material, covering the basic ideas of blind source separation (linear instantaneous and convolutive mixtures) and some basic principles of the human auditory system (perceptual masking). It also discusses the suitability of using psychoacoustic models to implement perceptually motivated solutions for reducing the computational complexity of solving the permutation problem.

- Chapter 3 describes the experimentation undertaken on convolutive mixtures of speech signals. It investigates the exploitation of the perceptual irrelevancy of some of the observed input speech signal spectrum before applying the complex FDICA algorithm (multiple time-delayed decorrelation and Infomax) and the effect that this perceptual masking has on the separation performance of BSS system. This approach will then be compared with BSS system which do not take perceptual masking into account.

- Chapter 4 examines whether perceptual masking criteria, which takes into account the process whereby one auditory stimulus prohibits the detection of another speech signal, can enhance the separation performance of existing BSS system. The perceptual solution proposed in this Chapter, is a variation of that described in Chapter 3; the alternation is that perceptual masking is applied to the separated signals (obtained by the complex Infomax algorithm) before solving the permutation ambiguity problem.

- Chapter 5 extends the work described in Chapter 3 by developing a combined approach of the perceptual masking and the subspace filtering method to enhance the performance of blind separation of speech signals in a noisy and highly reverberant environment. In this approach, two important signal processing techniques namely audio masking and the subspace method are utilised respectively for suppressing the perceptually irrelevant components of some of the input speech signal spectrum and to reduce the effect of room reflections prior to the application of the complex Infomax algorithm.

- Chapter 6 presents a summary of the research undertaken, and the conclusions that can be drawn from each of the previous chapters. It also identifies future directions of work that could be developed from the research presented in this thesis.

Tables of psychoacoustic model 2 and critical band analysis are given in Appendix A and copies of papers published from work undertaken in the thesis are reproduced in Appendix B.

Chapter 2
Background

This chapter provides a general background for the research that has been carried out in this thesis by presenting a complete literature review of the underlying principles upon which the research is built. An introduction to blind source separation (BSS) of instantaneous mixtures is provided, followed by a complete description of time-domain and the frequency-domain approaches for BSS of convolutive mixtures. Finally, some basic principles of the human auditory system are presented, covering both psychoacoustic models that have been considered for reducing the computational complexity of solving the blind signal processing problems in general and the permutation ambiguity problem in particular, and the aims of the thesis are restated on the basis of this background information.

2.1 BSS of Instantaneous Mixtures

Blind source separation is the process wherein source signals are separated from a finite set of observed or sensor signals without prior knowledge of the source signals and the type of mixing environment. Although this lack of prior knowledge may be considered as a drawback, it is the actual strength of BSS methods, making them a versatile tool for exploiting the spatial diversity provided by an array of microphones (sensors)[14–16]. BSS has various levels of difficulty, mainly according to three features. The first feature is the type of mixing. The easiest case to deal with is *time-invariant linear instantaneous* mixtures, when the observations at time t are only a linear combination of the sources at the same time t.

However, in practice, the observed mixtures are usually *convolutive*, possibly *time-variant*, and sometimes *non-linear*. The real *cocktail party problem* [22], for example, consists of separating speech emitted by people speaking simultaneously from the signals recorded by a few sensors located in the room. The room reflection/reverberation creates convolutive mixtures and the movement of the speakers creates time-varying mixtures (because room transfer functions are space dependent). Non-linearity can be caused by possible saturation at the microphones.

Another important feature of BSS problems is the number of sources D with respect to the number of microphones M. In the overdetermined case ($M \geq D$) it is usually sufficient to estimate the mixing system and apply its pseudoinverse to the observations to recover the sources. In the opposite case, the underdetermined case ($M < D$) is an ill-posed problem because the mixing system can no longer be inverted and prior information about the sources is required to allow for their reconstruction. The third feature is the nature of the source signals (stationary or non-stationary) that usually determines the type of method to be employed.

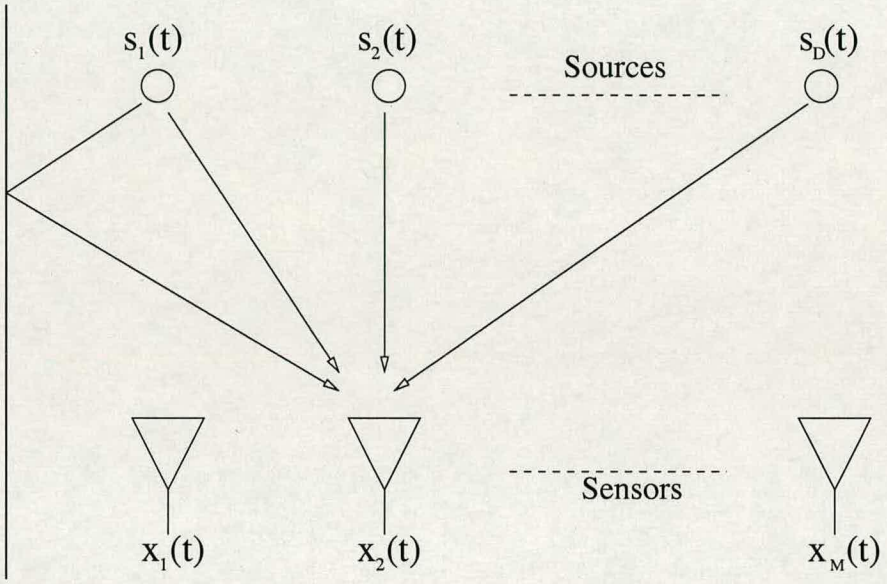


Figure 2.1: *The Basic Blind Source Separation Model*

The basic source separation model shown in Fig. 2.1 deals with recovery of original sources from a finite set of *linear time-invariant instantaneous mixtures*. Assume there are D sources transmitting the signals in the vector form, $\mathbf{s}(t) = [s_1(t), s_2(t), \cdots, s_D(t)]^T$ through the medium (air, cable, network, etc), where $.^T$ denotes the transpose. At different locations of this medium, there are M sensors that capture these observed signals in the vector form, $\mathbf{x}(t) = [x_1(t), x_2(t), \cdots, x_M(t)]^T$. Let us model this simple and easiest case of *linear time-invariant instantaneous mixing* procedure with a matrix operator $\mathbf{A}$.

Then, the observed signals $\mathbf{x}(t)$ in the presence of additive noise $\mathbf{n}(t)$ can be expressed in the matrix-vector representation as:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \tag{2.1}$$

The main aim of BSS system is to reverse this mixing process, given only the set of observed mixtures $\mathbf{x}(t)$, and determine an unmixing (separating) matrix $\mathbf{B}$ such that:

$$\mathbf{y}(t) = \mathbf{B}\mathbf{x}(t) \tag{2.2}$$

where $\mathbf{y}(t) = [y_1(t), y_2(t), \cdots, y_D]^T$ is the separated output signal vector which is an estimate of the original source signal vector $\mathbf{s}(t)$ in the absence of additive noise $\mathbf{n}(t)$. It implies that for source separation, the matrix $\mathbf{A}$ must be invertible. Further, the number of sources must be less than or equal to the number of sensors ($D \leq M$) so that the system is not underdetermined. This is the standard ICA framework used in basic blind source separation problem [14–16]. However, there are two major ambiguities inherent in blind source separation. These are that the order of the estimated sources is indeterminate, and the separated sources are scaled by a nonzero constant as shown in Fig. 2.2.

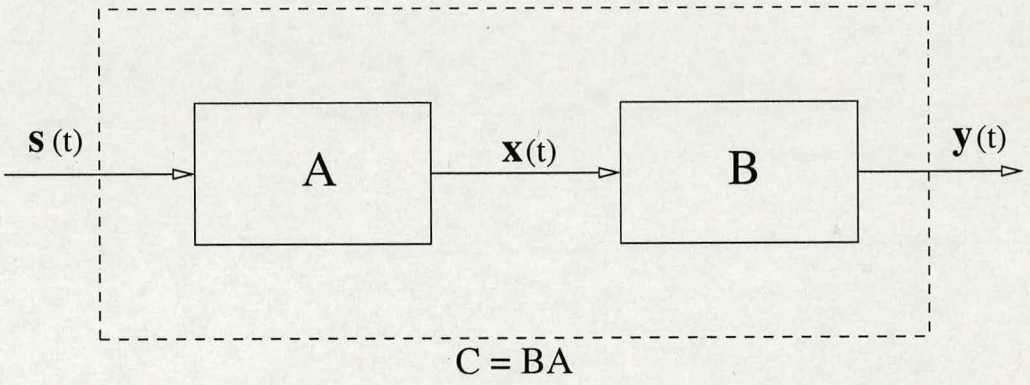


Figure 2.2: *The Principle of Global Separation System (Mixing and Unmixing)*

Thus the product of the unmixing matrix $\mathbf{B}$ and the mixing matrix $\mathbf{A}$ can be referred to as global separation system that will take the following form:

$$\mathbf{C} = \mathbf{B}\mathbf{A} = \mathbf{D}\mathbf{P} \tag{2.3}$$

where $\mathbf{C}$ is the matrix that characterizes the global separation system (mixing and unmixing) that eliminates the emphasis of particular values of the mixing matrix and the scaling factors (without scaling and permutation, $\mathbf{C}$ would converge to the identity matrix $\mathbf{I}$), $\mathbf{D}$ is a non-singular diagonal matrix that accounts for the scaling of each of the separated outputs, and $\mathbf{P}$ is a permutation matrix consists of only one non-zero element per each row and column.

We will now discuss the essentials of two techniques used to perform source separation of instantaneous mixtures: *Principal Component Analysis* (PCA) and *Independent Component Analysis* (ICA). PCA is essentially a *prewhitening or decorrelation* tool, however, not sufficient to perform source separation. On the other hand, ICA can perform source separation assuming much stronger criterion of statistical independence of the source signals.

2.1.1 Principal Component Analysis

Principal Components Analysis (PCA) is a statistical tool used in many applications, such as statistical data analysis, feature extraction and data compression. It is also proposed as a pre-processing tool to enhance the performance of *Gaussian Mixture Models* (GMM) [23]. Its objective is to find a smaller set of variables with less redundancy that would represent the original source signal as accurately as possible [11, 12]. In PCA, the redundancy is measured in terms of correlation between the observed signals. For the rest of the thesis, the first analysis step in PCA will be to remove possible bias (DC offset from microphones in the case of audio signals) from the observed data. This procedure can be referred to as *centering*.

The second analysis step in PCA will be to find the eigenvalues and eigenvectors of the covariance matrix of the centered observations using the *Singular Value Decomposition* (SVD) method [12]. The third analysis step in PCA will be to transform the centered observations into a set of *orthogonal* (decorrelated) signals. This can be achieved by multiplying the centered observations with a matrix containing the eigenvectors.

The final analysis step in PCA will be to transform *orthogonal* signals into *orthonormal* signals (unit variance). This can be done by multiplying the orthogonal signals with a diagonal matrix containing the inverse square root of the corresponding eigenvalues. Thus the components obtained by this PCA method are uncorrelated. Further, it is well known that lack of correlation is not a sufficient criterion for performing blind source separation.

Thus the entire PCA procedure can be summarised, as follows:

1. Find the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_D$ and the eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_D$ of the covariance matrix of the centered observations. Ensure that $\lambda_1, \lambda_2, \cdots, \lambda_D > 0$.
2. Form the matrices $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_D]^T$ and $\mathbf{\Lambda} = \text{diag}[\lambda_1, \lambda_2, \cdots, \lambda_D]$.
3. Apply PCA by $\mathbf{y}_{PCA}(t) = \mathbf{\Lambda}^{-1/2}\mathbf{E}\mathbf{x}(t)$.

2.1.2 Independent Component Analysis

Independent Component Analysis (ICA) was firstly introduced as a concept in the early 1980s by J. Herault and C. Jutten under a different name [24]. Since then many researchers worked on BSS and contributed to this field. However, it was not until P. Comon published a paper describing the essentials of this technique and gave it its final name [14]. ICA has been applied in many diverse fields, as a tool that can separate linearly mixed independent components. *The general ICA framework* assumes linear instantaneous mixtures model, as described in (2.1) and makes the following assumptions:

1. The original source signals $\mathbf{s}(t)$ are assumed to be statistically independent. It implies that: $p(\mathbf{s}(t)) = p(s_1(t), s_2(t), \cdots, s_D(t)) = p(s_1(t))p(s_2(t)), \cdots, p(s_D(t))$.

2. At most one of the independent components can have Gaussian statistics. Since the mixing matrix $\mathbf{A}$ is not identifiable for more than one Gaussian independent components.

In addition, there are certain *ambiguities* that characterise all ICA methods:

1. *ICA cannot determine the exact order of the independent components.* This is also known as the *permutation ambiguity.* In the instantaneous mixtures case, this is not a great problem, it becomes rather serious in other cases.

2. *ICA cannot determine the actual energies (variance) of the independent components.* This is also known as the *scale ambiguity.* As both $\mathbf{A}$ and $\mathbf{s}(t)$ are unknown, any scalar multiplication on $\mathbf{s}(t)$ will be lost in the mixing.

In instantaneous ICA, these ambiguities are not so important, however, we will see that there are some applications, where these ambiguities need to be addressed. As explained earlier, prewhitening by PCA is considered to be the first stage in ICA. Prewhitening orthogonalises the sources present in the mixtures, using the second-order statistics (SOS). However, PCA cannnot separate the sources, as non Gaussian signals are not identifiable using SOS only. The unitary rotation transform needed to separate the decorrelated mixtures is achieved by ICA.

There are many approaches to solve the source separation problem using the general framework of ICA [11]. Some approaches perform separation using *Maximum Likelihood* (ML) estimation to separate the sources, imposing *probabilistic priors* on the sources [25, 26]. Other approaches

try to separate the source signals by *entropy maximisation* that minimises the *mutual information* conveyed by the separated sources [27]. Some other approaches perform separation by estimating the directions of most non Gaussian components (*maximisation of non Gaussianity*) using non Gaussianity measures such as *kurtosis or negentropy* [28, 29]. Another approach for separating the sources is by performing *nonlinear decorrelation* of the observed mixtures [30]. Finally, another approach for source separation is by *tensorial methods* for performing the *joint approximate diagonalization of a cumulant tensor* of the observed mixtures [31, 32].

A lot of literature has been directed towards the simple case of instantaneous mixing; i.e., when the observed signals are a linear combination of the sources and no time delays are involved in the mixing model [14, 27, 33]. The current literature on BSS can be divided into the higher order statistics (HOS) [33–35] and second order statistics (SOS) methods [36–38]. The criterion used most often in the SOS category is minimising the correlation function of the observed signals subject to a constraint on the separating network or the power of the output signals. The main motivation behind the use of SOS methods is that estimating the correlation functions is easier and more robust compared to estimating higher order cumulants, required in most HOS methods. Further, the SOS methods have a simple implementation, require fewer data samples, and unlike HOS methods, they can handle Gaussian distributed inputs.

In recent years a few blind source separation methods have been proposed for instantaneous and convolutive mixing cases by exploiting the nonstationarity of the source signals [39–42]. A nonstationarity assumption can be justified by realizing that most real world signals are inherently nonstationary (e.g., speech or biological signals) [43]. A few blind source separation algorithms have been proposed for the instantaneous mixing case based on reducing the degree of nonstationarity by silence removal techniques that offer notable improvement over the standard separation algorithms [44–46]. A more challenging case is the convolutive mixing; i.e., when the sources are mixed through a linear filtering operation and the observed signals are linear combinations of the sources and their corresponding delayed versions [47–50].

A more difficult practical example of a convolutive BSS problem is separation of audio signals mixed or recorded in a real room with noisy and highly reverberant environment [17, 22, 51, 52]. Unfortunately, the linear instantaneous mixture model is rather incomplete in the case of sound sources recorded in a real room environment with bad acoustics. Previous research showed that the signal captured by the microphones can be well represented by convolution of the sources with FIR filters, modelling the room acoustics between the sources and sensors [53–55].

2.2 BSS of Convolutive Mixtures

As explained in the previous section, different approaches based on the general ICA framework can perform high-quality source separation of linear instantaneous mixtures. However, if we try to apply these techniques on observation signals acquired from microphones in a real room environment with bad acoustics, we will see that all actually fail to separate the audio sources. The main reason is that the instantaneous mixtures model does not hold in the real room recording scenario with highly reverberant environment as shown Fig. 2.3.

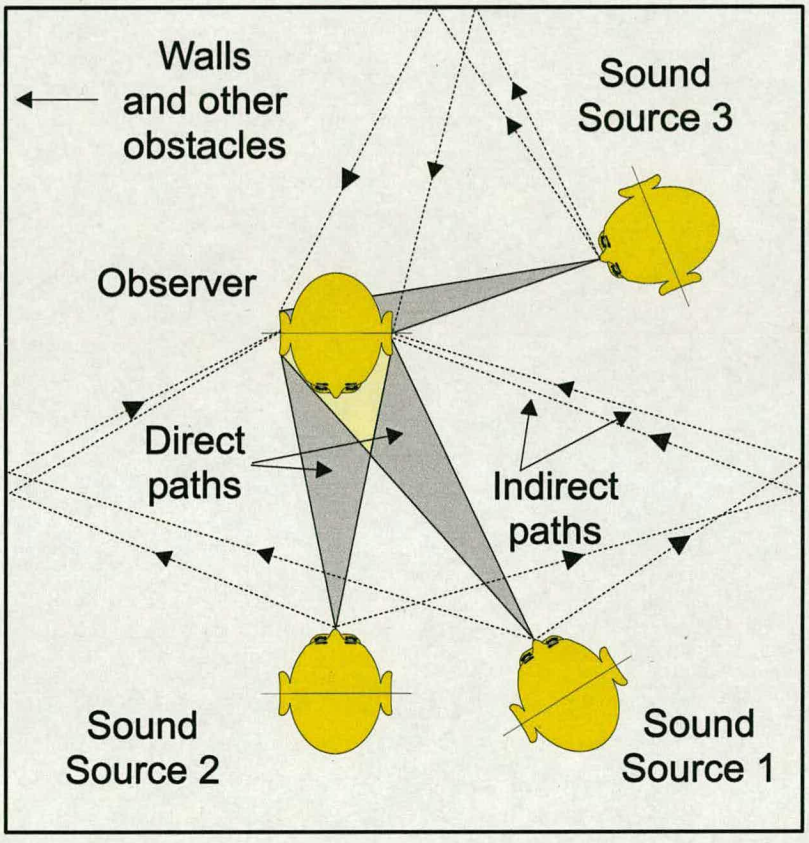


Figure 2.3: *The Real Room Recording Scenario (after [1])*

From Fig. 2.3, it can be clearly seen that in a real recording environment, microphones (sensors) record delayed and attenuated versions of the source signals, apart from direct path signals. This is mainly due to reflections/reverberations on the surfaces inside the room (multipath signals). Therefore, the observation mixtures $\mathbf{x}(t)$ in this real room recording scenario can be more accurately modelled by a nonlinear combination of the signals captured by each microphone so

that we have a Volterra time series as follows:

$$\mathbf{x}(t) = \int_{-\infty}^{\infty} \mathbf{A}_1(\tau_1)\mathbf{s}(t+\tau_1)d\tau_1 + \int_{-\infty}^{\infty} \mathbf{A}_2(\tau_1,\tau_2)\mathbf{s}(t+\tau_1)\mathbf{s}(t+\tau_2)d\tau_1 d\tau_2 + \cdots \quad (2.4)$$

where $\mathbf{A}_i$ is a nonlinear filter operator, which models the reverberation and mixing. In most of the models considered for describing the *cocktail party* problem, it is generally assumed that the propagation of sound is linear. At normal sound pressure level (SPL) the linear FIR is a good approximation to actual room acoustics.

Thus the above Volterra series can be modified to suit the requirements of linear convolutive mixing case modelled by

$$\mathbf{x}(t) = \int_{-\infty}^{\infty} \mathbf{A}(\tau)\mathbf{s}(t+\tau)d\tau \quad (2.5)$$

where $\mathbf{A}$ is a linear time-domain filter operator. Since the BSS algorithms (time and frequency domain) implement FIR filters which are always stable, these FIR filters must be considered for modeling the linear convolutive mixtures. Based on the ICA operating domain, convolutive BSS methods can be generally classified into the time-domain ICA (TDICA) method and the frequency-domain ICA (FDICA) method.

2.2.1 Time-Domain ICA

The linear convolutive mixing described by (2.5) is written in filter matrix and vector form as

$$\mathbf{x}(t) = \mathbf{A} * \mathbf{s}(t) \quad (2.6)$$

where $*$ denotes the convolution operator. Similarly, the separated output is represented by

$$\mathbf{y}(t) = \mathbf{B} * \mathbf{x}(t) \quad (2.7)$$

where $\mathbf{B}$ is the demixing filter matrix.

The scaling and permutation ambiguities still exist due to lack of information about the amplitude and the order of the sources that satisfies

$$\mathbf{C} = \mathbf{B} * \mathbf{A} = \mathbf{PD} \quad (2.8)$$

where $\mathbf{C}$, $\mathbf{P}$ and $\mathbf{D}$ are global separation, permutation and diagonal filter matrices respectively.

Several researchers have proposed methods for solving TDICA. Torkkola proposed a feedback architecture to solve the delay compensation problem of TDICA [49]. He also generalised the feedback architecture to remove temporal dependencies. In a similar sense to Bell-Sejnowski's Infomax principle [27], taking into account Amari's natural gradient approach [25, 56–58], Lee [50] proposed an infinite impulse response (IIR) separation structure, assuming that this structure can only invert minimum phase acoustic room environments (all zeros of the mixing filter system and consequently all poles of the unmixing filter system are inside the unit circle).

There are certain drawbacks in using TDICA methods for the source separation problem. From filter theory, we know that time domain algorithms are very efficient for small mixing filters (communication channels etc), however they can be rather computationally expensive for long room filter transfer functions [59]. The solution of using smaller IIR filters, instead of long FIR filters, may be numerically unstable and the inability to invert nonminimum phase filters [60]. In addition, the problem of spectral whitening introduced by a feedforward architecture, was observed and solved by Torkkola [49] using a feedback architecture, however, it showed there are interdeterminacies in the time-domain ICA methods. All these led researchers to search for a new domain to work on the convolutive mixtures problem.

2.2.2 Frequency-Domain ICA

Several researchers proposed different methods for solving the convolutive mixtures in the frequency-domain. Smaragdis [61], Lee et al [50], Parra and Spence [42] proposed moving to the frequency domain, in order to reduce the computational complexity of the convolution problem. Looking at the FIR feedforward convolutive mixing model, one can use the model based on FIR matrix algebra [53]. From filter theory, it is known that such problems can be addressed with a general multichannel, subband filterbank. However, there are certain motivations (benefits) by choosing a Fourier basis filter bank, i.e. the Fourier transform.

The first motivation is that the source signals become more super Gaussian [18] in the frequency domain, which will be more beneficial for any ICA learning algorithm. The second motivation is that by applying the fast Fourier transform (FFT), we can approximate the time-domain convolutive mixing problem into multiple, complex instantaneous mixing problems in the frequency-domain. As a result, the time domain convolutive mixing problem with a large number of estimated parameters is decomposed into multiple but complex instantaneous mixing problems, each with a small number of parameters to be estimated.

The short time fast Fourier Transform (STFFT) is used instead of the FFT, in order to divide the signal into shorter overlapping frames and preserve the signal's stationarity [60, 61]. Using the STFFT and assuming statistical independence between frequency bins, we have transformed a convolutional problem into several instantaneous mixtures problems, i.e. an instantaneous mixtures problem for each frequency bin. In order to transform the convolution into multiplication, one has to use windows larger than the maximum length of the transfer functions. Hence, we can use the very well established theory of instantaneous mixtures to solve this problem. However, this case is not as simple as the general ICA framework used for the source separation of linear instantaneous mixtures due to the following reasons:

1. As the dataset in this case are instantaneous mixtures of complex numbers, we have to ensure the stability and convergence of the original algorithms with complex data.
2. The *scale and permutation ambiguity*, which had negligible effect in the linear instantaneous mixtures case, now plays a very important role in this FDICA approach.

2.2.2.1 Signal Model

Let us consider the convolutive mixing case when there are D sound sources in the environment. By observing this sound field with M microphones and taking the STFFT of the microphone inputs, the convolutive mixing problem is reduced to complex but multiple instantaneous mixing problems. Therefore, the observed input signal vector in the frequency-domain $\mathbf{x}(\omega, t)$ is

$$\mathbf{x}(\omega, t) = [X_1(\omega, t), ..., X_M(\omega, t)]^T. \tag{2.9}$$

Where $\mathbf{X}_m(\omega, t)$ is the STFFT of the input signal $\mathbf{x}(t)$ in the tth time frame at the mth microphone. The symbol $.^T$ denotes the transpose. Further, $\mathbf{x}(\omega, t)$ is modeled as [62]

$$\mathbf{x}(\omega, t) = \overbrace{\mathbf{A}(\omega)\mathbf{s}(\omega, t)}^{Direct} + \overbrace{\mathbf{A}_{er}(\omega)\mathbf{s}(\omega, t)}^{Early-Reflection} + \overbrace{\mathbf{n}(\omega, t)}^{Reverb.}. \tag{2.10}$$

The first term in (2.10), $\mathbf{A}(\omega)\mathbf{s}(\omega, t)$, express the *directional* components. The second term, $\mathbf{A}_{er}(\omega)\mathbf{s}(\omega, t)$, denotes the *early reflection*, which is defined as a portion of the reflections whose delay relative to the direct sound is within the window length of the STFFT. Finally, the third term, $\mathbf{n}(\omega, t)$, will be considered as the *reverberation* which is a mixture of less directional components, that includes the room reflections and ambient noise.

Matrix $\mathbf{A}(\omega)$ is termed the mixing matrix, its (m,n) element, $A_{m,n}(\omega)$, being the transfer function of the direct path from the nth source to the mth microphone as

$$A_{m,n}(\omega) = H_{m,n}(\omega)e^{-j\omega\tau_{m,n}}. \tag{2.11}$$

Where $H_{m,n}(\omega)$ is the magnitude of the transfer function and $\tau_{m,n}$ is the propagation time from the nth source to the mth microphone. Vector $\mathbf{s}(\omega,t)$ consists of the source spectra as

$$\mathbf{s}(\omega,t) = [S_1(\omega,t),...,S_D(\omega,t)]^T \tag{2.12}$$

where $S_n(\omega,t)$ denotes the spectrum of the source. Thus, $\mathbf{A}(\omega)\mathbf{s}(\omega,t)$, has only the directional components. On the other hand, $\mathbf{A}_{er}(\omega)\mathbf{s}(\omega,t)$ has less directional components due to multiple reflection paths. Therefore, the element of $\mathbf{A}_{er}(\omega)$ will take the following form

$$A^{er}_{m,n}(\omega) = \sum_i H^{er}_{m,n,i}(\omega)e^{-j\omega\tau_{m,n,i}} \tag{2.13}$$

where the subscript i denotes the path number. Since $\mathbf{A}_{er}(\omega)\mathbf{s}(\omega,t)$ is a filtered replica of $\mathbf{s}(\omega,t)$, it is highly correlated with the direct sound $\mathbf{A}(\omega)\mathbf{s}(\omega,t)$.

We are treating the rest of the reflections with a delay greater than the window length of the STFFT as *reverberation*. Based on this definition, the reverberation term, $\mathbf{n}(\omega,t)$, can be expressed as

$$\mathbf{n}(\omega,t) = \sum_d \mathbf{A}_r(\omega,d)\mathbf{s}(\omega,d) \tag{2.14}$$

where d is the delay associated with the mixing matrix for the reverberation term $\mathbf{A}_r(\omega,d)$.

Hence, $\mathbf{n}(\omega,t)$ consists of the filtered replica of the signal in the previous frames and, thus, has small or zero coherence with the direct sound and the early reflection. A typical example of a situation with small coherence is a consonant frame overlapped by the reflection of a previous vowel. Therefore, $\mathbf{n}(\omega,t)$ functions rather as random additive noise. Also, since $\mathbf{n}(\omega,t)$ includes a large number of reflections, its directivity and, hence, the coherency between the microphones is assumed to be low. Further, early reflections can be classified into weak early reflections and the strong early reflections depend upon the strength of the reflections that are coming from the walls and other obstacles of the acoustic room. Generally, weak early reflections are due to the soft nature of the walls and other obstacles. On the other hand, strong early reflections are generated by the hard walls, tables etc., of the room.

In most of the practical applications, weak early reflections are considered as an integral part of the reverberation term $\mathbf{n}(\omega, t)$. This implies that the impact of weak early reflections on the observed input signal spectrum is very small. So, $\mathbf{x}(\omega, t)$ can be written as

$$\mathbf{x}(\omega, t) \simeq \overbrace{\mathbf{A}(\omega)\mathbf{s}(\omega, t)}^{Direct} + \overbrace{\mathbf{n}(\omega, t)}^{Reverb.}. \tag{2.15}$$

On the other hand, the presence of strong early reflections will definitely affect the overall mixing system $\mathbf{A}'(\omega)$ described by the following model

$$\mathbf{x}(\omega, t) = \mathbf{A}'(\omega)\mathbf{s}(\omega, t) + \mathbf{n}(\omega, t). \tag{2.16}$$

Where $\mathbf{A}'(\omega) = \mathbf{A}(\omega) + \mathbf{A}_{er}(\omega)$. In other words, the unmixing system $\mathbf{B}(\omega)$, obtained by FDICA learning algorithm, has not only the direct sound but also the early reflection.

2.2.2.2 Lee et al's Approach of FDICA

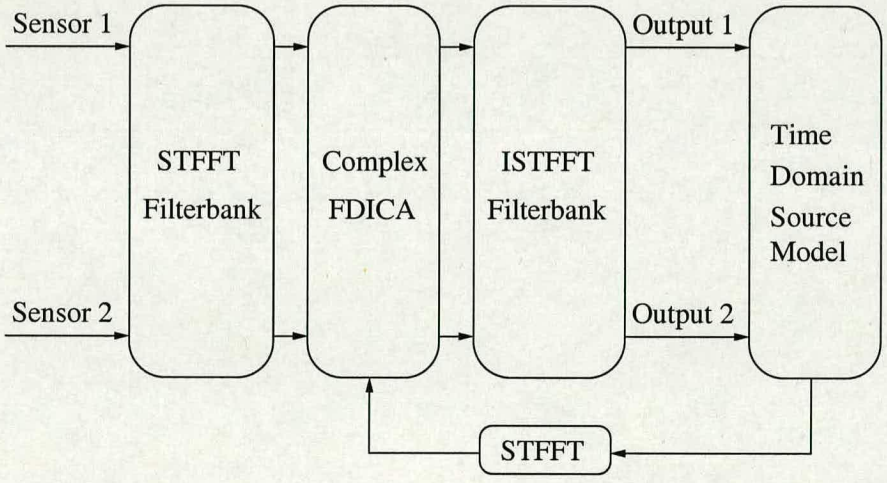


Figure 2.4: *FDICA Method Proposed by Lee et al; Unmixing in the frequency-domain*

Based on the time-domain ICA approach, Lee at al [50] argued that a FIR unmixing structure would be more beneficial in the audio source separation case, mainly because real room acoustics usually involve non-minimum phase mixing (zeros outside the unit circle). In addition, they proposed a FDICA method shown in Fig. 2.4 for unmixing the sources in the frequency-domain, in order to avoid the convolution in the time-domain.

In their approach, the source signals are modelled by the *sigmoid tanh nonlinearity function* in the time-domain. There is a benefit from imposing time-domain source models: the permutation problem does not seem to exist. When we apply the source model in the time-domain, we do not assume that the signals are statistically independent along each frequency bin. As a result, the permutations are coupled due to the source model applied to the whole signal and not to its independent decompositions. However, there is evidence reported that problems similar to the permutation problem do exist [63]. This method is computationally expensive, due to the mapping back and forth between the frequency and time domains and do not take advantage of the strong non Gaussianity in the frequency domain.

2.2.2.3 Smaragdis's Approach of FDICA

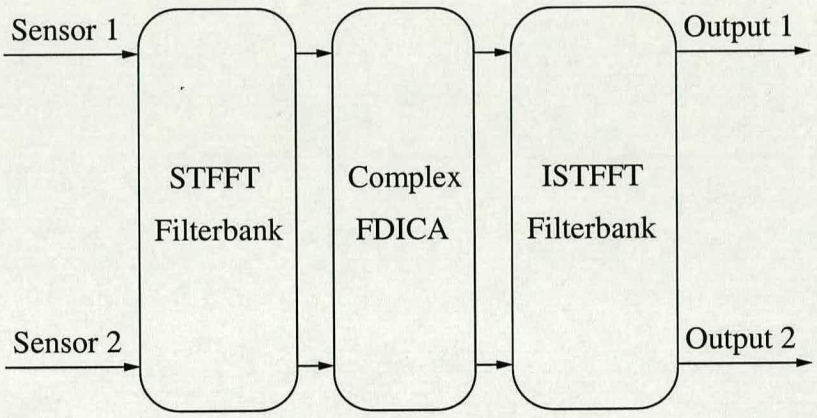


Figure 2.5: *Smaragdis's FDICA: Unmixing and Source Modeling in the frequency-domain*

As shown in Fig. 2.5, Smaragdis [61] worked solely in the frequency domain for the convolutive mixing problem, i.e. he performs both the unmixing and the source modelling in the frequency domain, in order to avoid the extra complexity of moving from the frequency to the time domain and vice versa. Therefore, the system is adapting solely in the frequency domain, independently for each frequency bin. Further, it is well known that the statistical properties of an audio (e.g. speech) signal over short quasistationary periods in the time-domain (frames of the STFFT) are not always well modelled as super Gaussian. On the other hand, the statistical properties of these speech segments in the frequency domain can be better modelled as super Gaussian as these sections have very heavy tailed distributions. This implies that the frequency domain is a better candidate for source modelling.

This will provide a better achievable performance, since as noted by various researchers, the Cramer-Rao bound (the performance bound for an estimator) for the estimation of the unmixing matrix in FDICA algorithms is related to how close the source distributions are to Gaussian. That is that the more non Gaussian the distributions are, the better the achievable performance of the FDICA algorithm. Further, Smaragdis observed that the *sigmoid tanh nonlinearity function* cannot be applied for complex data under any circumstances, as it has singularities for certain values of the separated signal given by

$$y(\omega,t) = j\pi\,(k+1/2)\,, \quad \text{for} \quad k = 0,1,2,\cdots\,. \tag{2.17}$$

These singularities can cause instability to the natural gradient rule. As a result, Smaragdis proposed the following split-complex *sigmoid tanh nonlinearity function* which is smooth, bounded and differentiable in the complex domain.

$$\varphi(y(\omega,t)) = \tanh(\Re(y(\omega,t))) + j\tanh(\Im(y(\omega,t))). \tag{2.18}$$

Where $\varphi(y(\omega,t))$ is the nonlinear score function or activation function for the complex data.

The two typical FDICA approaches, proposed by Lee et al [50] and Smaragdis [61], have considered Bell-Sejnowski's Infomax principle [27] for solving the convolutive mixing problem by taking Amari's natural gradient learning [25] into account. In general, the FDICA methods based on the natural gradient algorithm have a simpler implementation and better convergence properties with respect to their time domain counterparts when the mixing is reverberant [64]. On the whole, the FDICA framework proposed by Smaragdis using the natural gradient algorithm seems to be a robust, general solution for the convolutive mixtures problem. Therefore, several researchers considered this framework using the natural gradient algorithm (based on the complex Infomax principle) for solving the cocktail party problem.

Several algorithms have been developed for FDICA using HOS. Some have used the natural gradient algorithm and others have used the fixed point solutions for FDICA [65–67]. Nishikawa et al [68–71] proposed a multi-stage ICA (MSICA) algorithm in which FDICA and TDICA methods are combined to achieve a superior performance of BSS under reverberant conditions. Further, they developed a new algorithm for overdetermined BSS of linear convolutive mixtures based on MSICA [72]. Recently Mitianoudis and Davies proposed a complex fixed point (Fast FDICA) algorithm for the separation of audio sources [65, 66, 73].

A few BSS methods based on SOS were proposed that exploit the nonwhiteness property by simultaneous diagonalization of output correlation matrices over multiple time-lags [74, 75]. Other SOS methods were proposed that exploit the nonstationarity property of the sources by simultaneous diagonalization of short-time output correlation matrices at different time intervals [20, 76–79]. Parra and Spence [42] considered coloured nonstationary signals. Rahbar and Reilly [80, 81] proposed a new algorithm based on the joint diagonalization using alternating least squares (ALS) optimization methods. Recently, Ikram and Morgan [82] proposed a multistage frequency-domain (MSFD) algorithm based on SOS for blind separation of speech signals in a reverberant environment. Very recently Wang et al [83, 84] proposed a robust and faster converging FDICA algorithm using a penalty function-based joint diagonalization approach by explicitly exploiting the second-order nonstationarity of the sources and obtained a better performance in terms of shape preservation and amplitude ambiguity reduction.

Based on the above discussion, we have considered two different FDICA approaches based on SOS and HOS respectively for solving the cocktail party problem in this thesis. The first FDICA approach is based on the multiple time-delayed (lagged) decorrelation algorithm that exploits the second order statistics of the source signals [75]. The second FDICA approach is based on the complex Infomax algorithm (taking Amari's natural gradient learning rule into account) that exploits the higher order statistics of the sources [27, 64, 85]. As we are solving the BSS problem for multiple, complex linear instantaneous mixtures based on the general ICA framework, we cannot avoid the inherent arbitrary *scaling and the permutation ambiguities* of the estimated frequency response of the unmixing system at each frequency bin.

The *scaling ambiguity* means that the scaling for each frequency bin can be different and this leads to a spectral deformations of the separated signals. In addition, it is not guaranteed that the scaling will be uniformly distorted with frequency, changing the signal envelope after separation. On the other hand, the permutation ambiguity in the frequency-domain ICA is much more difficult and complicated problem than the permutation (ordering) ambiguity in the general ICA, since the ordering of the sources should remain the same along the frequency axis. As a result, the FDICA algorithm produces different permutations of separated sources along the frequency axis, and therefore the separated output signals remains mixed. Hence, it is essential to keep the same permutation to avoid the mixed frequency content for each frequency bin of the separated signals. Many solutions have been proposed for solving the scaling and permutation problem and these will be discussed in the next subsection.

2.2.3 Solutions for Scaling and Permutation Ambiguity

2.2.3.1 Solutions for Scaling Problem

Several methods have been proposed for solving the scaling ambiguity problem of FDICA. Smaragdis [61] solved the scaling ambiguity problem by the normalization of the unmixing matrices i.e. applying a *constraint on the unmixing matrix.* This normalization procedure helps the convergence of the algorithm by preventing overshooting. Parra and Spence [42] constrained the diagonal elements of the unmixing matrix to unity and thereby constrained the scaling of the unmixing matrix in similar manner to the method proposed by Smaragdis. Another approach would be to constrain the variance of the signal. In the frequency domain framework, the signal will have different signal levels at each frequency bin. The unmixing matrix updates are calculated for each signal frame at each frequency bin. Thus, different energy levels may lead the unmixing matrix to different scaling. Normalising the signal to unit variance can enforce uniform scaling of the unmixing matrix along frequency axis.

Cardoso [86] proposed a valid approach to solve the scaling ambiguity by *mapping the separated sources back to the observation space*, i.e. the sensor space. In this method, Cardoso explained that instead of focusing on the columns of the mixing matrix, we can focus on the observation spaces containing each component and then we can get the same separation result, without the ambiguity of scale (sign and magnitude). In other words, by mapping the separated sources back to the observation space of the microphones, we can undo an arbitrary scaling deformation, performed by the unmixing matrix. Recently Murata et al [87] proposed a method similar to that of Cardoso for correcting the scaling ambiguity problem, in which the separated output is filtered by the inverse of the unmixing (separation) filter and showed good performance of blind separation of acoustic signals in a reflective/reverberant environment.

Based on the above discussion, we have considered a method proposed by Murata et al [87] for solving the scaling ambiguity problem of FDICA when the mixing is highly reverberant. The argument proposed by Murata et al [87] can be well supported mathematically with the following simple analysis. At first, let us assume that the permutation ambiguity problem is sorted completely. The 2×2 case will be used for simplicity, but it is straightforward to generalise the analysis to the $D \times D$ case (where D is the number of sources). By using the unmixing matrix $\mathbf{B}(\omega)$ and its inverse matrix $\mathbf{B}^{-1}(\omega)$, the observed signals in the frequency-domain $\mathbf{x}(\omega, t)$ are decomposed such that the decomposed components are mutually independent.

The physical meaning of each component is a signal vector generated by one independent component which is observed on sensors.

$$\begin{aligned}
\mathbf{x}(\omega,t) &= \mathbf{B}^{-1}(\omega)\mathbf{B}(\omega)\mathbf{x}(\omega,t) && (2.19)\\
&= \mathbf{B}^{-1}(\omega)\mathbf{I}\mathbf{B}(\omega)\mathbf{x}(\omega,t) && (2.20)\\
&= \mathbf{B}^{-1}(\omega)(\mathbf{E}_1+\cdots+\mathbf{E}_n)\mathbf{B}(\omega)\mathbf{x}(\omega,t) && (2.21)\\
&= \mathbf{B}^{-1}(\omega)\mathbf{E}_1\mathbf{B}(\omega)\mathbf{x}(\omega,t)+\cdots+\mathbf{B}^{-1}(\omega)\mathbf{E}_n\mathbf{B}(\omega)\mathbf{x}(\omega,t) && (2.22)\\
&= \sum_{i=1}^{n}\mathbf{B}^{-1}(\omega)\mathbf{E}_i\mathbf{B}(\omega)\mathbf{x}(\omega,t) && (2.23)
\end{aligned}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{E}_i$ is a matrix with 1 for the i th diagonal element and 0 for the other elements and satisfy $\mathbf{E}_1+\cdots+\mathbf{E}_n=\mathbf{I}$.

Therefore, the problem of amplitude ambiguity can be solved by putting back the separated independent components to the sensor input with the inverse of the demixing matrix $\mathbf{B}(\omega)$. The nth component of $\mathbf{y}(\omega,t), y_n(\omega,t)$ is filtered by $\mathbf{B}^{-1}(\omega)$ separately as

$$\widetilde{\mathbf{y}}_n(\omega,t)=\mathbf{B}^{-1}(\omega)[0,\cdots,0,y_n(\omega,t),0,\cdots,0]^T \tag{2.24}$$

where $\widetilde{\mathbf{y}}_n(\omega,t)=[\widetilde{y}_{1,n}(\omega,t),\cdots,\widetilde{y}_{M,n}(\omega,t)]^T$ and $\widetilde{y}_{\widetilde{m},n}(\omega,t)$ corresponds to the recovered signal of the nth source observed at the $\widetilde{m}$th arbitrary microphone.

Eq. (2.24) can be written in the matrix-vector notation as

$$\widetilde{\mathbf{y}}(\omega,t)=\widetilde{\mathbf{B}}_{\widetilde{m}}^{-1}(\omega)\mathbf{y}(\omega,t) \tag{2.25}$$

where $\widetilde{\mathbf{y}}(\omega,t)=[\widetilde{y}_{\widetilde{m},1},\cdots,\widetilde{y}_{\widetilde{m},D}]^T$ and the arbitrary scaling matrix $\widetilde{\mathbf{B}}_{\widetilde{m}}^{-1}(\omega)$ is defined as

$$\widetilde{\mathbf{B}}_{\widetilde{m}}^{-1}(\omega)=\mathrm{diag}[B_{\widetilde{m},1}^{-1},\cdots,B_{\widetilde{m},D}^{-1}]. \tag{2.26}$$

Based on the above analysis, we can easily remove the arbitrary scaling ambiguity by projecting the signals back to the microphone's space and still have the unmixed signal characteristics. Similarly, we can prove that this scheme can also remove the arbitrary scaling ambiguity, even when the permutation ambiguity problem is not sorted.

Thus, moving (mapping) the separated signals back to the microphones' (observation) space, we can undo or avoid the so-called scaling ambiguity problem of FDICA.

2.2.3.2 Solutions for Permutation Problem

Solving the convolutive mixtures in the frequency domain, independently for each frequency bin generates the permutation ambiguity problem and thus the separated signals remain mixed. In order to solve this permutation problem, we need to impose some sort of coupling between the independent unmixing algorithms, so that they converge to the same order of sources. Based on different approaches, many soultions have been proposed to solve the permutation ambiguity problem. Some methods based on integration of computational auditory scene analysis (CASA) and BSS techniques have been proposed to solve the permutation problem. In general, these frequency-domain permutation solving strategies can be classified into: *the source modelling, the channel modelling and the hybrid modelling solutions.*

(i) Source Modelling Solutions

In source modelling solutions, the main aim is to exploit the coherence (continuity) and the information between frequency bands, in order to identify the correct alignment between the subbands. In fact, audio signals can rarely be considered independent between frequency bands due to the audio structure (harmonics and temporal) in both music and speech. As a result, any clustering rule that can group similar objects will align the permutations. Murata et al [87] have exploited the nonstationarity of source signals using SOS in order for solving the permutation problem. They have developed a method using the correlation between the spectral envelopes at different frequencies (denoted as inter frequency spectral envelope correlation (IFSEC)), but has been reported to sometimes fail when the input signals have similar envelopes [88].

Recently Mitianoudis and Davies [73, 89] solved the permutation problem using the likelihood ratio jump (probabilistically justified) with two fixed point fast FDICA methods while exploiting the time-frequency spectral envelope, but works only in batch mode and becomes more complicated for more than two sources. In a similar effort, Rahbar and Reilly [80, 81] developed an efficient diadic algorithm to resolve the frequency dependent permutation ambiguities to an arbitrary number of sources while exploiting the inherent nonstationarity of the sources. Very recently Hu et al [90, 91] proposed a new method which is similar but different from that of Murata et al [87]. They assumed that there exists the continuity in power between the waveforms of adjacent frequency components of same source. Further, they used the distance between the signals at adjacent frequencies to align the separated signals while implicitly utilizing the information of inter frequency correlation for a better performance.

(ii) Channel Modelling Solutions

In channel modelling solutions, the main objective is to exploit additional information about the room transfer functions, in order to select the correct permutations. These room transfer functions have certain properties. In source separation, we usually employ long FIR filters to estimate the room transfer functions, as their stability is guaranteed. In addition, most room transfer functions have a dominant first delay (direct path) term that can be used to identify the angular position of each source signal to the sensor array.

Smaragdis [61] proposed a method in which some heuristic coupling is applied between adjacent frequency bins in order to solve the permutation problem. This method forces the separation matrices at each frequency bin to have a similar permutation using an influence factor. However, it had limited effect, as it has been reported to fail in several cases [18]. Recently Asano et al [9, 10] proposed a combined approach of array processing (using the subspace method) and ICA for solving the permutation problem by utilizing the coherency of the mixing matrices in several adjacent frequencies (denoted as inter frequency coherency (IFC)). Further, they reported a low permutation error rate in the frequency range over 1 kHz and very high permutation error rate below 1 kHz when the mixing is highly reverberant.

In recent years a few BSS methods based on SOS have been proposed for solving the permutation ambiguity problem while exploiting the nonstationarity of the source signals. Parra and Spence [42] have also exploited nonstationarity of the signals to perform source separation in the frequency-domain. Their solution to the problem was to impose a constraint on the unmixing filter length. In other words, it imposes a smooth constraint on the unmixing filters, as they are modelled as FIR filters. Again mixed success has been reported for this method, as it became trapped in local minima [92]. Very recently Serviere and Pham [93] proposed a novel technique to solve the permutation problem of FDICA. They exploited the nonstationarity and the presence of pauses in the speech signals and thereby separating the convolutive mixtures with fairly long impulse responses containing strong echoes.

Recently, the relationship between convolutive BSS and beamforming has been highlighted. In the context of FDICA, at a given frequency bin, the unmixing matrix can be interpreted as a nullsteering beamformer that uses a blind algorithm (ICA) to place nulls on the interfering sources. Based on this beamforming, channel modelling solutions have been proposed to align the permutations along the frequency axis. All BSS methods make no assumptions about the

position of the sound sources in the 3D space. However, beamforming estimates the directions of signal's arrival (DOA) in order to steer the beam of an array of sensors to target a specific source. An additional geometrical information employed by the beamforming technique is the sensors' configuration, which is assumed to be fixed for a particular experiment.

Saruwatari et al [94–97] estimated the DOA by taking the statistics with respect to the direction of the nulls in all frequency bins and then tried to align the permutations by grouping the nulls that exist in the same DOA neighbourhood. On the other hand, Ikram and Morgan [98] proposed to estimate the sources DOA in the lower frequencies, as they don't contain multiple nulls. Parra and Alvino [63] used more sensors than sources along with known source locations and added this information as a geometric constraint to their unmixing algorithm.

Very recently Ikram and Morgan [99] proposed a permutation alignment scheme based on microphone array directivity patterns. After properly aligning the permutations, they showed that the blind speech separation method outperforms the nonblind beamformer in a highly reverberant environment. Furthermore, by exploring the tradeoff between the permutation alignment and the spectral resolution of the unmixing filters, they proposed a multistage frequency-domain (MSFD) algorithm [82] for aligning the permutations of the unmixing filters without sacrificing the spectral resolution and obtained a better performance than the single-stage system.

(iii) Hybrid Modelling Solutions

Each of the above source and the channel modelling solutions has different characteristics, and may perform well under certain specific conditions but not under all conditions. Researchers around the world have considered hybrid modelling approaches by integrating some of the source and the channel modelling solutions in order to improve the separation performance of FDICA system. Ciaramella et al [100] proposed a method based on combined approaches of cross correlation of the separated signals and the frequency coupling of unmixing matrices at each frequency bin to have a similar permutation.

Recently Sawada et al [101, 102] proposed a robust and precise method based on the combined approach of direction of arrival (DOA) estimation using beamforming and inter frequency correlation of output spectral envelopes (IFSEC) for solving the permutation ambiguity problem. Furthermore, by utilizing the harmonic structure of signals, they proposed a more robust and precise method even for low frequencies where DOA estimation is inaccurate and thereby solving the permutation ambiguity problem almost perfectly [103].

Based on this robust and precise idea, Wang et al [104] proposed a novel hybrid method based on the combined approaches of exploiting spectral continuity (performing filter constraint), exploiting the time envelope structure and beamforming alignment for solving the permutation ambiguity problem. Their method is based on a robust and faster converging FDICA algorithm using a penalty function-based joint diagonalization approach by exploiting the second-order nonstationarity of the signals [83]. However, they have used the subspace method instead of conventional beamforming technique for the accurate estimation of DOAs of the source components and the frequency dependent distance for the correlation of time envelopes.

Furthermore, they have used the psychoacoustic model as a postprocessing filter for the misaligned permutations unable to be sorted out by the aforementioned combined approaches. This psychoacoustic model exploits the two properties of the human auditory system: absolute threshold of hearing (ATH) (also known as threshold in quiet) and auditory masking (AM). The human auditory system and the relevant psychoacoustic models used for exploting the perceptual irrelevancy of the separated signals will be discussed in the following section.

In the recent years some researchers have started working on methods in which two powerful techniques namely computational auditory scene analysis (CASA) and blind source separation (BSS) were integrated and thereby solving the convolutive mixtures problem in a noisy and highly reverberant environment. Rutkowski et al [105, 106] proposed a novel biologically plausible model for segregation of one dominant speaker from the other concurrent speakers and environmental noise in real cocktail party scenario. However, they have used the gammatone filterbank for bandpass preprocessing. Further, they reported that computer simulation results showed good performance under noisy and highly reverberant environment.

Recently Barros et al [107–110] proposed more biologically plausible algorithm for extracting one speaker signal out of a mixture of reverberated sounds by mimicking some properties of the human auditory system: 1) *mimicking the inner ear through the use of a bank of self-adaptive band-pass wavelet filters*; 2) *tracking of speech fundamental frequency* f_0; 3) *separating the mixed signals in each subband through an FDICA algorithm based on SOS which tracks the signals related to the voice pitch (*f_0 *and its harmonics) during learning*; and 4) *mimicking the temporal masking characteristic of the auditory system by a switch which is one for the voiced part and zero for the silent (unvoiced) part of the speech.* However, they have avoided the permutation problem by focusing the attention on only one speaker signal.

2.3 Human Auditory System

In the pursuit of the best possible performance for a convolutive BSS system, it should be noted that the separated speech is judged by the human ear according to the perception of the sound. Consequently, it is very advantageous to develop perceptually motivated BSS system that exploits the irrelevancy associated with the observed speech in advance before separating the signals by ICA. This creates the need for the modelling of the human auditory system. The human auditory system consists of the ear, auditory nerve fibres, and a section of the brain. The ear is the outer peripheral system, which converts sound waves into electrical impulses that are picked up by the auditory nerve. Even a sound wave that has a magnitude of only a ten millionth part of the atmospheric pressure is sufficient to be perceived by the human ear.

The energy of the sound waves captured by the ear cavity travels to the inner ear and goes through frequency analysis which is generally assumed to be affected by very sophisticated active and nonlinear processes. In the final conception of the sound, cognitive effects also play a role. During the last three decades, remarkable progress has been made within the research of the human hearing but many details especially the higher order brain functions in connection with information processing still need to be explained. The field of psychoacoustics has an interesting standpoint in the research since it examines directly the relationships between acoustic stimuli and the associated sensations. The results of the psychoacoustic research form the basis of the work presented in this thesis.

In the study of the human auditory system, it is a normal practice to make a distinction between the peripheral part, and the part that contains the nervous system and leads to the final auditory sensation. The peripheral part of the human auditory system refers to the elements in which the oscillations due to a sound stimulus retain their original character. Zwicker [111] designates the function of the peripheral system as preprocessing of sound. In contrast, the neural processing takes place in the second region of the hearing system that consists of the auditory sensation area of the brain together with the nerve fibres. However, the nervous system and its functions are beyond the scope of this thesis and they are not discussed here. Rather, emphasis is placed on the structure and operation of the peripheral part, as the aim is to introduce the properties influencing the perception of sounds. This will serve as a basis for the psychoacoustic models, some of which will be presented with the theory by giving detailed descriptions of the two most essential psychoacoustic models considered in this research work and by also providing a view on some other existing models in the following section.

The frequency analysis performed by the ear has a certain finite resolution, leading to the basic concept of masking. It is a phenomenon in which one sound can drown out another sound either partially or totally. The relative levels and frequencies of these sounds determine for the most part the degree of masking, but temporal factors also have some influence. The quantitative effect of masking can be depicted by means of a masking threshold. It shows the sound pressure level that a test tone must have in order to be just audible in the presence of a masker. The masking threshold is one of the main concepts in this thesis and it is viewed more precisely and analytically in the next section.

2.3.1 Structure of Human Ear

The purpose of the human ear is to capture sound waves and to convert the acoustic energy of these small pressure fluctuations into electrical nerve impulses. The nerve fibres convey the information to the brain in which it is perceived as sounds. Reciprocally, the brain sends information to the ear, thus actively controlling some of the functions of the so-called sound preprocessing [2, 112]. The ear also contains the vestibular organ that contributes to balancing the body, but it has no effect in the perception of sounds. A simplified structure of the human ear can be described by three parts: *the outer, middle and inner ear.*

2.3.1.1 Outer Ear

The outer ear is composed of the pinna (the visible part, ear cavity), the meatus (auditory canal) and the tympanic membrane (eardrum). The pinna collects the sound energy which then travels down the meatus and makes the eardrum vibrate. The eardrum is a hermetic membrane whose function is to convey the acoustical energy to the middle ear. The pinna and the outer ear canal have a strong influence on the incoming sound. The pinna filters the sound in a way that depends on its inlet angle, thus providing cues to the localisation of sound [113]. This works especially at high frequencies where the shadowing effect of the pinna attenuates the sounds that come from behind the listener. At low frequencies, this does not take effect because the wavelengths are too large compared to the dimensions of the pinna. The meatus, acting like an open pipe with a length of approximately 2 cm, has a resonant frequency at about 4 kHz. Consequently, the meatus is responsible for the high sensitivity of hearing around this frequency. In addition to the sound wave modifications performed by the pinna and the meatus, the head and shoulders of the subject have the effect of shadowing and reflection.

2.3.1.2 Middle Ear

The middle ear is a chamber that contains the auditory ossicles: malleus (hammer), incus (anvil) and stapes (stirrup), the smallest bones in the human body. The middle ear provides the two important functions of impedance transformation and amplitude limiting to ensure the efficient transfer of the acoustical energy, avoiding large reflections. The impedance transformation is based on both the lever system constructed from the ossicles and the ratio of the area of the eardrum to that of the small oval window. The amplitude limiting is made possible by the tiny inner ear muscles that are attached to the auditory ossicles. When the subject is exposed to very intense sound pressure level (SPL) above 85-90 dB, these muscles contract and thereby limit the transmission of sound through the ossicles (the ossicles act as a low pass filter with a cutoff frequency of around 1 kHz). This operation, called the middle ear reflex, may help to protect the vulnerable structure of the inner ear. The reflex also decreases the audibility of self-generated sounds by getting activated at the starting time of vocalisation [2, 111, 114].

2.3.1.3 Inner Ear

The inner ear, comprising the cochlea and the semi-circular canals of the vestibular organ, is the most complicated part of the ear. The vestibula is the organ that helps balance the body with no apparent role in the hearing process. The cochlea is the most dominant organ in the physiology of the mammalian ear. The shape of the cochlea resembles a snail and it is filled with nearly incompressible fluids and surrounded by very hard bone. Uncoiled it measures about 32 mm in length. It is mechanically connected to vibrating parts of the ear and it is responsible for the transduction of physical energy to electrical impulses to be detected by the auditory nerve. Inside it lies the basilar membrane floating in the cochlear fluids.

The basilar membrane extends throughout the length of the cochlea, starting out as being narrow at the beginning and gradually becoming three to four times wider at the other end. As vibrations caused by incoming sounds excite the basilar membrane, it tends to resonate with the higher frequencies near its base at the beginning of the cochlea, while progressively lower frequencies create displacements towards its apex. The cochlea thus performs the very important function of analysing the frequency content of the incoming sound. The frequency of the stimulus that causes maximum response at a given point on basilar membrane is called the characteristic frequency for that point. However, the situation gets somewhat complicated with other than pure sinusoidal signals.

If two frequency components of the stimulus are sufficiently close to each other, basilar membrane fails in the exact frequency-to-place conversion and only a single, broader maximum can be observed in its response. These vibrations of the basilar membrane are detected by a series of hair cells inside the cochlea, that upon stimulation release chemical transmitters through a connection with the nervous system and cause neural pulses in proportion to the detected activity. There are two different kinds of hair cells, the inner hair cells (IHC) and outer hair cells (OHC), both having their own special functions. It seems that IHC convey most, or even all, information about the sounds to the brain. OHC receive messages from the brain through several descending nerve fibres. These messages are most likely used for active processes affecting the high sensitivity and sharp tuning of basilar membrane.

Computational Modelling of Cochlea

As the cochlear function suggests, it performs a decomposition very similar to a harmonic analysis (time-frequency analysis). This observation was noted by many researchers who have worked on computational audition models and spawned an entire culture of research dealing with front end design for audition. The visual appeal of harmonic analyses and the further justification that our hearing system includes one, have been catalysts for their adoption in audio analysis systems. The short time fast Fourier transform (STFFT) has been, and still is, a dominant model for the front end. It is easy to manipulate, efficient and well understood. A later model, closer to the cochlear function, as well as a better estimator of time-frequency analysis, are constant-Q transforms (harmonic transforms in which the frequency spread versus the time spread are constant throughout the bases). They were used as front ends for audio analysis systems, reporting a better analysis performance as compared to STFFT [6, 115].

Other models made use of the sinusoidal analysis technique in conjunction with a constant-Q transform to mimic the behavior of the human auditory system and provide a perceptual representation. Due to its unique multiresolution properties wavelet transforms are preferred to STFFT for very accurate analyses with superior results [116]. Additional transform methods such as the cepstrum and linear predictive coding (LPC) have been employed [117], their uses however are specifically application based. Although the aforementioned decompositions are inspired by the function of the cochlea, they were by no means meant to be biologically plausible (accurate) models. The accurate reconstruction of the cochlear function has been extensively studied and has become a field of study on its own. Today, the dominant model employs a gammatone filterbank to approximate the function of the cochlea [118].

A gammatone filterbank is composed of basis functions which are sinusoidal tones modulated by gamma distributions. Further, researchers have employed even more complex front ends, more notably the correlogram and its derivatives [119]. The use of correlogram generally consists of a cochlear like filterbank extended by an additional dimension, which represents the lag time of autocorrelations applied on the energies of every frequency channel. Building on that model, the weft was introduced as an element for decompositions of primarily harmonic sounds. The weft is defined by its periodic track (time varying excitation period) and smooth spectrum (energy in each frequency channel for each time frame). The weft is extracted from the correlogram data and is inherently connected to the common modulation characteristics between frequency bands. Wefts form a compact and biologically plausible representation of periodic sounds as part of the vocabulary of CASA system [120–124].

2.3.2 Properties of Human Hearing

The concept of hearing area refers to the ranges of frequency and sound pressure values within which the human ear generally perceives sound. Reviewing the limits of the hearing area is the first step in studying the properties of the human hearing. Another very commonly discussed property of the auditory system is the masking phenomenon. Masking is a process in which the threshold of audibility of a sound is raised due to the presence of another sound. These two sounds are referred to as the maskee and the masker, the latter representing the one that causes the shift in the threshold. Masking as a whole is a very complicated phenomenon, containing some details that are still not completely understood.

However, masking effects can be experienced in everyday life. For example, it is more difficult to hear what the person in the next room is saying when the television or music system is blaring out, compared to the situation in which the interfering sounds from the television or music system are muted. The reason for not hearing what the person tries to tell is most often based on the inevitable masking property of the hearing system and not the lack of interest. In other words, one cannot hear the talking even by trying harder unless the television or music system is turned down. In order to be able to evaluate which parts of the incoming sound are masked, the main psychoacoustic properties of the human hearing system are to be clearly understood and for this we need an appropriate mathematical model. The field of psychoacoustics has made significant progress toward characterizing the human auditory perception and particularly the time-frequency analysis capabilities of the inner ear.

Although applying perceptual rules to signal coding is not a new idea, most current audio coders (MPEG) achieve compression by exploiting the fact that irrelevant signal information is not detectable by even a well trained or sensitive listener. Irrelevant information is identified during signal analysis by incorporating into the coder several psychoacoustic principles, including *the absolute threshold of hearing, critical band frequency analysis, simultaneous frequency masking, the spread of masking along the basilar membrane, and the temporal masking.*

2.3.2.1 Absolute Threshold of Hearing

The *absolute threshold of hearing* (ATH), also known as the *threshold in quiet*, characterizes the amount of energy needed in a pure tone such that it can be detected by a listener in a noiseless environment. The frequency dependence of ATH shown in Fig. 2.6 is typically expressed in terms of sound pressure level (dB SPL). The audible frequency range is from about 20 Hz to 16 kHz but the frequency limits can be even 16 Hz and 20 kHz for healthy children. In case of elderly people, the upper frequency limit can be dropped to as low as 10 kHz.

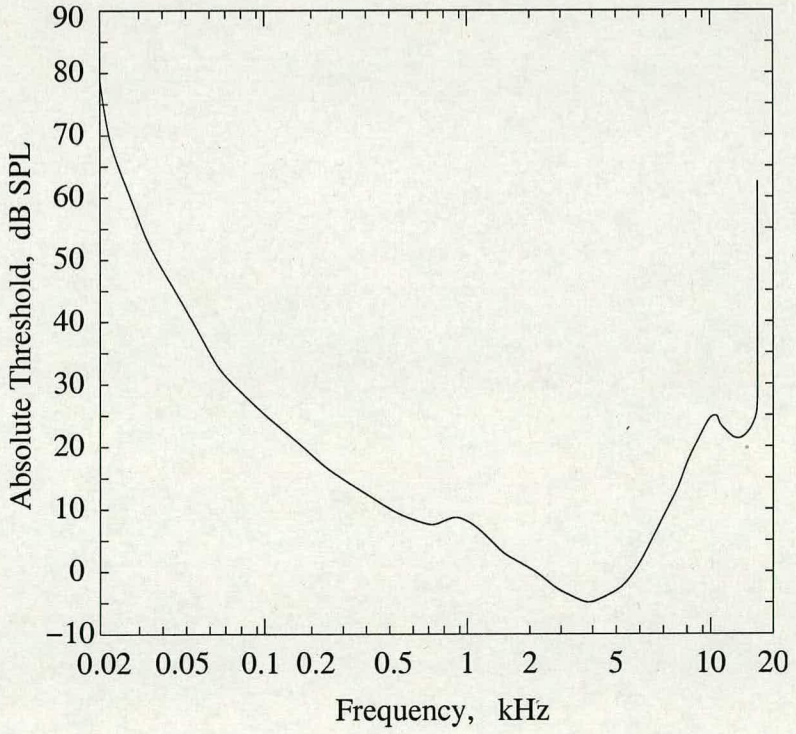


Figure 2.6: *Absolute Threshold of Hearing (ATH) as a Function of Frequency (after [2])*

A widely known approximation for the absolute threshold of hearing or threshold in quiet (T_q) as a function of frequency is modelled as

$$T_q(f) = 3.64(f)^{-0.8} - 6.5e^{0.6(f-3.3)^2} + 10^{-3}(f)^4 \quad \text{(dB SPL)} \tag{2.27}$$

where the frequency, f, is expressed in kHz. The dip in the absolute threshold curve in the neighbourhood of 4 kHz indicates the high sensitivity of hearing (about -5 dB SPL) and also the high susceptibility to hearing impairment in this region. Sound pressure levels just detectable at 4 kHz are not detectable at other frequencies. In general, two frequency tones of equal power but different frequencies will not sound equally loud. The perceived loudness of a sound can be expressed in sones, where 1 sone is defined as the loudness of a 40 dB tone at 1 kHz.

2.3.2.2 Critical Bands of Hearing

Fletcher introduced the concept of critical bandwidth (CB), denoting the noise bandwidth limit at which the detection threshold of the signal (tone) ceased to increase. For simplicity, he thought that the auditory filter could be approximated as having a rectangular shape and a passband width equal to CB. The shape of the auditory filter is not really rectangular, as Fletcher also knew, but this kind of a model can be useful in evaluating the masking effects in many applications such as those presented in this thesis. Fletcher suggested that, with this rectangular model, CB could be evaluated by measuring the threshold of a sinusoidal signal in broadband white noise. In this method, the power of the tone and the power spectral density of the noise masker are first measured. The noise power within the same critical band with the signal is then equal to the product of the measured power spectral density and the CB in question [2].

Fletcher also suggested that the ratio of the signal power to the noise power inside the critical band is equal to unity. In the described conditions, the tone would be just masked by the noise. Zwicker and Fastl [111] have later presented several methods for finding the values of CB and concluded that the threshold is reached when the ratio of the signal power to the power of the noise lies between 0.25 and 0.5. The critical bandwidth can also be explained based on the physical structure of the inner ear. Each point on the basilar membrane (BM) responds only to a certain range of frequencies, which leads to the idea that these different points correspond to auditory filters with different centre frequencies. When the bandwidth of the auditory filter is expressed by the equivalent rectangular bandwidth (ERB), the relation to BM is very simple: each ERB corresponds to a distance of about 0.89 mm on the basilar membrane.

ERB is defined so that the power of a signal inside the rectangular band equals the power of the same signal in the passband of the auditory filter. The ERB value increases with increasing centre frequency. A commonly used scale for signifying the critical bands is the Bark scale that divides the audible frequency range of 16 kHz into 24 critical bands. A distance of one critical band is commonly referred to as *one Bark* and it can be interpreted as the bandwidth at which subjective responses of human ear change abruptly. This critical band analysis conveniently simplifies the calculation of the spread of masking, i.e., the effect of adjacent critical bands on the amount of masking in a particular band. A complete list of the discrete set of critical bands is shown in "Critcal Bands Table" given in Appendix A [111].

From this Table, it is evident that the critical bands have constant width of 100 Hz for center frequencies up to 500 Hz, and the width of critical bands increase as the center frequency increases further. Since location on the basilar membrane has an approximately linear relationship to the frequency scale for low frequencies but a logarithmic relationship at higher frequencies, the linear frequency scale is inadequate for representing the auditory system. The relationship between the frequency in Hz and the critical band rate in Bark, both in proportion to the length of the unwound cochlea is illustrated in Fig. 2.7 [111].

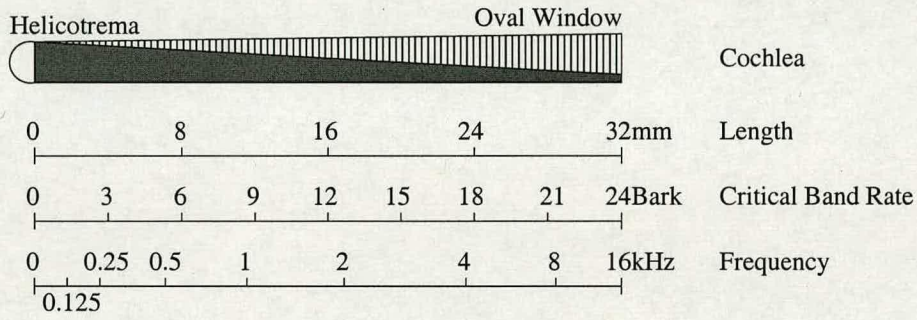


Figure 2.7: *Relation of Frequency, Critical Band Rate and Length of Unwound Cochlea*

An approximate analytical expression to describe the conversion from linear frequency f in Hz into the critical band number Z_c in Barks is given by [111]

$$Z_c(f) = 13\arctan(.00076f) + 3.5\arctan\left[\left(\frac{f}{7500}\right)^2\right] \quad \text{(Bark)} \tag{2.28}$$

and the critical bandwidth, BW_c for a given frequency f in Hz, can be expressed as [111]

$$BW_c(f) = 25 + 75\left[1 + 1.4(f/1000)^2\right]^{0.69} \quad \text{(Hz)} \tag{2.29}$$

2.3.2.3 Perceptual Masking Techniques

The human auditory system has some interesting properties, which are exploited in perceptual audio coding. We have a dynamic frequency range from about 20 to 20000 Hz, and we hear sounds with intensity varying over many magnitudes. The hearing system may thus seem to be a very wide range instrument, which is not altogether true. To obtain those characteristics , the hearing is very adaptive; what we hear depends on what kind of audio environment we are in. In the presence of a strong white noise, for example, many weaker sounds get masked and thus we cannot hear them at all. Some of these masking characteristics are due to the physical ear, and some are due to the processing in the brain. Masking effects occur in the frequency domain as well as in the time domain. There are two types of masking effects: simultaneous masking, and nonsimultaneous masking (also known as temporal masking).

Simultaneous Masking

Simultaneous masking refers to a frequency domain phenomenon that can be observed whenever two or more stimuli are simultaneously presented to the auditory system. Depending on the shape of the magnitude spectrum, the presence of certain spectral energy will mask the presence of other spectral energy. Although arbitrary audio spectra may contain complex simultaneous masking scenarios, for the purposes of shaping coding distortions it is convenient to distinguish between only two types of simultaneous masking, namely tone-masking-noise, and noise-masking-tone. In the first case, a tone occurring at the center of a critical band masks noise of any sub critical bandwidth or shape, provided the noise spectrum is below a predictable threshold directly related to the strength of the masking tone.

The second type of masking follows the same pattern with the roles of masker and maskee reversed. The masking effect of a tone (or noise) is not confined to within the critical bands. The presence of a strong noise or tone masker creates an excitation of sufficient strength on the basilar membrane at the critical band location to block effectively detection of a weaker signal. Inter-band masking has also been observed, i.e., a masker centered within one critical band has some predictable effect on detection thresholds in other critical bands. This is also known as the spread of masking and often modeled by an approximately triangular spreading function that has slopes of +25 dB and -10 dB per Bark. After tone (or noise) like masker types have been identified, their individual masking thresholds are combined to form a global masking threshold that estimates the level at which quantization noise becomes just noticeable.

Consequently, the global masking threshold is sometimes referred to as the level of just noticeable distortion (JND). The standard practice in perceptual coding involves first classifying masking signals as either noise or tone, next computing appropriate thresholds, then using this information to shape the noise spectrum beneath JND. Note that the absolute threshold of hearing (T_q) is also considered when shaping the noise spectra, and that MAX(JND,T_q) is most often used as the permissible distortion threshold.

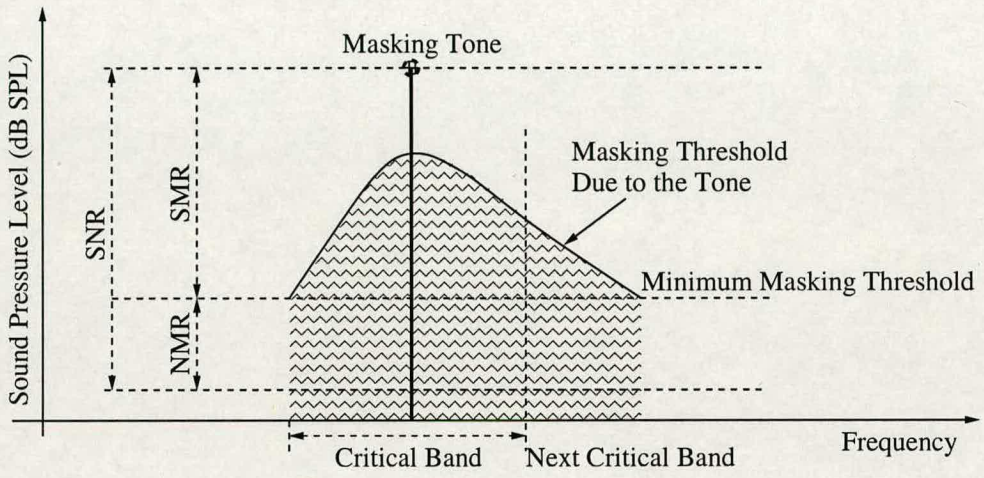


Figure 2.8: *Illustration of Simultaneous Masking Effects of a Tone*

Notions of critical bandwidth and simultaneous masking in the audio coding context give rise to some convenient terminology such as signal-to-mask ratio (SMR), noise-to-mask ratio (NMR) and signal-to-noise ratio (SNR) illustrated in Fig. 2.8, where the case of a single masking tone occurring at the center of a critical band has been considered. A hypothetical masking tone occurs at some masking level. This generates an excitation pattern along the basilar membrane that is modeled by a spreading function and a corresponding masking threshold [125–127].

The excitation pattern can be interpreted as an internal representation of the spectrum of the sound, i.e., a representation of the amount of activity evoked by a sound as a function of the characteristic frequency of the excited neuron. Since the upper slope of the excitation pattern, and hence also that of the masking pattern, is determined by the lower part of the auditory filter and vice versa, the spread of masking towards upper frequencies occurs. For the critical band under consideration, the minimum masking threshold can be referred to as the spreading function in-band minimum. Methods for calculating the spread of masking in the Bark domain for a particular psychoacoustic model will be presented in the following section.

Temporal Masking

Masking can also occur when the masker and the maskee are presented consecutively in time, without any overlapping time section. This phenomenon, called non-simultaneous masking, also known as temporal masking, is even more poorly understood than simultaneous masking and it is often considered to be of less importance when the masking effects are estimated on a coarse level. Temporal masking is typically divided into two different cases: backward and forward masking. In the former, the maskee appears before the masker and in the latter, the temporal positions are reversed. The backward and forward masking are also known as prestimulatory masking and poststimulatory masking respectively.

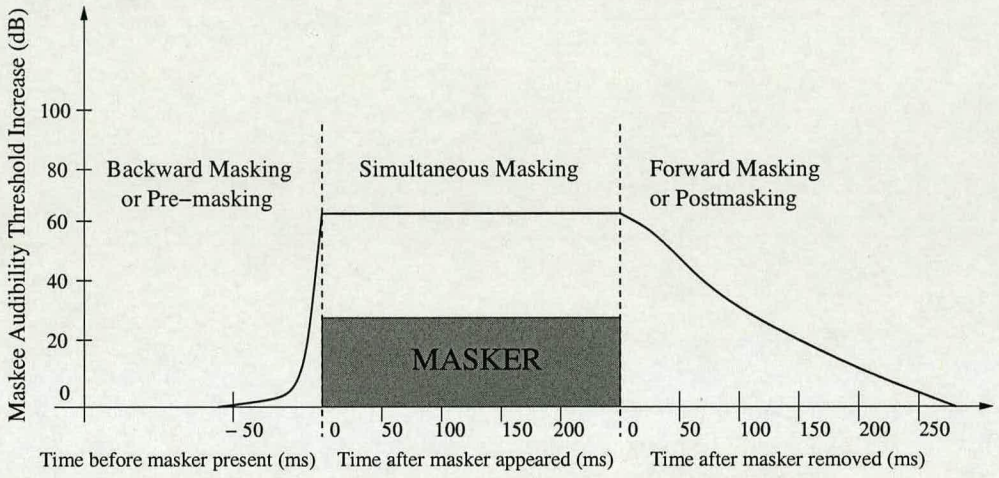


Figure 2.9: *Illustration of Temporal Masking Effects*

Fig. 2.9 shows the relevant time scale for masking effects with a rather long masker duration of 250 ms. The amount of forward masking depends strongly on the duration of the masker, a shorter duration causes the masking threshold to drop faster. Generally, forward masking will extend anywhere from 50 to 300 ms, whereas backward masking tends to last only about 5 ms, depending upon the strength and duration of the maskér. Temporal masking has been used in several audio coding algorithms. Backward masking in particular has been exploited in conjunction with adaptive block size transform coding to compensate for pre-echo distortions. In psychoacoustics, forward masking has long been regarded as an indication of the decay of the hearing system's internal loudness. So, it is often modeled using psychoacoustic specific loudness versus critical band rate and time [2, 111].

Trained listeners often show considerably less backward masking than those who have got little or no practice [2]. In any event, the effect of backward masking is minor and henceforth, it will not be discussed further. A typical example of forward masking in speech is the situation in which a plosive follows a loud vowel and gets masked. This is a very customary situation to which the human communication has been adapted, and therefore, it does not usually hinder intelligibility of the information. However, this kind of a forward masking phenomenon may be advantageous from the viewpoint of speech preprocessing.

2.4 Psychoacoustic Masking Models

The need for psychoacoustic masking models arises from the objective of developing audio codecs that preserve a good perceptual quality of the output signal despite significant reduction of bit rate. Masking models are utilised, for example, for shaping the noise introduced in the coding process such that it is masked as effectively as possible by the signal of interest, e.g. speech while preserving a good perceptual quality. Utilising the human auditory properties and the derived masking models is not a new idea. Schroeder et al [128] have developed a method of exploiting the auditory masking effects in speech coders. Their approach contained an auditory model that was used to evaluate the loudness of the quantisation noise and that of the signal, providing an objective measure of speech signal degradation caused by the coder. Based on their method, various psychoacoustic masking models have been proposed with different levels of accuracy and computational complexity.

Two well known psychoacoustic models published by the moving picture expert group (MPEG) to estimate the masking threshold are only considered in this thesis. These are ISO/MPEG-1 psychoacoustic model 1 and the ISO/MPEG-1 psychoacoustic model 2. Many other models have also been published, for example, the advanced audio coding (AAC) standard in MPEG-2 uses an auditory model derived from the MPEG-1 model 2, Johnston's [129] masking model, the perceptual evaluation of audio quality (PEAQ) [130, 131] masking model etc. Rather, the emphasis is on the two masking models that have been considered for preprocessing of speech signals. However, some properties are very different from one model to another. MPEG-1 psychoacoustic model 1 is simple and it estimates the masking threshold rather coarsely, while MPEG-1 psychoacoustic model 2 has somewhat higher computational complexity and better frequency resolution. Furthermore, the consideration of temporal masking makes model 2 more sophisticated than model 1 since in the latter, this part is omitted [126, 127, 132, 133].

2.4.1 ISO/MPEG-1 Psychoacoustic Model 1

The ISO/MPEG-1 psychoacoustic model 1 [126, 127] uses a 512 point FFT for high resolution spectral analysis, then estimates for each input frame individual simultaneous frequency masking thresholds due to the presence of tone-like and noise-like maskers in the input signal spectrum. A global masking threshold is then estimated for a subset of the original 256 frequency bins by (power) additive combination of the tonal and non-tonal individual masking thresholds. This thresholding is designed to select the perceptually relevant spectral components in each frame of the input speech. This model assumes masking effects are additive.

The five steps leading to computation of global masking threshold are as follows:

Step 1: Spectral Analysis and SPL Normalization

First, incoming digital audio samples, $\widetilde{\mathbf{x}}(n)$, are normalized according to the FFT length, N, and the number of bits per sample, nb, using the relation

$$\mathbf{x}(n) = \frac{\widetilde{\mathbf{x}}(n)}{N(2^{nb-1})} \tag{2.30}$$

The normalized input, $\mathbf{x}(n)$, is then segmented into frames of size of 512 samples using an appropriate time shift and window function. A power spectral density (PSD) estimate, *P(k)*, is then obtained using a 512-point FFT as

$$P(k) = PN + 10\log_{10}\left|\sum_{n=0}^{N-1} w(n)x(n)e^{-j\frac{2kn\pi}{N}}\right|^2 \text{ when } \left(0 \leq k \leq \frac{N}{2}\right) \tag{2.31}$$

where the power normalization term, *PN*, is fixed at 96 dB and the Hamming window, $w(n)$, is defined as

$$w(n) = \left[0.54 - 0.46\cos\left(\frac{2n\pi}{N}\right)\right] \tag{2.32}$$

Because playback levels are unknown during psychoacoustic signal analysis, the normalization procedure and the parameter *PN* are used to estimate SPL conservatively from the input signal.

Step 2: Identification of Tonal and Noise Maskers

After PSD estimation and SPL normalization, tonal and non-tonal masking components are identified. Local maxima in the sample PSD which exceed neighbouring components within a certain Bark distance by at least 7 dB are classified as tonal components. The tonal set, S_T, is

defined as

$$S_T = \left. P(k) \right| \begin{array}{l} P(k) > P(k \pm 1), \\ P(k) > P(k \pm \Delta_k) + 7\text{dB} \end{array} \tag{2.33}$$

where

$$\Delta_k \in \begin{cases} 2 & (2 < k < 63) \quad (0.17 - 5.5 \quad \text{kHz}) \\ [2,3] & (63 \leq k < 127) \quad (5.5 - 11 \quad \text{kHz}) \\ [2,6] & (127 \leq k \leq 256) \quad (11 - 20 \quad \text{kHz}) \end{cases} \tag{2.34}$$

Tonal maskers, $P_{TM}(k)$, are computed from the spectral peaks listed in S_T as follows

$$P_{TM}(k) = 10 \log_{10} \sum_{j=-1}^{1} 10^{0.1P(k+j)} \quad \text{(dB)} \tag{2.35}$$

a single noise masker for each critical band, $P_{NM}(\overline{k})$, is then computed from the remaining spectral lines not within the $\pm\Delta_k$ neighborhood of a tonal masker using the sum

$$\begin{gathered} P_{NM}(\overline{k}) = 10 \log_{10} \textstyle\sum_j 10^{0.1P(j)} \quad \text{(dB)}, \\ \forall P(j) \notin \{P_{TM}(k, k \pm 1, k \pm \Delta_k)\} \end{gathered} \tag{2.36}$$

where $\overline{k}$ is defined to be the geometric mean spectral line of the critical band, i.e.,

$$\overline{k} = \left(\prod_{j=l}^{u} j \right)^{\frac{1}{l-u+1}} \tag{2.37}$$

and l and u are the lower and upper spectral line boundaries of the critical band, respectively.

Step 3: Decimation and Reorganization of Maskers

In this step, the number of maskers is reduced using two criteria. First, any tonal or noise maskers below the absolute threshold of hearing, T_q, are discarded, i.e., only maskers which satisfy

$$P_{TM,NM}(k) \geq T_q(k) \tag{2.38}$$

are retained, where $T_q(k)$ is the SPL of the threshold in quiet at spectral line k. Next, a sliding 0.5 Bark-wide window is used to replace any pair of maskers occuring within 0.5 Bark distance by the stronger of the two. After the sliding window procedure, masker frequency bins are

reorganized according to the subsampling scheme

$$P_{TM,NM}(i) = P_{TM,NM}(k) \tag{2.39}$$

$$P_{TM,NM}(k) = 0 \tag{2.40}$$

where

$$i = \begin{cases} k & 1 \leq k \leq 48 \\ k + (k.\text{mod}2) & 49 \leq k \leq 96 \\ k + 3 - ((k-1) \bmod 4) & 97 \leq k \leq 232 \end{cases} \tag{2.41}$$

The net effect of (2.41) is 2:1 decimation of masker bins in critical bands 18-21 and 4:1 decimation of masker bins in critical bands 22-24, with no loss of masking components.

Step 4: Calculation of Individual Masking Thresholds

Having obtained a decimated set of tonal and noise maskers, individual tone and noise masking thresholds are computed next. Each individual threshold represents a masking contribution at frequency bin i due to the tone or noise masker located at bin j (reorganized during step 3). Tonal masker thresholds, $T_{TM}(i,j)$ are given by

$$T_{TM}(i,j) = P_{TM}(j) - 0.275Z_c(j) + SF(i,j) - 6.025 \quad \text{(dB SPL)} \tag{2.42}$$

where $P_{TM}(j)$ denotes the SPL of the tonal masker in frequency bin j, $Z_c(j)$ denotes the Bark frequency of bin j, and the spread of masking from masker bin j to maskee bin i, $SF(i,j)$, is modeled by the expression in (dB SPL)

$$SF(i,j) = \begin{cases} 17\Delta_{Z_c} - 0.4P_{TM}(j) + 11, & -3 \leq \Delta_{Z_c} < -1 \\ (0.4P_{TM}(j) + 6)\,\Delta_{Z_c}, & -1 \leq \Delta_{Z_c} < 0 \\ -17\Delta_{Z_c}, & 0 \leq \Delta_{Z_c} < 1 \\ (0.15P_{TM}(j) - 17)\,\Delta_{Z_c} - 0.15P_{TM}(j), & 1 \leq \Delta_{Z_c} < 8 \end{cases} \tag{2.43}$$

Individual noise masker thresholds, $T_{NM}(i,j)$, are given by

$$T_{NM}(i,j) = P_{NM}(j) - 0.175Z_c(j) + SF(i,j) - 2.025 \quad \text{(dB SPL)} \tag{2.44}$$

where $P_{NM}(j)$ denotes the SPL of the noise masker in frequency bin j and $SF(i,j)$ is obtained by replacing $P_{TM}(j)$ with $P_{NM}(j)$ everywhere in (2.43).

Step 5: Calculation of Global Masking Threshold

In this step, individual masking thresholds are combined to estimate a global masking threshold for each frequency bin in the subset given by (2.41). The model assumes that masking effects are additive. The global masking threshold, $T_g(i)$, is therefore obtained by computing the sum

$$T_g(i) = 10\log_{10}\left(10^{0.1T_q(i)} + \sum_{l=1}^{L_T} 10^{0.1T_{TM}(i,l)} + \sum_{m=1}^{M_N} 10^{0.1T_{NM}(i,m)}\right) \quad \text{(dB SPL)} \quad (2.45)$$

where $T_q(i)$ is the absolute hearing threshold for frequency bin i, $T_{TM}(i,l)$ and $T_{NM}(i,m)$ are the individual masking thresholds from step 4, and L_T and M_N are the number of tonal and noise maskers , respectively, identified during step 3.

2.4.2 ISO/MPEG-1 Psychoacoustic Model 2

The ISO MPEG-1 psychoacoustic model 2 evaluates the maximum inaudible distortion energy for the coding of a frame of audio using a 1024 point FFT for high resolution spectral analysis. Psychoacoustic model 2 never actually separates tonal and non-tonal components. Instead, it computes a tonality index as a function of frequency that gives a measure of whether the component is more tone-like or noise-like. Model 2 uses this tonality index which is based on a measure of predictability to interpolate between pure tone-masking-noise and noise-masking-tone values. Furthermore, model 2 uses data from the previous two analysis windows to predict, via linear extrapolation, the component values for the current window. Tonal components are more predictable and thus will have higher tonality indices. Because this process relies on more data, it is more likely to better discriminate between tonal and non-tonal components than the ISO/MPEG-1 psychoacoustic model 1 method [132, 133].

Then, the model 2 determines the noise masking thresholds by first applying an empirically determined spreading function to the signal components. Further, it includes an empirically determined absolute masking threshold, the threshold in quiet. This threshold is the lower bound on the audibility of sound signal. Next, model 2 selects the minimum of the masking thresholds covered by the subband only where the band is wide relative to the critical band in that frequency region. It uses the average of the masking thresholds covered by the subband where the band is narrow relative to the critical band. Finally, the masking threshold is computed for each subband in the uniform frequency domain [133].

The following are the necessary steps for computing the masking threshold [133]:

1. Reconstruct 1024 samples of the input signal

The FFT shift, *iblen*, must remain constant over any particular application of the threshold calculation process. The newest *iblen* samples of the signal are made available at every call to the threshold generator. The threshold generator must store 1024 - *iblen* samples, and concatenate those samples to accurately reconstruct 1024 consecutive samples of the input signal, $\mathbf{x}_i(n)$, where i represents the index, $1 < i < 1024$ of the current input stream. Then, apply an appropriate windowing function (Hamming) to the reconstructed 1024 samples of the input signal given by

$$\mathbf{x}_i(n)\left[0.54 - 0.46\cos\left(\frac{2n\pi}{N}\right)\right]. \tag{2.46}$$

2. Calculate the complex spectrum of the input signal

First, apply FFT to the windowed input signal. Then, express the complex spectrum of the input signal in the polar representation in terms of the magnitude (r_w) and the phase (f_w). Here, w indicates that the calculation is indexed by frequency in the FFT spectral line domain. An index of 1 corresponds to the DC term and an index of 513 corresponds to the spectral line at the Nyquist frequency.

3. Calculate a predicted magnitude and phase

A predicted magnitude, $\hat{r}_w$, and phase, $\hat{f}_w$ are calculated from the preceding two threshold calculation blocks' r_w and f_w:

$$\hat{r}_w = 2.0r_w(t-1) - r_w(t-2) \tag{2.47}$$

$$\hat{f}_w = 2.0f_w(t-1) - f_w(t-2) \tag{2.48}$$

where t represents the current block number, $t-1$ indexes the previous block's data, and $t-2$ indexes the data from the threshold calculation block before that.

4. Calculate the unpredictability measure

The unpredictability measure, c_w is defined as:

$$c_w = (((r_w\cos(f_w) - \hat{r}_w\cos(\hat{f}_w))^2 + (r_w\sin(f_w) - \hat{r}_w\sin(\hat{f}_w))^2)^{0.5})/(r_w + \text{abs}(\hat{r}_w)) \tag{2.49}$$

By sacrificing performance, this unpredictability measure can be calculated on only a lower portion of the frequency lines. Calculations should be done from DC to at least 3 kHz and preferably to 7 kHz. An upper limit of less than 5.5 kHz may considerably reduce performance from that obtained during the subjective testing of the audio algorithm. The c_w values above this limit should be set to 0.3. Best results will be obtained by calculating c_w up to 20 kHz.

5. Calculate the energy and unpredictability in the threshold calculation partitions

The energy in each partition, e_b, can be expressed as:

$$e_b = \sum_{w=wlb}^{w=whb} r_w^2 \tag{2.50}$$

where b denotes the index of the calculation partition, wlb is the lowest frequency line in the partition and whb represents the highest frequency line in the partition.

Further, the weighted unpredictability, c_b, is defined as:

$$c_b = \sum_{w=wlb}^{w=whb} r_w^2 c_w \tag{2.51}$$

The threshold calculation partitions provide a resolution of approximately either one FFT line or 1/3 critical band, whichever is wider. At low frequencies, a single line of the FFT will constitute a calculation partition. At high frequencies, many lines will be combined into one calculation partition. A set of partition values is provided for the sampling rate (32 kHz has been considered for the work presented in the thesis) in "Calculation Partition Table" given in Appendix A. There are several table elements will be used in the threshold calculation process and these are: the index of the calculation partition, b, the lowest frequency line in the partition, wlb, the highest frequency line in the partition, whb, the median bark value of the partition, bvb, a lower limit for the SNR in the partition that controls stereo unmasking effects, mvb and the value for tone masking noise (in dB) for the partition, $TMNb$.

6. Convolve the partitioned energy and unpredictability with the spreading function

The spreading function, $SF(i,j)$, is calculated by the following method:

$$SF(i,j) = \begin{cases} 10^{(x+y)/10}, & \text{if } y > -100 \\ 0 & \text{if } y < -100 \end{cases} \tag{2.52}$$

where $x = 8\min[(1.05(j-i)-0.5)^2 - 2(1.05(j-i)-0.5), 0]$, i is the Bark value of the signal being spread, j is the Bark value of the band being spread into and the value of y is given by: $y = 15.811389 + 7.5(1.05(j-i) + 0.474) - 17.5(1.0 + (1.05(j-i) + 0.474)^2)^{0.5}$.

Then, the convolved partitioned energy and the weighted unpredictability are given by

$$ec_{bb} = \sum_{bb=1}^{bmax} e_{bb} * SF(bvbb, bvb) \tag{2.53}$$

$$ct_b = \sum_{bb=1}^{bmax} c_{bb} * SF(bvbb, bvb) \tag{2.54}$$

In the case where the calculation includes a convolution or sum in the threshold calculation partition domain, bb will be used as the summation variable. Partition numbering starts at 1 and the largest value of b, $bmax$, equal to the largest index, exists for each sampling rate.

Because ct_b is weighted by the signal energy, it must be renormalized to c_{bb}, i.e. $c_{bb} = ct_b/ec_{bb}$. At the same time, due to the non-normalized nature of the spreading function, ec_{bb} should be renormalized and the calculated normalized energy en_b, is $en_b = ec_{bb} * rn_b$. Where rn_b is the normalization coefficient given by: $rn_b = 1/(\sum_{bb=0}^{bmax} SF(bvbb, bvb))$.

7. Convert c_{bb} to t_{bb}, the Tonality Index

$$t_{bb} = -0.299 - 0.43\log_e(c_{bb}) \tag{2.55}$$

Each t_{bb} is limited to the range of $0 < t_{bb} < 1$.

8. Calculate the required signal to noise ratio in each partition

The required signal to noise ratio, SNR_b, is:

$$SNR_b = \max(mvb, t_{bb} * TMNb + (1 - t_{bb})NMTb) \tag{2.56}$$

where $NMTb$ is the value for noise-masking-tone (5.5 dB) for each of the partition index b.

9. Calculate the power ratio

The power ratio, bc_b, is calculated as:

$$bc_b = 10^{-SNRb/10} \tag{2.57}$$

10. Calculation of actual energy threshold

The actual energy threshold, nb_b, is calculated as:

$$nb_b = en_b bc_b \tag{2.58}$$

11. Spread the threshold energy over FFT spectral lines

The threshold energy that is spread over FFT spectral line, nb_w, is expressed as:

$$nb_w = nb_b/(whb - wlb + 1) \tag{2.59}$$

12. Include absolute thresholds, yielding the final energy threshold of audibility

The final energy threhold of audibility, Th_w, is calculated after including the absolute threshold of hearing (threshold in quiet, ATH_w,) and is given by

$$Th_w = \max(nb_w, ATH_w) \tag{2.60}$$

The dB values of ATH_w shown in "Absolute Threshold Table" of Appendix A. These values must be converted into the energy domain after taking the FFT normalization into account.

2.5 Aims

The research in this thesis focuses on the linear convolutive mixing of speech signals in a noisy and highly reverberant environment. The real cocktail party problem cannot be directly solved by general framework of ICA because of two important reasons: first, there is a reverberation effect due to actual observation of sound sources in a room that is no longer modeled by a linear instantaneous mixing process, but this can be modeled by a linear convolutive mixing process; second, in practice there are fewer microphones or sensors than unknown acoustic source signals. However, humans deal with this cocktail party effect very effectively and easily by using 2 dynamic sensors (ears), because the sounds are filtered by thousands of band pass filters in the cochlea of auditory system based on critical band analysis. Moreover, the higher brain functions at the auditory cortex take care of tracking the signal related to the voice pitch and thereby segregating the sound source which is of our interest.

Perceptual audio coders use human hearing models to determine perceptual relevance and then eliminate redundancy with the minimal degradation of relevant information. When the sources are stationary, perceptual audio coders are frame or packet based because masking thresholds are computed for finite input blocks (1024 or 512 samples). This framework is known as block based perceptual masking approach. Similarly, when the sound sources are dynamic then we need to consider the sequential or adaptive perceptual masking approach.

The primary objective of this thesis is to reduce the computational complexity of solving the permutation ambiguity problem of FDICA using block based and sequential based perceptual masking approaches. In the block based approach, simultaneous frequency masking is applied between adjacent frequency components of the input speech block at the same time to remove the frequency components that are perceptually irrelevant. On the other hand, a sequential based perceptual approach deals with temporal masking. This framework will be used to construct blind source separation algorithms that exploit audio masking prior to the source separation (preprocessor) and after the source separation (postprocessor) in the frequency-domain.

The final objective of this thesis will be to develop an intelligent FDICA system that extracts a single source of interest from the mixtures. This can be achieved by exploiting the irrelevancy of some of the input speech spectrum using perceptual masking techniques before utilizing the subspace filtering method as a preprocessor of FDICA which reduces the effect of room reflections in advance and the remaining direct sounds then being separated by ICA.

2.6 Summary

In this chapter, we have analysed some of the techniques that have been developed to solve the instantaneous and convolutive mixtures. The frequency-domain approach for BSS of convolutive mixtures is the emphasis of this thesis. Some basic principles of the human auditory system, covering both psychoacoustic models that were considered for reducing the computational complexity of solving the permutation ambiguity problem of FDICA, have also been presented in this chapter. Finally, the aims of the thesis were accordingly restated on the basis of this background information. Thus, the objective of this chapter was not to perform a thorough review of the methods developed on the subject but on the other hand, give a complete overview of the area, emphasizing the different approaches that influenced our research work.

Chapter 3
BSS of Convolutive Audio Mixtures with Perceptual Preprocessing Filter

Preprocessing of speech signals typically aims at facilitating the ICA process before initiating the actual source separation. This objective is pursued in this work by taking advantage of the properties of the human auditory system. The idea is to make the speech signal applicable for more efficient FDICA by removing the perceptually irrelevant components of the signal.

The main objective of this Chapter is to investigate whether perceptual criteria, which takes into account the process whereby one auditory stimulus prohibits the detection of another signal (perceptual masking), can enhance the separation performance of existing BSS system when the mixing is noisy and highly reverberant. In this Chapter, a perceptually motivated FDICA system with preprocessor is proposed for solving the permutation ambiguity problem when the mixing is noisy and highly reverberant.

This preprocessing filter will be utilized to reduce the overall computational complexity of BSS system by exploiting the perceptual irrelevancy of the input speech spectrum using block based perceptual masking (simultaneous frequency masking) and the sequential perceptual masking (temporal masking) approaches before separating the signals by ICA.

The motivation for attempting the approach of psychoacoustic model based preprocessor is that different solutions proposed by several researchers as discussed in Chapter 2 (background) failed to solve the permutation ambiguity problem of FDICA completely. Most existing BSS techniques, however, apply an independence criterion to the speech signals directly and do not take the human auditory system into account.

Henceforth, we are proposing an optimal blind speech separation system with a perceptually motivated preprocessor in this Chapter for improving the overall performance of FDICA system while exploiting the perceptual irrelevancy of some of the observed speech signal spectrum before applying the complex FDICA algorithm. This approach will then be compared with existing BSS techniques which do not take perceptual masking into account.

3.1 Entire System

The FDICA system with perceptually motivated preprocessor is explained in the form of a block diagram as shown in Fig. 3.1.

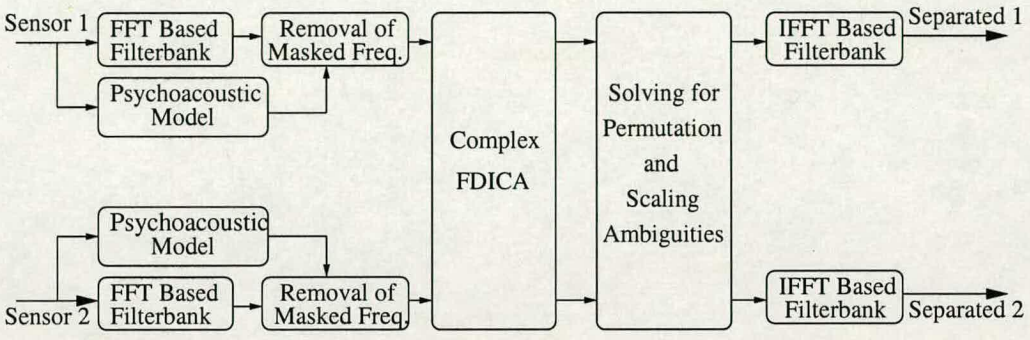


Figure 3.1: *Block Diagram of FDICA System with Perceptually Motivated Preprocessor*

First, the STFFT of the multichannel input signal, $\mathbf{x}(\omega, t)$, is obtained with an appropriate time shift and window function. Once the STFFT is obtained, the Fourier coefficients at each frequency are treated as complex time series. By doing this, the convolutive mixture problem is reduced to complex but several instantaneous mixture problems.

Next, the psychoacoustic model is used as a preprocessor in order to determine the perceptual masking threshold for each segment of speech and thereby exploiting the perceptual irrelevancy of some of the input speech spectrum.

The PCA filtering method is then applied to the perceptually observed input speech vector $\mathbf{x}_f(\omega, t)$ to orthogonalize its output, $\mathbf{y}(\omega, t)$ and thereby finding PCA filter matrix $\mathbf{W}(\omega)$.

Then, the complex FDICA is applied to the PCA filter output to obtain the ICA filter matrix $\mathbf{U}(\omega)$ while making the separated output as independent as possible. The product of $\mathbf{U}(\omega)$ and $\mathbf{W}(\omega)$ can be referred to as the separation filter matrix $\mathbf{B}(\omega)$.

After obtaining the separation filter, the permutation and the scaling ambiguity problem is solved by processing the output of the separation filter with the permutation matrix $\mathbf{P}(\omega)$ and the scaling matrix $\widetilde{\mathbf{B}}_{\widetilde{m}}(\omega)$.

Finally, the filter matrices are transformed into the time domain, and the input speech signal is processed with the reconstructed time-domain filters.

3.2 Implementation of Preprocessor

The main objective of this section is to introduce an implementation of perceptually motivated preprocessing filter that performs the perceptual irrelevancy removal. The procedure starts with the selection of the masking model which is considered in the next subsection. Using the resulting model, the masking threshold is calculated in order to determine the perceptually relevant components of the input speech spectrum.

3.2.1 Choosing of Masking Models

The speech preprocessing block implemented in this work was designed around the human auditory model. From different masking models presented in detail in Chapter 2, two independent models were chosen for the implementation of the speech preprocessor. Being such an essential part of the preprocessor, the selection of the model inevitably affects many of the properties of the final implementation.

Naturally, the overall modelling of the human auditory system is done with variable accuracy; for example, the temporal masking is totally omitted in psychoacoustic model 1 of MPEG-1. Thus, the auditory models can be considered from several different operating environments and the choice is finally made according to the particular situation, often ending up in a compromise. In this work, the auditory models under consideration were basically ISO/MPEG-1 psychoacoustic model 1 and model 2 [126, 127, 132, 133].

3.2.2 Removal of Masked Components

Masked components are referred to as irrelevant spectral components. If these irrelevant spectral components can be identified when the speech is recorded (observed), they can be thrown away and do not need to be considered for further signal processing. This process of dumping (throwing) the perceptually irrelevant frequency components of the input speech signal spectrum is referred to as the spectral modification.

The modification of the speech signal spectrum can be simply performed by multiplying, or masking, the spectral values of the components to selectively eliminate, reduce, or enhance them. For example, if we wanted to eliminate one spectral component, we would simply multiply it by zero and mask it out.

In general, the spectral masking or filter spectrum is a fixed pattern of masking values determined by an algorithm, where a set of harmonics in the frequency-domain speech spectrum is kept or masked out, based on the consideration of adjacent (neighbouring) frequencies in the input speech spectrum. This spectral masking is dependent on the masking threshold that has been calculated for the input speech frame which is compared with the input power spectrum in the corresponding frame to produce a perceptual binary mask [134, 135].

The mask is set to a value of zero at those frequency bins where the power spectrum is below the masking threshold and a value of one is used elsewhere as shown in Fig. 3.2. A straightforward means to remove the masked frequency bins would be the multiplication of the complex spectrum of the input speech frame by the binary mask at each frequency bin. This corresponds to an adaptive filtering of the input speech since the mask changes from frame to frame.

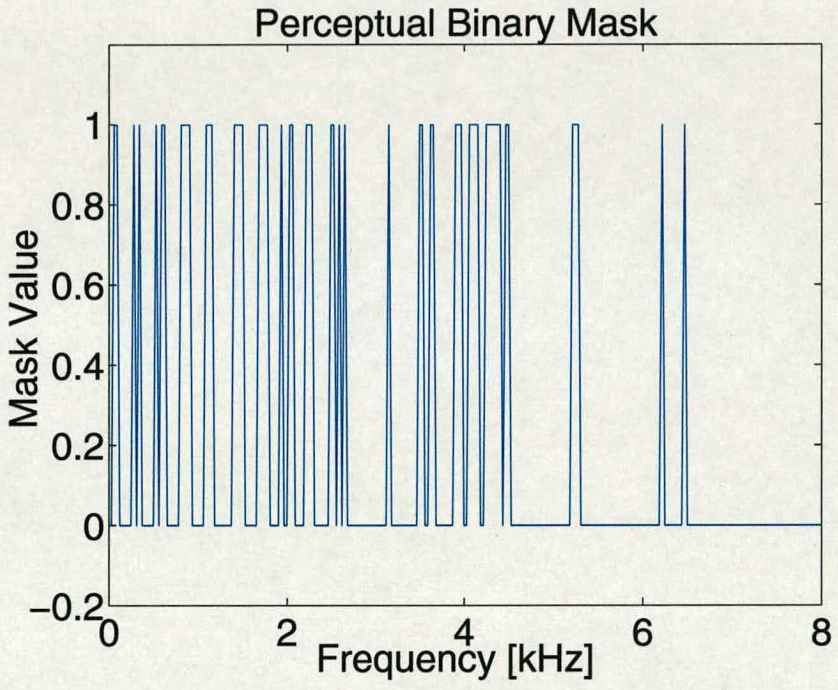


Figure 3.2: *An Example of Perceptual Binary Mask*

Therefore, this spectral masking or filter spectrum is multiplied against the input speech spectrum on an individual element by element basis, which is described by a simple multiplication operation. Hence, multiplying the complex spectrum of the input speech frame by the perceptual binary mask at each frequency bin would remove the masked frequency bins [134, 135].

This procedure of removing masked frequency bins by multiplication operation can also be extended to the stereo environment where we have more than one microphone. Hence, the thresholding in the stereo environment is described by logical AND operation.

In perceptual audio coding, thresholding sets the quanization level, here we set a threshold for further processing of the frequencies by ICA according to their psychoacoustic relevance and thereby reducing the computational complexity of similarity measure among spectral envelopes of the separated output signals at several adjacent frequencies for solving the permutation ambiguity problem of the FDICA system. While this perceptual thresholding is a nonlinear activity which might at first sight appear to destroy the linear convolutive properties of the BSS system, it can also be viewed as an irregular sampling rate strategy which is linear.

For almost all audio signals many spectral components are below the the masking threshold and can be discarded. From our analysis, on average for voiced speech signals, more than 50% of spectral components are masked out. Therefore, the number of spectral components available at the output of perceptually motivated preprocessing filter are also changed. Thus, this thresholding will alter the probability density function (pdf) of the perceptually masked input speech signals presented to the complex FDICA.

3.3 Perceptually Motivated Time-Delayed Decorrelation Algorithm

After suppressing the irrelevant frequency components from the observed input speech spectrum by a perceptually motivated preprocessor, we applied the multiple time-delayed decorrelation algorithm [75]. This algorithm is considered because of the following two advantages:

- It uses only the second order statistics, hence the estimation is generally robust, and
- this procedure does not include iterative operations.

The principle of multiple time-delayed decorrelation algorithm that consists of two stages, *sphering* and *rotation*, is explained in the form of a block diagram as shown in Fig. 3.3. Sphering is a procedure to obtain whitening matrix, $\mathbf{W}(\omega)$, whereas rotation is a procedure to remove off digonal elements of correlation matrices with an orthogonal transformation. This concept is simplified as the simultaneous diagonalization of the correlation matrix of perceptually masked observations at several time-lags.

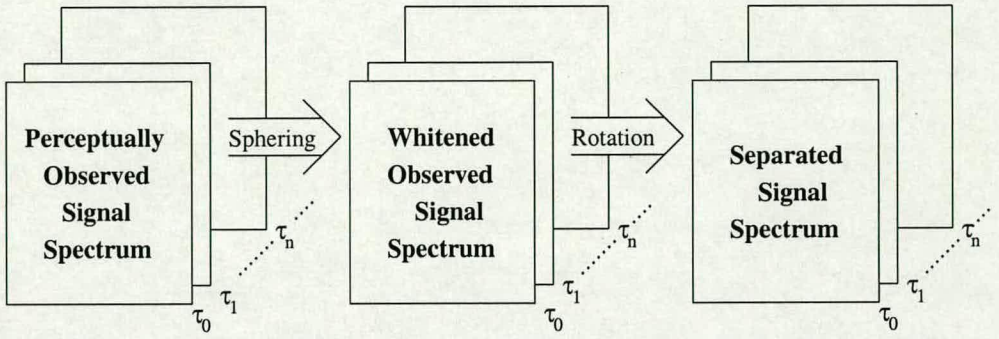


Figure 3.3: *Principle of Perceptually Motivated Time-Delayed Decorrelation Algorithm*

To solve this simultaneous diagonalization problem, we use a Jacobi like algorithm proposed by Cardoso and Souloumiac [136]. It is an extension of Givens unitary rotation transform and the problem is reduced to combination of subproblems of the 2×2 case that can be solved analytically and thereby separating the signal for each frequency bin independently.

In the first instance, we cannot directly compute the decorrelation matrices (rotational matrices) at multiple time-lags. The reason is very simple. Whenever the perceptually masked input speech $\mathbf{x}_f(\omega, t)$ in one of the channels contains no values, the rotational matrix $\mathbf{U}(\omega)$ is singular, resulting in rank deficiency problem. This is mainly due to very low eigenvalues of decorrelation matrix of perceptually masked input speech spectrum.

Without loss of generality, we assumed identity matrix of order M as the rank of rotational matrix $\mathbf{U}(\omega)$ to avoid this rank deficiency problem while retaining the whitening properties of perceptually masked input speech signal even after rotational procedure.

Thus, we are considering only the perceptually relevant frequency components of the observed input speech signal spectrum for computing the multiple time-lagged decorrelation matrices and thereby reducing the computational complexity of decorrelation algorithm.

Let us assume that source signals are weakly stationary and perceptually masked observed signals are a non-convolutive or instantaneous mixture of complex-valued time series $\mathbf{s}(\omega, t)$. For a fixed frequency ω, the relationship between sources and perceptually masked observations are written in the matrix-vector notation as

$$\mathbf{x}_f(\omega, t) = \mathbf{A}(\omega)\mathbf{s}(\omega, t)\mathbf{\Phi}_{th} \tag{3.1}$$

where $\mathbf{A}(\omega)$ is the Fourier transform of room filter matrix $\mathbf{A}(t)$, $\mathbf{s}(\omega, t)$ is the windowed Fourier transform of sources $\mathbf{s}(t)$ and $\mathbf{\Phi}_{th}$ is the perceptual masking threshold matrix.

The correlation matrix of perceptually masked observations at several time-lags is

$$\begin{aligned}
\mathbf{R}_{xx}(\omega,\tau) &= E[\mathbf{x}_f(\omega,t)\mathbf{x}_f(\omega,t+\tau)^H] && (3.2)\\
&= \mathbf{A}(\omega)E[\mathbf{s}(\omega,t)\mathbf{s}(\omega,t+\tau)^H]\mathbf{A}(\omega)^H\mathbf{\Phi} && (3.3)\\
&= \mathbf{A}(\omega)\begin{pmatrix} \mathbf{R}_{s_1}(\omega,\tau) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mathbf{R}_{s_n}(\omega,\tau) \end{pmatrix}\mathbf{A}(\omega)^H\mathbf{\Phi} && (3.4)
\end{aligned}$$

where $.^H$ denotes the Hermitian transpose, $E[.]$ denotes taking the average, $\mathbf{R}_{s_1}(\omega,\tau)$ is the auto-correlation function of $\mathbf{s}(\omega,t)$ and $\mathbf{\Phi}$ is the perceptual gain given by $\mathbf{\Phi}_{th}\mathbf{\Phi}_{th}^H$.

With a desired separation filter matrix $\mathbf{B}(\omega)$, the reconstructed signals are represented by

$$\begin{aligned}
\mathbf{y}(\omega,t) &= \mathbf{B}(\omega)\mathbf{x}_f(\omega,t) && (3.5)\\
&= \mathbf{B}(\omega)\mathbf{A}(\omega)\mathbf{s}(\omega,t)\mathbf{\Phi}_{th} && (3.6)\\
&= \mathbf{P}(\omega)\mathbf{D}(\omega)\mathbf{s}(\omega,t)\mathbf{\Phi}_{th} && (3.7)
\end{aligned}$$

The correlation matrix of the reconstructed signal becomes

$$\begin{aligned}
\mathbf{R}_{yy}(\omega,\tau) &= E[(\mathbf{PDs}(\omega,t))(\mathbf{PDs}(\omega,t+\tau))^H]\mathbf{\Phi} && (3.8)\\
&= \begin{pmatrix} |\lambda_{1'}|^2\mathbf{R}_{s_{1'}}(\omega,\tau) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & |\lambda_{n'}|^2\mathbf{R}_{s_{n'}}(\omega,\tau) \end{pmatrix}\mathbf{\Phi} && (3.9)
\end{aligned}$$

where $1', 2', \cdots, n'$ denotes a permutation of the indices $1, 2, \cdots, n$ determined by matrix $\mathbf{P}(\omega)$ and λ_i is the ith diagonal element of matrix $\mathbf{D}(\omega)$.

Hence, except for the ambiguity of permutation $\mathbf{P}(\omega)$ and scaling $\mathbf{D}(\omega)$, an optimal $\mathbf{B}(\omega)$ can be characterized as a matrix that diagonalizes the correlation matrices at any time lag τ. With these two operations, i.e. sphering (pre-whitening) and rotation, one can find $\mathbf{B}(\omega)$ in a certain class of matrices, which satisfies

$$\mathbf{B}(\omega)\mathbf{M}_i\mathbf{B}(\omega)^H = \mathbf{\Lambda}_i(\omega), \quad i = 1, \cdots, r \tag{3.10}$$

where $\mathbf{\Lambda}_i(\omega)$'s are diagonal matrices, r is the number of matrices to be simultaneously diagonalized and $\mathbf{M}_i$'s are time-delayed correlation matrices given by $E[\mathbf{x}_f(\omega,t)\mathbf{x}_f(\omega,t+\tau_i)^H]$.

Finally, the separation filter matrix $\mathbf{B}(\omega)$ is determined by

$$\mathbf{B}(\omega) = \mathbf{U}(\omega)\mathbf{W}(\omega) = \mathbf{U}(\omega)\sqrt{\mathbf{V}^{-1}(\omega)} = \mathbf{U}(\omega)\mathbf{V}^{-1/2}(\omega) \tag{3.11}$$

where $\sqrt{\mathbf{V}^{-1}(\omega)}$ is the inverse square root of the matrix $\mathbf{V}(\omega)$ given by

$$\mathbf{V}(\omega) = \mathbf{R}_{xx}(\omega,\tau)_{\tau=0} \tag{3.12}$$

Generally, the square root of $\mathbf{V}(\omega)$ is referred to as the *Cholesky decomposition* and sometimes it is also known as the *Cholesky square root matrix* [137, 138].

Further, the inverse square root matrix $\mathbf{V}^{-1/2}(\omega)$ can be expressed as [11]

$$\mathbf{V}^{-1/2}(\omega) = \mathbf{E}\mathbf{\Lambda}^{-1/2}\mathbf{E}^H \tag{3.13}$$

where matrices, $\mathbf{E} = [\mathbf{e}_1, \cdots, \mathbf{e}_M]$ and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \cdots, \lambda_M)$, are corresponding to the eigenvector $\mathbf{e}_m$ and the eigenvalue λ_m, respectively. The symbol $.^H$ denotes the Hermitian transpose.

$\mathbf{U}(\omega)$ is the unitary matrix obtained by minimizing

$$\sum_{i=1}^{r}\sum_{j\neq k}|(\mathbf{U}\mathbf{M}_i\mathbf{U}^H)_{jk}|^2 \tag{3.14}$$

where $(\mathbf{U}\mathbf{M}_i\mathbf{U}^H)_{jk}$ denotes the jk-element of matrix $\mathbf{U}\mathbf{M}_i\mathbf{U}^H$.

3.3.1 Method of Solving the Permutation and Scaling

3.3.1.1 Scaling Problem

As explained in Chapter 2 (see 2.2.3.1 for details), the scaling problem can be solved by filtering individual outputs of the separation filter by $\mathbf{B}^{-1}(\omega)$ separately [87]. This procedure will lead towards finding the scaling matrix $\widetilde{\mathbf{B}}_{\widetilde{m}}^{-1}(\omega)$.

After solving the scaling ambiguity problem, the separated output can be written in the matrix-vector notation as

$$\widetilde{\mathbf{y}}(\omega,t) = \widetilde{\mathbf{B}}_{\widetilde{m}}^{-1}(\omega)\mathbf{y}(\omega,t) \tag{3.15}$$

where $\widetilde{m}$ denotes an arbitrary microphone number.

3.3.1.2 Permutation Problem

Based on the nonstationarity of the speech signals, we have assumed that components at different frequencies from the same source signals are under the influence of a similar modulation in amplitude. Let us define $\widetilde{s}_i(\omega,t)$ in terms of magnitude and phase as

$$\widetilde{s}_i(\omega,t) = a_i(\omega,t)\exp(j\phi_i(\omega,t)) \tag{3.16}$$

Because of nonstationarity, the magnitude $a_i(\omega,t)$ changes in time, and it corresponds to the envelope of $s_i(\omega,t)$. As $s_i(\omega,t)$ and $s_j(\omega,t)$ are independent, the correlation between envelopes $a_i(\omega,t)$ and $a_j(\omega,t)$ vanishes.

$$\text{Corr}(a_i(\omega,t), a_j(\omega,t)) = 0, \quad i \neq j \tag{3.17}$$

Similarly, correlation between different frequency components, ω and ω', from different source signals also vanishes.

$$\text{Corr}(a_i(\omega,t), a_j(\omega',t)) = 0, \quad i \neq j, \quad \omega \neq \omega' \tag{3.18}$$

However, for different frequency components from the same source signal, we can assume that

$$\text{Corr}(a_i(\omega,t), a_j(\omega',t)) \neq 0, \quad i = j, \quad \omega \neq \omega' \tag{3.19}$$

When adjacent frequency components from the same source signal are zero, it is assumed that

$$\text{Corr}(a_i(\omega,t), a_j(\omega',t)) = 0, \quad i = j, \quad \omega \neq \omega' \tag{3.20}$$

It implies that frequency components of speech signals will not change the super Gaussian distributions drastically in time, but they are similarly affected by the amplitude modulation of

the vocal chords. Therefore, the correlation coefficient ρ of their envelopes

$$\rho(\mathrm{Corr}(a_i(\omega,t),a_j(\omega',t))) = \frac{\mathrm{Corr}(a_i(\omega,t),a_j(\omega',t))}{\sqrt{\mathrm{Corr}(a_i(\omega,t),a_i(\omega',t))\mathrm{Corr}(a_j(\omega,t),a_j(\omega',t))}} \tag{3.21}$$

would be a natural measure for estimating appropriate combination of frequency components.

With the help of moving average operator ε, we can estimate the envelope of time series as

$$\varepsilon\widetilde{\mathbf{y}}(\omega,t;i) = \frac{1}{2L+1}\sum_{t'=t-L}^{t+L}\left(\sum_{j=1}^{n}|\widetilde{y}_j(\omega,t';i)|\right) \tag{3.22}$$

Where L is a positive constant that gives an idea about the size of the problem ($L = 1000$ for the spectrogram of 2 sec with the STFFT shift of 16 samples at 16 kHz) and $\widetilde{y}_j(\omega,t;i)$ is the jth component of $\widetilde{\mathbf{y}}(\omega,t;i)$.

The permutation is then solved by sorting based on the inter frequency spectral envelope correlations (IFSEC) of separated signals (Fig. 3.4) as per the following procedure [87]:

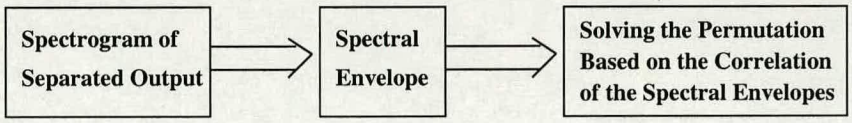


Figure 3.4: *Principle of Solving the Permutation Problem by IFSEC Method*

1) Sort ω in order of low correlation between independent components in each frequency bin. This is done by sorting in increasing order of similarity defined by

$$sim(\omega) = \sum_{i\neq j}\rho(\varepsilon\widetilde{\mathbf{y}}(\omega,t;i),\varepsilon\widetilde{\mathbf{y}}(\omega,t;j)), \tag{3.23}$$

$$sim(\omega_1) \leq sim(\omega_2) \leq \cdots \leq sim(\omega_r). \tag{3.24}$$

2) For ω_1, assign $\widetilde{\mathbf{y}}(\omega_1,t;i)$ to $\mathbf{y}(\omega_1,t;i)$ as it satisfies

$$\mathbf{y}(\omega_1,t;i) = \widetilde{\mathbf{y}}(\omega_1,t;i), \qquad i = 1,\cdots,n. \tag{3.25}$$

3) For ω_r, find a permutation $\mathbf{P}(i)$ which maximizes the correlation between the envelope of ω_r and the aggregated envelope from ω_1 to ω_{r-1}. This can be achieved by maximizing

sum of correlation coefficients

$$\sum_{i=1}^{n} \rho \left(\epsilon\widetilde{\mathbf{y}}(\omega_r, t; \mathbf{P}(i)), \sum_{j=1}^{r-1} \epsilon\mathbf{y}(\omega_j, t; i) \right) \tag{3.26}$$

within all the possible permutations of $i = 1, \cdots, n$.

4) Assign the appropriate permutation to $\mathbf{y}(\omega_r, t; i)$ as it is

$$\mathbf{y}(\omega_r, t; i) = \widetilde{\mathbf{y}}(\omega_r, t; \mathbf{P}(i)), \qquad i = 1, \cdots, n. \tag{3.27}$$

5) Go to step no. 3 until number of simultaneous matrices to be diagonalized, $r = 40$.

3.3.2 Reconstructed Signals

After solving the scaling and the permutation ambiguities problem, the separated spectrograms are obtained as

$$\mathbf{y}(\omega, t; i), \qquad i = 1, \cdots, n. \tag{3.28}$$

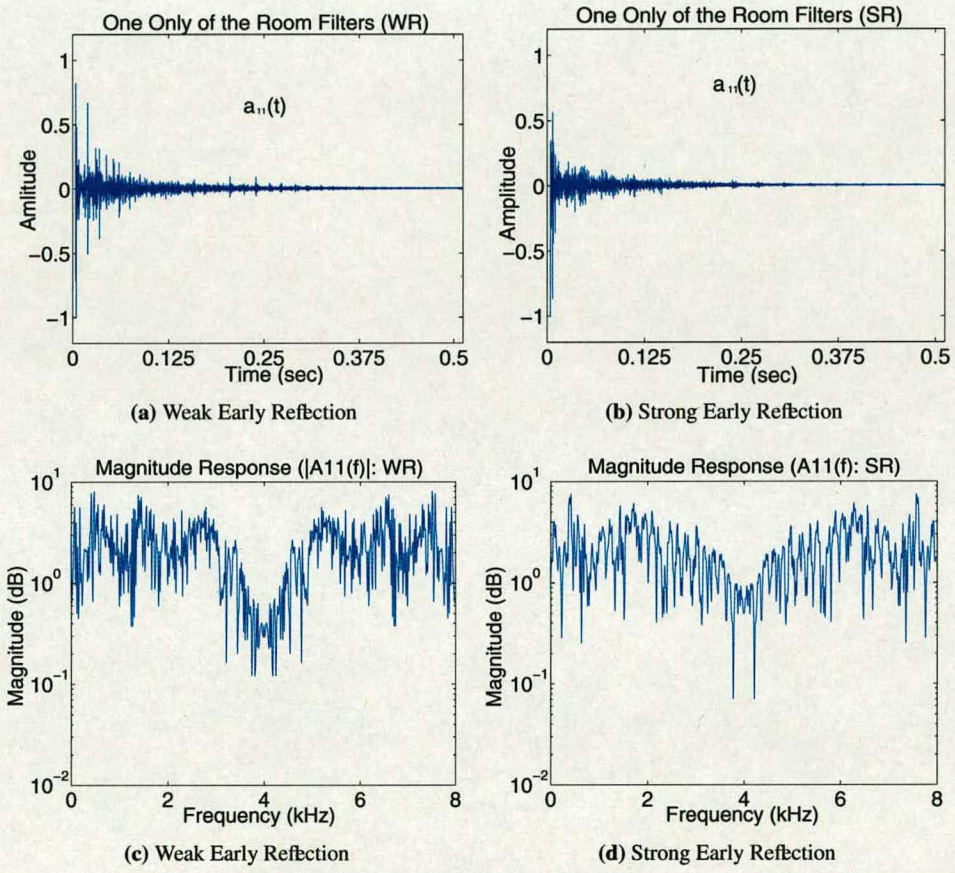
Applying the inverse Fourier transform to separated spectrograms $\mathbf{y}(\omega, t; i)$, we obtain a set of time-domain reconstructed signals, $\mathbf{y}(t; i)$.

Further, each component of the reconstructed signal, $\mathbf{y}_m(t; i)$ represents a separated independent component i on sensor m and $\sum_i \mathbf{y}(t; i) = \mathbf{x}_f(t)$ holds.

3.3.3 Experimental Results

3.3.3.1 Synthetic Room Mixing Scenario

In this experiment, we created a synthetic room mixing of two speech sources (4 s at 16 kHz) using Asano's [139] conference room (5.9 m×8.7 m×4.1 m) with a reverberation time of 0.5 sec for both the weak and strong early reflection cases of mixing environment. The corresponding room filters and their magnitude frequency responses are shown in Fig. 3.5.



(a) Weak Early Reflection **(b)** Strong Early Reflection

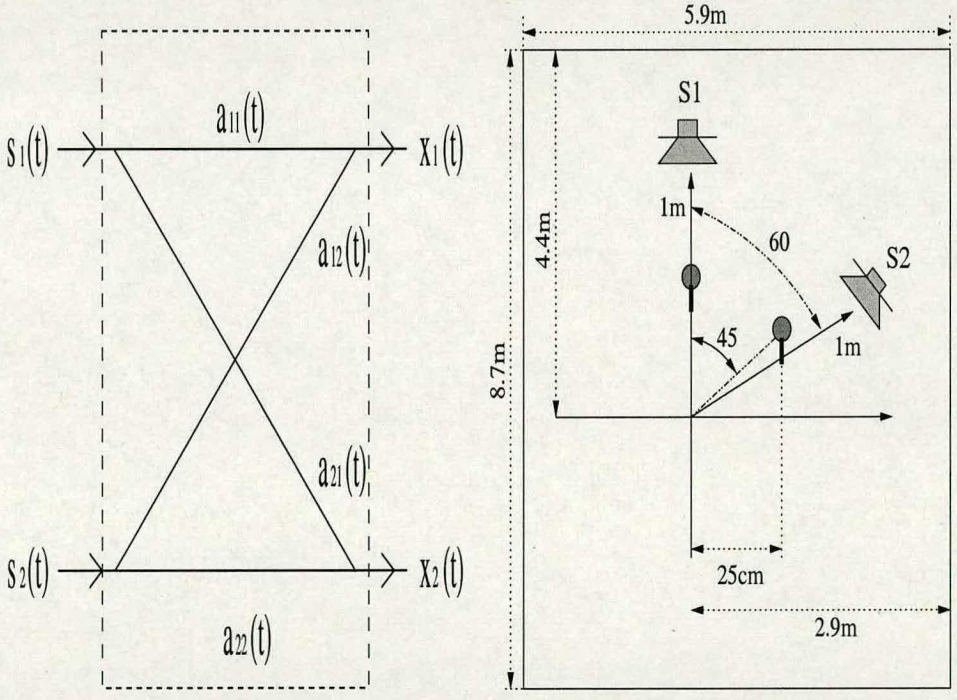
(c) Weak Early Reflection **(d)** Strong Early Reflection

Figure 3.5: *One Only of the Room Filters Used and Their Magnitude Frequency Responses (Synthetic Room Mixing Scenario)*

The system configuration and the experimental setup of the sound sources (loudspeakers) and the microphones are shown in Fig. 3.6. The impulse responses from the sound sources to the microphones were used to convolve with the source signal to generate the observed input speech signal. We applied the time-delayed decorrelation algorithm for both early reflection cases, using the parameters of the proposed BSS system that are summarized in Table 3.1.

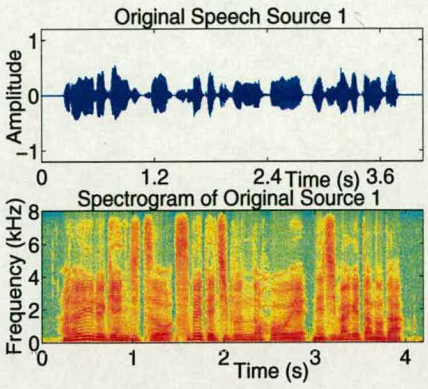
Sampling Frequency	16 kHz
STFFT Frame Length	512
Shift of STFFT	20
Window Function	Hamming
Normalized Sound Pressure Level, SPL	96 dB
Number of Microphones, M	2
Number of Sources, D	2
Number of Matrices for Diagonalization, r	40

Table 3.1: ***Proposed BSS System-1 Parameters (TDDA: Perceptual Preprocessing)***

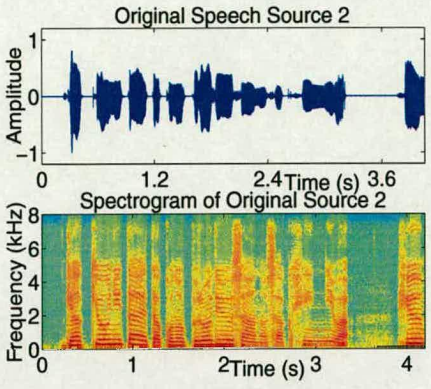


(a) BSS System **(b)** Experimental Setup

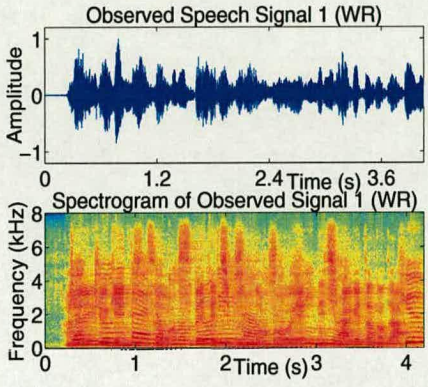
Figure 3.6: *System Configuration and the Experimental Setup (Synthetic Room Mixing)*



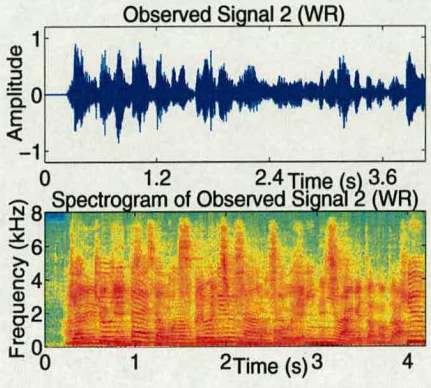
(a) Source 1 and its Spectrogram



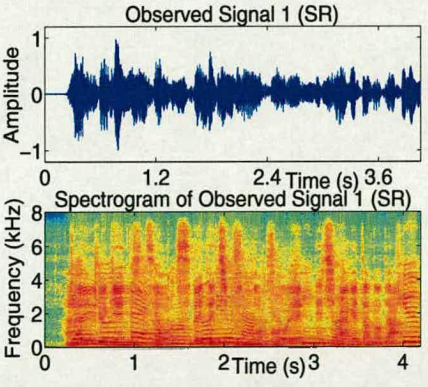
(b) Source 2 and its Spectrogram



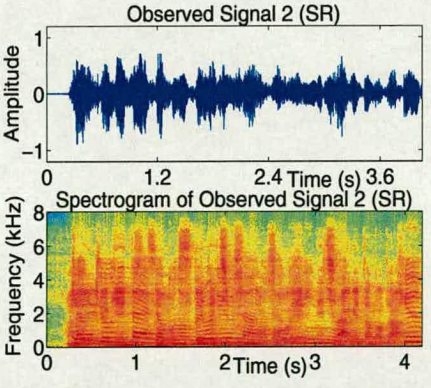
(c) Observed 1 and its Spectrogram (WR)



(d) Observed 2 and its Spectrogram (WR)



(e) Observed 1 and its Spectrogarm (SR)



(f) Observed 2 and its Spectrogram (SR)

Figure 3.7: *Speech Sources, Observed Signals and the Corresponding Spectrograms (Synthetic Room Mixing Scenario)*

Since the *Hamming window* is the most often used windowing technique for power spectrum estimation based speech enhancement applications, we also considered the *Hamming window* for all the experimental work reported in this thesis. Further, it can be used to control the spectral leakage and also to minimise its effect on reducing the dynamic range capability of the transform output and thereby reconstructing the speech signal in the time domain.

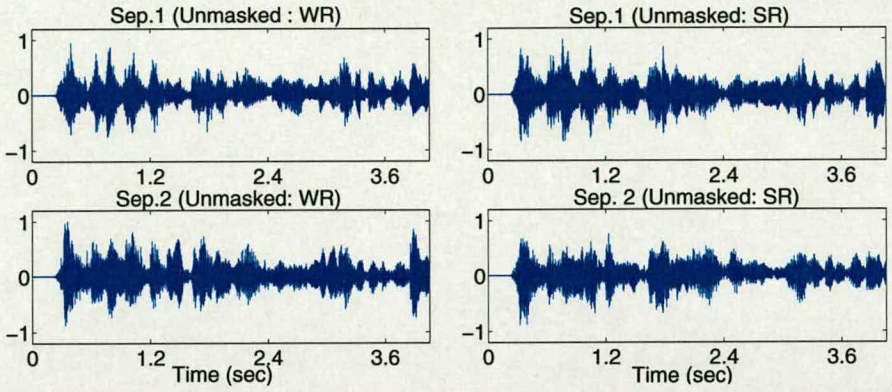
Further, the practical outcome of compact disc (CD) audio is the ability to represent from 20 to 20000 Hz, with a maximum theoretical dynamic range of 96.33 dB due to 16 bit resolution. Henceforth, a normalised sound pressure level (SPL) of 96 dB is considered for all the experimental work reported in this thesis.

Sound sources, observed signals and the corresponding spectrograms are shown in Fig. 3.7. The separated (reconstructed) signals obtained for both unmasked and masked FDICA systems (using perceptually motivated preprocessor) are shown in Fig. 3.8.

From Fig. 3.8, it is evident that the separated signals for both unmasked and masked systems (both models) are entirely different from original sources when the weak reflection case is considered. Further, the separated signals for both unmasked and masked systems (both models) are similar to observed input speech when the strong reflection case is considered.

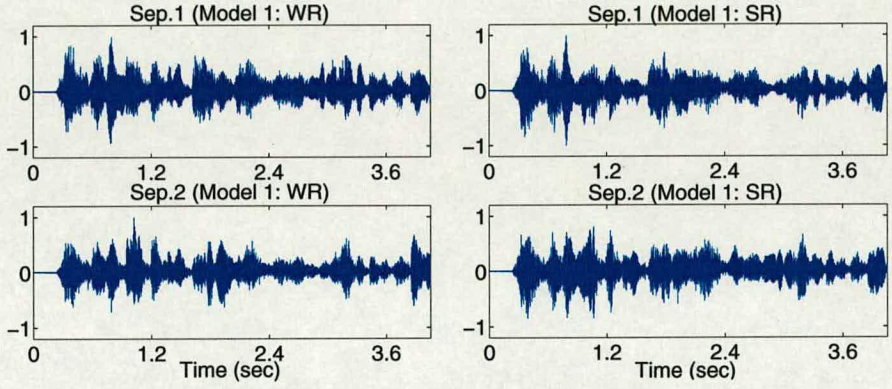
The permutation error is defined as the case when the result of inter frequency spectral envelope correlation (IFSEC) differs from that of source output crosscorrelation (SOC) (assumed as correct permutation). It is clearly evident from Fig. 3.9 that the measured permutation error is large for most of the frequencies when both unmasked and masked FDICA systems (using psychoacoustic model 1) are considered for both early reflection cases. On the other hand, masked FDICA system using model 2 reduces the measured permutation error for most of the frequencies except in the range 3-5 kHz when the weak early reflection case is considered. Whereas, the measured permutation error is small for the frequencies over 5 kHz when the strong early reflection case is considered.

Further, it is well known that the decorrelation algorithm fails when the speech sources have identical spectral envelopes [88] and even if one spectral component does not have any power for both unmasked and masked cases. Based on the above discussion and experimental results (synthetic room mixing environment), we conclude that the perceptually motivated time-delayed decorrelation algorithm cannot solve the permutation ambiguity problem of FDICA. Therefore, we did not consider the real room recording scenario for further experimentation.



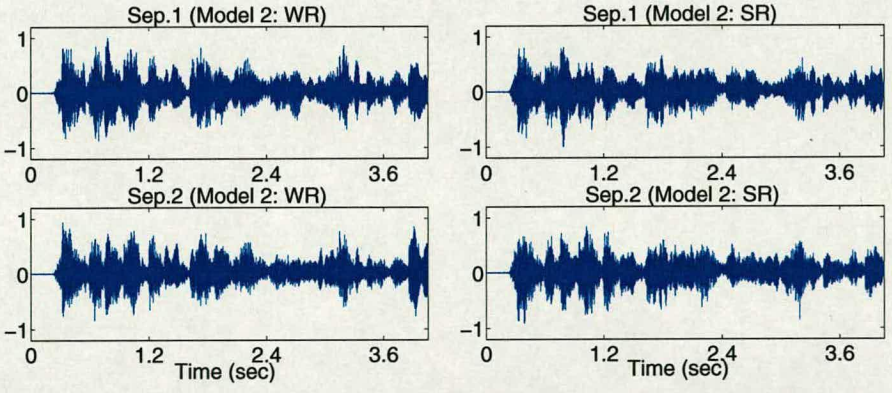
(a) Unmasked (Weak Reflection)

(b) Unmasked (Strong Reflection)



(c) Masked (Model 1:Weak Reflection)

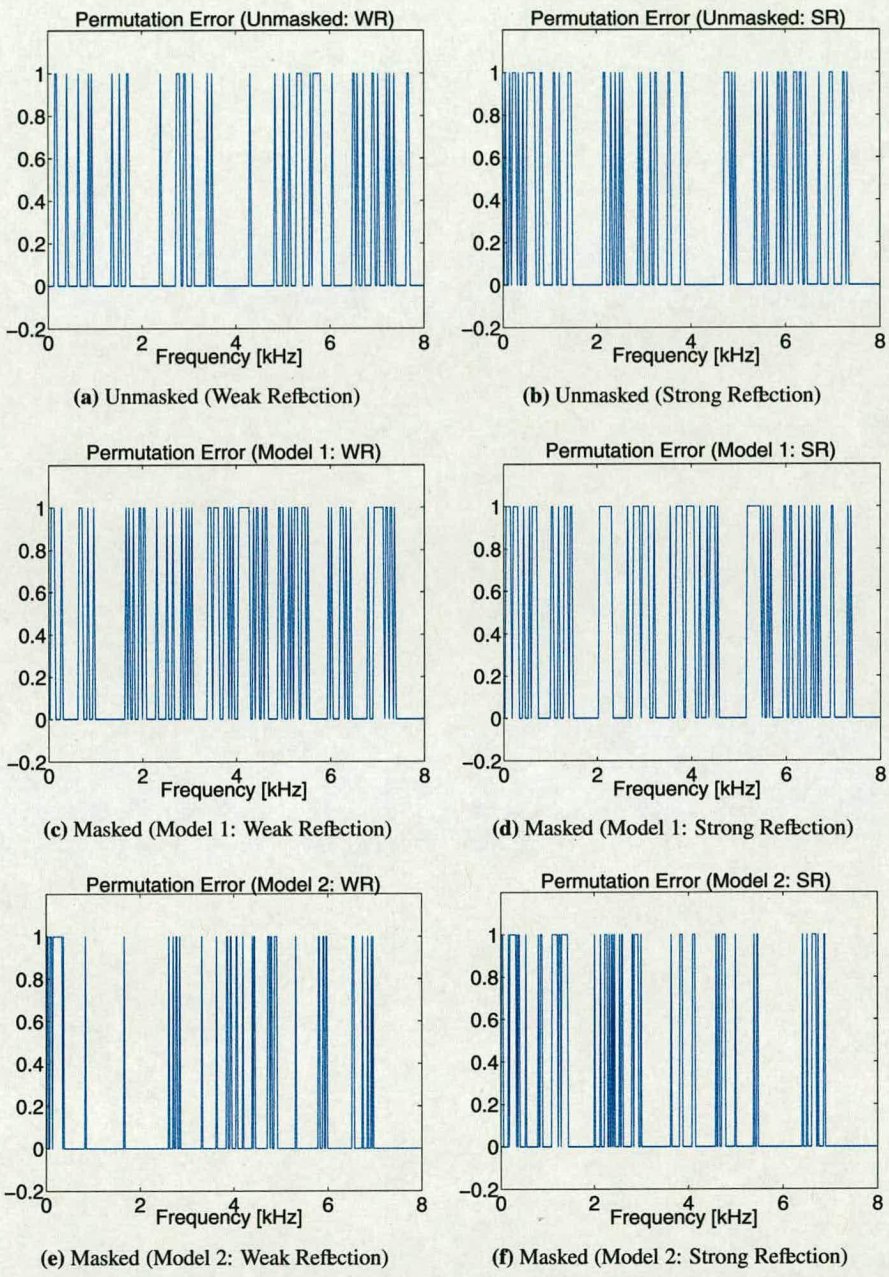
(d) Masked (Model 1:Strong Reflection)



(e) Masked (Model 2:Weak Reflection)

(f) Masked (Model 2:Strong Reflection)

Figure 3.8: *Separated Signals for Unmasked and Masked FDICA Systems (TDDA: Perceptual Preprocessing: Synthetic Room Mixing Scenario)*



(a) Unmasked (Weak Reflection)

(b) Unmasked (Strong Reflection)

(c) Masked (Model 1: Weak Reflection)

(d) Masked (Model 1: Strong Reflection)

(e) Masked (Model 2: Weak Reflection)

(f) Masked (Model 2: Strong Reflection)

Figure 3.9: *Measured Permutation Error for Unmasked and Masked FDICA Systems (TDDA: Perceptual Preprocessing: Synthetic Room Mixing Scenario)*

Finally, this is our first attempt of studying the impact of perceptual masking techniques on the speech source separation in the frequency domain while solving the permutation ambiguity problem of the FDICA system. From this study, we observed that there is a positive impact of temporal masking which is used in model 2 on reducing the permutation to some extent. On the other hand, simultaneous frequency masking has total failure to aid the source separation using time-lagged decorrelation matrices.

However, it will be helpful to utilize the continuity properties for demixing filter matrices of adjacent frequency channels in addition to the spectral envelope correlation properties of the separated signals to master the full range of realistic application scenarios. Thus, we need an iterative algorithm based on higher order statistics for solving the permutation ambiguity problem. We will apply these techniques in the next section.

3.4 Perceptually Motivated Complex Infomax Algorithm

In this section, the Infomax algorithm with a feed-forward architecture extended to complex data [27, 64, 85] to suit the requirements of perceptual auditory masking is briefly described.

In the first instance, we cannot directly apply PCA to the perceptually masked input speech. The reason is very simple. Whenever the speech vector at the output of the perceptual preprocessor (psychoacoustic model), $\mathbf{x}_f(\omega, t)$, in one of the channels contains no values, the PCA filter matrix $\mathbf{W}(\omega)$ is singular, resulting in a rank deficiency problem.

This rank deficiency problem is mainly due to very low eigenvalues of the spatial correlation matrix of the perceptually masked input speech spectrum. Without loss of generality, we assumed an identity matrix of order M as its rank to avoid this problem while retaining the whitening properties of the perceptually masked input speech.

Then, the complex Infomax algorithm is applied to the perceptually relevant output of the PCA filter, $\mathbf{y}(\omega, t)$ to obtain the ICA filter $\mathbf{U}(\omega)$. For the sake of convenience, the product of $\mathbf{W}(\omega)$ and $\mathbf{U}(\omega)$ is termed the separation filter, hereafter.

$$\mathbf{B}(\omega) = \mathbf{U}(\omega)\mathbf{W}(\omega) \tag{3.29}$$

In the ICA stage, the input signal (the output of the PCA filter) $\mathbf{y}(\omega, t)$ is processed with the

filter matrix $\mathbf{U}(\omega)$ as

$$\mathbf{z}(\omega, t) = \mathbf{U}(\omega)\mathbf{y}(\omega, t) \tag{3.30}$$

The ICA learning rule is given by

$$\mathbf{U}(\omega, t+1) = \mathbf{U}(\omega, t) + \eta[\mathbf{I} - \varphi(\mathbf{z}(\omega, t))\mathbf{z}^H(\omega, t)]\mathbf{U}(\omega, t) \tag{3.31}$$

where the score function for the complex data $\varphi(\mathbf{z})$ is defined as

$$\varphi(\mathbf{z}) = [\varphi(z_1), \cdots, \varphi(z_d), \cdots, \varphi(z_D)]^T \tag{3.32}$$

$$\varphi(z_d) = 2\tanh(G\Re(z_d)) + 2j\tanh(G\Im(z_d)). \tag{3.33}$$

Where z_d is the dth element of the vector $\mathbf{z}(\omega, t)$, $\mathbf{I}$ is an identity matrix, $.^H$ denotes the Hermitian transpose, η is the learning rate parameter and G is the gain constant for the nonlinear score function, assuming that the magnitude of $\mathbf{y}(\omega, t)$ is normalized.

3.4.1 Method of Solving Scaling and Permutation

3.4.1.1 Scaling Problem

As explained in Chapter 2 (see 2.2.3.1 for details), the scaling problem can be solved by filtering individual outputs of the separation filter by $\mathbf{B}^{-1}(\omega)$ separately [87]. This procedure will lead towards finding the scaling matrix $\widetilde{\mathbf{B}}_{\widetilde{m}}^{-1}(\omega)$.

After solving the scaling ambiguity problem, the separated output can be written in the matrix-vector notation as

$$\widetilde{\mathbf{z}}(\omega, t) = \widetilde{\mathbf{B}}_{\widetilde{m}}^{-1}(\omega)\mathbf{z}(\omega, t) \tag{3.34}$$

where $\widetilde{m}$ denotes an arbitrary microphone number.

3.4.1.2 Permutation Problem

Here, we propose a perceptually relevant method for solving the permutation problem robustly and precisely by integrating two of the approaches based on the coherency (continuity) properties of both the mixing matrix and the separated signals and hereafter denoted as the combined inter frequency correlation (CIFC) method. The first approach is the inter frequency coherency

of the mixing matrix at several adjacent frequencies (IFC), which is discussed below, as this will provide the proposed method with robustness. The IFC method is robust since its performance is not only independent of the source spectra but also associated with fixing the permutations at some frequencies where the confidence measure is sufficiently high [9, 10].

The second approach is based on the inter frequency correlations of separated output signal envelopes (IFSEC) [87], which is discussed earlier (see 3.3.1.2 for details), and will make the proposed method precise and perceptually relevant. The IFSEC method is precise and perceptually relevant as long as the masked input speech signals are well separated by ICA since the measurement is based on separated output signals. Thus, the IFSEC method is associated with deciding the permutations for the remaining frequencies based on neighboring correlations without changing the permutations fixed by the IFC method.

Hence, the proposed CIFC method has benefited from both advantages of IFC and IFSEC approaches for solving the permutation ambiguity problem of FDICA and thereby obtaining better performance of BSS system when the mixing is noisy and highly reverberant.

(i) Permutation by IFC Method

Let us consider the structure of the mixing matrix $\mathbf{A}(\omega)$ modeled as [9, 10]

$$\mathbf{A}_{m,n}(\omega) = \mathbf{H}_{m,n}(\omega)e^{-j\omega\tau_{m,n}} \tag{3.35}$$

where $\mathbf{A}_{m,n}(\omega)$ is the transfer function from the nth source to the mth microphone, $\mathbf{H}_{m,n}(\omega)$ is the magnitude of the transfer function and $\tau_{m,n}$ denotes the propagation time from the nth source to the mth microphone.

From (3.35), the nth column vector (location vector of the nth source) in $\mathbf{A}(\omega)$ at the frequency ω and that at the adjacent frequency $\omega_0 = \omega - \Delta\omega$ are

$$\mathbf{a}_n(\omega) = \begin{pmatrix} e^{-j\omega\tau_{1n}} \\ \vdots \\ e^{-j\omega\tau_{Mn}} \end{pmatrix}, \quad \mathbf{a}_n(\omega_0) = \begin{pmatrix} e^{-j(\omega-\Delta\omega)\tau_{1n}} \\ \vdots \\ e^{-j(\omega-\Delta\omega)\tau_{Mn}} \end{pmatrix}. \tag{3.36}$$

Here, $H_{m,n}(\omega) = 1$ in (3.35) is assumed for the sake of simplicity. From (3.36), the location vector $\mathbf{a}_n(\omega)$ at ω_0 is $\mathbf{a}_n(\omega_0)$ which is rotated by the angle θ_n.

Based on this coherency (relation) of the location vectors at the adjacent frequencies, the mixing matrix can be expressed as

$$\mathbf{A}(\omega) = \mathbf{T}_c(\omega, \omega_0)\mathbf{A}(\omega_0) \tag{3.37}$$

where the matrix $\mathbf{T}_c(\omega, \omega_0)$ is the complex rotation matrix.

When the difference in frequency $\Delta\omega$ (frequency resolution of STFFT) is sufficiently small,

$$\mathbf{A}(\omega) \simeq \mathbf{A}(\omega_0), \qquad \mathbf{T}_c(\omega, \omega_0) \simeq \mathbf{I} \tag{3.38}$$

the angle between the location vectors at ω and ω_0, θ_n, is expected to be the smallest for the correct permutation as shown in Fig. 3.10. Based on the coherency of $\mathbf{A}(\omega)$, the permutation problem can be solved so that the sum of the angles $\{\theta_1, \cdots, \theta_D\}$ between the location vectors in the adjacent frequencies is minimized.

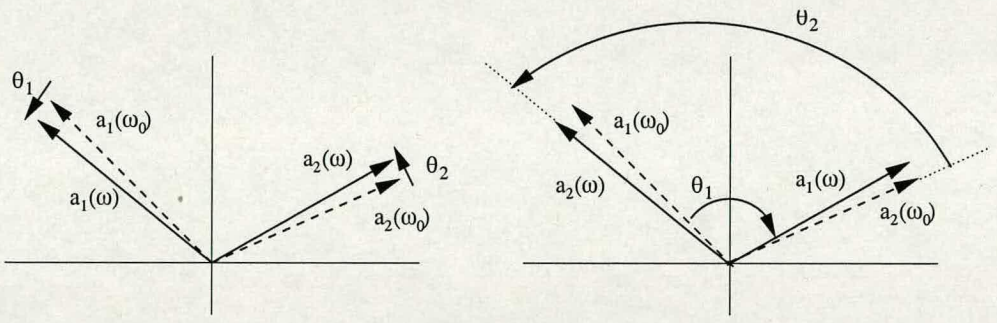
An estimate of the mixing matrix $\hat{\mathbf{A}}(\omega)$ can be obtained from $\mathbf{B}^{-1}(\omega)$ as

$$\hat{\mathbf{A}}(\omega) = \mathbf{B}^{-1}(\omega). \tag{3.39}$$

Let us denote the mixing matrix multiplied by the arbitrary permutation matrix $\mathbf{P}$ as

$$\bar{\mathbf{A}}^T(\omega) = \mathbf{P}\hat{\mathbf{A}}^T(\omega). \tag{3.40}$$

The permutation $\mathbf{P}\hat{\mathbf{A}}^T(\omega)$ exchanges the row vectors of $\hat{\mathbf{A}}^T(\omega)$ (the column vectors of $\hat{\mathbf{A}}(\omega)$). The column vectors of $\bar{\mathbf{A}}(\omega)$ are denoted as $\bar{\mathbf{A}}(\omega) = [\bar{\mathbf{a}}_1(\omega), \cdots, \bar{\mathbf{a}}_D(\omega)]$.



(a) Correct Permutation (b) Incorrect Permutation

Figure 3.10: *Rotation of the Location Vectors for Correct and Incorrect Permutations*

The cosine of the angle θ_n between the two vectors, $\bar{\mathbf{a}}_n(\omega)$ and $\bar{\mathbf{a}}_n(\omega_0)$, is defined as

$$\cos\theta_n = \frac{\bar{\mathbf{a}}_n^H(\omega)\bar{\mathbf{a}}_n(\omega_0)}{\|\bar{\mathbf{a}}_n(\omega)\|.\|\bar{\mathbf{a}}_n^H(\omega_0)\|}. \tag{3.41}$$

By doing this, the permutation matrix is determined as

$$\hat{\mathbf{P}} = \arg\max_{\mathbf{P}}[F(\mathbf{P})] \tag{3.42}$$

where the cost function $F(\mathbf{P})$ is defined as

$$F(\mathbf{P}) = \frac{1}{D}\sum_{n=1}^{D}\cos\theta_n. \tag{3.43}$$

The above method assumes that the estimate of the mixing matrix $\hat{\mathbf{A}}(\omega)$ is a good approximation of the true mixing matrix $\mathbf{A}(\omega)$. However, at some frequencies, this assumption may not hold due to the failure of ICA. Since the permutation at frequency ω is determined based on only the information of the two adjacent frequencies, ω and ω_0, and the permutation is solved iteratively with increasing frequency, once the permutation at the certain frequency fails, the permutation in the succeeding frequencies may also fail.

To prevent this, the reference frequency ω_0 is extended to the following frequency range:

$$\omega_0 = \omega - k.\Delta\omega, \qquad \text{for} \quad k = 1, \cdots, K. \tag{3.44}$$

The cost function $F(\mathbf{P})$ is calculated at all K frequencies in this range. Let us denote the value of the cost function at $\omega_0 = \omega - k.\Delta\omega$ as $F(\mathbf{P}, k)$. Next, a confidence measure for $F(\mathbf{P}, k)$ is considered. When the largest value of the cost function $\max F(\mathbf{P}, k)$ is close to $F(\mathbf{P}, k)$ with other permutations, it may be difficult to determine which permutation is correct, and the value of $F(\mathbf{P}, k)$ is not reliable. Based on this, the following confidence measure is defined as:

$$C(k) = \max_{\mathbf{P}\in\Omega}[F(\mathbf{P}, k)] - \max_{\mathbf{P}\in\Omega'}[F(\mathbf{P}, k)] \tag{3.45}$$

Here, Ω denotes the set of all possible $\mathbf{P}$ while Ω' denotes Ω without $\hat{\mathbf{P}} = \arg\max_{\mathbf{P}\in\Omega}[F(\mathbf{P}, k)]$. The appropriate reference frequency ω_0 is determined as $\omega_0 = \omega - \hat{k}.\Delta\omega$ with

$$\hat{k} = \max_{\mathbf{P}}[C(k)]. \tag{3.46}$$

The permutation is then solved using the information at this reference frequency as

$$\hat{\mathbf{P}} = \arg\max_{\mathbf{P}}[F(\mathbf{P}, \hat{k})] \tag{3.47}$$

(ii) Permutation by IFSEC Method

As explained earlier, the IFSEC method [87] decides the permutations for the remaining frequencies where the permutation is not fixed by the IFC method. This IFSEC method does not cause a large misalignment as long as the permutations fixed by the IFC method are correct.

As indicated in (3.41) and (3.43), the cost function of the IFC method is given by

$$F_{IFC}(\mathbf{P}) = \frac{1}{D}\sum_{n=1}^{D} \frac{\bar{\mathbf{a}}_n^H(\omega)\bar{\mathbf{a}}_n(\omega_0)}{\|\bar{\mathbf{a}}_n(\omega)\|.\|\bar{\mathbf{a}}_n^H(\omega_0)\|}. \tag{3.48}$$

On the other hand, the cost function of the IFSEC method can be expressed as

$$F_{IFSEC}(\mathbf{P}) = \frac{1}{D}\sum_{n=1}^{D} \frac{\bar{\mathbf{z}}_n^H(\omega)\bar{\mathbf{z}}_n(\omega_0)}{\|\bar{\mathbf{z}}_n(\omega)\|.\|\bar{\mathbf{z}}_n^H(\omega_0)\|}. \tag{3.49}$$

In the IFSEC method also, the cost function is maximized in the same manner as that of the IFC method for solving the permutation. The vector $\bar{\mathbf{z}}_n$ is the nth column vector of the following matrix similar to that of (3.40)

$$\bar{\mathbf{Z}}^T(\omega) = \mathbf{P}\hat{\mathbf{Z}}^T(\omega). \tag{3.50}$$

The matrix $\hat{\mathbf{Z}}^T(\omega)$ has the estimated spectral envelope of the separated output (smoothed by the moving-average) obtained by ICA and expressed as a column vector as

$$\hat{\mathbf{Z}}(\omega) = [\hat{\mathbf{z}}_1(\omega), \cdots, \hat{\mathbf{z}}_D(\omega)] \tag{3.51}$$

where $\hat{\mathbf{z}}_n(\omega) = [\hat{z}_n(\omega, t_1), \cdots, \hat{z}_n(\omega, t_2)]$, $\hat{z}_n(\omega, t)$ is the estimated spectral envelope at the nth channel, frequency ω and tth time frame and $[t_1, t_2]$ describe the period of the spectrogram.

Thus by considering the merits of both the inter-frequency coherency (IFC) and the inter-frequency spectral envelope correlation (IFSEC) methods, we realize a new and the perceptually relevant method denoted as combined inter-frequency correlation (CIFC) for solving the permutation ambiguity problem of the proposed FDICA system.

3.4.2 Overall Filtering System

After solving the permutation and scaling ambiguities, the overall (final) reconstructed filtering matrix in the frequency domain can be written as

$$\mathbf{F}(\omega) = \mathbf{P}(\omega)\widetilde{\mathbf{B}}_{\widetilde{m}}^{-1}(\omega)\mathbf{B}(\omega). \tag{3.52}$$

Thus, the reconstructed time domain filters are obtained as the inverse Fourier transform of $\mathbf{F}(\omega)$ as

$$f_{n,m}(i) = \mathbf{IFFT}[F_{n,m}(\omega)]w(i) \tag{3.53}$$

where **IFFT**[.] operator denotes the inverse **FFT**. The symbols $F_{n,m}(\omega)$ and $f_{n,m}(i)$ denote the (n,m)th element of the frequency domain filter $\mathbf{F}(\omega)$ and its time domain correspondence, respectively. The symbol $w(i)$ denotes the windowing function.

The multiplication by $w(i)$ is necessary to control the spectral leakage and also to minimise its effect on reducing the dynamic range capability of the transform output and thereby reconstructing the speech signal in the time domain.

3.4.3 Experimental Results

3.4.3.1 Synthetic Room Mixing Scenario

As explained earlier (see 3.3.3.1 for details), we created a synthetic room mixing of two speech sources (4 s at 16 kHz) using Asano's [139] conference room (5.9 m×8.7 m×4.1 m) with a reverberation time of 0.5 sec for both weak and strong reflections of room filters (Fig. 3.5). We applied the complex Infomax algorithm for both reflection cases, using the parameters of the proposed BSS system that are summarized in Table 3.2. For achieving statistical reliability with a better convergence, the experiment is repeated over 100 times of the data.

Sampling Frequency	16 kHz
STFFT Frame Length	512
Shift of STFFT	16
Window Function	Hamming
Learning Rate, η	0.0001
Gain for Score Function, G	100
Normalized Sound Pressure Level, SPL	96 dB
Number of Microphones, M	2
Number of Sources, D	2
Reference Range in Permutation, K	5

Table 3.2: ***Proposed BSS System-2 Parameters (Infomax: Perceptual Preprocessing)***

Original speech sources, observed signals and the separated signals are shown in Fig. 3.11. Further, each of the original source is divided into eight segments ($A1, B1, \cdots, H1$ in the case of the first source and $A2, B2, \cdots, H2$ in the case of the second source) for simplifying the comparitive analysis of each category of the above mentioned speech signals. These signal segments will help us to compare each of the separated (reconstructed) speech signals as to whether they resemble the shape of the original speech sources or the observed signals.

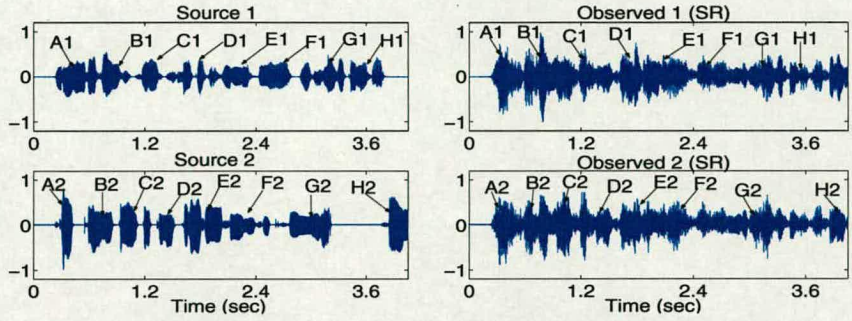
From Fig. 3.11(c), it is evident that the segments i.e., A1, C1, D1, F1, G1 and H1; A2, B2, C2, D2, E2 and H2 of the first and the second separated speech signals obtained by unmasked FDICA system respectively are similar to the original sources when the weak early reflection case is considered. The remaining segments i.e., B1, E1, F2 and G2 are remain mixed. However, these separated signals have some crosstalk whenever the original speech sources have zero or minmium signal strength (see Fig. 3.11(a)). On the other hand, the separated signals obtained by unmasked FDICA system (shown in Fig. 3.11(d)) are similar to the observed signals (shown in Fig. 3.11(b)) when the strong reflection case is considered.

From Fig. 3.11(e), it is also evident that the separated speech signals in one of the channels (second output signal) obtained by masked FDICA system (using psychoacoustic model 1) is different from the original source when weak early reflection case is considered. However, the segments A2, C2, D2, E2 and H2 of the separated speech signal are slightly different from those of the original speech signal. The remaining segments i.e., B2, F2 and G2 still appear as a mixed signal. From Fig. 3.11(f), it can be seen that the second separated speech signal obtained by masked FDICA system (using psychoacoustic model 1) is slightly different from the observed speech signal when strong early reflection case is considered. However, the segments A2, D2, F2 and G2 of the separated speech signal are slightly different from those of the original speech signal. The remaining segments i.e., B2, C2, E2 and H2 still appear as a mixed signal due to the presence of strong reflections in the separated output.

From Figs. 3.11(g) and 3.11(h), it is also evident that the separated speech signals in one of the channels (second output in this case) obtained by masked FDICA system (using psychoacoustic model 2) is slightly different from the original speech source (with reference to most of the signal segments i.e., $A2, B2, \cdots, H2$) when both weak and strong early reflection cases are considered. Though, these segments $A2, B2, \cdots, H2$ of the separated speech signal are better than those obtained by the psychoacoustic model 1 under similar experimental conditions, still there is some crosstalk due to reflective environment. However, the separated signal in the first channel is similar to the observed signal when both reflection cases of unmasked and masked systems (using either model) are taken into account.

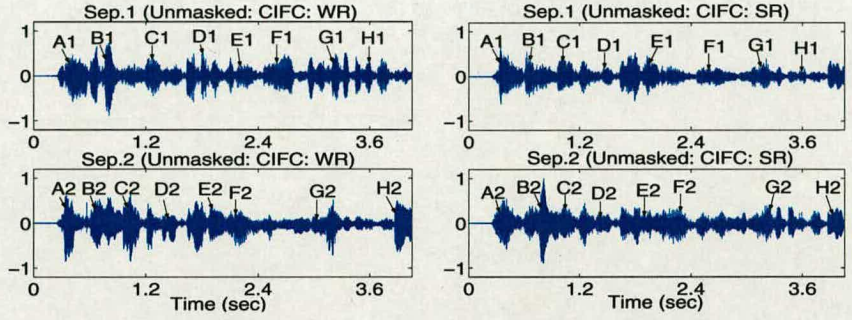
The spectrograms of original speech sources, observed speech signals and the separated speech signals are shown in Fig. 3.12. Further, the speech signal frequency range (with a bandwidth of 5 kHz) is divided into two frequency bands namely F_1 (0-3 kHz) and F_2 (3-5 kHz) to simplify the comparative analysis of the above mentioned spectrograms. From Figs. 3.12(e) and 3.12(f), it is clearly observed that most of the higher frequency components ($>$ 3 kHz) of the second separated speech signal spectrum are masked when the perceptual preprocessor (using the psychoacoustic model 1) is used for both early reflections of the FDICA system.

However, the psychoacoustic model 2 based preprocessing filter not only masks the higher frequencies ($>$ 3 kHz) but also masks some frequencies in the range of 1-3 kHz when both weak and strong reflection cases of the masked FDICA system are taken into account (shown in Figs. 3.12(g) and 3.12(h)). Results are highlighted in Figs. 3.13, 3.14, 3.15 and 3.16 for the worst case scenario of highly reverberant environemnt.



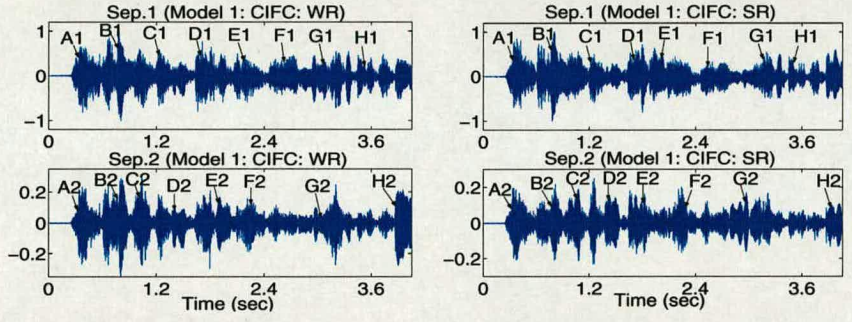
(a) Original Speech Signals

(b) Observed Speech Signals



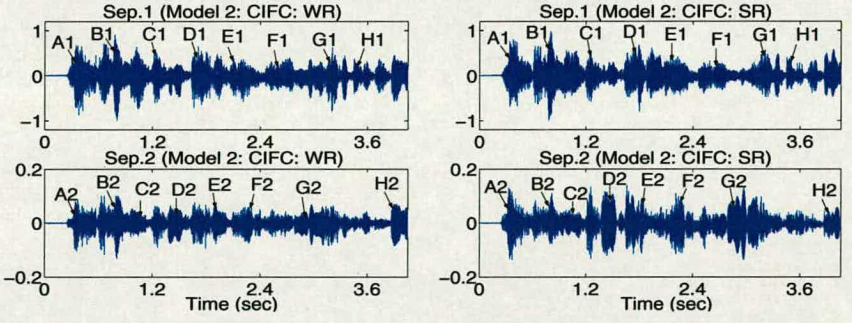
(c) Unmasked (Weak Reflection)

(d) Unmasked (Strong Reflection)



(e) Masked (Model 1:Weak Reflection)

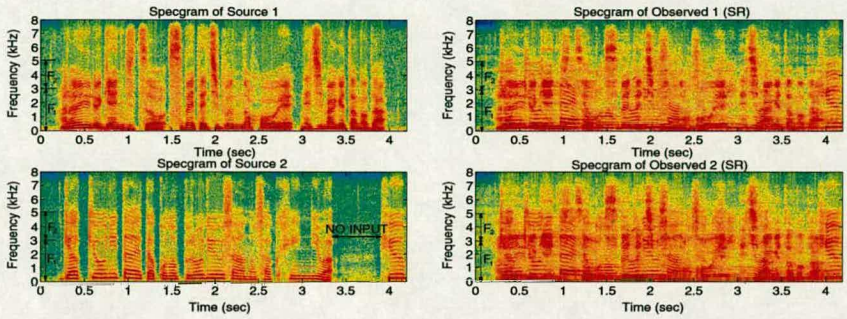
(f) Masked (Model 1:Strong Reflection)



(g) Masked (Model 2:Weak Reflection)

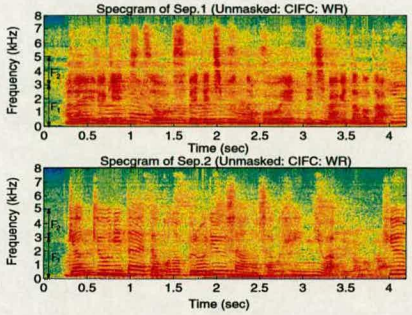
(h) Masked (Model 2:Strong Reflection)

Figure 3.11: *Original Sources, Observed and Separated Signals for Unmasked and Masked Systems (Infomax: Perceptual Preprocessing: Synthetic Room Mixing Scenario)*

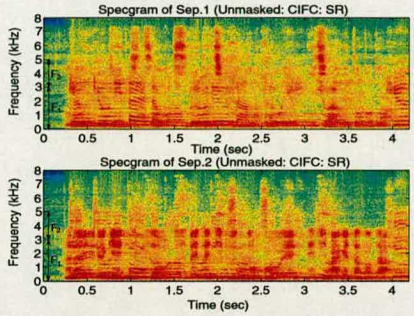


(a) Original Speech Signals

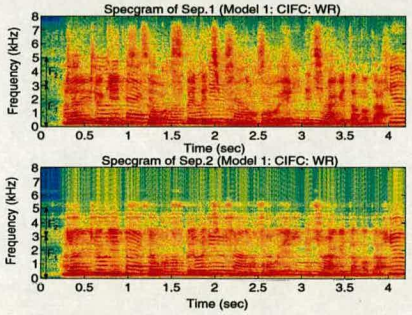
(b) Observed Speech Signals



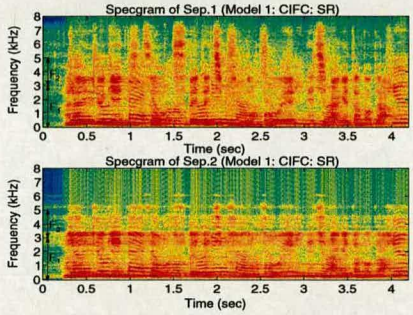
(c) Unmasked (Weak Reflection)



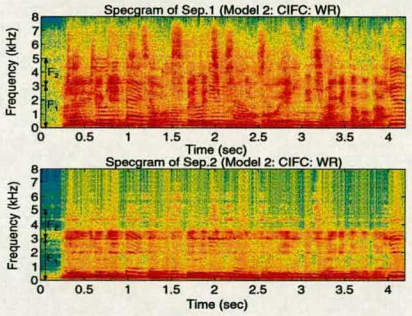
(d) Unmasked (Strong Reflection)



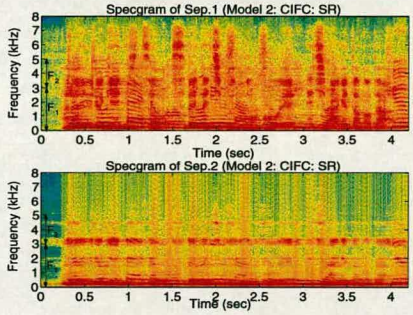
(e) Masked (Model 1:Weak Reflection)



(f) Masked (Model 1:Strong Reflection)

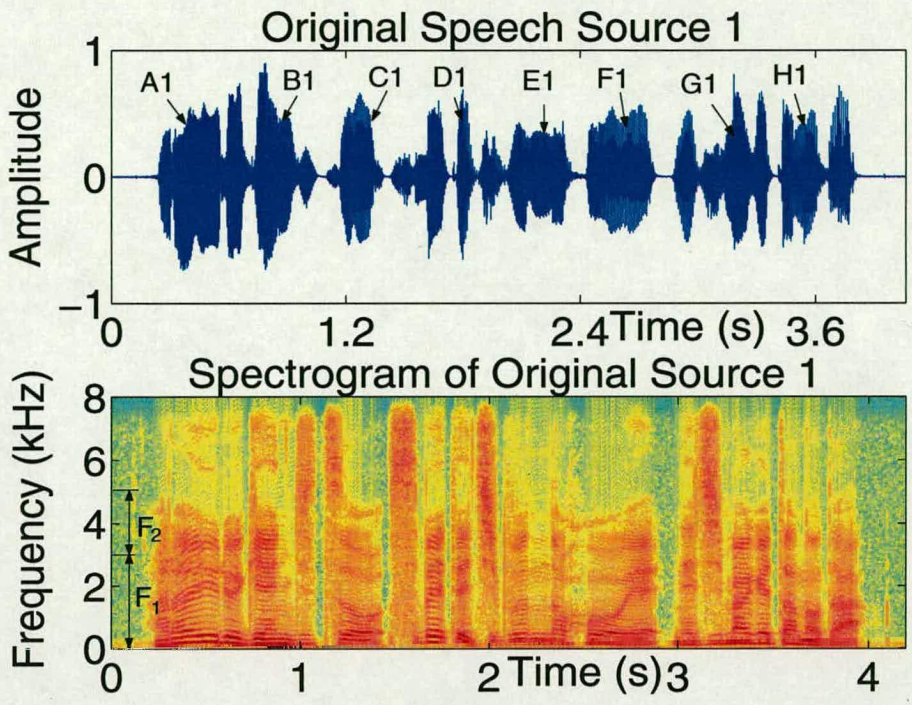


(g) Masked (Model 2:Weak Reflection)

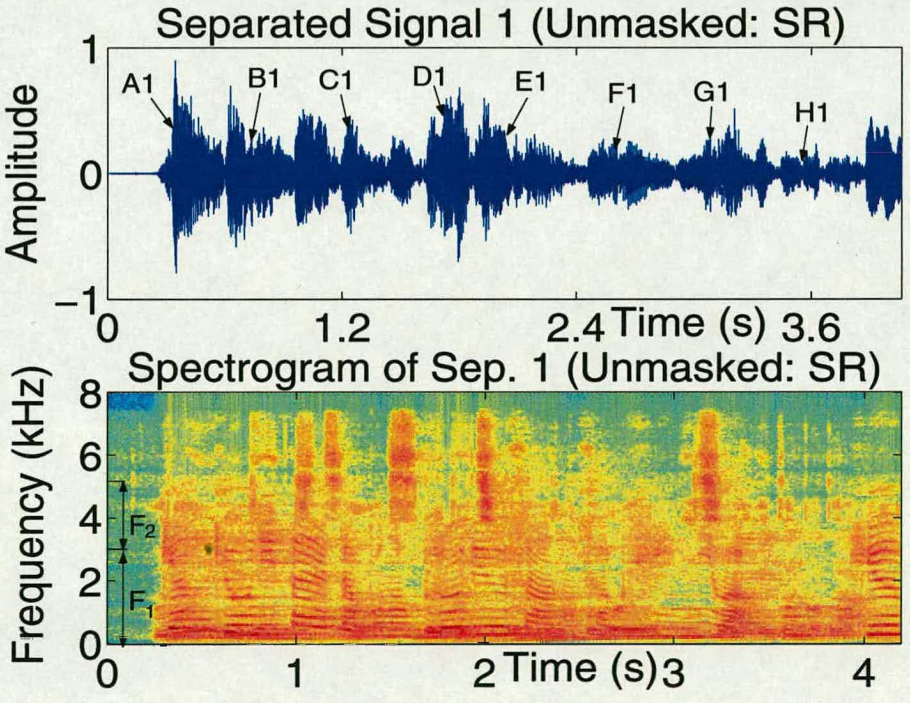


(h) Masked (Model 2:Strong Reflection)

Figure 3.12: *Specgrams of Sources, Sensors and Separated Signals for Unmasked and Masked Systems (Infomax: Perceptual Preprocessing: Synthetic Room Mixing Scenario)*

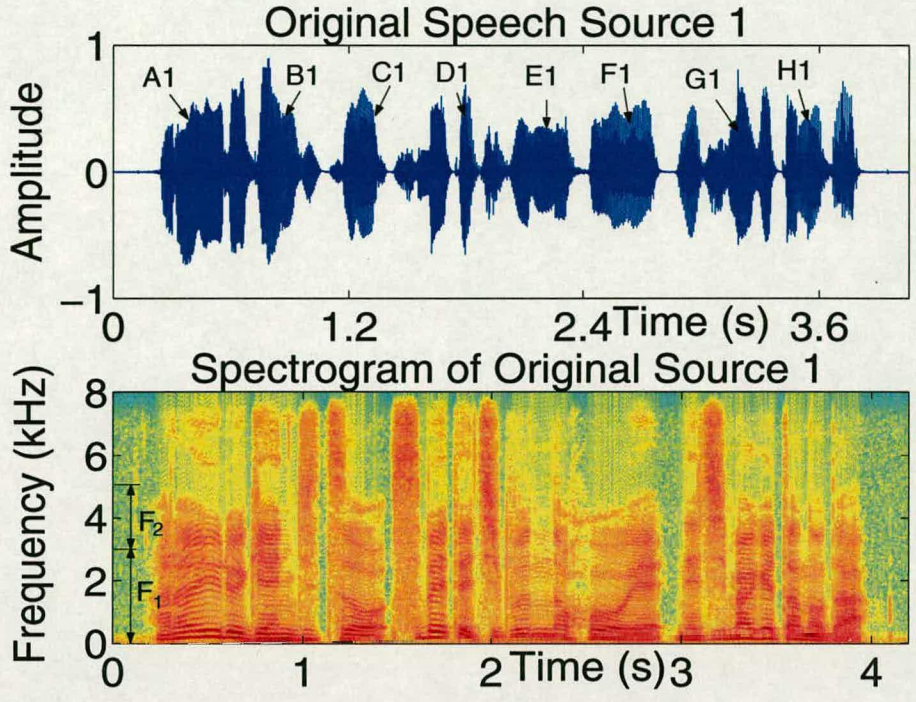


(a) Source 1 and its Spectrogram

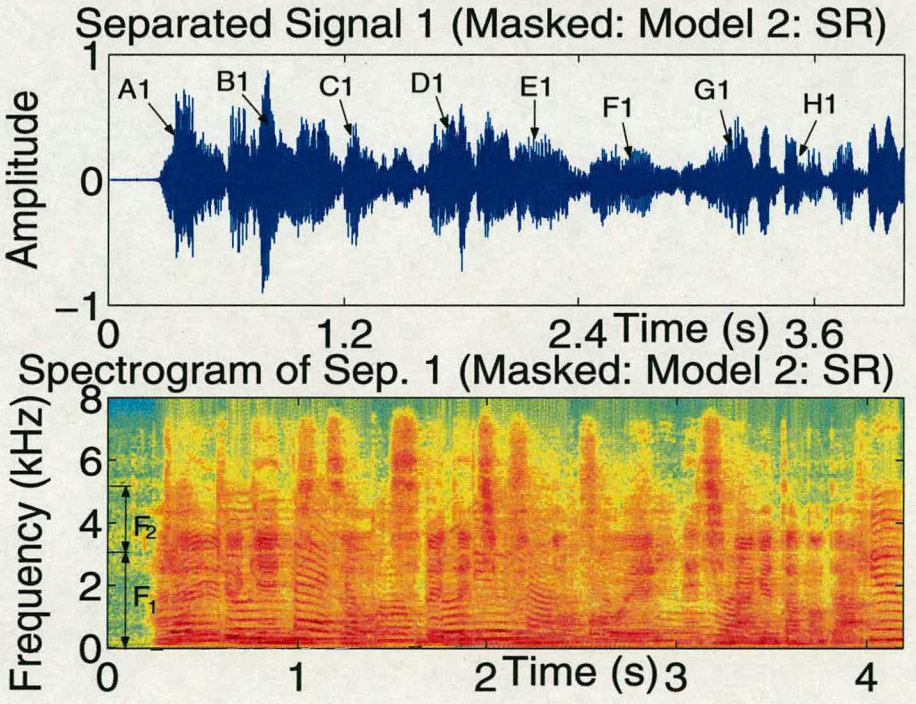


(b) Sep. 1 and its Spectrogram (Unmasked)

Figure 3.13: *Separated Signal 1 and its Spectrogram for the Unmasked FDICA System When Speech Source 1 and its Spectrogram are Known (Strong Reflection Case)*

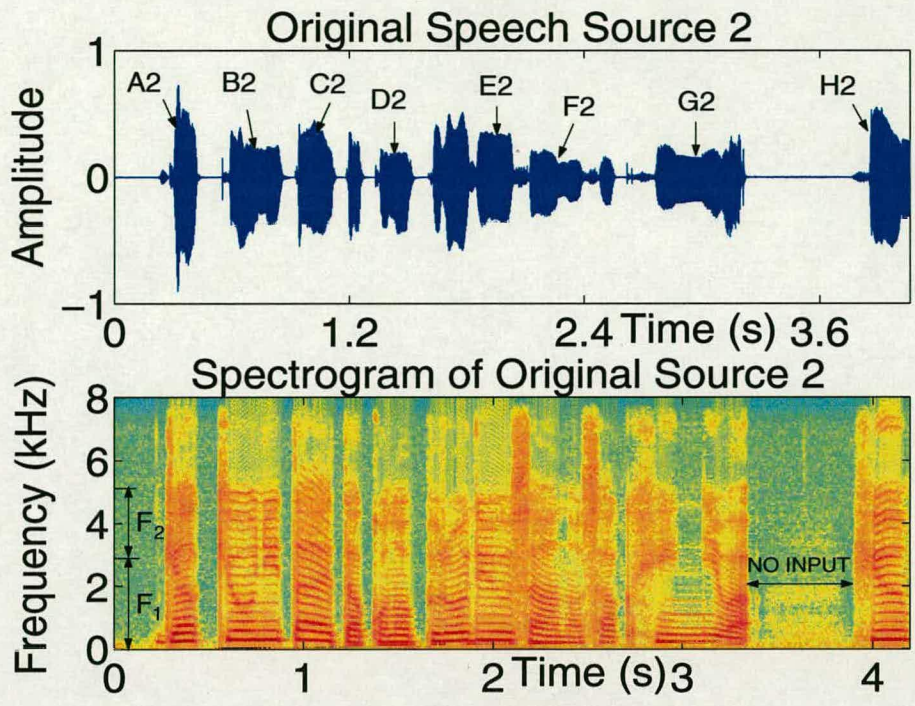


(a) Source 1 and its Spectrogram

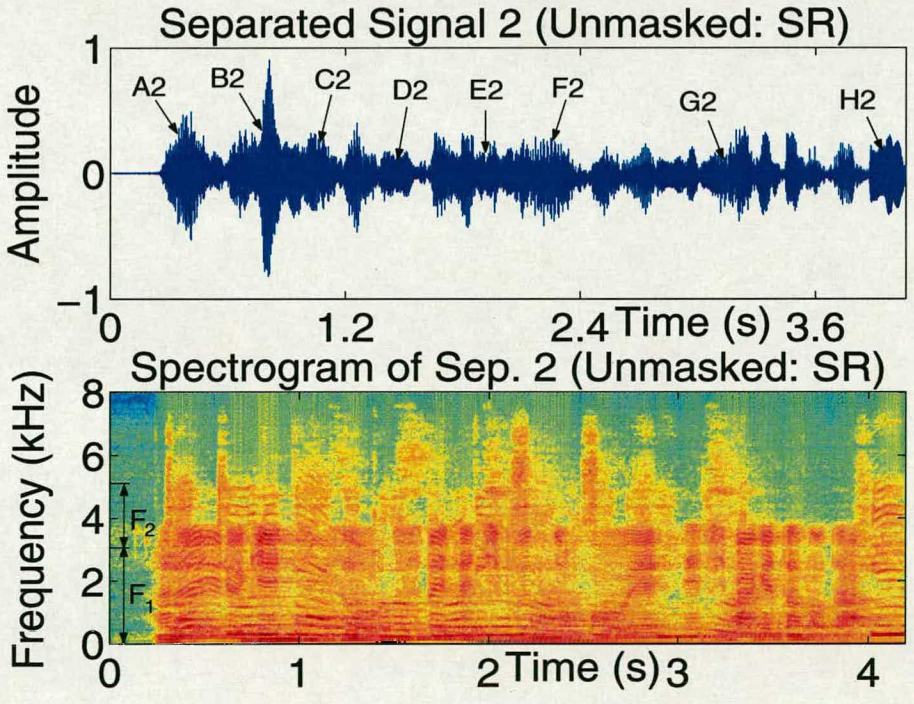


(b) Sep. 1 and its Spectrogram (Masked)

Figure 3.14: *Separated Signal 1 and its Spectrogram for the Masked FDICA System When Speech Source 1 and its Spectrogram are Known (Strong Reflection Case)*

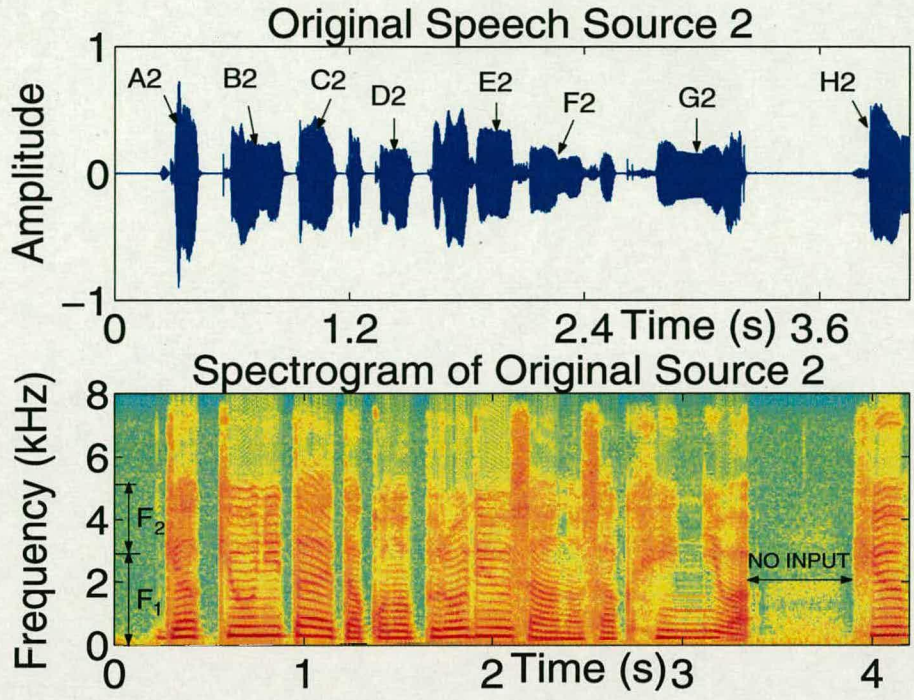


(a) Source 2 and its Spectrogram

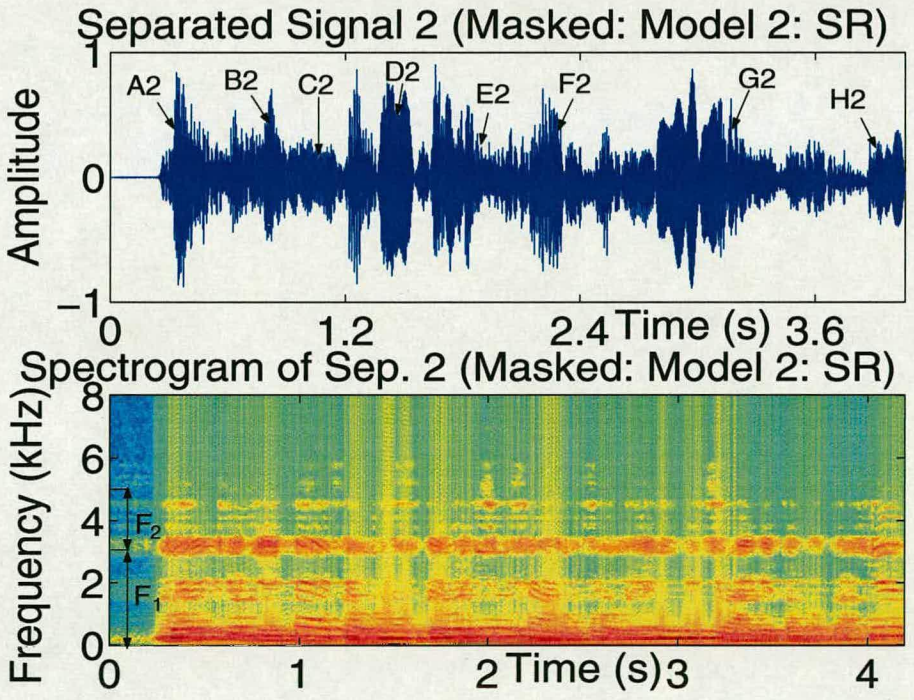


(b) Sep. 2 and its Spectrogram (Unmasked)

Figure 3.15: *Separated Signal 2 and its Spectrogram for the Unmasked FDICA System When Speech Source 2 and its Spectrogram are Known (Strong Reflection Case)*



(a) Source 2 and its Spectrogram



(b) Sep. 2 and its Spectrogram (Masked)

Figure 3.16: *Separated Signal 2 and its Spectrogram for the Masked FDICA System When Speech Source 2 and its Spectrogram are Known (Strong Reflection Case)*

From Fig. 3.17, it can be seen that the measured value of the cost function $F(\mathbf{P}, k)$ shows a smaller value at all frequencies except in the very few frequencies when the unmasked system is considered for both weak and strong reflection cases. These smaller values of the cost function are represented by vertical lines. These vertical lines show that it is necessary to exchange the output at those frequencies where the permutation problem still exists.

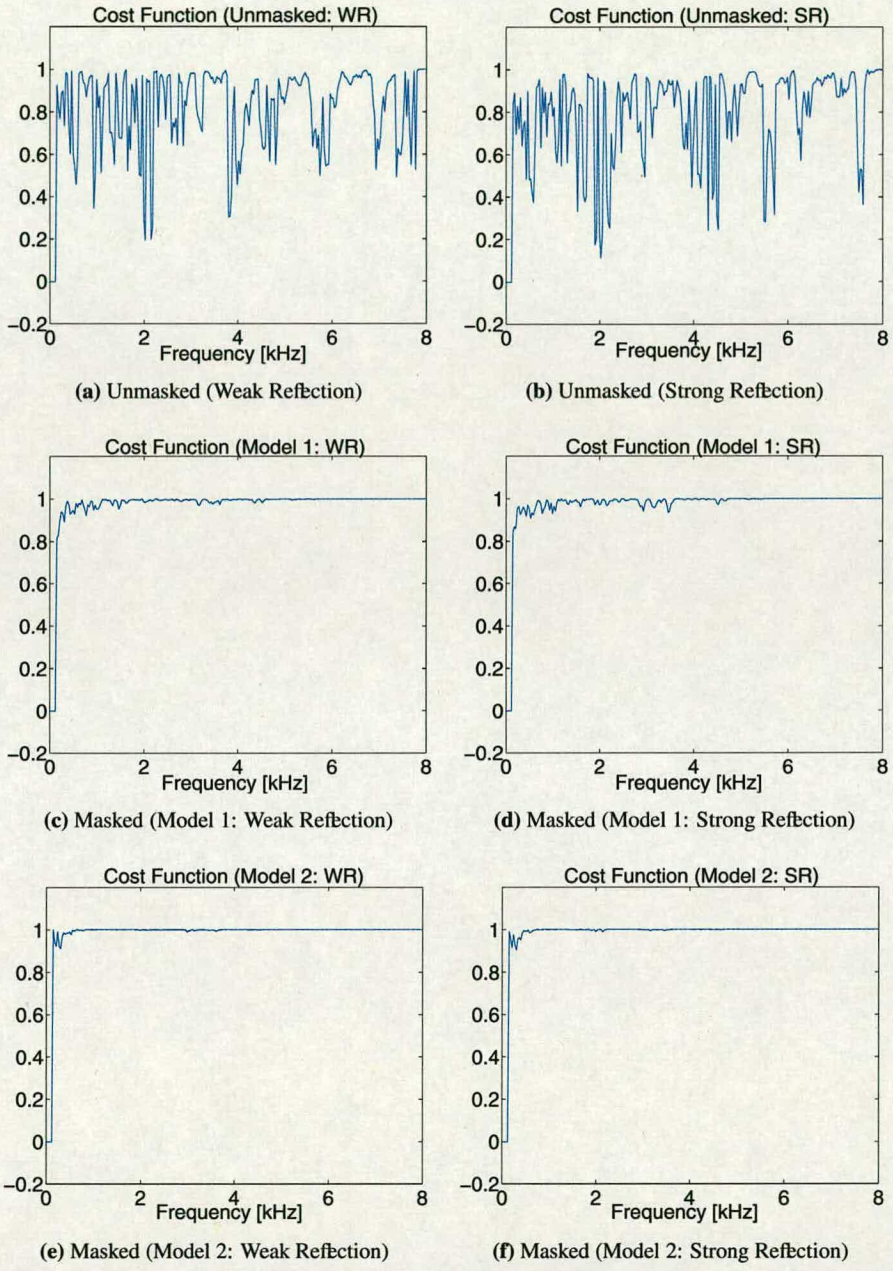
On the other hand, the cost function is close to unity for all the frequencies except the very low frequencies when the masked system using psychoacoustic model 1 is considered for both early reflection cases. Further, there is an improvement in the measured value of the cost function at these low frequencies also when the masked FDICA system (using the model 2) is considered for both early reflection cases. Thus, the permutation ambiguity problem encountered in an unmasked FDICA system at most of the frequencies is mitigated by the proposed FDICA system that employs both perceptual masking techniques.

The confidence measure $C(k)$ depicted in Fig. 3.18 has a smaller value for most of the frequencies when the unmasked system is used for both cases of reflection. Further, the measured value of confidence measure has high values at all frequencies except the low frequencies when the perceptually motivated FDICA system (using psychoacoustic model 1) is employed for both early reflection cases. However, there is an improvement in the value of the confidence measure at most of the low frequencies when the masked FDICA system (using the model 2) is considered for both reflection cases. Thus, the perceptually motivated FDICA system helps in increasing the confidence at those frequencies where the permutation actually occured.

Permutation error is defined as the case when the result of CIFC differs from that of source output crosscorrelation (SOC) (assumed as correct permutation). The measured permutation error (Fig. 3.19) is 7.4% and 46.7% for both weak and strong reflection cases of unmasked system respectively. On the other hand, the permutation error is zero for all the frequencies when a masked system (using either model) is used for both reflection cases.

3.4.3.2 Real Room Mixing Scenario

This experiment was chosen to test the algorithm's ability in a real room recording environment using recorded speech signals (6 s at 16 kHz). Real room mixing results for both unmasked and masked systems shown in Figs. 3.20 and 3.21 are similar to that of the synthetic mixing case. The permutation error cannot be computed in this case as original sources are unknown.



(a) Unmasked (Weak Reflection)

(b) Unmasked (Strong Reflection)

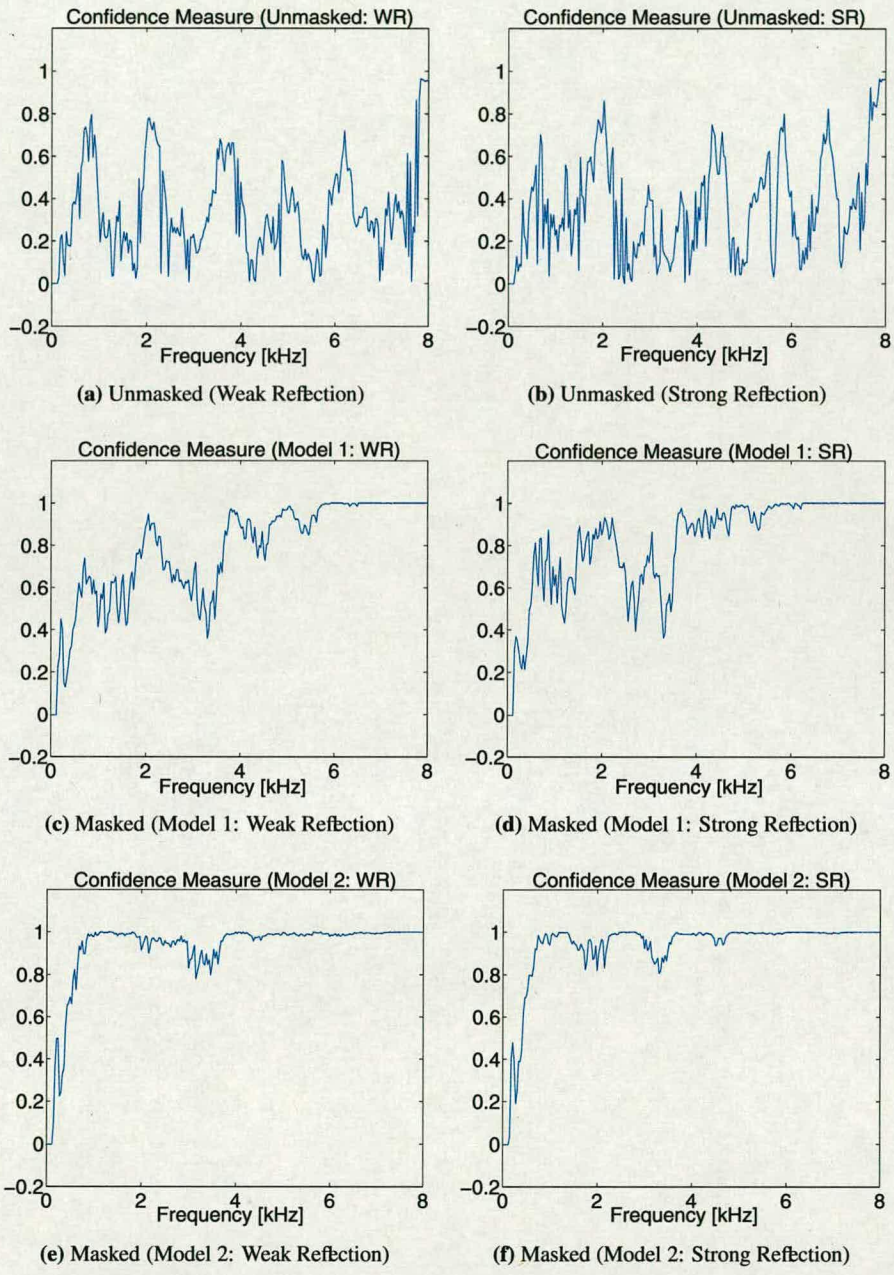
(c) Masked (Model 1: Weak Reflection)

(d) Masked (Model 1: Strong Reflection)

(e) Masked (Model 2: Weak Reflection)

(f) Masked (Model 2: Strong Reflection)

Figure 3.17: *Measured Cost Function for Unmasked and Masked FDICA Systems (Infomax: Perceptual Preprocessing: Synthetic Room Mixing Scenario)*



(a) Unmasked (Weak Reflection)

(b) Unmasked (Strong Reflection)

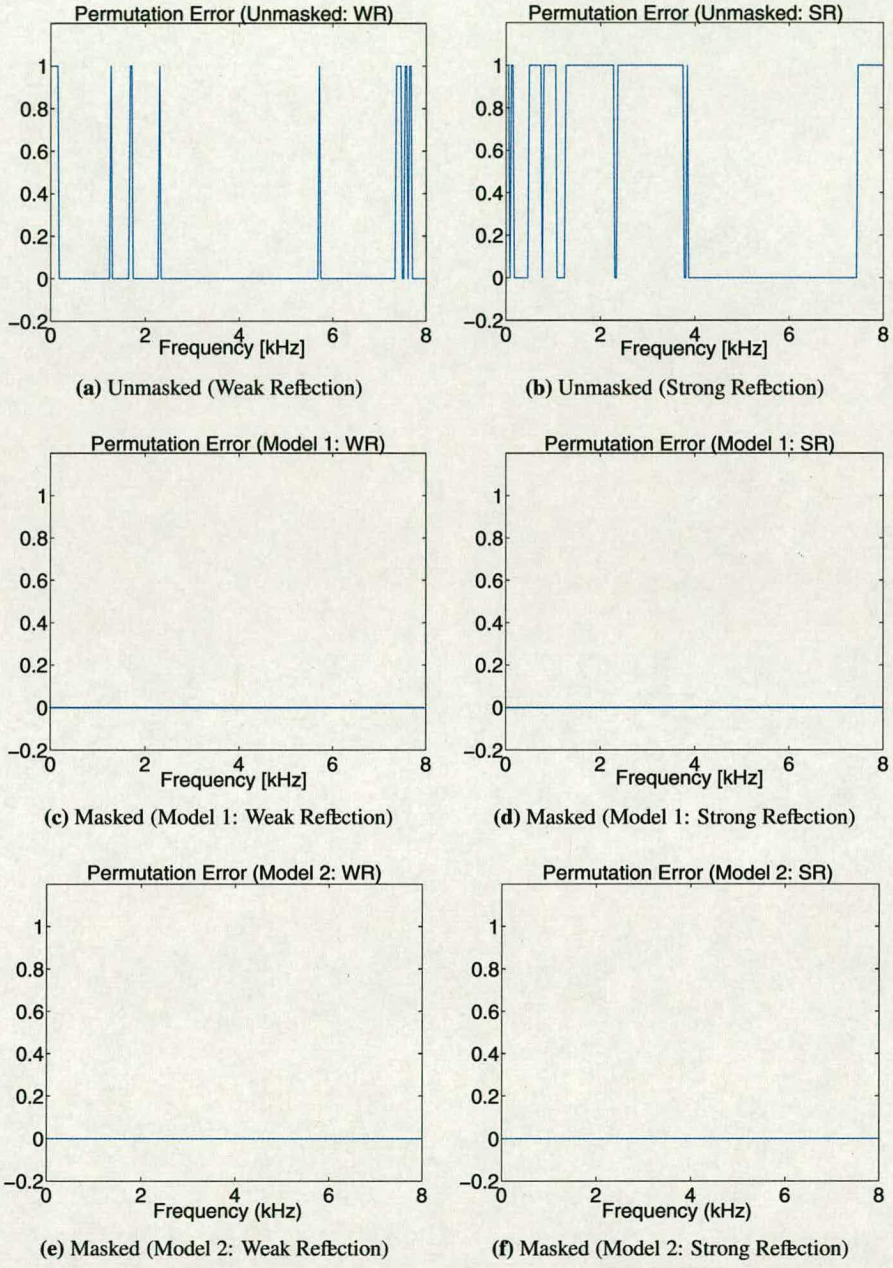
(c) Masked (Model 1: Weak Reflection)

(d) Masked (Model 1: Strong Reflection)

(e) Masked (Model 2: Weak Reflection)

(f) Masked (Model 2: Strong Reflection)

Figure 3.18: *Measured Confidence Measure for Unmasked and Masked FDICA Systems (Infomax: Perceptual Preprocessing: Synthetic Room Mixing Scenario)*



(a) Unmasked (Weak Reflection)

(b) Unmasked (Strong Reflection)

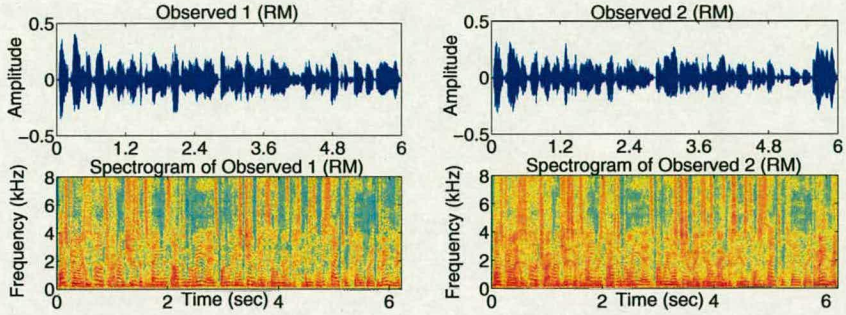
(c) Masked (Model 1: Weak Reflection)

(d) Masked (Model 1: Strong Reflection)

(e) Masked (Model 2: Weak Reflection)

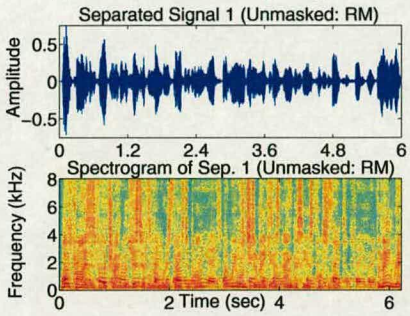
(f) Masked (Model 2: Strong Reflection)

Figure 3.19: *Measured Permutation Error for Unmasked and Masked FDICA Systems (Infomax: Perceptual Preprocessing: Synthetic Room Mixing Scenario)*

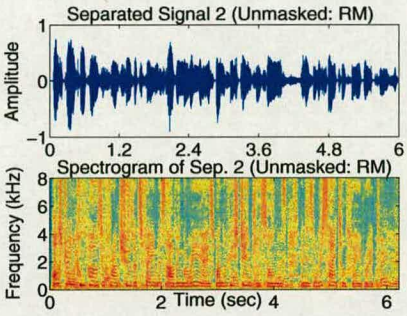


(a) Observed 1 and its Spectrogram

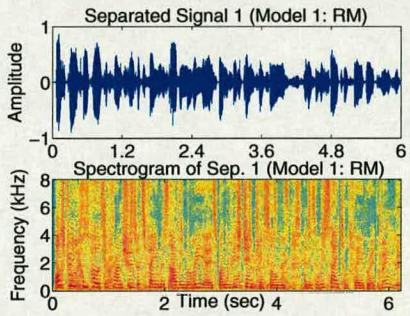
(b) Observed 2 and its Spectrogram



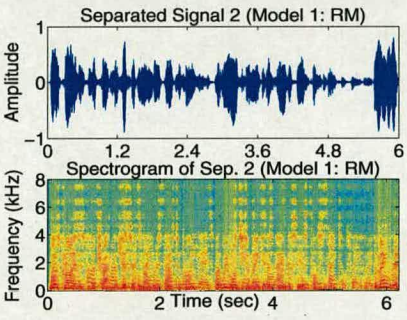
(c) Sep. 1 and its Spectrogram (Unmasked)



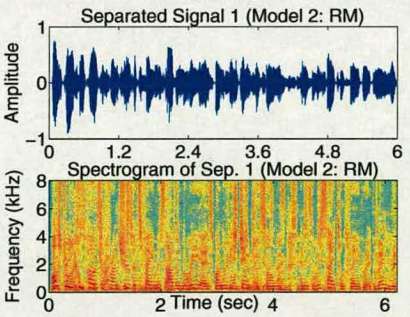
(d) Sep. 2 and its Spectrogram (Unmasked)



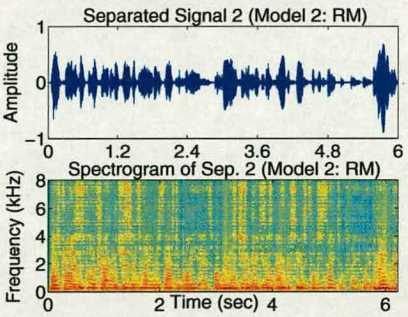
(e) Sep. 1 and its Spectrogram (Model 1)



(f) Sep. 2 and its Spectrogram (Model 1)

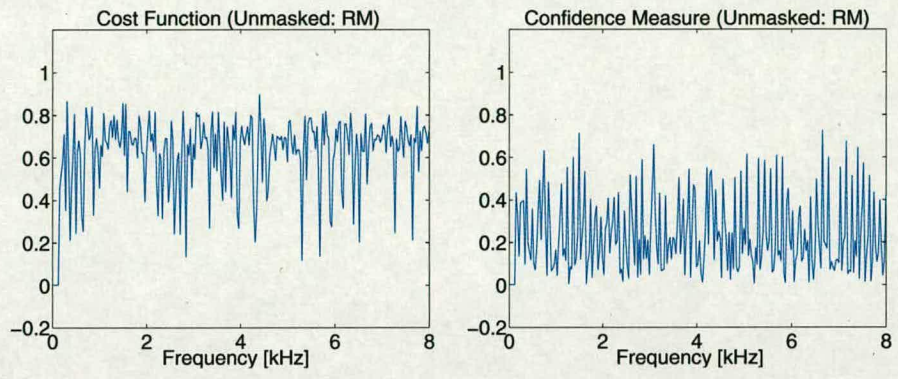


(g) Sep. 1 and its Spectrogram (Model 2)



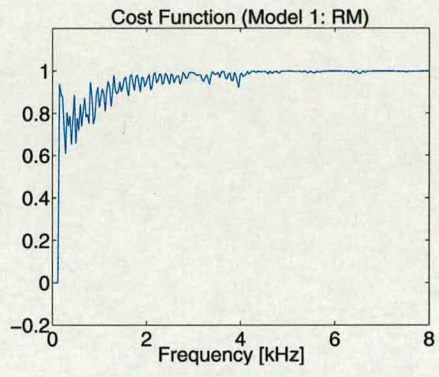
(h) Sep. 2 and its Spectrogram (Model 2)

Figure 3.20: *Observed Signals, Separated Signals and Spectrograms for both Unmasked and Masked Systems (Infomax: Perceptual Preprocessing: Real Room Recording)*

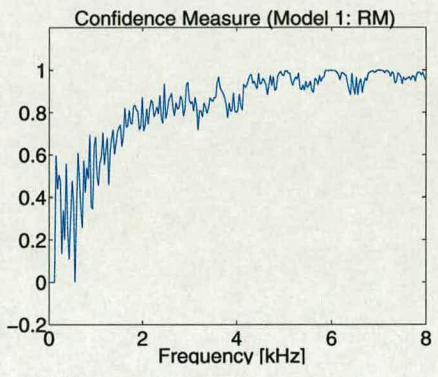


(a) Unmasked (Real Room)

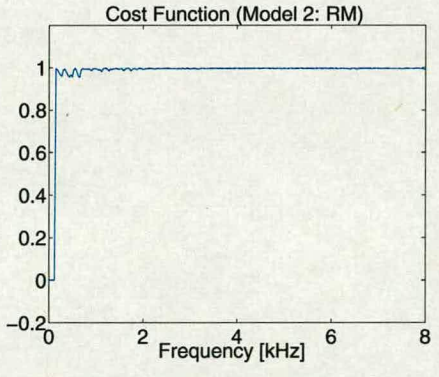
(b) Unmasked (Real Room)



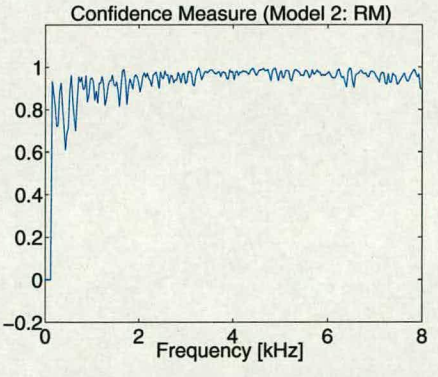
(c) Masked (Model 1: Real Room)



(d) Masked (Model 1: Real Room)



(e) Masked (Model 2: Real Room)



(f) Masked (Model 2: Real Room)

Figure 3.21: *Cost Function and Confidence Measure for Unmasked and Masked FDICA Systems (Infomax: Perceptual Preprocessing: Real Room Mixing Scenario)*

3.5 Performance Evaluation

The performance evaluation of a BSS system can be achieved by subjective and objective quality measures [140–147]. Subjective speech quality is performed by human listeners. Such listening tests are expensive, time-consuming and difficult to administer. Further, such tests seldom provide much insights into the factors which may lead to improvement in the evaluation. The Mean Opinion Score (MOS) has been the usual subjective speech quality test used to evaluate objective quality measures. In a MOS test, listeners are not provided with an original sample and rate the overall speech quality of the separated sample [145, 147].

However, objective speech quality measures estimate subjective scores by comparing the separated speech to the original speech, which has more in common with a Degradation Mean Opinion Score (DMOS) test in which listeners listen to an original speech sample prior to each separated speech sample [145, 147]. Although objective speech quality measures are not expected to completely replace subjective speech quality measures, a good objective measure would be a valuable assessment tool for BSS system.

Objective quality measures can be classified according to the domain in which they estimate the distortion: time domain, spectral domain and perceptual domain. Time domain measures are usually applicable to BSS systems in which the goal is to reproduce the signal waveform. Spectral domain measures are mainly based on speech production models and their performance is limited by the failure of speech production models to adequately describe the listener's auditory response. Perceptual domain measures transform the speech signal into a perceptually relevant domain incorporating human auditory models. Hence, perceptual domain measures would appear to have the best chance of predicting subjective quality of speech.

Here, we are mainly focusing on two important objective speech quality measures namely time-domain and perceptual domain measures and their metrics are signal-to-interference ratio (SIR) and Enhanced Modified Bark Spectral Distortion (EMBSD) respectively [42, 144].

Since the time-delayed decorrelation algorithm failed to solve the permuatation problem, we did not consider it for performance evaluation of both unmasked and masked FDICA systems. Henceforth, the complex Infomax algorithm with two permutation solving strategies namely: the source-output cross-correlation (SOC) method and a combined approach of inter frequency coherency of separated filter and output spectral envelope (CIFC) method are considered for evaluating the performance of both unmasked and masked FDICA systems.

3.5.1 Time-Domain Objective Quality Measure

When the original speech signals are $s_i(t)(i = 1, ..., D)$, the signals observed by sensor j are $x_j(t)(j = 1, ..., M)$ and the separated signals are $y_k(t)(k = 1, ..., D)$, the convolutive BSS model can be: $x_j(t) = \sum_{i=1}^{D}(a_{ji} * s_i)(t)$, $y_k(t) = \sum_{j=1}^{M}(f_{kj} * x_j)(t)$, where a_{ji} is the impulse response from source i to sensor j, f_{kj} are the separating filters and $*$ is the convolution operator. The portion of $y_k(t)$ that comes from $s_i(t)$ is calculated by $y_{ki}(t) = \sum_{j=1}^{M}(f_{kj} * a_{ji} * s_i)(t)$. After solving the permutation and scaling ambiguity we measure the SIR for $y_k(t)$ so that $s_i(t)$ is output to $y_i(t)$. The SIR is defined as [42, 104]:

$$\mathbf{SIR}_k = 10\log\left[\sum_t y_{kk}(t)^2 / \sum_t \left(\sum_{i\neq k} y_{ki}(t)\right)^2\right] (\text{dB}). \tag{3.54}$$

The results of performance evaluation based on this time-domain metric (SIR) are summarized in Tables 3.3 and 3.4.

	Unmasked FDICA		Masked FDICA (Model 1)		Masked FDICA (Model 2)	
Method	SIR_1	SIR_2	SIR_1	SIR_2	SIR_1	SIR_2
SOC	10.30	10.54	-1.48	12.69	-1.30	14.67
CIFC	10.10	10.52	-1.48	12.69	-1.30	14.67

Table 3.3: ***SIR (dB) for Unmasked and Masked FDICA Systems (Preprocessing: WR)***

	Unmasked FDICA		Masked FDICA (Model 1)		Masked FDICA (Model 2)	
Method	SIR_1	SIR_2	SIR_1	SIR_2	SIR_1	SIR_2
SOC	9.91	6.10	-1.64	10.93	-1.44	15.11
CIFC	1.88	1.57	-1.64	10.93	-1.44	15.11

Table 3.4: ***SIR (dB) for Unmasked and Masked FDICA Systems (Preprocessing: SR)***

From Table 3.3, it is clearly evident that SIR_1 is degraded by the masked FDICA system (using both psychoacoustic models) when the weak reflection case is considered. On the other hand, SIR_2 enhanced by 2.2 dB and 4.1 dB using models 1 and 2 respectively.

From Table 3.4, it is also evident that SIR_1 is degraded by the masked FDICA system (using both psychoacoustic models) when strong early reflections are considered. However, SIR_2 improved by 4.8 dB and 9.4 dB using SOC and CIFC methods respectively when psychoacoustic model 1 is employed. On the other hand, model 2 improves the SIR_2 by 9 dB and 13.5 dB using SOC and CIFC methods respectively.

3.5.2 Perceptual Domain Objective Quality Measure

The EMBSD is an enhancement of the Modified Bark Spectral Distortion Measure (MBSD) in which some procedures have been modified. The MBSD uses a simple cognition model to calculate the distortion for the entire separated speech by averaging over all non-silence frames identified from access to the source data. This model is based on two assumptions: (1) non-silence segments represent speech quality of the separated speech, and (2) the variance of distortion in the separated speech is small enough to be well represented by its mean. On the other hand, the EMBSD uses better cognition model based on two psychoacoustic results [Zwicker]: (1) the hearing system integrates the sound intensity over a period of 200 ms, and (2) premasking is very short, while postmasking can last longer than premasking [111].

Several terms were defined in the cognition model used by EMBSD. A cognizable segment is defined as set of consecutive frames correspnding to 200 ms. A cognizable unit is defined as the number of frames in a cognizable segment. Perceptual distortion is defined as a maximum distortion over a cognizable segment. Postmasking distortion is defined as the amount of the previous cognizable distortion masking the current perceptual distortion. Cognizable distortion is defined as the largest value between the current perceptual distortion and the postmasking distortion. Then, the final distortion of the separated speech is the average over the cognizable distortions. The cognizable distortion is assumed to contribute to listeners' response on speech quality even when there is no distortion at the current perceptual distortion.

The EMBSD computes the distortion frame by frame, with the frame length of 320 samples using 50% overlap. Each frame is weighted by a Hanning window, and $x(n)$ and $y(n)$ denote the nth frame of the original and separated speech, respectively. $L_x(n)$ and $L_y(n)$ are the normalized loudness vectors of the nth frame of the original and separated speech, respectively. $D_{xy}(n)$ is the loudness difference between $L_x(n)$ and $L_y(n)$ and *NMTh(n)* is the noise masking threshold calculated from the original speech without the spreading function. The new cognitive model uses the perceptual distortion of the nth frame, MBSD(n) to calculate the EMBSD value. In order to compute MBSD(n), an indicator of perceptible distortion of the nth frame ($M_d(n,i)$) is used in the ith critical band. $M_d(n,i)$ is obtained by comparing the ith loudness difference of the nth frame ($D_{xy}(n,i)$) to the noise masking threshold (*NMTh(n,i)*) as follows

$$M_d(n,i) = 0, \quad \text{if} \quad D_{xy}(n,i) \leq NMTh(n,i) \tag{3.55}$$

$$M_d(n,i) = 1, \quad \text{if} \quad D_{xy}(n,i) > NMTh(n,i) \tag{3.56}$$

MBSD(n) is defined as the sum of the loudness difference which is greater than the noise masking threshold and its value can be expressed as [144]

$$MBSD(n) = \sum_{i=1}^{N_c} (M_d(n,i) D_{xy}(n,i))^m \tag{3.57}$$

where N_c denotes the number of critical bands and m is the order of proper metric used for computing the MBSD(n) using the first 15 loudness components (critical bands) only.

The final EMBSD value can be expressed as [144]

$$EMBSD = \frac{1}{N_f} \sum_{j=1}^{N_f} C_d(j) \tag{3.58}$$

where N_f is the total number of cognizable segments and $C_d(j)$ is the cognizable distortion of jth cognizable segment given by

$$C_d(j) = \max(P_d(j), Q_d(j)) \tag{3.59}$$

where $P_d(j)$ is the perceptual distortion of the jth cognizable segment given by

$$P_d(j) = \max\left[MBSD(v(j-1)+1), \cdots, MBSD(v(j-1)+v)\right] \tag{3.60}$$

where v is the cognizable unit and $MBSD(i)$ is same as defined in (3.57).

The postmasking distortion ($Q_d(j)$) of the jth cognizable segment is defined as

$$Q_d(j) = \frac{\gamma}{100} C_d(j-1) \tag{3.61}$$

where γ is the post masking factor equal to 80.

The results of performance evaluation based on perceptual domain metric (EMBSD) are summarized in Tables 3.5 and 3.6. From these Tables, it is clearly evident that $EMBSD_2$ reduced by 3 dB and 3.5 dB for weak and strong reflection conditions respectively when masked FDICA system (using both psychoacoustic models) is considered. Further, it can be seen that $EMBSD_1$ increased by 3.3 dB and 2.4 dB for weak and strong reflection cases of masked FDICA system (using both models) respectively. Thus, the EMBSD obtained by percptually motivated preprocessor is more effective in one of the separated signals.

Method	Unmasked FDICA		Masked FDICA (Model 1)		Masked FDICA (Model 2)	
	$EMBSD_1$	$EMBSD_2$	$EMBSD_1$	$EMBSD_2$	$EMBSD_1$	$EMBSD_2$
SOC	4.0	6.9	7.2	4.0	7.5	4.1
CIFC	4.1	7.0	7.2	4.0	7.5	4.1

Table 3.5: ***EMBSD (dB) for Unmasked and Masked FDICA Systems (Preprocessing: WR)***

Method	Unmasked FDICA		Masked FDICA (Model 1)		Masked FDICA (Model 2)	
	$EMBSD_1$	$EMBSD_2$	$EMBSD_1$	$EMBSD_2$	$EMBSD_1$	$EMBSD_2$
SOC	5.1	7.0	7.1	3.7	7.9	3.5
CIFC	5.2	7.1	7.1	3.7	7.9	3.5

Table 3.6: ***EMBSD (dB) for Unmasked and Masked FDICA Systems (Preprocessing: SR)***

3.6 Summary

In this study, we explored the Blind Source Separation problem of convolved speech mixtures (when the mixing environment is highly reverberant), in the case of an equal number of sources and sensors, proposing a perceptual solution. The key points are:

A perceptually motivated FDICA system using the complex Infomax algorithm, proposed in this chapter, reduces the frequency components that are perceptually irrelevant by exploiting the masking properties of the input speech. This system also reduces the computation complexity of a similarity measure among spectral envelopes of separated signals for solving the permutation. The measured permutation error is 7.40% and 46.70% for the unmasked FDICA system under both weak and strong reflection conditions respectively. On the other hand, the permutation error is zero for both reflection cases of the masked system (using either model).

Furthermore, SIR_2 improved by 4.8 dB and 9.4 dB using SOC and CIFC methods respectively when psychoacoustic model 1 is employed. On the other hand, model 2 improves the SIR_2 by 9 dB and 13.5 dB using SOC and CIFC methods respectively. It can be seen that $EMBSD_2$ reduced by 3 dB and 3.5 dB for weak and strong reflections respectively when the masked FDICA system (using both psychoacoustic models) is considered. Though, the SIR_2 and the $EMBSD_2$ obtained by the masked FDICA system are better than those of the unmasked system, but the informal listening test confirms the poor performance of the FDICA system.

Chapter 4
BSS of Convolutive Audio Mixtures with Perceptual Postprocessing Filter

Postprocessing of separated signals typically aims at reducing the computational complexity of the similarity measure for solving the permutation problem of FDICA. The idea of perceptual postprocessing is to make the separated signal applicable for more efficient permutation solving strategy by removing the irrelevant components from the separated signal spectrum by taking advantage of the perceptual masking properties of the human auditory system.

The main objective of this Chapter is to investigate whether perceptual criteria, which take into account the process whereby one auditory stimulus prohibits the detection of another speech signal (perceptual masking), can enhance the separation performance of existing BSS system when the mixing is noisy and highly reverberant.

The perceptual FDICA system proposed in this Chapter, is a variation of that already described in Chapter 3; the alteration is that the perceptual masking is applied to the separated speech signals before computing the similarity measure for solving the permutation problem. Then, the coherency property of both the mixing matrix and the spectral envelope correlations corresponding to the perceptually relevant output in several adjacent frequencies is utilized to solve the permutation ambiguity problem of the FDICA system.

The motivation for attempting the approach of a psychoacoustic model based postprocessor is that perceptual solution proposed in Chapter 3 (perceptually motivated preprocessor) failed to reduce the effect of the room reflections/reverberations. The presence of room reverberations force the separated signals to be seen as sensor/observed speech signals and thereby degrading the overall separation performance of FDICA system.

Therefore, we are proposing a perceptually optimal blind speech separation system in this Chapter for improving the overall performance of the BSS system while exploiting the perceptual irrelevancy of some of the separated output speech signal spectrum before solving the scaling and permutation ambiguity problems of the FDICA system.

4.1 Entire System

The FDICA system with perceptually motivated postprocessor is explained in the form of a block diagram as shown in Fig. 4.1.

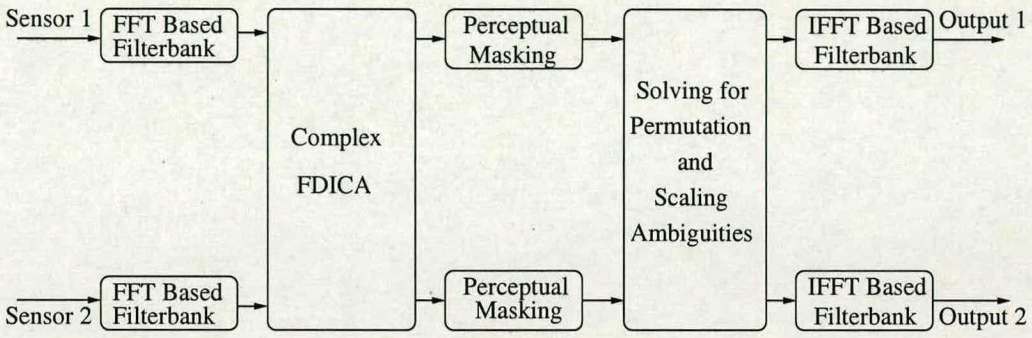


Figure 4.1: *Block Diagram of FDICA System with Perceptually Motivated Postprocessor*

First, the STFFT of the multichannel input signal, $\mathbf{x}(\omega, t)$, is obtained with an appropriate time shift and window function. Once the STFFT is obtained, the Fourier coefficients at each frequency are treated as a complex time series. By doing this, the convolutive mixture problem is reduced to complex but several linear instantaneous mixture problems.

Next, the PCA filtering method is applied to the input speech vector $\mathbf{x}(\omega, t)$ to orthogonalize its output $\mathbf{y}(\omega, t)$ and thereby obtaining the PCA filter matrix $\mathbf{W}(\omega)$.

Then, the complex Infomax algorithm with a feed-forward architecture is applied to the output of the PCA filter stage, $\mathbf{y}(\omega, t)$ to obtain the ICA filter matrix $\mathbf{U}(\omega)$. The product of $\mathbf{U}(\omega)$ and $\mathbf{W}(\omega)$ can be referred to as the separation filter matrix $\mathbf{B}(\omega)$.

After source separation by ICA, a perceptually motivated postprocessor is used in order to determine the masking threshold for each segment of the separated speech spectrum and thereby removing the perceptually irrelevant frequency components from the separated spectrum.

After obtaining $\mathbf{B}(\omega)$, the permutation and scaling problems are solved by processing only the perceptually relevant frequency components of the separation filter output $\mathbf{z}_f(\omega, t)$ with the permutation matrix $\mathbf{P}(\omega)$ and the scaling matrix $\widetilde{\mathbf{B}}_{\widetilde{m}}(\omega)$.

Finally, the filter matrices are transformed into the time domain, and the input speech signal is processed with the reconstructed time-domain filters.

4.2 Complex Infomax Algorithm

As described in Chapter 3, the complex Infomax algorithm with a feed-forward architecture is applied to the output of the PCA filter to obtain the ICA filter matrix $\mathbf{U}(\omega)$ [27, 64, 85].

In the ICA stage, the input signal (the output of PCA filter) $\mathbf{y}(\omega, t)$ is processed with the ICA filter matrix $\mathbf{U}(\omega)$ as

$$\mathbf{z}(\omega, t) = \mathbf{U}(\omega)\mathbf{y}(\omega, t) \tag{4.1}$$

4.3 Implementation of Postprocessor

As explained in Chapter 3, the perceptual postprocessing filter implemented in this work was designed around the human auditory system. Here also, two independent masking models were chosen for implementing the postprocessor, namely MPEG-1 psychoacoustic model 1 and MPEG-1 psychoacoustic model 2 [126, 127, 132, 133]. After computing the masking threshold for each separated speech signal frame, it is compared with the power spectrum in the corresponding frame to produce a perceptual binary mask [134, 135].

In perceptual audio coding, thresholding sets the quanization level, here we set a threshold for computing the similarity measure according to their psychoacoustic relevance and thereby reducing the computational complexity of solving the permutation problem.

4.4 Method of Solving Scaling and Permutation

4.4.1 Scaling Problem

As explained earlier, the scaling problem can be solved by filtering the perceptually relevant output of the separation filter $\mathbf{z}_f(\omega, t)$ by the inverse of $\mathbf{B}(\omega)$ separately [87].

The nth component of $\mathbf{z}_f(\omega, t), z_{nf}(\omega, t)$ is filtered by $\mathbf{B}^{-1}(\omega)$ separately as

$$\widetilde{\mathbf{z}}_{nf}(\omega, t) = \mathbf{B}^{-1}(\omega)[0, \cdots, 0, z_{nf}(\omega, t), 0, \cdots, 0]^T \tag{4.2}$$

Eq. (4.2) is equivalent to

$$\widetilde{z}_{\widetilde{m},nf}(\omega, t) = B^{-1}_{\widetilde{m},n}(\omega) z_{nf}(\omega, t) \tag{4.3}$$

where $B^{-1}_{\widetilde{m},n}(\omega)$ denotes the $(\widetilde{m},n)$th element of $\mathbf{B}^{-1}(\omega)$. The symbol $\widetilde{m}$ denotes an arbitrary microphone number. Eq. (4.3) can be written in the matrix-vector notation as

$$\widetilde{\mathbf{z}}_f(\omega,t) = \widetilde{\mathbf{B}}^{-1}_{\widetilde{m}}(\omega)\mathbf{z}_f(\omega,t) \tag{4.4}$$

4.4.2 Permutation Problem

As described in Chapter 3, the permutation problem can be solved by minimizing the sum of the angles between the location vectors in the adjacent frequencies and thereby computing the permutation matrix $\mathbf{P}(\omega)$. For solving the permutation problem, a method utlizing both the coherency of mixing matrices [9, 10] and the spectral envelope correlations [87] corresponding to perceptually relevant output at several adjacent frequencies has been considered. This method is denoted as the combined inter frequency correlation (CIFC).

The crosscorrelation between output spectral envelopes at adjacent frequencies is assumed to be equal to zero when any one of the separated output signals under the influence of perceptual masking is zero. This is essential to avoid the rank deficiency of the permutation matrix.

4.5 Final Filtering

After solving the permutation and scaling ambiguities, the final reconstructed filtering matrix in the frequency domain can be obtained as

$$\mathbf{F}(\omega) = \mathbf{P}(\omega)\widetilde{\mathbf{B}}^{-1}_{\widetilde{m}}(\omega)\mathbf{B}(\omega). \tag{4.5}$$

Thus, the reconstructed time domain filters are obtained as the inverse Fourier transform of $\mathbf{F}(\omega)$ as

$$f_{n,m}(i) = \mathbf{IFFT}[F_{n,m}(\omega)]w(i) \tag{4.6}$$

where **IFFT**[.] operator denotes the inverse **FFT**. The symbols $F_{n,m}(\omega)$ and $f_{n,m}(i)$ denote the (n,m)th element of the frequency domain filter $\mathbf{F}(\omega)$ and its time domain correspondence, respectively. The symbol $w(i)$ denotes the windowing function.

4.6 Experimental Results

4.6.1 Synthetic Room Mixing Scenario

In this experiment, we created a synthetic room mixing of two speech sources (4 s at 16 kHz) and we used Asano's [139] conference room (5.9 m×8.7 m×4.1 m) with a reverberation time of 0.5 sec to simulate highly reverberant room conditions for both weak and strong reflection cases of room filters as shown in Fig. 3.4 of Chapter 3. We applied the complex Infomax algorithm for both reflection cases of mixing environment, using the parameters of the proposed BSS system that are summarized in Table 4.1.

Sampling Frequency	16 kHz
STFFT Frame Length	512
Shift of STFFT	16
Window Function	Hamming
Learning Rate, η	0.0001
Gain for Score Function, G	100
Normalized Sound Pressure Level, SPL	96 dB
Number of Microphones, M	2
Number of Sources, D	2
Reference Range in Permutation, K	5

Table 4.1: ***Parameters of the Proposed BSS System (Perceptual Postprocessing)***

Original speech sources, observed signals and the separated signals are shown in Fig. 4.2. Further, each of the original source is divided into eight segments ($A1, B1, \cdots, H1$ in the case of the first source and $A2, B2, \cdots, H2$ in the case of the second source) for simplifying the comparitive analysis of each category of the above mentioned speech signals. These signal segments will help us to compare each of the separated speech signals as to whether they resemble the shape of the original speech sources or the observed signals.

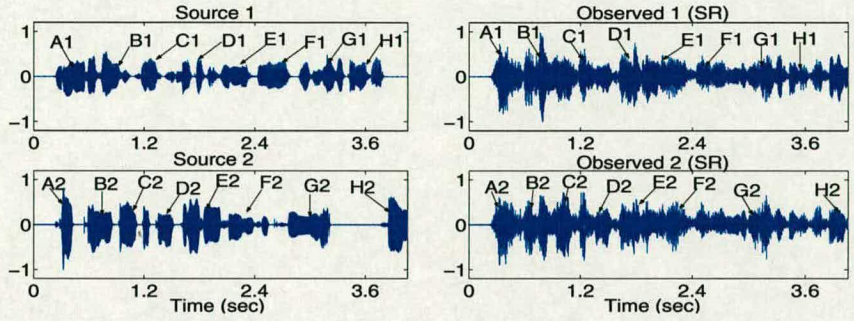
From Fig. 4.2(c), it is evident that the segments i.e., A1, C1, D1, F1, G1 and H1; A2, B2, C2, D2, E2 and H2 of the first and the second separated signals obtained by the unmasked FDICA system, respectively are similar to the original sources when the weak early reflection case is considered. The remaining segments i.e., B1, E1, F2 and G2 are remain mixed. However, these separated signals have some crosstalk whenever the original sources have zero or minmium signal strength (see Fig. 4.2(a)). On the other hand, the separated signals obtained by the unmasked system (shown in Fig. 4.2(d)) are similar to the observed signals (shown in Fig. 4.2(b)) when the strong reflection case is considered.

From Figs. 4.2(e) and 4.2(g), it is also evident that the separated speech signals in one of the channels (second output signal) obtained by the masked FDICA system (using either model) is slightly different from the original source (with reference to most of the segments namely A2, B2, C2, D2, F2 and H2) when the weak reflection case is considered. However, the separated speech signal is a permuted (positions are changed due to the inherent permutation ambiguity problem of FDICA) version of the original speech source.

From Fig. 4.2(f), it can be seen that the second separated signal obtained by the masked FDICA system (using psychoacoustic model 1) is entirely different from the original source when the strong reflection case is considered. However, there is no change in the position of the separated signal (not permuted) and further, the segments A2, G2 and H2 of the separated signal are slightly different from those of the original signal. The remaining segments i.e., B2, C2, D2, E2 and F2 are still appeared as a mixed signal due to the presence of strong early reflections in the separated output. On the other hand, the first separated output signal is almost identical to the observed speech signal for both early reflection cases.

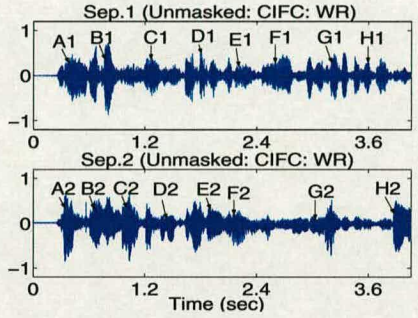
From Fig. 4.2(h), it is observed that the second separated signal obtained by the masked FDICA system (using psychoacoustic model 2) is entirely different from the original source (with reference to most of the segments i.e., $A2, B2, \cdots, H2$) when the strong reflection case is considered. Though these segments $A2, B2, \cdots, H2$ of the separated signal are slightly better than those obtained by the psychoacoustic model 1 under similar experimental conditions, still there is heavy crosstalk due to the strong reflective environment. However, the separated signal in the first channel is similar to the observed signal.

The spectrograms of original speech sources, observed speech signals and the separated speech signals are shown in Fig. 4.3. Further, the speech signal frequency range (with a bandwidth of 5 kHz) is divided into two frequency bands namely F_1 (0-3 kHz) and F_2 (3-5 kHz) to simplify the comparative analysis of the above mentioned spectrograms. From Figs. 4.3(e), 4.3(f), 4.3(g) and 4.3(h), it is clearly observed that some of the higher frequency components ($>$ 3 kHz) of the second separated signal spectrum are masked when the perceptual postprocessor (using both psychoacoustic models) is used for both weak and strong early reflections of FDICA system. On the other hand, the first separated speech signal spectrum is not masked by the perceptual postprocessing filter. Based on the above discussion, we conclude that a perceptually motivated postprocessor has a very little influence on the overall performance of FDICA system.

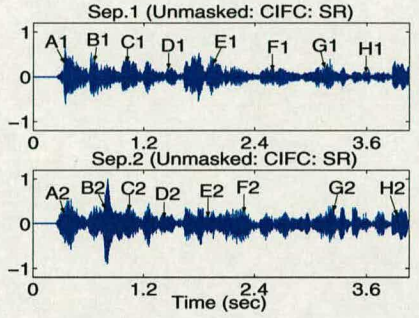


(a) Original Speech Signals

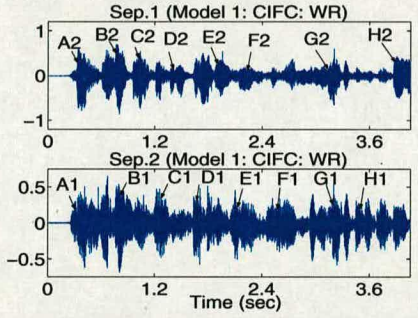
(b) Observed Speech Signals



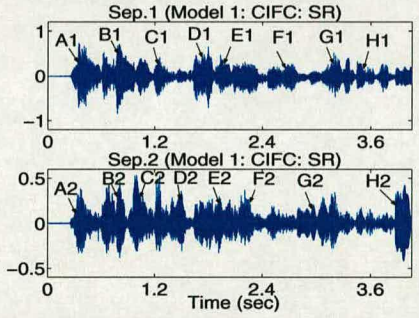
(c) Unmasked (Weak Reflection)



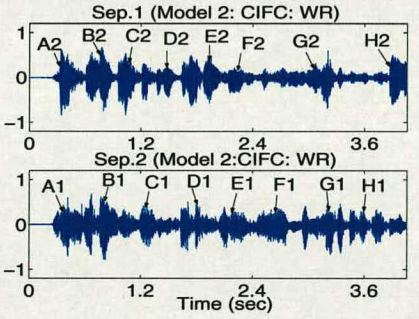
(d) Unmasked (Strong Reflection)



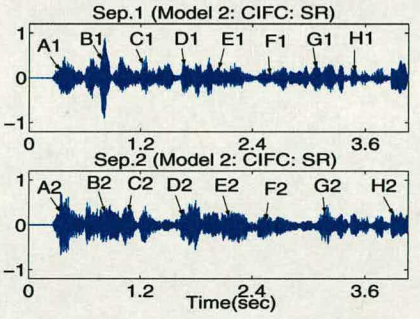
(e) Masked (Model 1:Weak Reflection)



(f) Masked (Model 1:Strong Reflection)

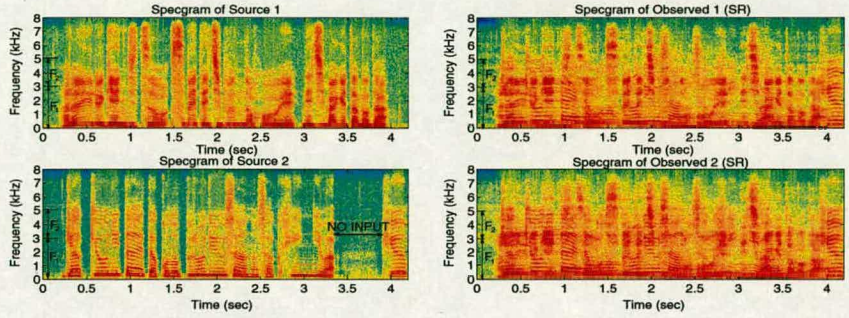


(g) Masked (Model 2:Weak Reflection)



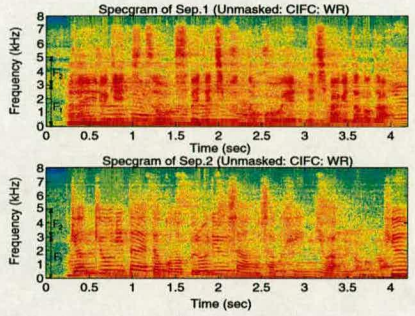
(h) Masked (Model 2:Strong Reflection)

Figure 4.2: *Original Sources, Observed Signals and Separated Signals for Unmasked and Masked Systems (Synthetic Room Mixing Scenario: Perceptual Postprocessing)*

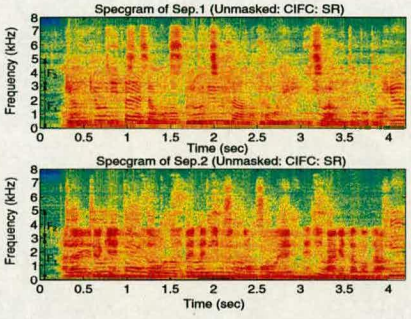


(a) Original Speech Signals

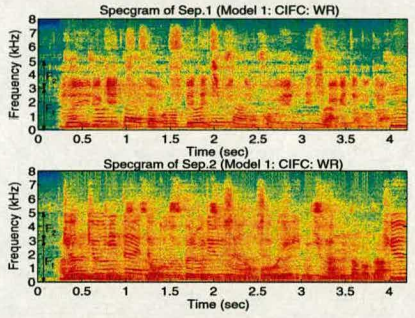
(b) Observed Speech Signals



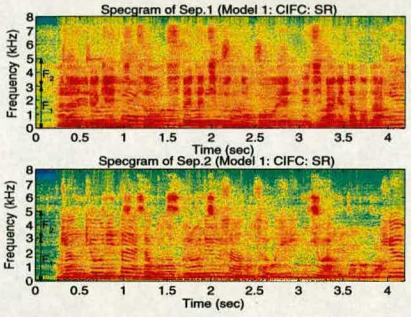
(c) Unmasked (Weak Reflection)



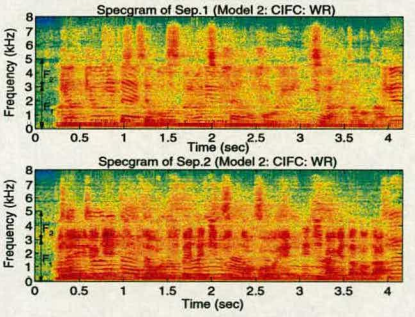
(d) Unmasked (Strong Reflection)



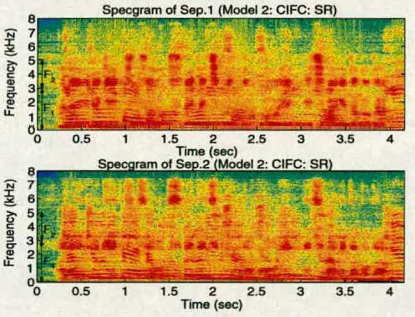
(e) Masked (Model 1:Weak Reflection)



(f) Masked (Model 1:Strong Reflection)



(g) Masked (Model 2:Weak Reflection)



(h) Masked (Model 2:Strong Reflection)

Figure 4.3: *Specgrams of Sources, Sensors and Separated Signals for Unmasked and Masked FDICA Systems (Synthetic Room Mixing Scenario: Perceptual Postprocessing)*

From Fig. 4.4, it can be seen that the measured cost function $F(\mathbf{P}, k)$ with $k = 5$ shows a smaller value at all frequencies except in the very few frequencies for both unmasked and masked system (with psychoacoustic model 1) under both weak and strong reflection cases. However, there is a slight improvement in the measured value of the cost function for most of the frequencies in the range over 2 kHz when psychoacoustic model 2 is used.

The confidence measure $C(k)$ depicted in Fig. 4.5 has a smaller value for most of the frequencies when both the unmasked and the masked FDICA systems (using both perceptual models) are considered for both cases of early reflections.

Fig. 4.6 shows the permutation error for $K = 5$. It is observed that the permutation error is zero for all frequencies except for a very few frequencies for the weak reflection case of the unmasked FDICA system. The average permutation error is found to be 7%.

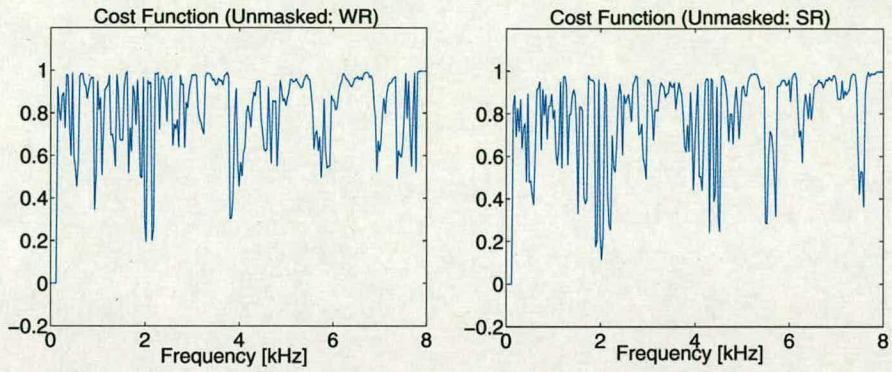
On the other hand, the permutation error is unity for the frequency range of 1 to 4 kHz and zero in the range of 4 to 7.5 kHz for the strong reflection case of the unmasked FDICA system. Therefore, the average permutation error is measured at 47%.

It is clearly evident that the permutation error is unity for most of the frequencies in the range of 4 to 6 kHz and zero for most of the frequencies in the range of 1 to 3 kHz when the FDICA system with the psychoacoustic model 1 based postprocessor is used for the weak reflection case. The average permutation error is found to be 42%.

Further, the permutation error is zero for most of the frequencies in the range of 1 to 4 kHz and unity for the frequency range of 5 to 6 kHz when the masked FDICA system with the psychoacoustic model 1 based postprocessor is used for the strong reflection case. Therefore, the average value of the permutation error is measured at 49%.

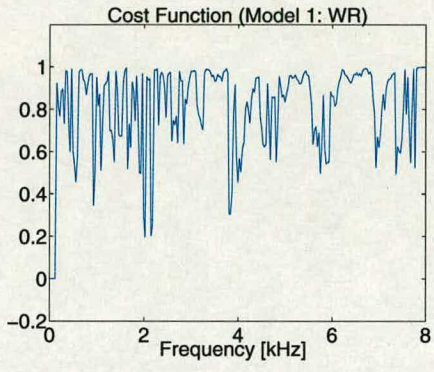
It is clearly observed that the permutation error is unity for very few frequencies in the range of 1 to 4 kHz and zero in the range of 4 to 7.5 kHz when the FDICA system with the psychoacoustic model 2 based postprocessor is considered for both early reflection cases. The measured average permutation error is 37% and 29% for weak and strong reflection cases respectively when the FDICA system with the model 2 based postprocessor is used.

Based on the results of synthetic room mixing, we conclude that the perceptual postprocessing filter using both psychoacoustic models can neither improve the separation performance nor solve the permutation problem. Hence, we did not consider the real room mixing scenario.

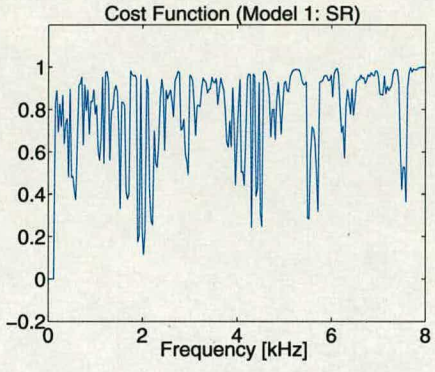


(a) Unmasked (Weak Reflection)

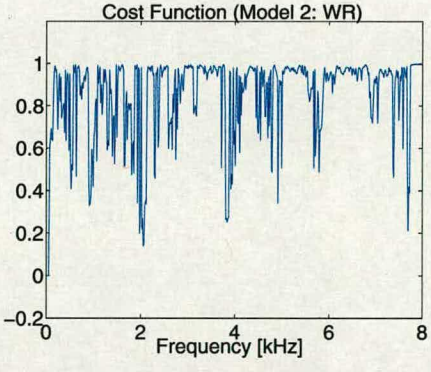
(b) Unmasked (Strong Reflection)



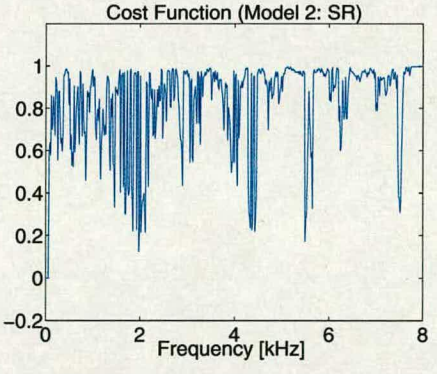
(c) Masked (Model 1: Weak Reflection)



(d) Masked (Model 1: Strong Reflection)

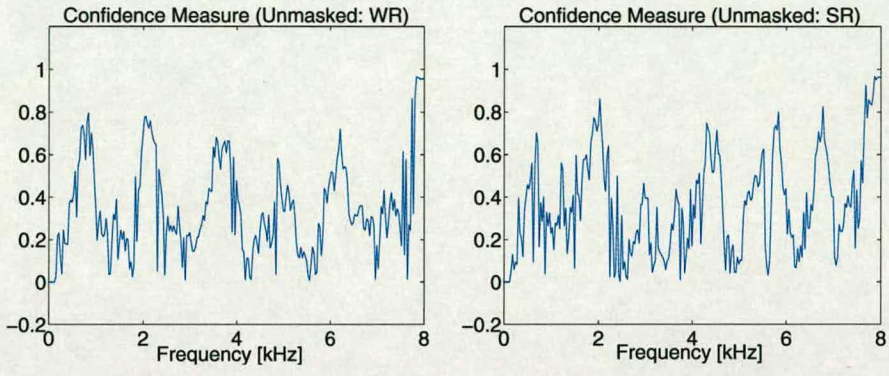


(e) Masked (Model 2: Weak Reflection)



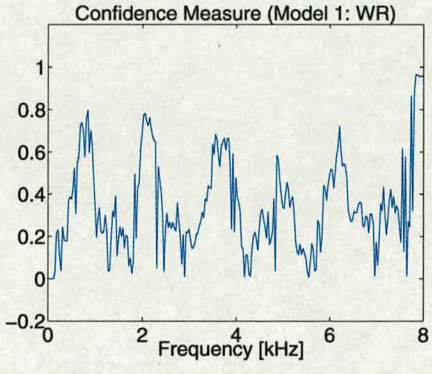
(f) Masked (Model 2: Strong Reflection)

Figure 4.4: *Measured Cost Function for Unmasked and Masked FDICA Systems (Synthetic Room Mixing Scenario: Perceptual Postprocessing)*

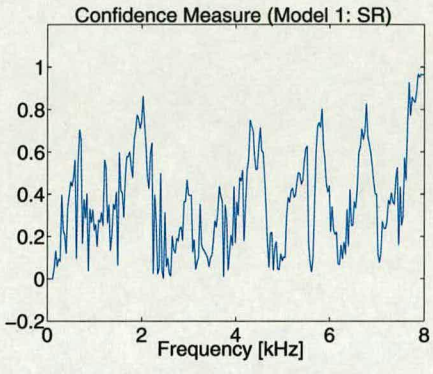


(a) Unmasked (Weak Reflection)

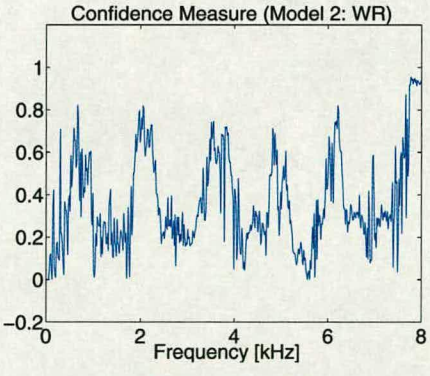
(b) Unmasked (Strong Reflection)



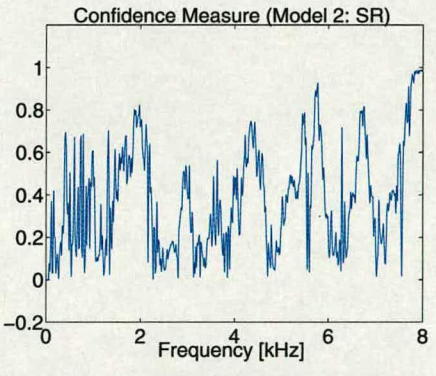
(c) Masked (Model 1: Weak Reflection)



(d) Masked (Model 1: Strong Reflection)

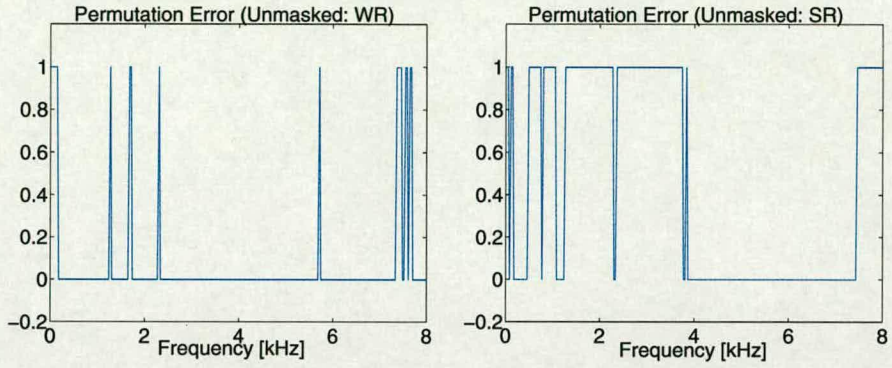


(e) Masked (Model 2: Weak Reflection)



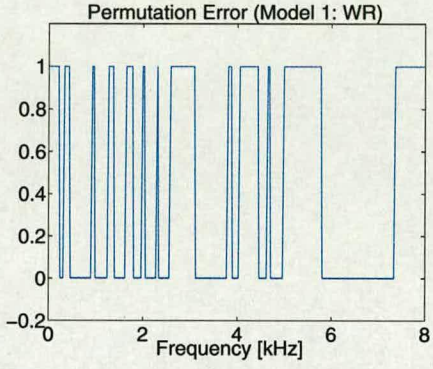
(f) Masked (Model 2: Strong Reflection)

Figure 4.5: *Confidence Measure for Unmasked and Masked FDICA Systems (Synthetic Room Mixing Scenario: Perceptual Postprocessing)*

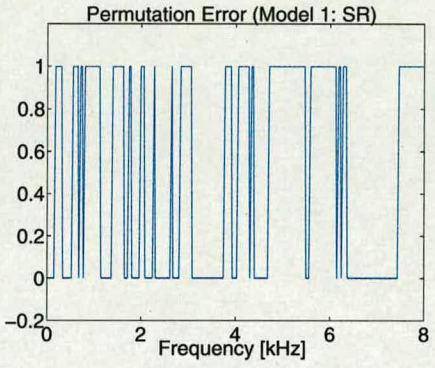


(a) Unmasked (Weak Reflection)

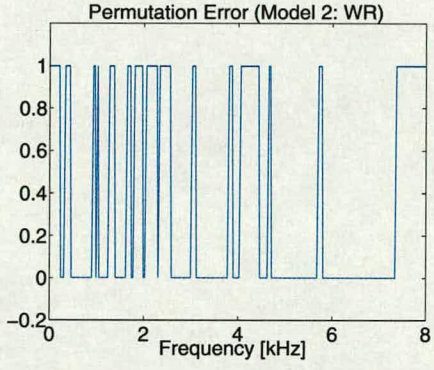
(b) Unmasked (Strong Reflection)



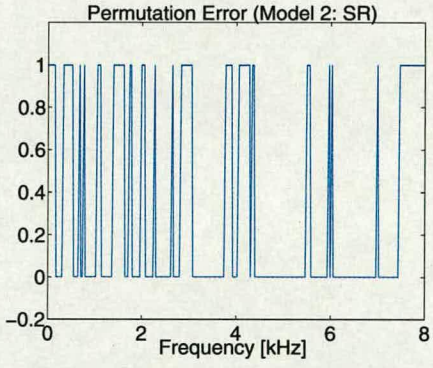
(c) Masked (Model 1: Weak Reflection)



(d) Masked (Model 1: Strong Reflection)



(e) Masked (Model 2: Weak Reflection)



(f) Masked (Model 2: Strong Reflection)

Figure 4.6: *Measured Value of Permutation Error for Unmasked and Masked FDICA Systems (Synthetic Room Mixing Scenario: Perceptual Postprocessing)*

4.7 Performance Evaluation

As explained in Chapter 3, two important objective speech quality measures, namely time-domain and perceptual domain measures, were considered for evaluating the performance of the proposed perceptually motivated FDICA system. Time-domain performance is evaluated by signal-to-interference ratio (SIR) [42] and the perceptual domain performance is evaluated by Enhanced Modified Bark Spectral Distortion (EMBSD) measures [144].

4.7.1 Time-Domain Objective Quality Measure

The results of performance evaluation based on the time-domain metric (**SIR**) are summarized in Tables 4.2 and 4.3.

	Unmasked FDICA		**Masked FDICA (Model 1)**		**Masked FDICA (Model 2)**	
Method	SIR_1	SIR_2	SIR_1	SIR_2	SIR_1	SIR_2
SOC	10.30	10.54	-1.83	-3.46	-1.17	-3.69
CIFC	10.10	10.52	-1.87	-2.69	-1.31	-2.87

Table 4.2: ***SIR (dB) for Unmasked and Masked FDICA Systems (Postprocessing: WR)***

	Unmasked FDICA		**Masked FDICA (Model 1)**		**Masked FDICA (Model 2)**	
Method	SIR_1	SIR_2	SIR_1	SIR_2	SIR_1	SIR_2
SOC	9.91	6.10	-7.15	-8.68	-8.16	-9.46
CIFC	1.88	1.57	-7.37	-8.96	-7.81	-9.80

Table 4.3: ***SIR (dB) for Unmasked and Masked FDICA Systems (Postprocessing: SR)***

From Tables 4.2 and 4.3, it is clearly evident that the measured values of SIR for the masked FDICA system (using both psychoacoustic models) are very poor when compared to that of the unmasked FDICA system for both early reflection cases of the mixing environment.

Based on the results of the time-domain performance, we conclude that neither psychoacoustic model can improve the values of SIR_1 and SIR_2 for either reflection case of the masked FDICA system due to the presence of reverberations in the separated signals.

Thus, the proposed FDICA system with a perceptually motivated postprocessing filter is highly ineffective in enhancing the performance of separated output signals that are perceptible to the human listener when the mixing is noisy and highly reverberant.

4.7.2 Perceptual Domain Objective Quality Measure

Performance evaluation results based on the perceptual domain metric (**EMBSD**) are summarized in Tables 4.4 and 4.5.

Method	Unmasked FDICA		Masked FDICA (Model 1)		Masked FDICA (Model 2)	
	$EMBSD_1$	$EMBSD_2$	$EMBSD_1$	$EMBSD_2$	$EMBSD_1$	$EMBSD_2$
SOC	4.0	6.9	8.3	8.2	7.3	7.1
CIFC	4.1	7.0	8.5	8.1	7.4	7.2

Table 4.4: ***EMBSD (dB) for Unmasked and Masked Systems (Postprocessing: WR)***

Method	Unmasked FDICA		Masked FDICA (Model 1)		Masked FDICA (Model 2)	
	$EMBSD_1$	$EMBSD_2$	$EMBSD_1$	$EMBSD_2$	$EMBSD_1$	$EMBSD_2$
SOC	5.1	7.0	8.4	8.8	7.6	7.8
CIFC	5.2	7.1	8.6	8.7	7.5	7.9

Table 4.5: ***EMBSD (dB) for Unmasked and Masked Systems (Postprocessing: SR)***

From Table 4.4, it is clearly evident that $EMBSD_1$ and $EMBSD_2$ increased by 4.4 dB and 1.2 dB respectively when the masked FDICA system (using psychoacoustic model 1) is considered for the weak reflection case. On the other hand, the psychoacoustic model 2 increases $EMBSD_1$ and $EMBSD_2$ by 3.3 dB and 0.2 dB respectively.

From Table 4.5, it can be seen that $EMBSD_1$ and $EMBSD_2$ increased by 3.4 dB and 1.7 dB respectively when the masked FDICA system (using psychoacoustic model 1) is considered for the strong reflection case. On the other hand, the psychoacoustic model 2 increases $EMBSD_1$ and $EMBSD_2$ by 2.4 dB and 0.8 dB respectively.

Thus, the EMBSD obtained by the proposed FDICA system (with a perceptual postprocessor using both models) are poor when compared to that of the unmasked FDICA system for both reflection cases. Based on these results of the perceptual-domain performance evaluation, we conclude that neither psychoacoustic model can improve the values of $EMBSD_1$ and $EMBSD_2$ for both reflection cases of the masked FDICA system. This is due to the presence of room reverberations in the separated output signals obtained by ICA.

Hence, the proposed FDICA system with a perceptually motivated postprocessing filter is highly ineffective in enhancing the performance of separated output signals that are perceptible to the human listener when the mixing is noisy and highly reverberant.

4.8 Summary

In this study, we explored the Blind Source Separation problem of convolved speech mixtures (when the mixing environment is noisy and highly reverberant), in the case of an equal number of sources and sensors, proposing a perceptual solution. The key points are:

The FDICA system with a perceptually motivated postprocessor, proposed in this chapter, reduces the frequency components that are perceptually irrelevant by exploiting the masking properties of separated speech signals and thereby reducing the computation complexity of a similarity measure among spectral envelopes of separated speech signals for solving the permutation ambiguity problem.

Further, the measured permutation error is 7% and 47% for the unmasked FDICA system under both weak and strong reflection conditions respectively. On the other hand, the permutation error is 42% and 49% for the FDICA system with the psychoacoustic model 1 under both weak and strong early reflection cases respectively.

The FDICA system with the psychoacoustic model 2 gives the permutation error of 37% and 29% for both weak and strong reflections respectively. Though, there is a net reduction of 18% in the measured average permutation error for the strong reflection case, but the informal listening test confirms the poor performance of the proposed perceptual solution for solving the permutation ambiguity problem of FDICA.

Thus, the measured values of the permutation error, SIR and EMBSD obtained by proposed FDICA system (with perceptual postprocessor using both psychoacoustic models) are inferior when compared to that of unmasked FDICA system for both reflection cases.

Hence, we conclude that the proposed FDICA scheme with a perceptually motivated postprocessing filter can neither solve the permutation problem nor enhance the performance of a BSS system due to the presence of room reflections. Therefore, it is strongly felt that the removal of room reflections prior to the application of the ICA algorithm might help the overall performance of a BSS system while solving the permutation ambiguity problem.

Chapter 5

A Combined Approach of Perceptual Preprocessing and Subspace Filtering for Blind Separation of Audio Signals

In this Chapter, a combined approach of perceptual preprocessing and subspace filtering is proposed to enhance the performance of blind separation of speech signals in a noisy and highly reverberant environment. In this combined approach, two important signal processing techniques, namely perceptual auditory masking and subspace filtering, are utilised as preprocessors of the complex FDICA system.

The main objective of this Chapter is to exploit the perceptual irrelevancy of some of the input speech spectrum using the perceptual masking techniques before utilising the subspace method as a preprocessor of FDICA which reduces the effect of room reflections in advance and the remaining direct sounds then being separated by FDICA. This objective can be achieved by taking both the advantages of the properties of the human auditory system and the subspace method for realizing the more efficient FDICA system.

The motivation for attempting the combined approach of a psychoacoustic model based preprocessor and the subspace method is that the perceptual solutions proposed in Chapter 3 (perceptually motivated preprocessor) and Chapter 4 (perceptually motivated postprocessor) failed to reduce the effect of room reflections/reverberations. The presence of room reflections force the separated signals to be seen as sensor/observed speech signals and thereby degrading the separation performance of FDICA system.

Therefore, we are proposing a perceptually motivated subspace filtering method in this Chapter for improving the separation performance of FDICA system while exploiting both perceptual irrelevancy of some of the observed input speech signal spectrum and the properties of the subspace filtering method for suppressing the perceptually relevant room reflections in advance before applying the complex Infomax algorithm.

5.1 Entire System

The entire system is described with the help of a block diagram as shown in Fig. 5.1.

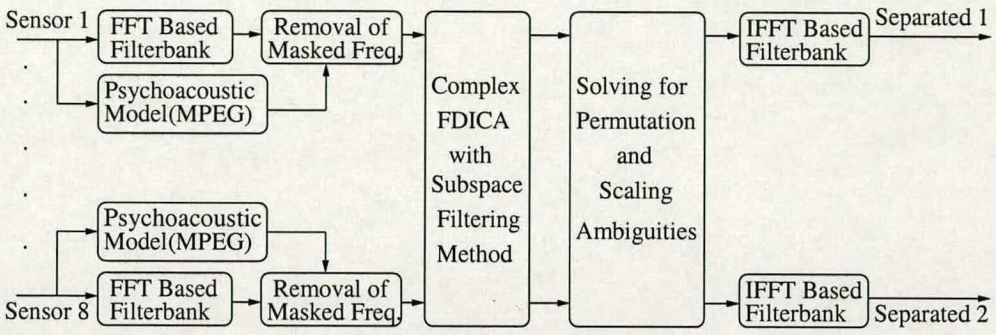


Figure 5.1: *Perceptually Motivated FDICA System with the Subspace Filtering Method*

First, the STFFT of the multichannel input speech signal, $\mathbf{x}(\omega, t)$, is obtained with an appropriate time shift and window function. Once STFFT is obtained, the Fourier coefficients at each frequency are treated as a complex time series. By doing this, the convolutive mixture problem is reduced to complex but several linear instantaneous mixture problems.

Next, the ISO/MPEG-1 psychoacoustic model is used as a preprocessor in order to estimate the auditory masking threshold for each segment of speech and thereby suppressing the perceptually irrelevant frequencies of the input speech signal.

The subspace method is then applied to the perceptually relevant spectral components of the input speech signal. In this stage, perceptually relevant room reflections and ambient noise are further reduced in advance of the application of ICA. By reducing the perceptually relevant room reflections, the output node of the subspace filter network is reduced from M to D. Thus, the subspace filtering method has the effect of both orthogonalizing the output and reducing the room reflections that are perceptually relevant.

Then, the complex Infomax algorithm with feed-forward architecture is applied to the output of the subspace filter stage to obtain the separation filter. After obtaining the separation filter, the permutation and scaling ambiguity problem is solved by processing the output of the separation filter with the permutation and the scaling matrices.

Finally, the filter matrices obtained in the above stages are transformed into the time domain and the input speech signal is processed with the time-domain filter network.

5.2 Perceptual Preprocessor

Here also, two independent masking models namely ISO/MPEG-1 psychoacoustic model 1 and psychoacoustic model 2 [126, 127, 132, 133] have been considered for computing the masking threshold for each frame of the input speech signal. After computing the masking threshold for each input speech signal frame, it is compared with the power spectrum in the corresponding frame to produce a perceptual binary mask [134, 135]. In perceptual audio coding, thresholding sets the quanization level, here we set a threshold for further processing by subspace method and FDICA according to the psychoacoustic relevance and thereby reducing the computational complexity of solving the permutation ambiguity problem.

5.3 Perceptually Motivated Subspace Method

The subspace filtering method (works as a self-organizing beamformer focusing on the target sources and does not require any previous knowledge of the sensor array or sound field) can be utilized as a preprocessor of FDICA which reduces the effect of room reflections in advance, the remaining direct sounds then being separated by FDICA. In the subspace method, perceptually relevant components of room reflections are separated from direct components in the eigenvalue domain of the spatial correlation matrix based on the spatial extent of the speech signals.

Then, the eigenvectors corresponding to the eigenvalues of the direct components are used as a filter which selects the subspace in which the direct components lie and discards the subspace filled with the energy of reflections. Further, it is well known that the subspace method is a special case of principal component analysis (PCA) with $M \gg D$, where M and D are the number of nodes (channels) of the input and the output of PCA, respectively [148, 149].

In the first instance, we cannot directly apply the subspace method to the perceptually processed input speech. The reason is very simple. Whenever the perceptually masked input speech $\mathbf{x}_f(\omega, t)$ in one of the channels contains no values, the subspace filter matrix $\mathbf{W}(\omega)$ is singular, resulting in a rank deficiency problem. This is mainly due to very low eigenvalues of spatial correlation matrix of perceptually masked input speech spectrum. Without loss of generality, we have assumed an identity matrix of order D for each pair of the input nodes as the rank of the subspace filter matrix $\mathbf{W}(\omega)$ to avoid this rank deficiency problem while retaining the whitening properties of the speech signal at the output of the subspace filter.

5.3.1 Spatial Correlation Matrix

The spatial correlation matrix is defined as

$$\mathbf{R}(\omega) = E[\mathbf{x}_f(\omega,t)\mathbf{x}_f^H(\omega,t)]. \tag{5.1}$$

Where $\mathbf{x}_f(\omega,t) = \mathbf{A}(\omega)\mathbf{s}(\omega,t)\mathbf{\Phi}_{th} + \mathbf{n}(\omega,t)\mathbf{\Phi}_{th}$. The first term, $\mathbf{A}(\omega)\mathbf{s}(\omega,t)\mathbf{\Phi}_{th}$, expresses the directional components of $\mathbf{x}_f(\omega,t)$. On the other hand, the second term, $\mathbf{n}(\omega,t)\mathbf{\Phi}_{th}$, is a mixture of less-directional components of $\mathbf{x}_f(\omega,t)$, which includes the room reflections and an ambient noise. The symbol $\mathbf{\Phi}_{th}$ denotes the perceptual masking threshold matrix.

Since the subspace method is performed for each frequency bin independently, the frequency index ω is omitted here for the sake of simplicity in notation. Assuming that sources $\mathbf{s}(t)$ and noise $\mathbf{n}(t)$ are uncorrelated, $\mathbf{R}$ can be expressed as

$$\mathbf{R} = (\mathbf{A}\mathbf{Q}\mathbf{A}^H + \mathbf{K})\mathbf{\Phi}. \tag{5.2}$$

Where $\mathbf{Q} = E[\mathbf{s}(t)\mathbf{s}^H(t)]$ and $\mathbf{K} = E[\mathbf{n}(t)\mathbf{n}^H(t)]$ are cross-spectrum matrix of $\mathbf{s}(t)$ and the correlation matrix of $\mathbf{n}(t)$ respectively and $\mathbf{\Phi}$ is the perceptual gain given by $\mathbf{\Phi}_{th}\mathbf{\Phi}_{th}^H$.

When $\mathbf{n}(t)$ includes the room reflections of $\mathbf{s}(t)$, $\mathbf{s}(t)$ and $\mathbf{n}(t)$ are highly correlated and the above assumption does not hold. However, when the time interval between the direct sound and the reflection exceeds the short window length of STFFT, this assumption holds.

5.3.2 Properties of the Perceptually Motivated Subspace Method

By taking the generalized eigenvalue decomposition of $\mathbf{R}$ as

$$\mathbf{R} = \mathbf{K}\mathbf{E}\mathbf{\Lambda}\mathbf{E}^{-1}\mathbf{\Phi} \tag{5.3}$$

we obtain the matrices $\mathbf{E} = [\mathbf{e}_1, \cdots, \mathbf{e}_M]$ and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_M)$, where $\mathbf{e}_m$ and λ_m are the eigenvector and the eigenvalue, respectively. Since $\mathbf{K}$ cannot be observed separately, we assumed $\mathbf{K} = \mathbf{I}$ to employ the standard eigenvalue decomposition, $\mathbf{R} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^{-1}\mathbf{\Phi}$.

The generalized eigenvalue decomposition whitens the non-directional components of $\mathbf{x}_f(t)$. Even when $\mathbf{K}$ is unknown, if the correlation is small, as in the case of room reverberation, the standard eigenvalue decomposition works considerably well in many cases.

Therefore, based on the well defined structure of $\mathbf{R}$ and the assumptions described previously, the eigenvalues and eigenvectors of $\mathbf{R}$ have the following properties:

1. The energy of D directional signals $\mathbf{s}(t)$ is concentrated on D dominant eigenvalues.
2. The energy of $\mathbf{n}(t)$ is equally spread over all eigenvalues.
3. $\text{span}(\mathbf{A}) = \text{span}(\mathbf{E}_s)$, where $\mathbf{E}_s = [\mathbf{e}_1, \cdots, \mathbf{e}_D]$ denotes the eigenvectors corresponding to the D dominant eigenvalues.
4. $\text{span}(\mathbf{A}) = \text{span}(\mathbf{E}_n)^{\perp}$, where $\mathbf{E}_n = [\mathbf{e}_{D+1}, \cdots, \mathbf{e}_M]$ denotes the eigenvectors corresponding to the other M - D eigenvalues.

Where $\text{span}(\mathbf{A})$ denotes the space spanned by the column vectors of $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_D]$, i.e., $\text{span}(\mathbf{A}) = \text{span}(\mathbf{a}_1, \cdots, \mathbf{a}_D)$ and $\text{span}(\mathbf{E}_n)^{\perp}$ denotes the orthogonal complement of $\text{span}(\mathbf{E}_n)$ and the subspaces namely $\text{span}(\mathbf{E}_s)$ and $\text{span}(\mathbf{E}_n)$ are referred to as the signal subspace and the noise subspace, respectively.

The relation of eigenvectors that reflects properties 3 and 4 is depicted in Fig. 5.2 (M = 3 and D = 2 is assumed).

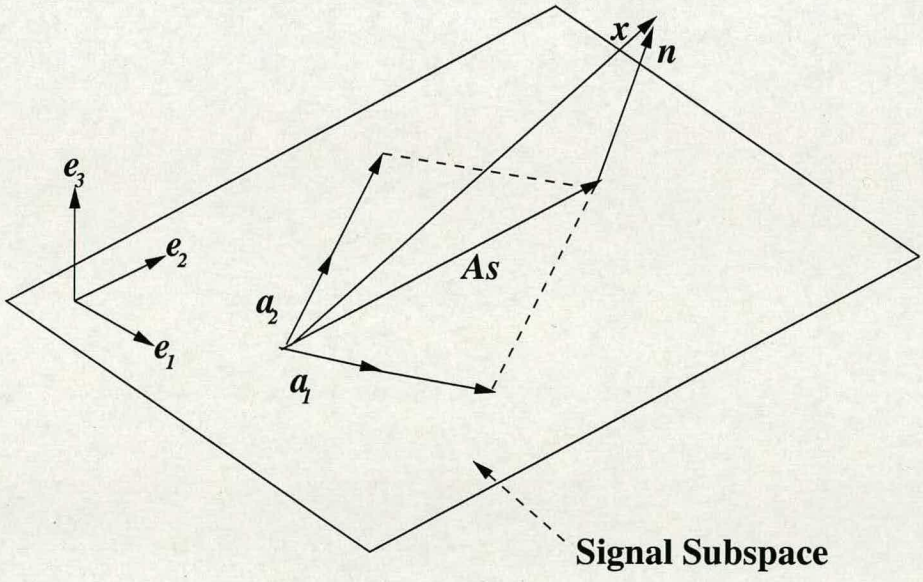


Figure 5.2: *The Relation of Eigenvectors of a Perceptually Motivated Subspace Method*

5.3.3 Perceptually Motivated Subspace Filter

In the subspace method, the input signal $\mathbf{x}_f(t)$ is processed as

$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}_f(t) = \mathbf{W}[\mathbf{A}\mathbf{s}_f(t) + \mathbf{n}_f(t)] = \mathbf{W}[\mathbf{A}\mathbf{s}_f(t) + \mathbf{n}_s(t) + \mathbf{n}_n(t)] \tag{5.4}$$

where the subspace filter is defined as

$$\mathbf{W} = \mathbf{\Lambda}_s^{-1/2}\mathbf{E}_s^H \tag{5.5}$$

where $\mathbf{\Lambda}_s = \text{diag}(\lambda_1, \cdots, \lambda_D)$. The term $\mathbf{\Lambda}_s^{-1/2}$ is a normalization factor, the same as that used in PCA. The term $\mathbf{E}_s^H$ plays an important role in the perceptually motivated subspace filter that reduces the energy of $\mathbf{n}_f(t)$ in the noise subspace as per the following analysis.

According to properties 1, 2 and 3, we have the following expressions:

$$\mathbf{A}\mathbf{s}_f(t) = \sum_{i=1}^{D} \alpha_i(t)\mathbf{e}_i\mathbf{\Phi}_{th} \tag{5.6}$$

$$\mathbf{n}_f(t) = \sum_{i=1}^{M} \beta_i(t)\mathbf{e}_i\mathbf{\Phi}_{th} \tag{5.7}$$

where $\alpha_i(t)$ is the projection coefficient of $\mathbf{A}\mathbf{s}_f(t)$ onto the basis vector (eigenvector) $\mathbf{e}_i$ and $\beta_i(t)$ is a projection coefficient of $\mathbf{n}_f(t)$ onto the basis vector $\mathbf{e}_i$. Equations (5.6) and (5.7) can be written in a matrix vector notation as

$$\mathbf{A}\mathbf{s}_f(t) = \mathbf{E}_s\boldsymbol{\alpha}(t)\mathbf{\Phi}_{th} \tag{5.8}$$

$$\mathbf{n}_f(t) = \mathbf{E}_s\boldsymbol{\beta}_s(t)\mathbf{\Phi}_{th} + \mathbf{E}_n\boldsymbol{\beta}_n(t)\mathbf{\Phi}_{th} \tag{5.9}$$

where projection coefficient vectors of $\mathbf{A}\mathbf{s}_f(t)$, $\mathbf{n}_s(t)$ and $\mathbf{n}_n(t)$ are $\boldsymbol{\alpha}(t) = [\alpha_1(t), \cdots, \alpha_D(t)]^T$, $\boldsymbol{\beta}_s(t) = [\beta_1(t), \cdots, \beta_D(t)]^T$ and $\boldsymbol{\beta}_n(t) = [\beta_{D+1}(t), \cdots, \beta_M(t)]^T$, respectively. From (5.9), it is observed that $\mathbf{n}_s(t) \in \text{span}(\mathbf{E}_s)$ and $\mathbf{n}_n(t) \in \text{span}(\mathbf{E}_n)$.

Using the properties of the eigenvectors, $\mathbf{E}_s^H\mathbf{E}_s = \mathbf{I}$ and $\mathbf{E}_s^H\mathbf{E}_n = 0$, the subspace filter output has two components only, i.e., $\mathbf{W}\mathbf{A}\mathbf{s}_f(t) = \mathbf{\Lambda}^{-1/2}\boldsymbol{\alpha}(t)\mathbf{\Phi}_{th}$ and $\mathbf{W}\mathbf{n}_s(t) = \mathbf{\Lambda}^{-1/2}\boldsymbol{\beta}_s(t)\mathbf{\Phi}_{th}$. Thus, by applying the subspace filter, the components in the subspaces $\mathbf{A}\mathbf{s}_f(t)$ and $\mathbf{n}_s(t)$ are preserved while the components in the subspace $\mathbf{n}_n(t)$ are suppressed. When $M \gg D$, it is expected that a large portion of $\mathbf{n}_n(t)$ can be cancelled by this subspace filter.

5.4 Complex Infomax Algorithm

After suppressing the perceptually relevant room reflections by means of the subspace filter, the complex Infomax algorithm is then applied to the remaining directional components of the subspace filter output to obtain the ICA filter matrix $\mathbf{U}(\omega)$ [27, 64, 85].

For the sake of convenience, the product of $\mathbf{W}(\omega)$ and $\mathbf{U}(\omega)$ is defined as the separation filter matrix $\mathbf{B}(\omega)$ given by

$$\mathbf{B}(\omega) = \mathbf{U}(\omega)\mathbf{W}(\omega) \tag{5.10}$$

In the ICA stage, the input signal (the output of the subspace filter) $\mathbf{y}(\omega, t)$ is processed with the filter matrix $\mathbf{U}(\omega)$ as

$$\mathbf{z}(\omega, t) = \mathbf{U}(\omega)\mathbf{y}(\omega, t) \tag{5.11}$$

5.5 Method of Solving Scaling and Permutation

The main objective of this section is to reduce the computational complexity of a similarity measure for solving the permutation problem of FDICA. The procedure starts with the solution for the scaling problem which is considered in the following subsection.

For solving the permutation ambiguity problem, a method utlizing both the coherency of the mixing matrices and the correlation between spectral envelopes of the separated speech signals at several adjacent frequencies has been considered [10, 87].

5.5.1 Scaling Problem

As explained earlier, the scaling problem can be solved by filtering individual outputs of the separation filter by the inverse of $\mathbf{B}(\omega)$ separately. In this analysis, the pseudoinverse of $\mathbf{B}(\omega)$, denoted as $\mathbf{B}^{+}(\omega)$, is used instead of the inverse of $\mathbf{B}(\omega)$ since $\mathbf{B}(\omega)$ is not square matrix due to the employment of the subspace method.

The pseudoinverse provides the least squares solution to a system of linear equations which are of overdetermined type [137, 138]. Furthermore, the pseudoinverse matrix $\mathbf{B}^{+}(\omega)$ can be expressed as

$$\mathbf{B}^{+}(\omega) = \left(\mathbf{B}^{H}(\omega)\mathbf{B}(\omega)\right)^{-1}\mathbf{B}^{H}(\omega). \tag{5.12}$$

The nth component of $\mathbf{z}(\omega,t)$, $z_n(\omega,t)$ is filtered by $\mathbf{B}^+(\omega)$ separately as [87]

$$\widetilde{\mathbf{z}}_n(\omega,t) = \mathbf{B}^+(\omega)[0,\cdots,0,z_n(\omega,t),0,\cdots,0]^T \tag{5.13}$$

Eq. (5.13) is equivalent to $\widetilde{z}_{\widetilde{m},n}(\omega,t) = B^+_{\widetilde{m},n} z_n(\omega,t)$. Where $B^+_{\widetilde{m},n}$ denotes the $(\widetilde{m},n)$th element of $\mathbf{B}^+(\omega)$. The symbol $\widetilde{m}$ denotes an arbitrary microphone number. Further, this can be written in the matrix-vector notation as

$$\widetilde{\mathbf{z}}(\omega,t) = \widetilde{\mathbf{B}}^+_{\widetilde{m}}\mathbf{z}(\omega,t) \tag{5.14}$$

where $\widetilde{\mathbf{z}}(\omega,t) = [\widetilde{z}_{\widetilde{m},1},\cdots,\widetilde{z}_{\widetilde{m},D}]^T$ and $\widetilde{\mathbf{B}}^+_{\widetilde{m}}(\omega)$ is a $D \times D$ diagonal matrix given by

$$\widetilde{\mathbf{B}}^+_{\widetilde{m}}(\omega) = \text{diag}[B^+_{\widetilde{m},1},\cdots,B^+_{\widetilde{m},D}]. \tag{5.15}$$

5.5.2 Permutation Problem

As explained in Chapter 3, the permutation problem can be solved by minimizing the sum of the angles between the location vectors in the adjacent frequencies using the combined inter frequency correlation (CIFC) method. In this analysis also, $\mathbf{B}^+(\omega)$ is used instead of the inverse of $\mathbf{B}(\omega)$. Accordingly, an estimate of the mixing matrix $\hat{\mathbf{A}}(\omega)$ can be obtained from $\mathbf{B}^+(\omega)$ to compute the permutation matrix $\mathbf{P}(\omega)$ [10, 87].

5.6 Final Filtering

After solving the permutation and scaling ambiguities, the final reconstructed filtering matrix in the frequency domain can be obtained as

$$\mathbf{F}(\omega) = \mathbf{P}(\omega)\widetilde{\mathbf{B}}^+_{\widetilde{m}}(\omega)\mathbf{B}(\omega). \tag{5.16}$$

Thus, the reconstructed time domain filters are obtained as the inverse Fourier transform of $\mathbf{F}(\omega)$ as

$$f_{n,m}(i) = \textbf{IFFT}[F_{n,m}(\omega)]w(i). \tag{5.17}$$

5.7 Experimental Results

5.7.1 Synthetic Room Mixing Scenario

In this experiment, we created a synthetic room mixing environment with a reverberation time of 0.5 sec for both weak and strong reflection cases. The configuration of the sound sources (loudspeakers) and the microphones is shown in Fig. 5.3. A microphone array with $M = 8$, was used [139]. The microphone array was circular in shape with a diameter of 0.5 m. The impulse responses from the sound sources to the microphones were used to convolve with the source signal to generate the observed input speech signal. We applied the complex Infomax algorithm for both early reflection cases, using the experimental parameters of the proposed BSS system that are summarized in Table 5.1.

Sampling Frequency	16 kHz
STFFT Frame Length	512
Shift of STFFT	16
Window Function	Hamming
Learning Rate, η	0.0001
Gain for Score Function, G	100
Normalized Sound Pressure Level, SPL	96 dB
Number of Microphones, M	8
Number of Sources, D	2
Reference Range in Permutation, K	5

Table 5.1: ***Proposed BSS System Parameters (Perceptually Motivated Subspace Method)***

The typical room filters used in this experiment for both weak and strong early reflection cases are shown in Fig. 5.4(a) and Fig. 5.4(b) respectively. Using these room filters, we have obtained the observed speech signals (using microphone array) as shown in Fig. 5.5.

Original speech sources, a pair of sensor signals and the separated speech signals are shown in Fig. 5.6. Further, each of the original source is divided into eight segments ($A1, B1, \cdots, H1$ in the case of the first source and $A2, B2, \cdots, H2$ in the case of the second source) for simplifying the comparitive analysis of each category of the above mentioned speech signals. These signal segments will help us to compare each of the separated (reconstructed) signals as to whether they resemble the shape of the original sources or the observed signals.

From Figs. 5.6(c) and 5.6(d), it is observed that the separated speech signals (with reference to the segments namely A1,C1, D1, F1, G1 and H1; A2, C2, E2, G2 and H2) obtained by

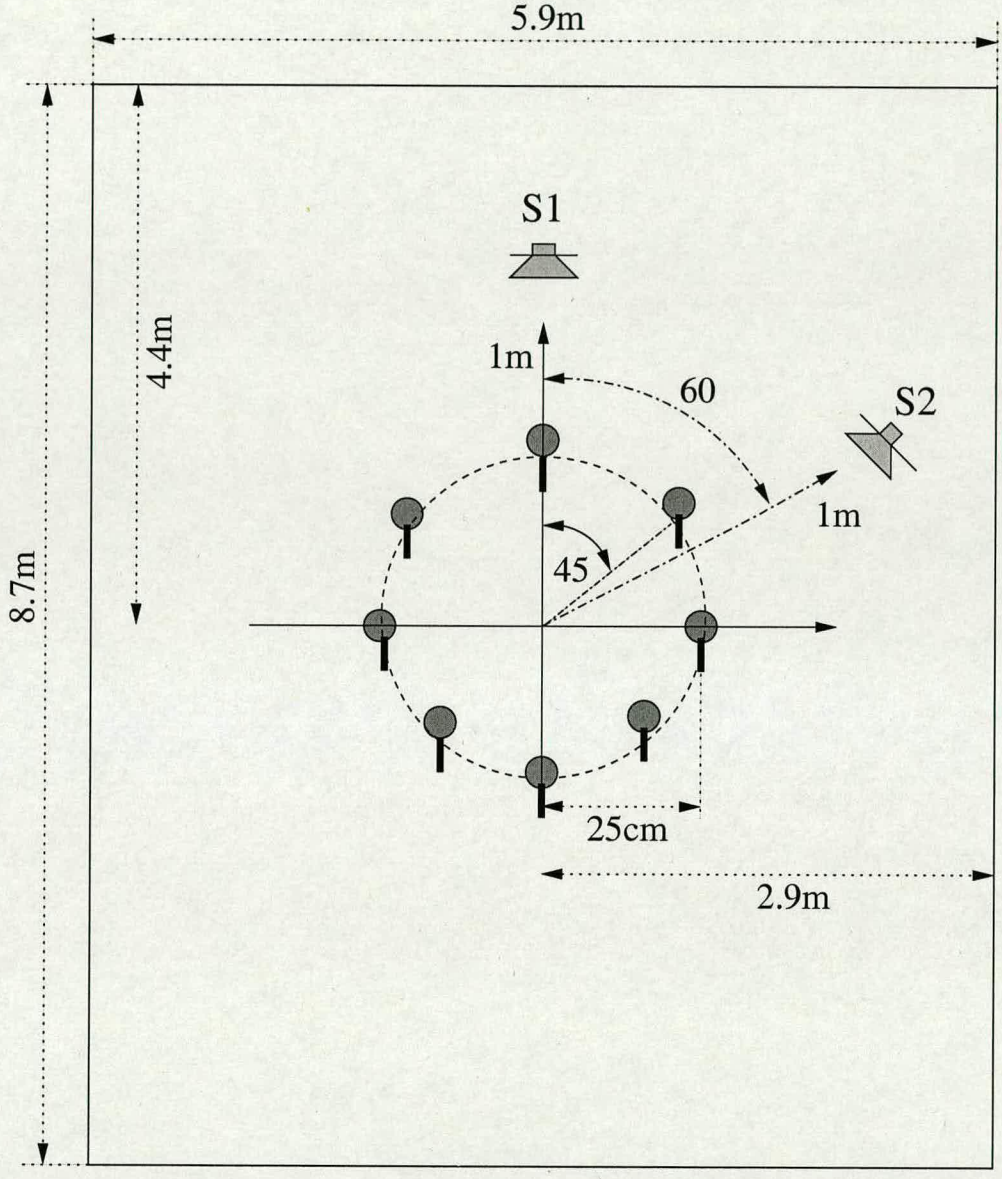
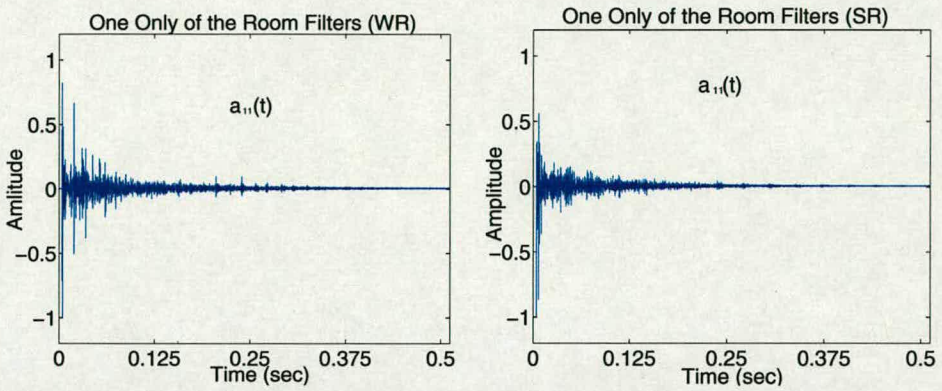


Figure 5.3: *Configuration of Microphone Array and Sound Sources (Subspace Method)*



(a) Room Filters (Weak Reflection)

(b) Room Filters (Strong Reflection)

Figure 5.4: *Room Filters for Weak and Strong Reflection Cases (Synthetic Room)*

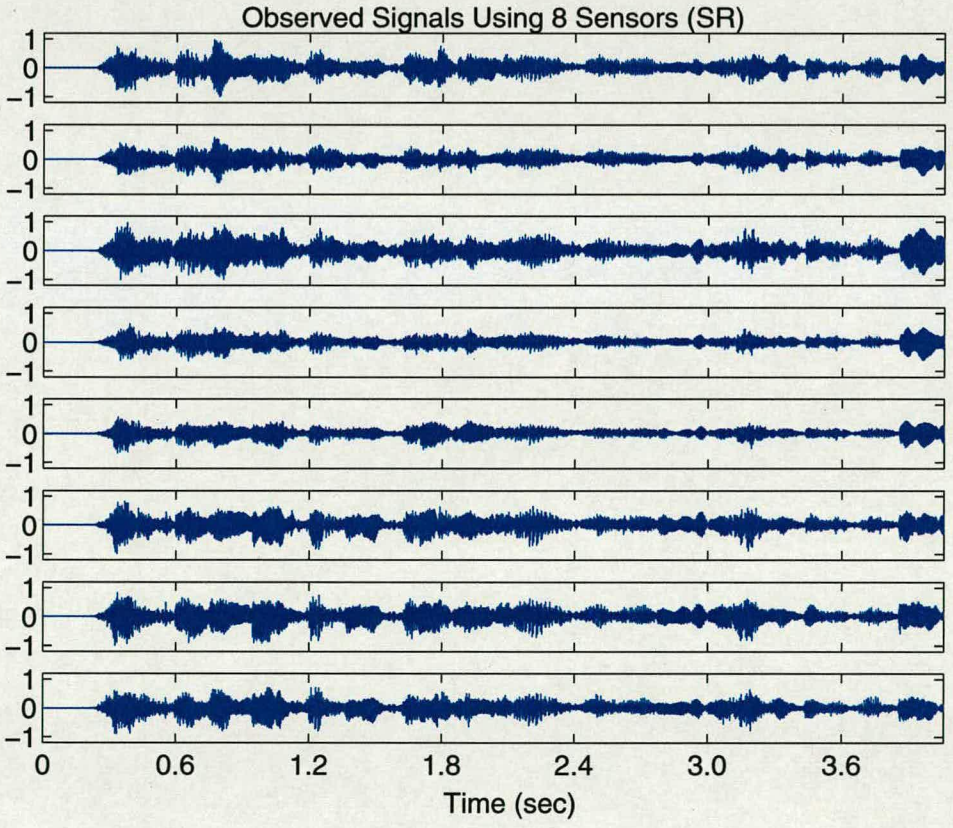


Figure 5.5: *Observed Signals Using Circular Microphone Array (Synthetic Room)*

the unmasked FDICA system are slightly different from the original speech sources when both weak and strong early reflection cases are considered. However, these separated speech signals have some crosstalk whenever the original speech sources have zero or minmium signal strength (see Fig. 5.6(a)) and also have some overlapping segments i.e., B2, C2, E2 and F2.

From Fig. 5.6(e), it is also evident that the separated speech signals in one of the channels (second output signal) obtained by the masked FDICA system (using the psychoacoustic model 1) is entirely different from the observed speech signal when the weak early reflection case is considered. However, the segments B2, D2, G2 and H2 of the separated speech signal are slightly different from those of the original speech signal. The remaining segments i.e., A2, C2, E2 and F2 still appear as a mixed signal.

From Fig. 5.6(f), it can be seen that the second separated signal obtained by the masked FDICA system (using model 1) is slightly different from the original speech signal when the strong reflection case is considered. However, the segments A2, B2, D2, E2 and F2 of the separated speech signal are similar to those of the original signal. The remaining segments i.e., C2, G2 and H2 still appear as a mixed signal. However, the first separated speech signal is similar to the observed speech signal (Fig. 5.6(b)) for both reflection cases.

From Figs. 5.6(g) and 5.6(h), it is also evident that the separated signals in one of the channels (second output in this case) obtained by the masked FDICA system (using psychoacoustic model 2) is similar to the original speech source (with reference to most of the signal segments i.e., $A2, B2, \cdots, H2$) when the strong reflection case is considered. On the other hand, the separated signal obtained by the weak reflection case still appears as the observed speech signal. However, the separated speech signal in the first channel is similar to the observed signal as shown in Fig. 5.6(b) for both reflection cases.

The spectrograms of original speech sources, observed speech signals and the separated speech signals are shown in Fig. 5.7. Further, the speech signal frequency range (with a bandwidth of 5 kHz) is divided into two frequency bands namely F_1 (0-3 kHz) and F_2 (3-5 kHz) to simplify the comparative analysis of the above mentioned spectrograms. From Figs. 5.7(e) and 5.7(f), it is clearly observed that most of the higher frequencies ($>$ 3 kHz) of the second separated output spectrum are masked when the psychoacoustic model 1 based preprocessor is used for both early reflection cases of the FDICA system. However, the psychoacoustic model 2 based preprocessing filter not only masks the most of the higher frequencies ($>$ 3 kHz) of the second

separated signal spectrum (shown in Figs. 5.7(g) and 5.7(h)) but also masks some frequencies in the range of 1-3 kHz when both reflection cases of the masked FDICA system are taken into account. Results are highlighted in Figs. 5.8, 5.9, 5.10 and 5.11 for the worst case scenario of highly reverberant environemnt.

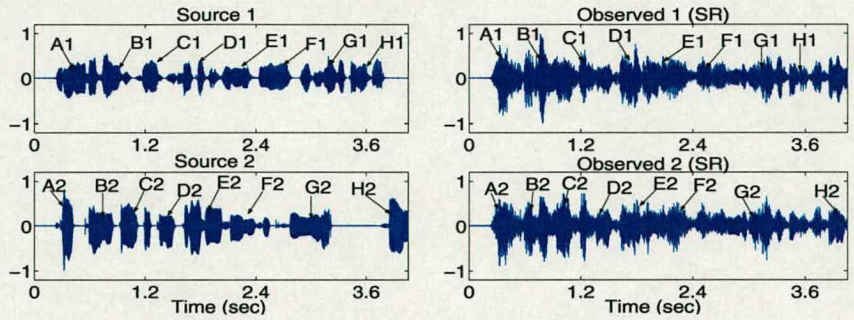
From Fig. 5.12, it can be seen that the measured value of the cost function $F(\mathbf{P}, k)$ with $k = 5$ has a smaller value at all frequencies except in the very few frequencies when the unmasked system is considered for both weak and strong reflection cases. On the other hand, the cost function is close to unity for all the frequencies except the very low frequencies when the masked FDICA system (using either model) is considered for both cases of early reflections. Further, the psychoacoustic model 2 drives the measured value of the cost function close to unity even at these very low frequencies also.

The confidence measure $C(k)$ depicted in Fig. 5.13 has a smaller value for most of the frequencies when the unmasked system is used for both cases of reflections. Further, the measured confidence measure has high values at all frequencies except at low frequencies when the perceptually motivated FDICA system (using models 1 and 2) is employed for both reflection cases. However, model 2 enhances its value for most of the low frequencies also.

Fig. 5.14 shows the measured value of the permutation error for $K = 5$. The permutation in this experiment is solved by using the crosscorrelation between the separated output spectrogram and the spectrogram of original speech source (unknown in a real room recording situation) as correct permutation for evaluating only the effect of the subspace filtering method. This method is referred to as source-output crosscorrelation (SOC).

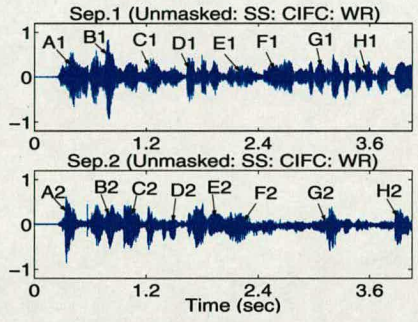
Therefore, the permutation error is defined as the case when the result of the combined inter frequency correlation (CIFC) differs from that of SOC (assumed as correct permutation). The measured permutation error is 7.8% and 6.6% for both weak and strong reflection cases of the unmasked FDICA system respectively. On the other hand, the permutation error is zero for all frequencies when the masked system (using either model) is used for both reflections.

Based on the above discussion, we conclude that strong early reflections are aiding the source separation by the masked FDICA system (using both perceptual models) in general and the second separated output signal in particular. Hence, it appears that a perceptually motivated subspace approach targets a specific source out of mixture of reverberated sounds.

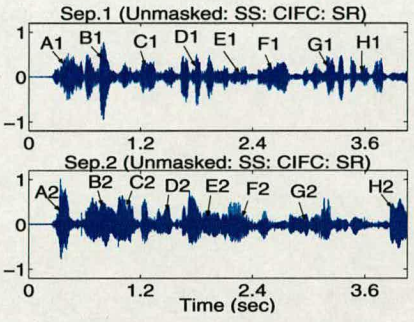


(a) Original Speech Signals

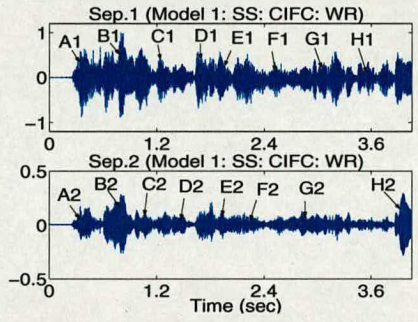
(b) A Pair of Observed Signals



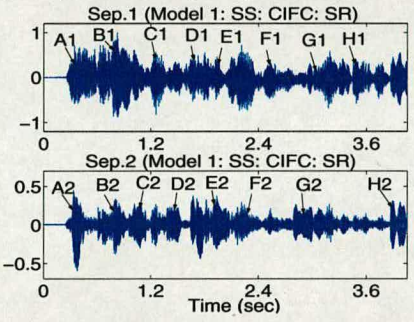
(c) Unmasked (Weak Reflection)



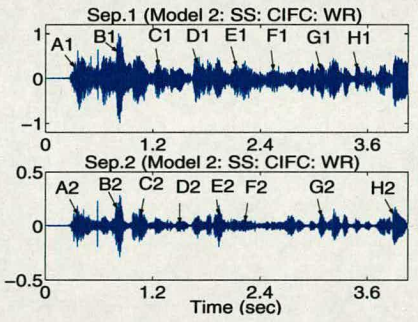
(d) Unmasked (Strong Reflection)



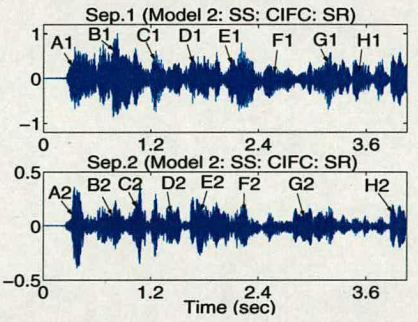
(e) Masked (Model 1:Weak Reflection)



(f) Masked (Model 1:Strong Reflection)

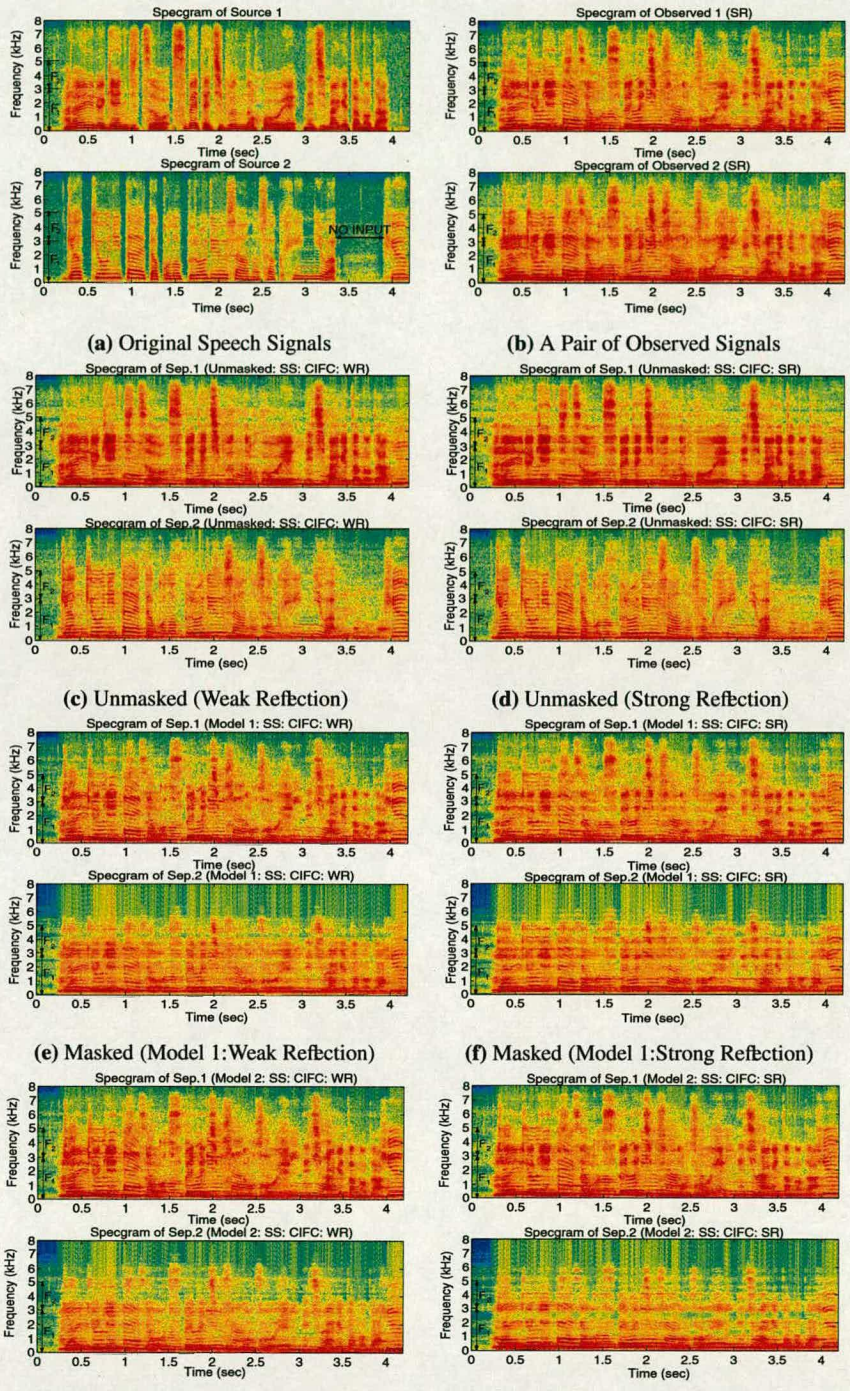


(g) Masked (Model 2:Weak Reflection)



(h) Masked (Model 2:Strong Reflection)

Figure 5.6: *Original Speech Sources, a Pair of Observed and Separated Speech Signals for Unmasked and Masked FDICA Systems (Perceptual Preprocessing: Subspace Method: Synthetic Room Recording Scenario)*



(a) Original Speech Signals

(b) A Pair of Observed Signals

(c) Unmasked (Weak Reflection)

(d) Unmasked (Strong Reflection)

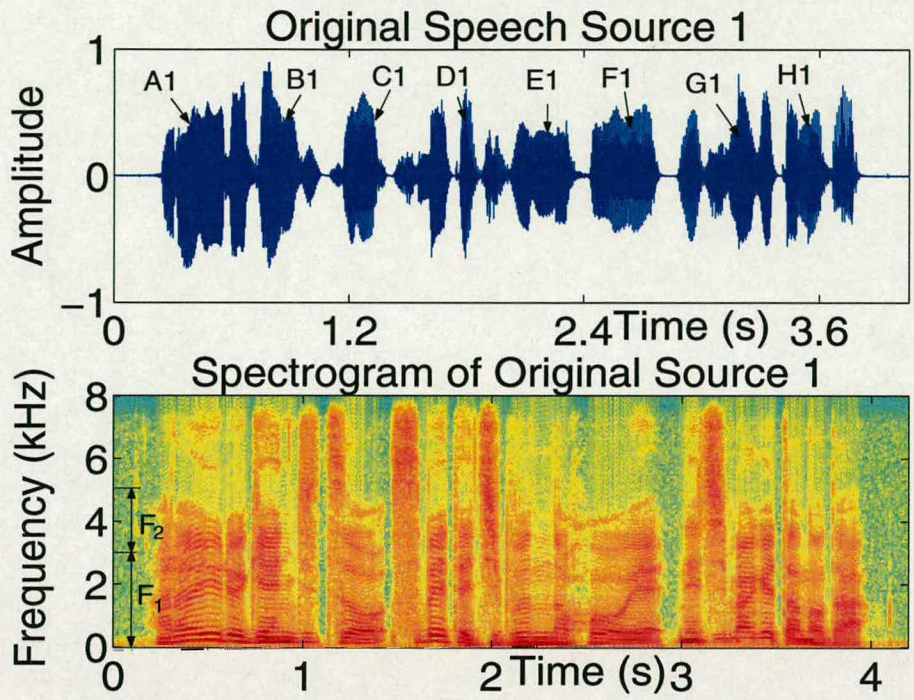
(e) Masked (Model 1:Weak Reflection)

(f) Masked (Model 1:Strong Reflection)

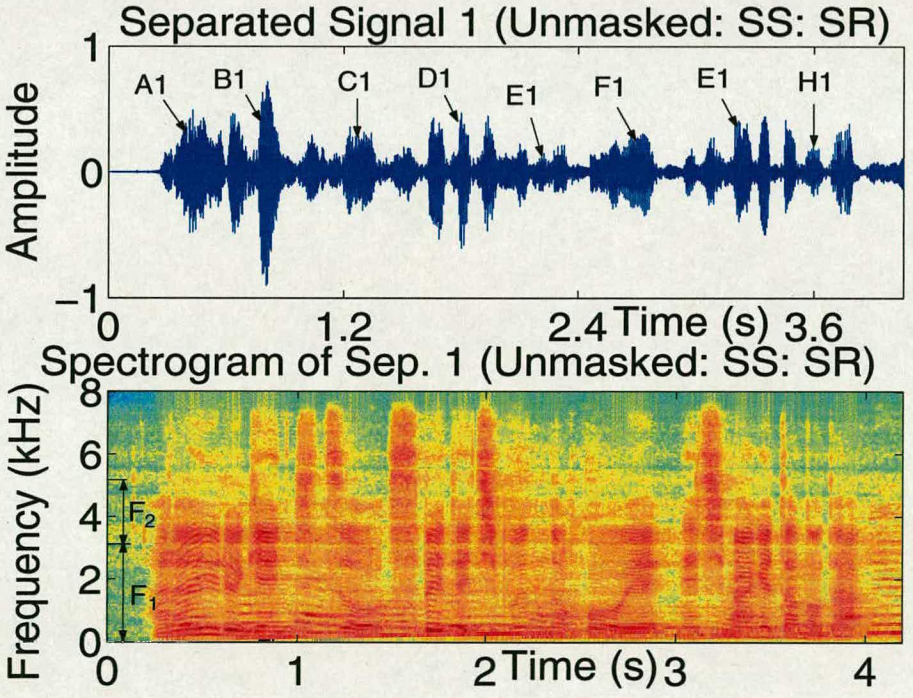
(g) Masked (Model 2:Weak Reflection)

(h) Masked (Model 2:Strong Reflection)

Figure 5.7: *Spectrograms of Original Speech Sources, a Pair of Observed and Separated Speech Signals for Unmasked and Masked FDICA Systems (Perceptual Preprocessing: Subspace Method: Synthetic Room Recording Scenario)*

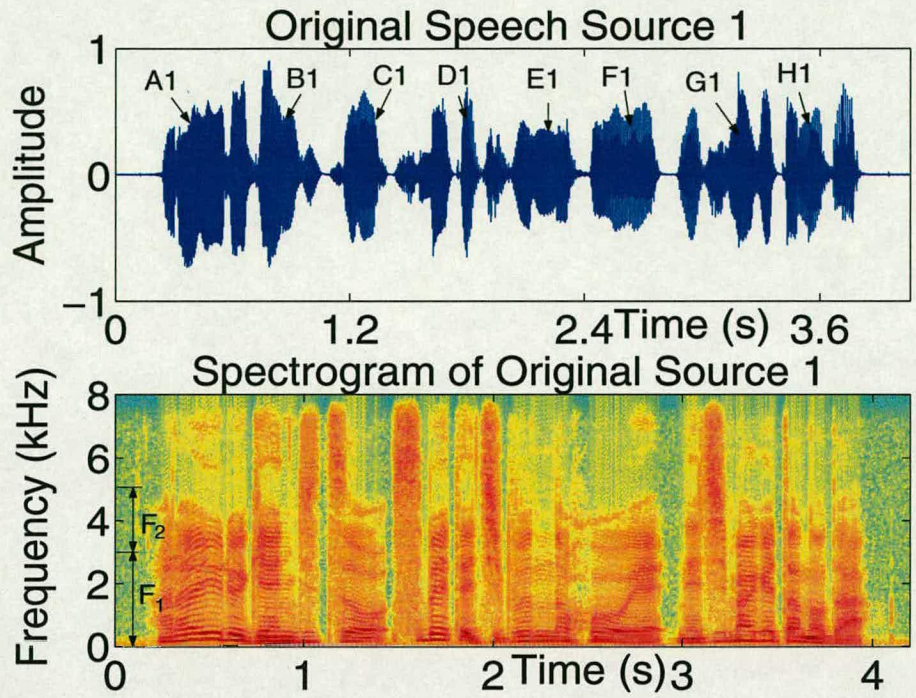


(a) Source 1 and its Spectrogram

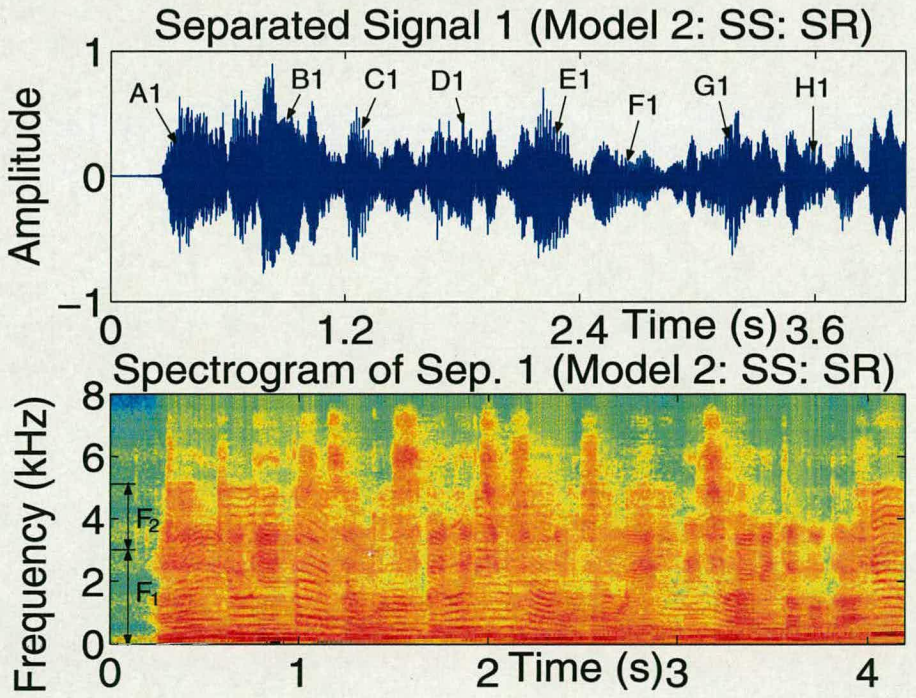


(b) Sep. 1 and its Spectrogram (Unmasked)

Figure 5.8: *Separated Signal 1 and its Spectrogram for the Unmasked FDICA System When Speech Source 1 and its Spectrogram are Known (SS: Strong Reflection Case)*

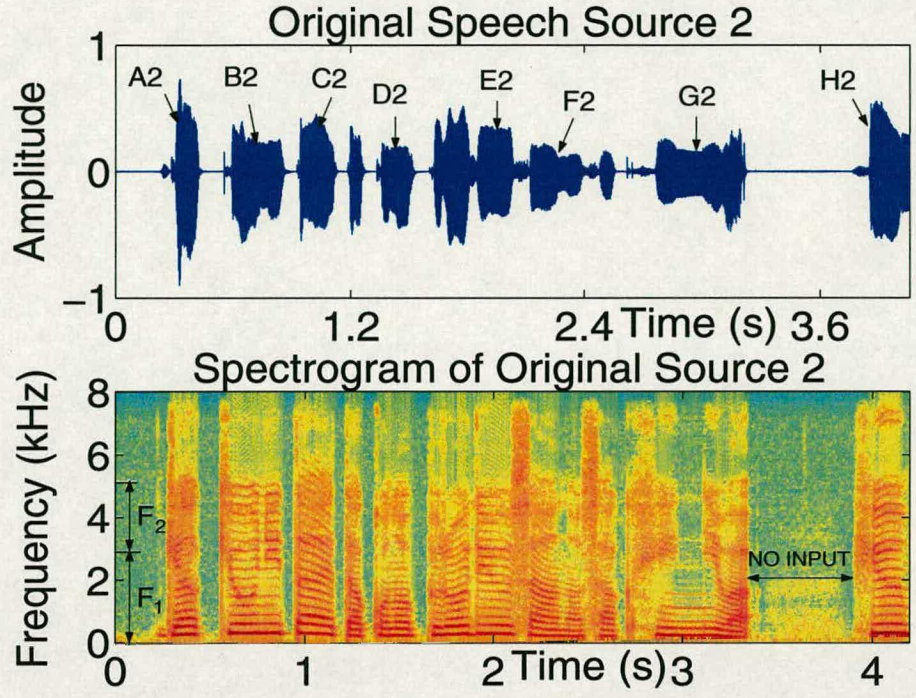


(a) Source 1 and its Spectrogram

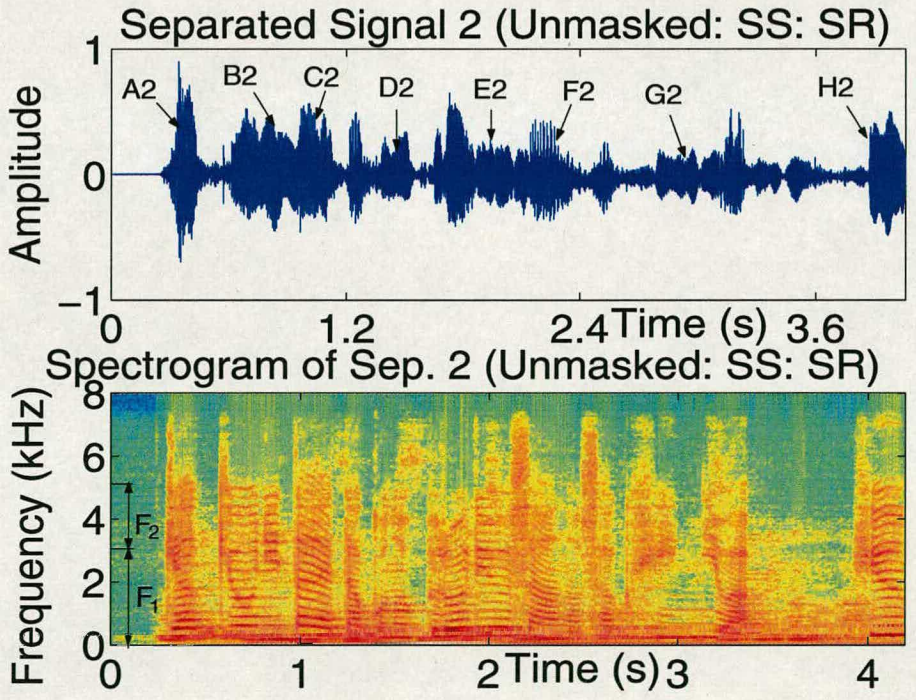


(b) Sep. 1 and its Spectrogram (Masked)

Figure 5.9: *Separated Signal 1 and its Spectrogram for the Masked FDICA System When Speech Source 1 and its Spectrogram are Known (SS: Strong Reflection Case)*

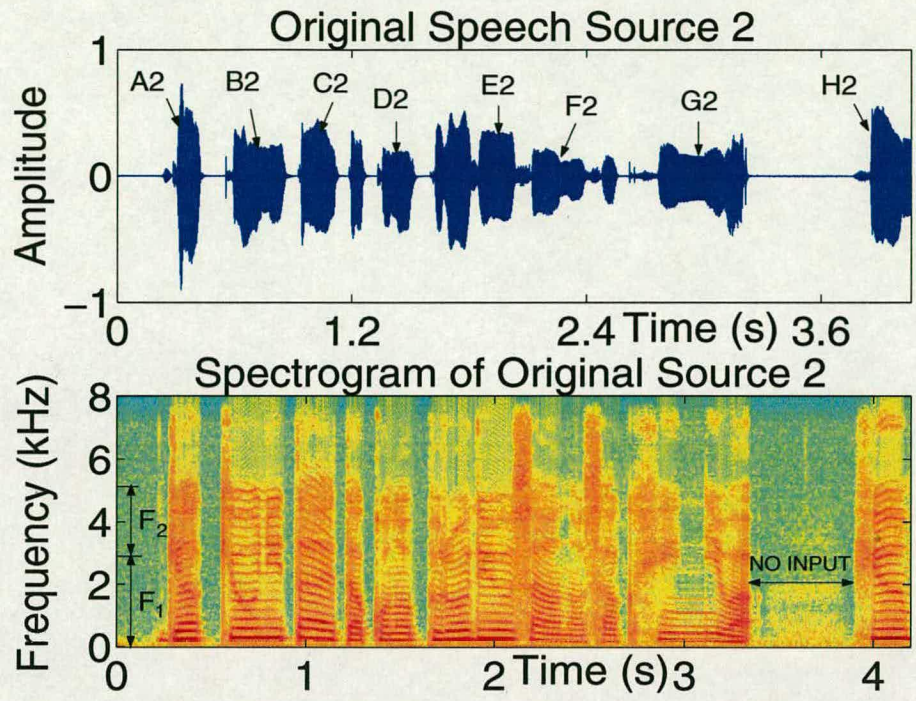


(a) Source 2 and its Spectrogram

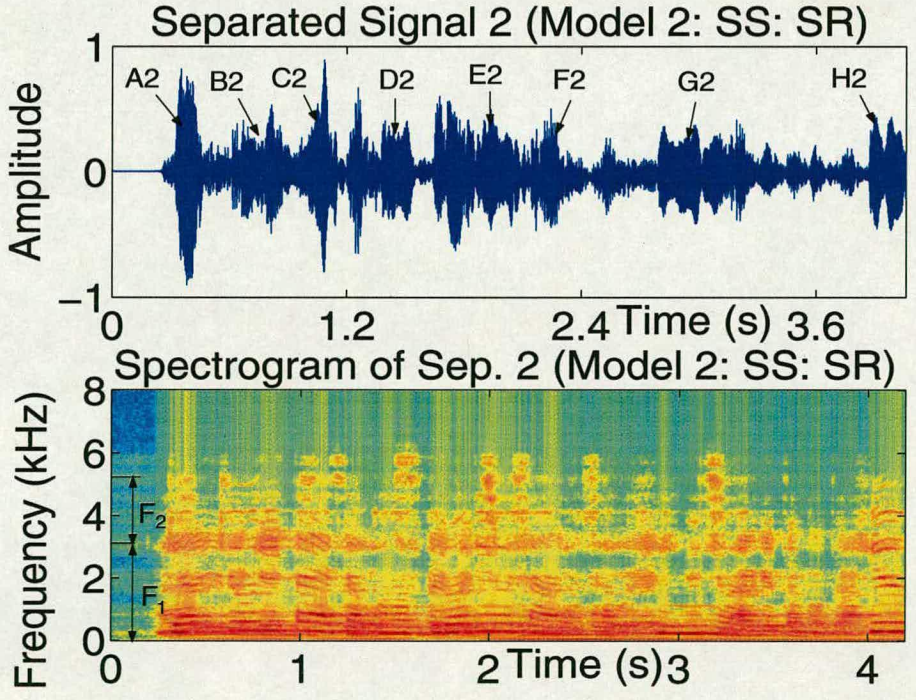


(b) Sep. 2 and its Spectrogram (Unmasked)

Figure 5.10: *Separated Signal 2 and its Spectrogram for the Unmasked FDICA System When Speech Source 2 and its Spectrogram are Known (SS: Strong Reflection Case)*

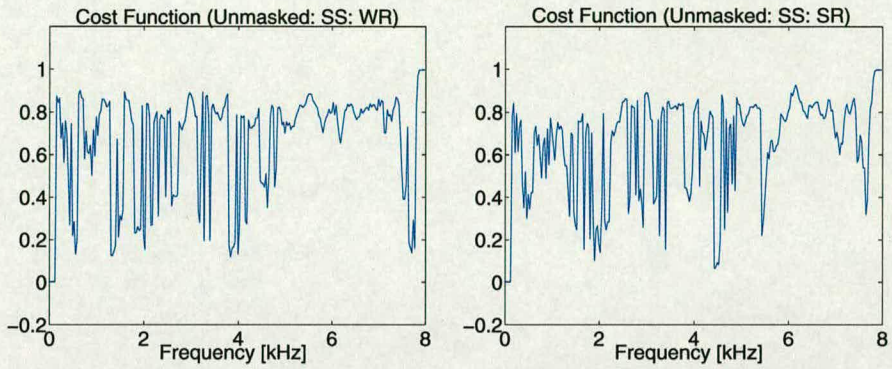


(a) Source 2 and its Spectrogram



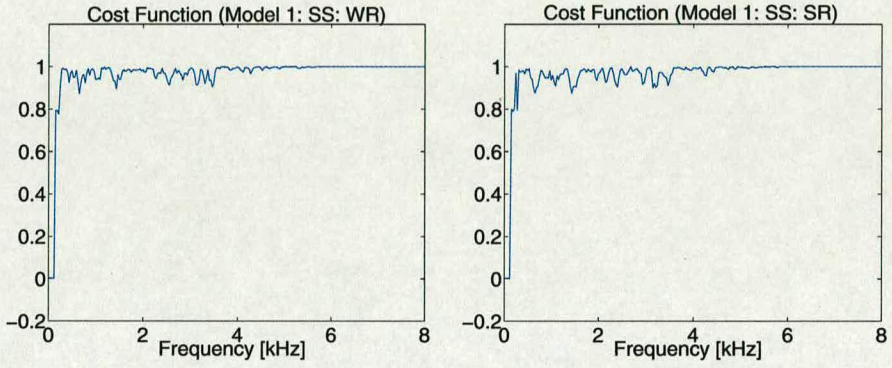
(b) Sep. 2 and its Spectrogram (Masked)

Figure 5.11: *Separated Signal 2 and its Spectrogram for the Masked FDICA System When Speech Source 2 and its Spectrogram are Known (SS: Strong Reflection Case)*



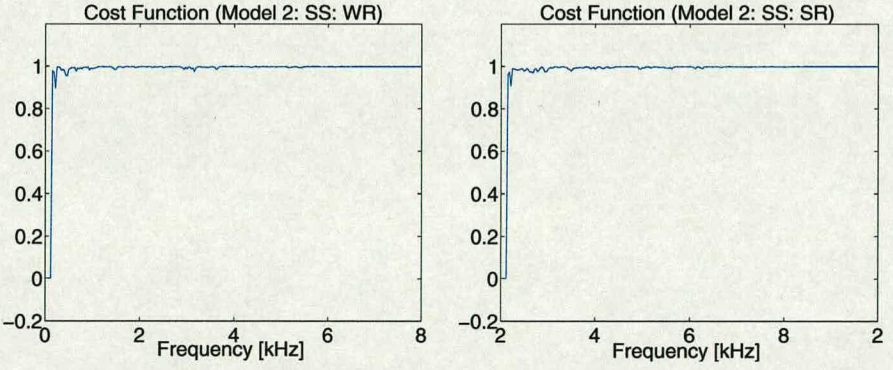
(a) Unmasked (Weak Reflection)

(b) Unmasked (Strong Reflection)



(c) Masked (Model 1: Weak Reflection)

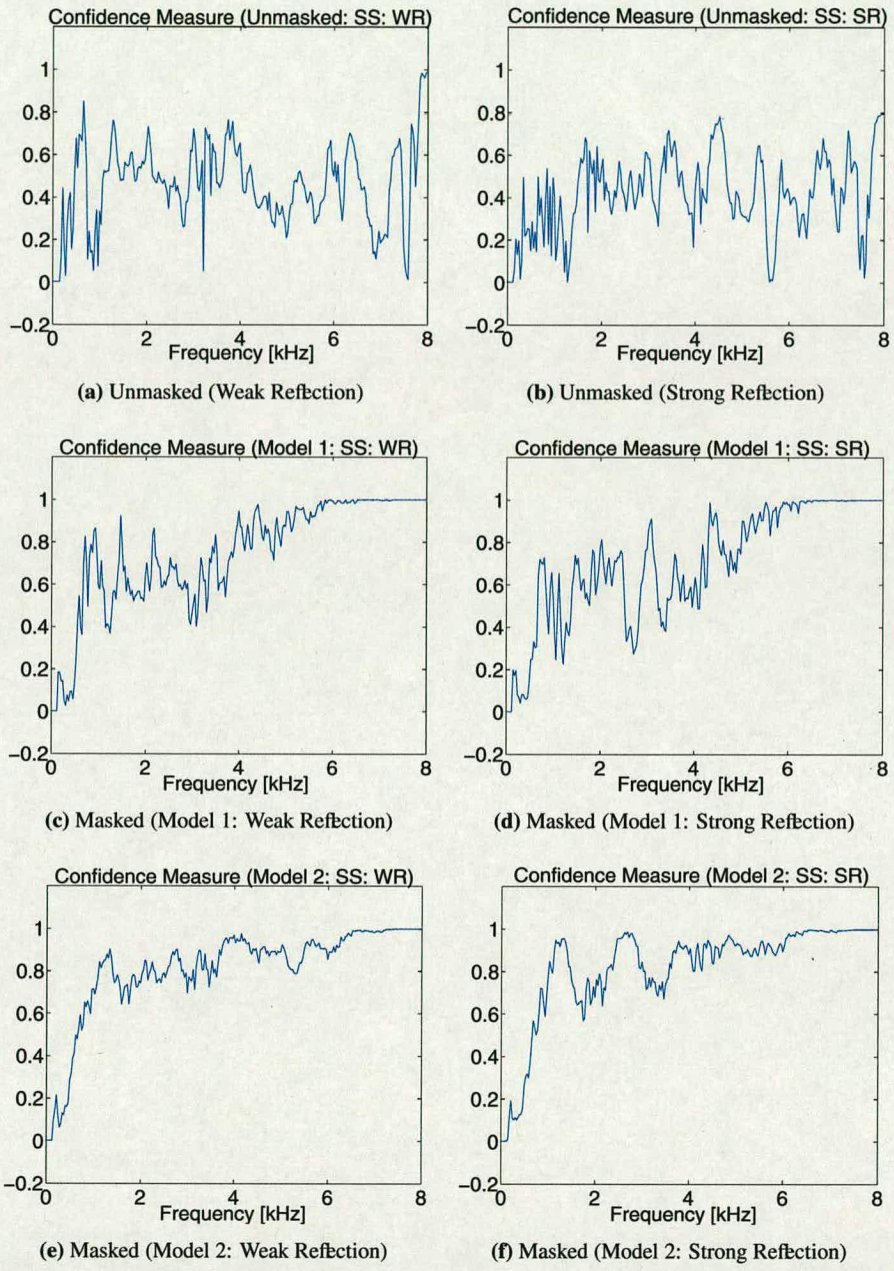
(d) Masked (Model 1: Strong Reflection)



(e) Masked (Model 2: Weak Reflection)

(f) Masked (Model 2: Strong Reflection)

Figure 5.12: *Measured Cost Function for Unmasked and Masked FDICA Systems (Perceptual Preprocessing: Subspace Method: Synthetic Room Recording Scenario)*



(a) Unmasked (Weak Reflection)

(b) Unmasked (Strong Reflection)

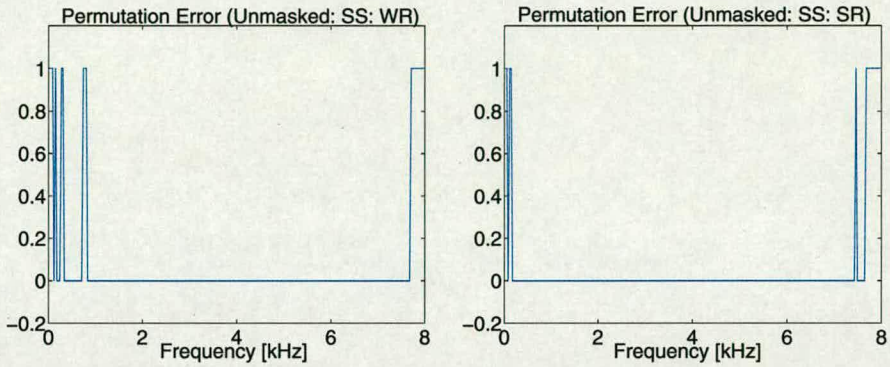
(c) Masked (Model 1: Weak Reflection)

(d) Masked (Model 1: Strong Reflection)

(e) Masked (Model 2: Weak Reflection)

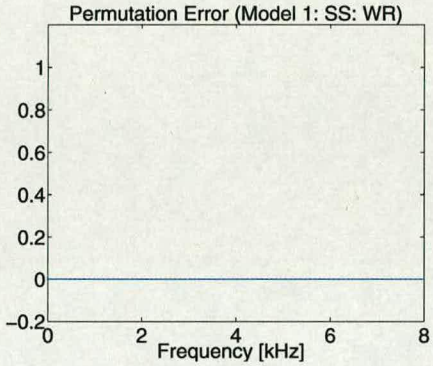
(f) Masked (Model 2: Strong Reflection)

Figure 5.13: *The Confidence Measure for Unmasked and Masked FDICA Systems (Perceptual Preprocessing: Subspace Method: Synthetic Room Recording Scenario)*

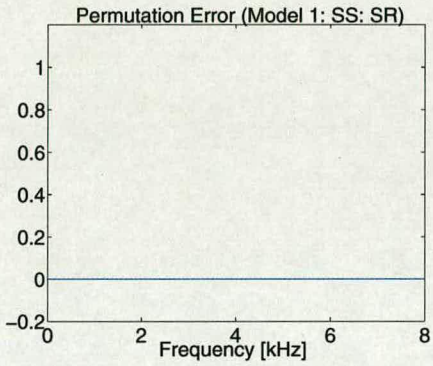


(a) Unmasked (Weak Reflection)

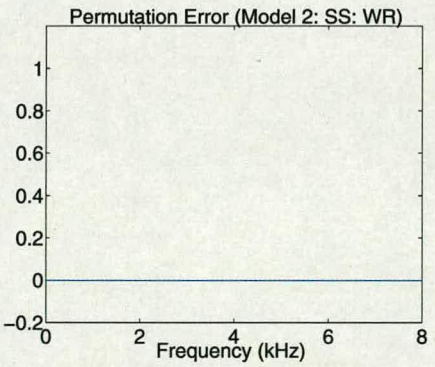
(b) Unmasked (Strong Reflection)



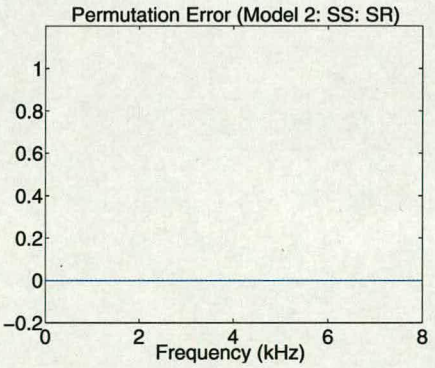
(c) Masked (Model 1: Weak Reflection)



(d) Masked (Model 1: Strong Reflection)



(e) Masked (Model 2: Weak Reflection)



(f) Masked (Model 2: Strong Reflection)

Figure 5.14: *The Permutation Error for Unmasked and Masked FDICA Systems (Perceptual Preprocessing: Subspace Method: Synthetic Room Recording Scenario)*

5.7.2 Real Room Mixing Scenario

This experiment was chosen to test the algorithm's ability in the real room recording situation. To do this, we used real room recorded speech signals (6 s at 16 kHz) as shown in Fig. 5.15.

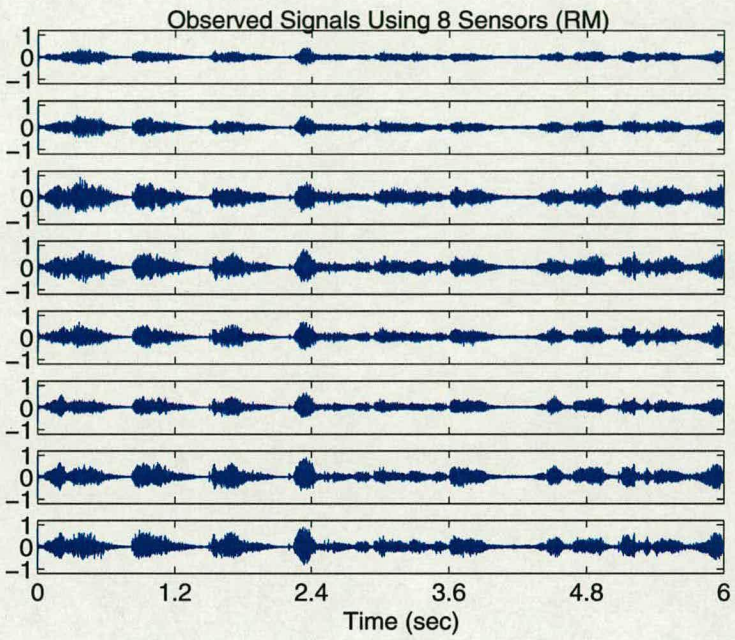


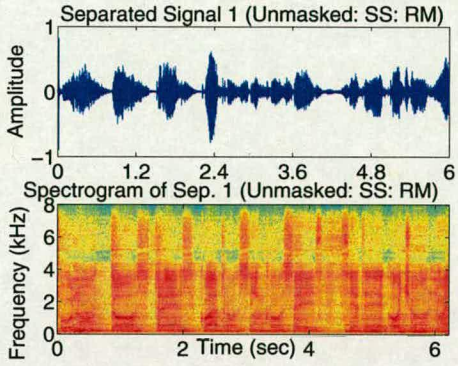
Figure 5.15: *Observed Signals Using Circular Microphone Array (Real Room Recording)*

Real room mixing results of separated signals and the corresponding spectrograms for both unmasked and masked systems (perceptual models 1 and 2) are shown in Fig. 5.16.

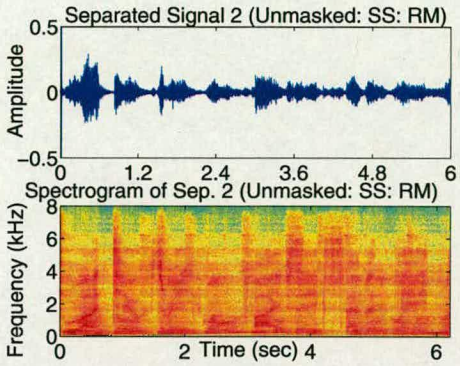
From Fig. 5.16, it is clearly evident that psychoacoustic model 2 helps to enhance the separation of targeted output signal (second output) when compared to that of psychoacoustic model 1. This is clearly observed in the corresponding spectrogram of the second separated output signal. Here, most of the higher frequencies (> 3 kHz) are completely suppressed by both simultaneous frequency masking and temporal masking techniques.

Further, other real room recording results are similar to that of previous experiment from the measured values of cost function and confidence measure point of view (Fig. 5.17).

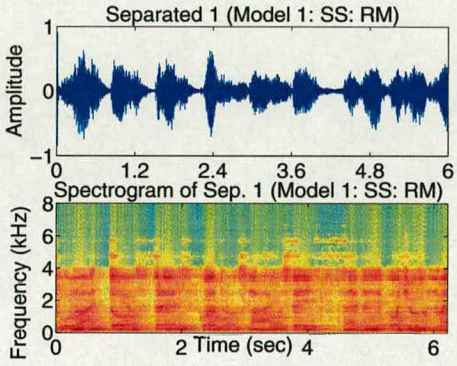
The permutation error cannot be computed in this real room recording case as the original speech sources are unknown.



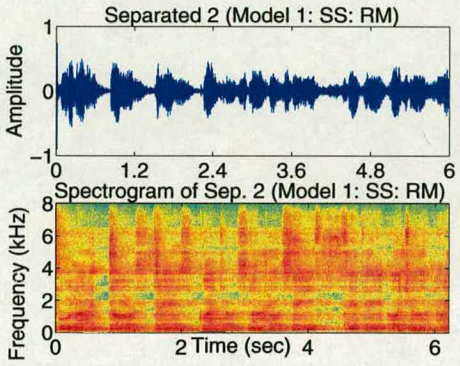
(a) Sep. 1 and its Spectrogram (Unmasked)



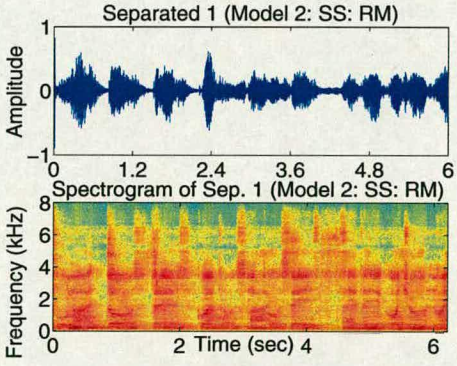
(b) Sep. 2 and its Spectrogram (Unmasked)



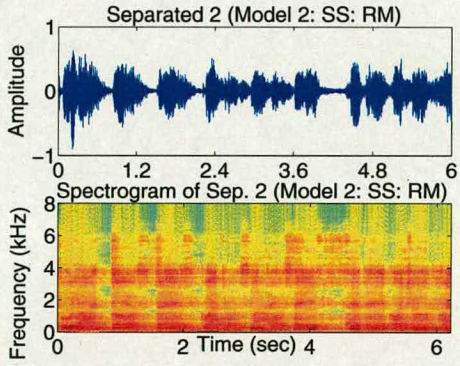
(c) Sep. 1 and its Spectrogram (Model 1)



(d) Sep. 2 and its Spectrogram (Model 1)

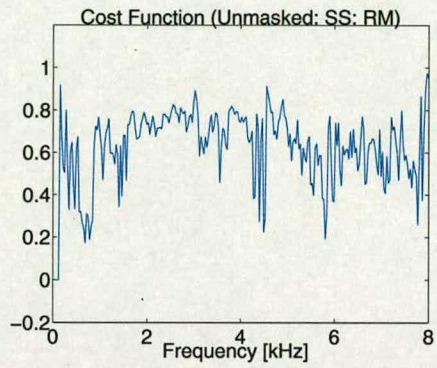


(e) Sep. 1 and its Spectrogram (Model 2)

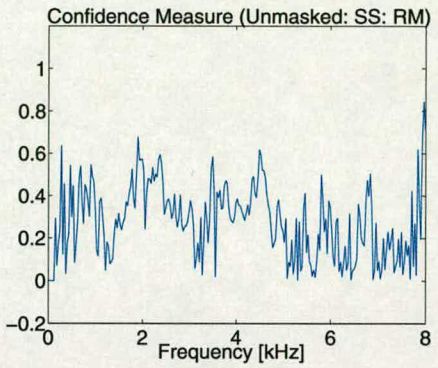


(f) Sep. 2 and its Spectrogram (Model 2)

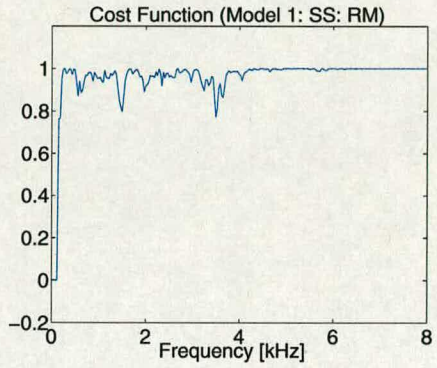
Figure 5.16: *Separated Signals and Spectrograms for Unmasked and Masked FDICA Systems (Perceptual Preprocessing: Subspace Method: Real Room Recording Scenario)*



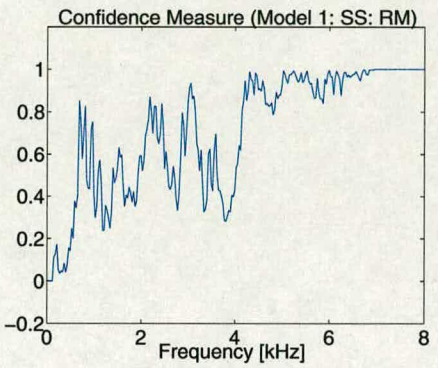
(a) Unmasked (Real Room)



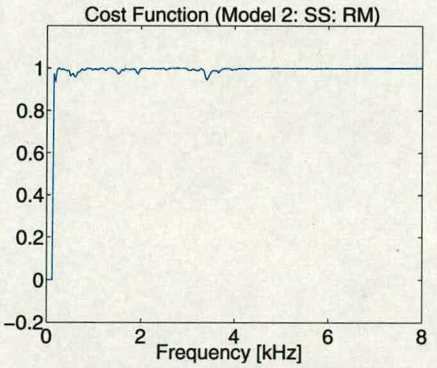
(b) Unmasked (Real Room)



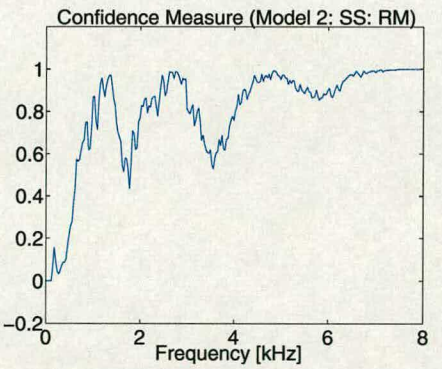
(c) Masked (Model 1: Real Room)



(d) Masked (Model 1: Real Room)



(e) Masked (Model 2: Real Room)



(f) Masked (Model 2: Real Room)

Figure 5.17: *Cost Function and Confidence Measure for Unmasked and Masked FDICA Systems (Perceptual Preprocessing: Subspace Method: Real Room Mixing Scenario)*

5.8 Performance Evaluation

5.8.1 Time-Domain Objective Quality Measure

The results of objective performance evaluation based on the time-domain metric (**SIR**) [42] are summarized in Tables 5.2 and 5.3.

	Unmasked FDICA		Masked FDICA (Model 1)		Masked FDICA (Model 2)	
Method	SIR_1	SIR_2	SIR_1	SIR_2	SIR_1	SIR_2
SOC	10.54	13.68	-1.07	14.18	-1.39	18.66
CIFC	10.00	14.19	-1.07	14.18	-1.39	18.66

Table 5.2: ***SIR (dB) for Unmasked and Masked FDICA Systems (Subspace Method: WR)***

	Unmasked FDICA		Masked FDICA (Model 1)		Masked FDICA (Model 2)	
Method	SIR_1	SIR_2	SIR_1	SIR_2	SIR_1	SIR_2
SOC	12.23	10.03	-1.75	14.96	-1.37	18.95
CIFC	11.68	10.22	-1.75	14.96	-1.37	18.95

Table 5.3: ***SIR (dB) for Unmasked and Masked FDICA Systems (Subspace Method: SR)***

From these Tables, it is clearly evident that SIR_1 obtained by the masked FDICA system (using both psychoacoustic models) is severely affected by room reverberations resulting in very poor performance when compared to that of the unmasked FDICA system under both weak and strong early reflection cases of the mixing environment.

Further, it is also evident from Table 5.2 that there is no significant improvement in SIR_2 when the masked FDICA system (using psychoacoustic model 1) is considered for the weak early reflection case. On the other hand, SIR_2 improved by 5 dB when the psychoacoustic model 2 is used for the same weak reflection case of the masked system.

From Table 5.3, it can be seen that SIR_2 improved by 5 dB when the masked FDICA system (using psychoacoustic model 1) is considered for the strong early reflection case. On the other hand, the psychoacoustic model 2 improves the SIR_2 by 9 dB for the same strong early reflection case of the masked FDICA system.

Thus, the value of SIR_2 obtained by the perceptually motivated FDICA system using the subspace method substantiates in improving the separation performance of a BSS system that targets the second output when the mixing is noisy and highly reverberant.

5.8.2 Perceptual Domain Objective Quality Measure

The objective performance evaluation based on the perceptual domain metric (**EMBSD**) [144] are summarized in Tables 5.4 and 5.5.

Method	Unmasked FDICA		Masked FDICA (Model 1)		Masked FDICA (Model 2)	
	$EMBSD_1$	$EMBSD_2$	$EMBSD_1$	$EMBSD_2$	$EMBSD_1$	$EMBSD_2$
SOC	3.9	7.3	8.3	3.9	6.9	3.6
CIFC	3.8	6.9	8.3	3.9	6.9	3.6

Table 5.4: ***EMBSD (dB) for Unmasked and Masked Systems (Subspace Method: WR)***

Method	Unmasked FDICA		Masked FDICA (Model 1)		Masked FDICA (Model 2)	
	$EMBSD_1$	$EMBSD_2$	$EMBSD_1$	$EMBSD_2$	$EMBSD_1$	$EMBSD_2$
SOC	3.9	7.4	7.5	3.3	6.9	3.1
CIFC	3.8	6.5	7.5	3.3	6.9	3.1

Table 5.5: ***EMBSD (dB) for Unmasked and Masked Systems (Subspace Method: SR)***

From Table 5.4, it is clearly evident that $EMBSD_2$ reduced by 3.4 dB and 3 dB using SOC and CIFC methods respectively when the masked FDICA system (using model 1) is considered for the weak reflection case of the mixing environment.

On the other hand, the psychoacoustic model 2 reduces the $EMBSD_2$ by 3.7 dB and 3.3 dB using SOC and CIFC permutation methods respectively when the masked FDICA system is considered for the same weak reflection case of the mixing environment.

From Table 5.5, it can be seen that the measured $EMBSD_2$ reduced by 4.1 dB and 3.2 dB using SOC and CIFC methods respectively when the masked FDICA system (using psychoacoustic model 1) is considered for the strong early reflection case.

However, the psychoacoustic model 2 reduces the value of $EMBSD_2$ by 4.4 dB and 3.4 dB using SOC and CIFC methods respectively when the masked FDICA system is considered for the same strong reflection case of the mixing environment.

Thus, the value of $EMBSD_2$ obtained by the perceptually motivated FDICA system using the subspace method substantiates in improving the separation performance of a BSS system that targets the second output when the mixing is noisy and highly reverberant.

5.9 Summary

In this study, we explored the Blind Source Separation problem of convolved speech mixtures (when the mixing environment is noisy and highly reverberant), in the case of more sensors than sources, proposing a perceptual solution. The key points are:

A perceptually motivated FDICA scheme with subsapce method, proposed in this chapter, not only reduces the perceptually irrelevant frequencies by exploiting the masking properties of the input speech spectrum but also reduces the perceptually relevant room reflections by utilising the properties of the subspace filtering. Further, this proposed system also reduces the computation complexity of a similarity measure among spectral envelopes of the separated output signals for solving the permutation. Finally, the proposed system completely avoids the permutation problem of FDICA while targetting the specific sound source.

The measured permutation error is 7.80% and 6.60% for the unmasked FDICA system under both weak and strong reflection conditions respectively. On the other hand, the permutation error is zero for the masked FDICA system under both reflection cases respectively.

Further, it is clearly observed that SIR_2 is enhanced by 5 dB and 9 dB using psychoacoustic models 1 and 2 respectively when the strong reflection case of the masked FDICA system is taken into account. On the other hand, it can be seen that SIR_2 is improved by 5 dB using model 2 when the mixing is noisy and lowly reverberant (weak reflection case).

On the other hand, the measured $EMBSD_2$ for the masked FDICA system (using either model) is more effective when compared to that of the unmasked system in improving the performance of a BSS system under both reflection conditions.

Thus, incorporating the proposed perceptual solution for the permutation problem of FDICA system produced good separation results in terms of the measured permutation error, SIR and EMBSD by exploiting both perceptual irrelevancy and statistical redundancy.

Chapter 6
Conclusions and Future Work

Each chapter in this thesis is largely self contained, providing a more detailed discussion of the topics and experimental results contained in that chapter. The following presents a summary of the work that has been conducted, highlighting the contributions to knowledge made during the study, drawing conclusions from the research and identifying some suggestions for future directions of research work.

6.1 Conclusions

In this thesis, we have explored many aspects in the field of *Perceptually Motivated Blind Source Separation of Convolutive Audio Mixtures*. The work was largely based around the natural gradient version of the complex Infomax algorithm for our investigation of a BSS system in the frequency-domain. Other BSS systems considered were a simple, non-iterative multiple time-delayed decorrelation algorithm, based on second order statistics. This section will conclude the main issues of the problem, giving specific emphasis on the observations and improvements introduced in this thesis.

We investigated the effect of perceptual irrelevancy removal techniques on the performance of blind source separation systems when the mixing is noisy and higly reverberant. We decomposed the source separation problem into two subproblems. Firstly, the subproblem of source separation of real world recordings is investigated in the case of an equal number of sources and sensors. Finally, the subproblem of more sensors than sources using a microphone array is then examined for real world audio recordings.

The FDICA framework has inherent scaling and the permutation ambiguity problems. For the scaling ambiguity problem, we considered the method proposed by Murata et al. On the other hand, for the permutation ambiguity problem, we proposed a perceptually relevant method denoted by CIFC based on the combined approaches of the inter-frequency coherency (IFC) of the mixing matrices and the inter-frequency spectral envelope correlation (IFSEC) of the separated speech signal at several adjacent frequencies.

6.1.1 Perceptual Preprocessing: Multiple TDD Algorithm

In the first instance, we have considered non-iterative multiple time-delayed decorrelation algorithm for our initial experimentation in a synthetic room mixing scenario in the case of an equal number of sources and sensors. Based on the experimental results, it is clearly observed that the measured permutation error is slightly reduced by the forward temporal masking which is used in model 2. However, this algorithm fails when the speech sources have an identical spectral envelopes and even if one spectral component of the input speech signal does not have any power for both unmasked and masked FDICA systems.

6.1.2 Perceptual Preprocessing: Complex Infomax Algorithm

Therefore, we considered an iterative procedure based on the complex Infomax algorithm with feed-forward architecture for our further investigation in the case of an equal number of sources and sensors. Based on the experimental results in a synthetic room mixing scenario, we observed that the measured permutation error is zero for both early reflection cases of the masked FDICA system (using either psychoacoustic model). Results from early studies by Guddeti and Mulgrew [150] suggested that by reducing perceptually irrelevant frequency components might enhance the separation performance.

However, the objective performance in one of the channels as evaluated by the time-domain metric (signal-to-interference ratio (SIR)) and the perceptual-domain metric (enhanced modified Bark spectral distortion (EMBSD)) are improved when either psychoacoustic model is employed for both early reflection cases. Though, SIR and EMBSD are better than those obtained by the unmasked FDICA system, but the informal listening test confirms the presence of the perceptually relevant room reflections in the separated speech signal obtained by ICA resulting in poor performance. Hence, we conclude that the proposed FDICA system with a perceptually motivated preprocessing filter reduced the computational complexity of a similarity measure by more than 50% while avoiding the permutation ambiguity problem.

6.1.3 Perceptual Postprocessing: Complex Infomax Algorithm

The perceptual irrelevancy reduction techniques were subsequently introduced after the source separation of the speech signals obtained from the complex Infomax algorithm in the case of an equal number of sources and sensors. Results from early experimentation in a synthetic room

mixing scenario suggested that by reducing the perceptually irrelevant frequency components using the perceptually motivated postprocessing filter using the psychoacoustic model 2 often appeared to reduce the permutation ambiguity problem partially. The more detailed analyses developed in this thesis have shown that the performance of the masked FDICA system (using either model) in both the channels as evaluated by SIR and EMBSD is poor when compared to that of the unmasked FDICA system for both reflection cases.

6.1.4 Perceptually Motivated Subspace Method: Complex Infomax Algorithm

Finally, we explored the idea of perceptually more efficient FDICA system in the case of more sensors than sources. By exploiting the perceptual irrelevancy of some of the input speech signal spectrum using perceptual masking techniques before utilizing the subspace method that reduces the effect of room reflections prior to ICA, we realize a perceptually more efficient FDICA system that targets a specific sound source (second source in our case).

Results from early studies by Guddeti and Mulgrew [151] suggested that by reducing both perceptually irrelevant frequency components and the effect of room reflections might help to enhance the separation performance by completely avoiding the permutation ambiguity problem of the FDICA system. However, the objective performance in one of the channels as evaluated by SIR and EMBSD are further improved when both psychoacoustic models are employed for both early reflection cases. Though, the values are better than those obtained by the unmasked FDICA system, but the informal listening test confirms the presence of the non-harmonic distortion in the separated speech signal resulting in poor audio quality.

6.1.5 Possible Reasons for Poor Audio Quality

- Due to the variation of the perceptual binary mask from frame to frame, the direct use of the mask would result in an output speech containing non-harmonic distortion caused by temporal aliasing and therby degrading separated speech quality [134, 135].
- For minimising this temporal aliasing, the perceptual binary mask should be smoothed by convolving it with an optimal digital prolate spheroidal window in the frequency-domain [152, 153]. The digital prolate spheroidal window is optimal in the sense that it concentrates most of its energy in the mainlobe and attenuates the aliasing components at its sidelobes, leading to a maximised signal to distortion power ratio.

- Further, by listening to the original speech source, we strongly felt that the original signal itself has aliasing due to non-harmonic components.
- Since EMBSD measure does not consider the relative significance of the loudness difference for spectral peaks and valleys above the noise masking threshold, the EMBSD measure is not the best objective speech quality measure [144].

6.2 Suggestions for Future Work

In addition to the work presented above, this study has identified other ideas of research which could not be addressed due to time constraints. Suggestions of areas worthy of further study to extend this interesting field of research are listed below. Some areas are of particular relevance to lines of research undertaken in this study, while others are more general in nature and may be of interest to the wider BSS / ICA research community.

- For minimising the temporal aliasing, the perceptual binary mask should be smoothed by convolving it with an optimal digital prolate spheroidal window in the frequency-domain and thereby improving separated speech signal quality while reducing the non-harmonic speech components from the separated signal.
- To confirm the promising SIR and EMBSD values, intensive subjective listening tests such as comparision category rating (CCR) procedure and absolute category rating (ACR) procedure to be performed. In the CCR test, listeners are presented with pairs of speech samples (sentences) and for each pair, they are asked to grade the quality of the latter sample with respect to the former. Each pair contains a processed sample and a quality reference that are presented in random order.

  In the ACR test, the listeners use a five-point scale to grade the quality of the samples that have been made processed with the different test conditions. The average of all scores given to a particular condition yields the corresponding mean opinion score (MOS).
- By integrating backward temporal masking technique with simultaneous frequency masking and forward temporal masking, a new perceptually motivated and more biologically plausible FDICA algorithm that tracks the speech signal related to the voice pitch (f_0 and its harmonics) during the learning might help the performance of the BSS system.

- In *real cocktail party* environment there are less number of microphones (sensors) than unknown sound sources resulting in the overcomplete (underdetermined) convolutive mixing problem. However, humans deal with this *real cocktail party problem* very easily and effectively by using only 2 dynamic sensors (ears). Hence, a possible extension of the proposed framework in this thesis is to adapt strategies for the case of more sources than sensors in the convolutive mixing case of audio source separation problem.

- Although the STFFT based signal decompositions are used in this thesis, they were by no means biologically accurate models for *cochlear function*. The accurate reconstruction of the *cochlear function* has been extensively modelled by the *gammatone filterbank*. A *gammatone filterbank* is composed of basis functions which are sinusoidal tones modulated by gamma distributions, might help to realize a more biologically plausible and intelligent BSS system that tracks the speech signal related to the voice pitch.

Appendix A
Tables of Human Auditory System

A.1 Critical Bands

Critical band numbers and the corresponding frequency limits in Hertz, are presented in the following Table as per the data published in [111].

Subband Number	Lower Edge [Hz]	Center [Hz]	Upper Edge [Hz]
0	0	50	100
1	100	150	200
2	200	250	300
3	300	350	400
4	400	450	510
5	510	570	630
6	630	700	770
7	770	840	920
8	920	1000	1080
9	1080	1170	1270
10	1270	1370	1480
11	1480	1600	1720
12	1720	1850	2000
13	2000	2150	2320
14	2320	2500	2700
15	2700	2900	3150
16	3150	3400	3700
17	3700	4000	4400
18	4400	4800	5300
19	5300	5800	6400
20	6400	7000	7700
21	7700	8500	9500
22	9500	10500	12000
23	12000	13500	15500
24	15500		

Table A.1: ***Critical Bands and Their Frequency Range***

A.2 Calculation Partition Table (Psychoacoustic Model 2)

Table A.2: **Calculation Partition at 32 kHz Sampling Rate**

Index	wlb	whb	bvb	mvb	TMNb
1	1	1	0.00	0.0	24.5
2	2	4	0.63	0.0	24.5
3	5	7	1.56	20.0	24.5
4	8	10	2.50	20.0	24.5
5	11	13	3.44	20.0	24.5
6	14	16	4.34	20.0	24.5
7	17	19	5.17	20.0	24.5
8	20	22	5.94	20.0	24.5
9	23	25	6.63	17.0	24.5
10	26	28	7.28	15.0	24.5
11	29	31	7.90	15.0	24.5
12	32	34	8.50	10.0	24.5
13	35	37	9.06	7.0	24.5
14	38	41	9.65	7.0	24.5
15	42	45	10.28	4.4	24.8
16	46	49	10.87	4.4	25.4
17	50	53	11.41	4.5	25.9
18	54	57	11.92	4.5	26.4
19	58	61	12.39	4.5	26.9
20	62	65	12.83	4.5	27.3
21	66	70	13.29	4.5	27.8
22	71	75	13.78	4.5	28.3
23	76	81	14.27	4.5	28.8
24	82	87	14.76	4.5	29.3
25	88	93	15.22	4.5	29.7
26	94	99	15.63	4.5	30.1
27	100	106	16.06	4.5	30.6
28	107	113	16.47	4.5	31.0

Continued on Next Page...

Table A.2 – Continued

Index	**wlb**	**whb**	**bvb**	**mvb**	**TMNb**
29	114	120	16.86	4.5	31.4
30	121	129	17.25	4.5	31.8
31	130	138	17.65	4.5	32.2
32	139	148	18.05	4.5	32.5
33	149	159	18.42	4.5	32.9
34	160	170	18.81	4.5	33.3
35	171	183	19.18	4.5	33.7
36	184	196	19.55	4.5	34.1
37	197	210	19.93	4.5	34.4
38	211	225	20.29	4.5	34.8
39	226	240	20.65	4.5	35.2
40	241	258	21.02	4.5	35.5
41	259	279	21.38	4.5	35.9
42	280	300	21.74	4.5	36.2
43	301	326	22.10	4.5	36.6
44	327	354	22.44	4.5	36.9
45	355	382	22.79	4.5	37.3
46	383	420	23.14	4.5	37.6
47	421	458	23.49	4.5	38.0
48	459	496	23.83	4.5	38.3
49	497	513	24.07	4.5	38.6

A.3 Absolute Threshold Table (Psychoacoustic Model 2)

A value of 0 dB represents a level in the absolute threshold calculation of 96 dB below the energy of a sine wave of amplitude ±32760.

Table A.3: **Absolute Threshold at 32 kHz Sampling Rate**

Lower Index	Higher Index	Absthrs (dB)
1	1	58.23
2	2	33.44
3	3	24.17
4	4	19.20
5	5	16.05
6	6	13.87
7	7	12.26
8	8	11.01
9	9	10.01
10	10	9.20
11	11	8.52
12	12	7.94
13	13	7.44
14	14	7.00
15	15	6.62
16	16	6.28
17	17	5.97
18	18	5.70
19	19	5.44
20	20	5.21
21	21	5.00
22	22	4.80
23	23	4.62
24	24	4.45
25	25	4.29
26	26	4.14

Continued on Next Page...

Table A.3 – Continued

Lower Index	**Higher Index**	**Absthrs (dB)**
27	27	4.00
28	28	3.86
29	29	3.73
30	30	3.61
31	31	3.49
32	32	3.37
33	33	3.26
34	34	3.15
35	35	3.04
36	36	2.93
37	37	2.83
38	38	2.73
39	39	2.63
40	40	2.53
41	41	2.42
42	42	2.32
43	43	2.22
44	44	2.12
45	45	2.02
46	46	1.92
47	47	1.81
48	48	1.71
49	50	1.49
51	52	1.27
53	54	1.04
55	56	0.80
57	57	0.55
59	60	0.29
61	62	0.02
63	64	-0.25

Continued on Next Page...

Table A.3 – Continued

Lower Index	**Higher Index**	**Absthrs (dB)**
65	66	-0.54
67	68	-0.83
69	70	-1.12
71	72	-1.43
73	74	-1.73
75	76	-2.04
77	78	-2.34
79	80	-2.64
81	82	-2.93
83	84	-3.22
85	86	-3.49
87	88	-3.74
89	90	-3.98
91	92	-4.20
93	94	-4.40
95	96	-4.57
97	100	-4.82
101	104	-4.96
105	108	-4.97
109	112	-4.86
113	116	-4.63
117	120	-4.29
121	124	-3.87
125	128	-3.39
129	132	-2.86
133	136	-2.31
137	140	-1.77
141	144	-1.24
145	148	-0.74
149	152	-0.29

Continued on Next Page...

Table A.3 – Continued

Lower Index	**Higher Index**	**Absthrs (dB)**
153	156	0.12
157	160	0.48
161	164	0.79
165	168	1.06
169	172	1.29
173	176	1.49
177	180	1.66
181	184	1.81
185	188	1.95
189	192	2.08
193	200	2.33
201	208	2.59
209	216	2.86
217	224	3.17
225	232	3.51
233	240	3.89
241	248	4.31
249	256	4.79
257	264	5.31
265	272	5.88
273	280	6.50
281	288	7.19
289	296	7.93
297	304	8.75
305	312	9.63
313	320	10.58
321	328	11.60
329	336	12.71
337	344	13.90
345	352	15.18

Continued on Next Page...

Table A.3 – Continued

Lower Index	**Higher Index**	**Absthrs (dB)**
353	360	16.54
361	368	18.01
369	376	19.57
377	384	21.23
385	392	23.01
393	400	24.90
401	408	26.90
409	416	29.03
417	424	31.28
425	432	33.67
433	440	36.19
441	448	38.86
449	456	41.67
457	464	44.63
465	472	47.76
473	480	51.03
481	488	51.03
489	496	51.03
497	504	51.03
505	513	51.03

Appendix B
Publications

This appendix contains re-prints of the papers [150, 151] published externally during the course of this research and these are as follows:

[150] R. R. Guddeti and B. Mulgrew, Perceptually motivated blind source separation of convolutive mixtures, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2005)*, vol. 5, (Philadelphia, PA, USA), pp. 273-276, 18-23 March, 2005.

[151] R. R. Guddeti and B. Mulgrew, Perceptually motivated blind source separation of convolutive audio mixtures with subspace filtering method, in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC2005)*, (High Tech Campus, Eindhoven, The Netherlands), 12-15 September, 2005.

The results published in the papers are from individual experiments performed in the early stages of the study. The corresponding results presented in the thesis, in Chapters 3 and 5, are from more extensive studies based on the earlier work and also using the results based on the objective performance evaluation of perceptually motivated BSS system of convolutive audio mixtures in terms of signal-to-interference ratio (SIR) (the time-domain metric) and enhanced modified Bark spectral distortion (EMBSD) (the perceptual-domain metric).

Furthermore, to accommodate these published papers in this thesis, their pages have been rescaled to reduce them slightly in size.

PERCEPTUALLY MOTIVATED BLIND SOURCE SEPARATION OF CONVOLUTIVE MIXTURES

Rammohana Reddy Guddeti and Bernard Mulgrew

Institute for Digital Communications
School of Engineering & Electronics
The University of Edinburgh
Edinburgh EH9 3JL U.K

ABSTRACT

A perceptually motivated method is proposed for solving the permutation ambiguity of frequency-domain independent component analysis when the mixing environment is noisy and reverberant. In this method, perceptually irrelevant frequencies are removed from the speech spectrum using block based perceptual masking (simultaneous frequency masking) and then independent component analysis is applied. After source separation in frequency domain, a physical property of the mixing matrix, i.e., the coherency in adjacent frequencies, is utilized to solve the permutation ambiguity. From the simulation results it appears that the perceptual masking avoids the permutation problem.

1. INTRODUCTION

The framework of blind source separation (BSS) based on independent component analysis (ICA) can be used to separate multiple signals without any previous knowledge of the sound sources and the mixing environment [1]. However, when applying to the cocktail party effect the performance of the BSS system is greatly reduced by the effect of the room reflections and ambient noise. Humans deal with this cocktail party effect very effectively by using only two ears (sensors). These perceptual masking techniques have been already exploited in successful development of MPEG audio coding standard which is the backbone of MP3 players.

In general, convolutive BSS methods can be classified into time domain ICA (TDICA) and frequency domain ICA (FDICA). TDICA has the disadvantage of being rather computational expensive due to computing many convolutions. The biggest obstacle in the FDICA is the permutation and scaling problem. For the scaling problem, the method proposed by Murata et al [2, 3], in which the separated output is filtered by the inverse of the separation filter.

For the permutation problem, Asano et al [4] have proposed a method that utilizes both the coherency of the mixing matrices and the correlation between spectral envelopes at several adjacent frequencies (denoted as inter frequency coherency (IFC)). In this paper, a perceptually motivated FDICA approach for solving the permutation problem is proposed. This method utilizes the block based perceptual masking for the complete omission of a signal at the given frequency that is perceptually irrelevant.

This paper is organized as follows: In Section 2, an outline of the proposed perceptually motivated FDICA system is presented in order to solve the permutation problem. In Section 3, simulation results of experiments using both synthetic and real data to evaluate the proposed perceptually motivated FDICA system are reported.

2. PERCEPTUAL FDICA SYSTEM

The flow of the proposed perceptual FDICA system is summarized in the form of block diagram as shown in Fig.1.

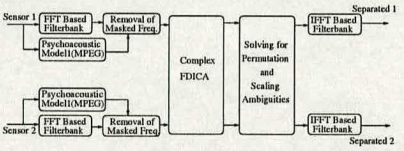


Fig. 1. Proposed Perceptually Motivated FDICA System

First, the short time Fourier transform (STFT) of the multichannel input signal, $\mathbf{x}(\omega, t)$, is obtained with an appropriate time shift and window function. Next, psychoacoustic model 1 (MPEG 1, layer I) [5] is used to determine the perceptual masking threshold for each segment of speech and thereby producing a binary mask for each frequency. A straightforward means to remove the masked frequency bins would be the multiplication of the complex spectrum of the input speech frame by the binary mask at each frequency bin. Thus, the thresholding in a stereo environment is described by logical AND operation.

Then, the FDICA algorithm (complex Infomax with feed-forward architecture [2,8,9]) is applied to the spectral components that are perceptually relevant for obtaining the separation filter. Next, the permutation and the scaling problem is solved by processing the output of the separation filter with the permutation and the scaling matrices. Finally, the filter matrices are transformed into the time domain and the input speech signal is processed with these filters.

2.1. Model of Signal

Let us consider the case when there are D sound sources in the mixing environment with M sensors. By taking STFT of the sensor inputs, we obtain the input vector

$$\mathbf{x}(\omega,t) = [X_1(\omega,t),...,X_M(\omega,t)]^T \quad (1)$$

Here, $\mathbf{X}_m(\omega,t)$ is STFT of the input signal in the tth time frame at the mth sensor. Further, the input signal is assumed to be modeled as

$$\mathbf{x}(\omega,t) = \mathbf{A}(\omega)\mathbf{s}(\omega,t) + \mathbf{n}(\omega,t) \quad (2)$$

$\mathbf{A}(\omega)$ is the mixing matrix and its (m,n) element, $A_{m,n}(\omega)$, being the transfer function from the nth source to the mth sensor as $A_{m,n}(\omega) = H_{m,n}(\omega)e^{-j\omega\tau_{m,n}}$. $\mathbf{s}(\omega,t)$ consists of the source spectra as $\mathbf{s} = [S_1(\omega,t),...,S_D(\omega,t)]^T$.

2.2. Psychoacoustic Model 1

The ISO MPEG-1 [5] psychoacoustic model 1 uses a 512 point FFT for high resolution spectral analysis, then selects the perceptually relevant spectral components in each frame of the input speech by means of thresholding. This model assumes masking effects are additive. In perceptual audio coding, thresholding sets the quanization level, here we set a threshold for further processing of the frequencies by ICA according to their psychoacoustic relevance and thereby reducing the computational complexity of solving the permutation problem. While this thresholding is a nonlinear activity which might at first sight appeared to destroy the linear convolutive properties of the BSS, but it can also be viewed as an irregular sampling rate strategy which is linear. It will however alter the pdf of the signals presented to ICA.

2.2.1. Power Spectrum

First, the sensor input, $\mathbf{x}(n)$, is segmented into frames of size of 512 samples using an appropriate time shift and Hann window function. A power spectral density (PSD), $P(k)$, for $\left(0 \le k \le \frac{N}{2}\right)$ is then obtained using a 512-point FFT as

$$P(k) = PN + 10\log_{10}\left|\sum_{n=0}^{N-1} w(n)x(n)e^{-j\frac{2kn\pi}{N}}\right|^2. \quad (3)$$

The power normalization term PN, fixed at 96 dB, is used to estimate the sound pressure level (SPL) conservatively from the input signal and $w(n)$ is Hann window function.

2.2.2. Global Masking Threshold

The absolute threshold of hearing is characterised by the amount of energy needed in a pure tone such that it can be detected by a listener in a noiseless environment. The quiet threshold is well approximated by

$$T_q(f) = \begin{array}{l} 3.64\left(\frac{f}{1000}\right)^{-0.8} \\ -6.5e^{0.6\left(\frac{f}{1000}-3.3\right)^2} \\ +10^{-3}\left(\frac{f}{1000}\right)^4 \end{array} \text{ (dB SPL)} \quad (4)$$

Simultaneous masking refers to a frequency domain phenomenon which has been observed within critical bands. Masking also occurs in the time domain. Sharp signal transients create pre- and post- masking regions in time during which a listener will not perceive signals beneath the elevated audibility thresholds produced by a masker. We didn't take into account temporal masking. This is due to the fact that our model is principally oriented to the speech signal that is stationary for a period shorter than 50 m sec.

Since masking refers to a psychoacoustic phenomenon, the masking threshold will be calculated in Barks. The Bark scale, in fact, refers to the critical bands of hearing. The conversion from frequency to Bark is given by

$$Bark(f) = \begin{array}{l} 13\arctan(.00076f) \\ +3.5\arctan\left[\left(\frac{f}{7500}\right)^2\right] \end{array} \quad (5)$$

From the PSD of equation 3 we detect all the local maxima, then we replace any two maxima in a 0.5 Bark sliding window by the stronger of the two. Once the tone and noise maskers are calculated, a decimation process takes place before calculating the global masking threshold according to the following scheme:

$$i = \begin{cases} k & 1 \le k \le 48 \\ k + (k.\mathrm{mod}2) & 49 \le k \le 96 \\ k + 3 - ((k-1)\,\mathrm{mod}4) & 97 \le k \le 232 \end{cases} \quad (6)$$

where k is the FFT index and i the decimation index. This process reduces the number of bins for the calculation of the global masking threshold, without loss of maskers. Having obtained a decimated set of tonal and noise maskers, individual tone and noise masking thresholds are computed next. Each individual threshold represents a masking contribution at frequency bin i due to the tone or noise masker located at bin j. Tonal masker thresholds, $T_{TM}(i,j)$ are expressed in (dB SPL) as

$$T_{TM}(i,j) = P_{TM}(j) - 0.275z(j) + SF(i,j) - 6.025 \quad (7)$$

where $P_{TM}(j)$ denotes the SPL of the tonal masker in frequency bin j, $z(j)$ denotes the Bark frequency of bin j, and the spread of masking from masker bin j to maskee bin i, $SF(i,j)$, is modeled by the expression in (dB SPL)

$$SF(i,j) = \tag{8}$$

$$\begin{cases} 17\Delta_z - 0.4P_{TM}(j) + 11, & -3 \le \Delta_z < -1 \\ (0.4P_{TM}(j) + 6)\,\Delta_z, & -1 \le \Delta_z < 0 \\ -17\Delta_z, & 0 \le \Delta_z < 1 \\ (0.15P_{TM}(j) - 17)\,\Delta_z - 0.15P_{TM}(j), & 1 \le \Delta_z < 8 \end{cases}$$

Individual noise masker thresholds (dB SPL) are given by

$$T_{NM}(i,j) = P_{NM}(j) - 0.175z(j) + SF(i,j) - 2.025 \tag{9}$$

where $P_{NM}(j)$ denotes the SPL of the noise masker in frequency bin j, $z(j)$ denotes the Bark frequency of bin j, and $SF(i,j)$ is obtained by replacing $P_{TM}(j)$ with $P_{NM}(j)$ everywhere in equation 8.

The global masking threshold, $T_g(i)$, is therefore obtained in dB by computing the sum

$$T_g(i) = 10\log_{10}\left(\begin{array}{l} 10^{0.1T_q(i)} + \sum_{l=1}^{L} 10^{0.1T_{TM}(i,l)} \\ + \sum_{m=1}^{M} 10^{0.1T_{NM}(i,m)} \end{array}\right) \tag{10}$$

where $T_q(i)$ is the absolute hearing threshold for frequency bin i, $T_{TM}(i,l)$ and $T_{NM}(i,m)$ are the individual masking thresholds and L and M are the number of tonal and noise maskers , respectively.

2.3. FDICA Algorithm

Whenever the perceptually masked input speech $\mathbf{x}(\omega,t)$ in one of the channels contains no values, the PCA filter matrix $\mathbf{W}(\omega)$ is singular, resulting in rank deficiency. Without loss of generality we have assumed identity matrix of order M as the rank of $\mathbf{W}(\omega)$ to avoid this problem. Then, the Infomax algorithm is applied to the output of the PCA filter, $\mathbf{y}(\omega,t)$ to obtain the ICA filter $\mathbf{U}(\omega)$. The separation filter $\mathbf{B}(\omega)$ is expressed as the product of $\mathbf{W}(\omega)$ and $\mathbf{U}(\omega)$. In the ICA stage, the input signal $\mathbf{y}(\omega,t)$ is processed with the filter matrix $\mathbf{U}(\omega)$ as $\mathbf{z}(\omega,t) = \mathbf{U}(\omega,t)\mathbf{y}(\omega,t)$.
The ICA learning rule is given by

$$\mathbf{U}(\omega,t+1) = \mathbf{U}(\omega,t) + \eta[\mathbf{I} - \varphi(\mathbf{z}(\omega,t))\mathbf{z}^H(\omega,t)]\mathbf{U}(\omega,t) \tag{11}$$

where the score function for $\varphi(\mathbf{z})$ is defined as

$$\varphi(\mathbf{z}) = [\varphi(z_1),\cdots,\varphi(z_d),\cdots,\varphi(z_D)]^T \tag{12}$$

$$\varphi(z_d) = 2\tanh(\Re(z_d)) + 2j\tanh(\Im(z_d)) \tag{13}$$

The symbol z_d is the dth element of the vector $\mathbf{z}(\omega,t)$. The matrix $\mathbf{I}$ is an identity matrix. The symbol $.^H$ denotes the Hermitian transpose. The constant η (.0001) is termed the learning rate. Here also we have avoided ICA filtering when the input of ICA filter in one of masked channels is zero in order to overcome the rank deficiency of ICA filter matrix.

The scaling problem can be solved by filtering individual output of the separation filter $\mathbf{B}(\omega)$ by its inverse [3]. The permutation problem can be solved by minimizing the sum of the angles $\{\theta_1,\cdots,\theta_D\}$ between the location vectors in the adjacent frequencies. The cosine of the angle θ_n between the two vectors, $\bar{\mathbf{a}}_n(\omega)$ and $\bar{\mathbf{a}}_n(\omega_0)$, of estimated mixing matrix is defined as [4]

$$\cos\theta_n = \frac{\bar{\mathbf{a}}_n^H(\omega)\bar{\mathbf{a}}_n(\omega_0)}{\|\bar{\mathbf{a}}_n(\omega)\|.\|\bar{\mathbf{a}}_n^H(\omega_0\|} \tag{14}$$

The cost function $F(\mathbf{P})$ is defined as

$$F(\mathbf{P}) = \frac{1}{D}\sum_{n=1}^{D}\cos\theta_n \tag{15}$$

In order to get reliable value of the cost function $F(\mathbf{P},k)$ at $\omega_0 = \omega - k.\Delta\omega$, for $k = 1,\cdots,K$, the confidence measure defined as [4]

$$C(k) = \max_{\mathbf{P}\in\Omega}[F(\mathbf{P},k)] - \max_{\mathbf{P}\in\Omega'}[F(\mathbf{P},k)] \tag{16}$$

Here, Ω denotes the set of all possible $\mathbf{P}$ while Ω' denotes Ω without $\hat{\mathbf{P}} = \arg\max_{\mathbf{P}\in\Omega}[F(\mathbf{P},k)]$. The permutation is then solved at $\omega_0 = \omega - \hat{k}.\Delta\omega \quad (\hat{k} = \max_{\mathbf{P}}[C(k)])$ as [4]

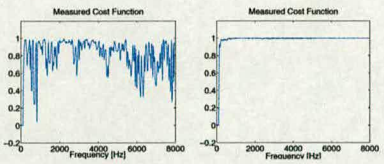
$$\hat{\mathbf{P}} = \arg\max_{\mathbf{P}}[F(\mathbf{P},\hat{k})] \tag{17}$$

The main contribution of this perceptual filtering is not only the reduction of frequencies that are processed by ICA, but also the reduction of frequencies where the similarity has to be checked for solving the permutation problem.

3. SIMULATION RESULTS

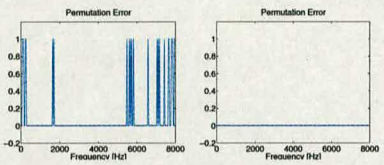
3.1. Experiment 1

In the first experiment, we created a synthetic convolutive mixture of two speech sources (7 s at 16 kHz) and we used Westner's [6] room acoustic data with reverberation time of 0.5 sec to simulate reverberant condition. From the Fig.2(a), it can be seen that there are many vertical lines in the measured value of the cost function when unmasked FDICA is considered. These vertical lines show that it is necessary to exchange the output at those frequencies where the permutation problem exists. From the Fig.2(b), it is clearly evident that the measured value of the cost function is almost unity for all the frequencies except for very low frequencies when the speech is perceptually masked. Permutation error is defined as the case when the result of IFC differs from that of



(a) Unmasked (b) Masked

Fig. 2. Measured Value of Cost Function for $k = 5$



(a) Unmasked (b) Masked

Fig. 3. Measured Value of Permutation Error for $k = 5$

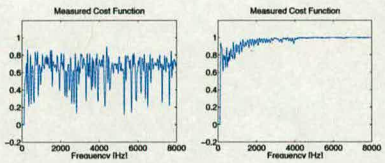
source output crosscorrelation (SOC) [4]. It is evident from this Fig.3(a) that there are many verticle lines in the measured permutation error when the speech is unmasked. It is clearly evident from the Fig.3(b) that the permutation error is zero for all the frequencies when the speech is masked.

3.2. Experiment 2

The second experiment was chosen to test the algorithm's ability in real room recording condition. To do this, we used real room recorded speech signals (6 s at 16 kHz). The permutation error cannot be computed in this real room recording case as the original sources are unknown. Real room recording results shown in Fig.4 are similar to that of previous experiment from the cost function point of view.

4. CONCLUSIONS

Incorporating the proposed perceptual solution for the permuation problem in the FDICA system produced good separation results in terms of the measured values of the cost function and the permutation error.



(a) Unmasked (b) Masked

Fig. 4. Measured Value of Cost Function for $k = 5$

5. REFERENCES

[1] A. Hyvarinen, J. Karhunen and E. Oja, "Independent Component Analysis,"*Wiley Inter-Science*, 2001.

[2] N. Murrata and S. Ikeda, "A Method of ICA in Time-Frequency Domain," *Proc. ICA'99*, pp. 365-370, Jan. 1999.

[3] N. Murrata, S. Ikeda and A. Ziehe, "An Approach to BSS Based on Temporal Structure of Speech Signals," *Proc. Neurocomputing*, vol. 41, pp. 1-24, Oct. 2001.

[4] F. Asano, S. Ikeda, M. Ogawa, H. Asoh and N. Kitawaki, "Combined Approach of Array Processing and ICA for Blind Separation of Acoustic Signals," *IEEE Trans. Speech and Audio Processing*, 11 (3), pp. 204-215, May 2003.

[5] T. Painter and A. Spanias, "A Review of Algorithms for perceptual Coding of Digital audio Signals," *Proc. DSP1997*, vol. 1, pp. 179-208, July 1997.

[6] A. G. Westner, "Object Based Audio Capture: Separating Acoustic Sounds," *M.S. Thesis, MIT Media Laboratory*, 1998.

[7] P. Smaragdis, "Blind Separation of Convolved Mixtures in the Frequency Domain," *Neurocomputing*, vol. 22, pp. 21-34, 1998.

[8] A. Bell and T. Sejnowski, "An Information-Maximization Approach to Blind Separation and Blind Deconvolution," *Neural Comput.*, vol. 7, pp. 1129-1159, 1995.

[9] S. Amari, A. Cichocki and A. A. Yang, "A New Learning Algorithm for Blind Signal Separation," *Proc. NIPS'95*, pp. 752-763, 1996.

PERCEPTUALLY MOTIVATED BLIND SOURCE SEPARATION OF CONVOLUTIVE AUDIO MIXTURES WITH SUBSPACE FILTERING METHOD

Rammohana Reddy Guddeti and Bernard Mulgrew

Institute for Digital Communications
School of Engineering & Electronics
The University of Edinburgh
Edinburgh EH9 3JL U.K
E-mail: ram.guddeti@ed.ac.uk

ABSTRACT

In this paper, a perceptually motivated subspace filtering method is proposed for solving the permutation ambiguity of frequency-domain independent component analysis when the mixing environment is noisy and highly reverberant. In this method, perceptually irrelevant frequencies are first removed from the speech spectrum using block based perceptual masking (simultaneous frequency masking) before applying the subspace method followed by frequency-domain independent component analysis. After source separation in frequency domain, a physical property of the mixing matrix, i.e., the coherency in adjacent frequencies while checking the similarity mesure among spectral envelopes of the separated output for reduced frequencies, is utilized as a post processing tool for solving the permutation ambiguity. From the simulation results it appears that the perceptual masking avoids the permutation problem.

1. INTRODUCTION

Blind source separation (BSS) aims to recover independent sources from their multiple observed mixtures using independent component analysis (ICA). However, when applying BSS to audio mixture problem such as a number of people talking in a room, the performance of the system is greatly reduced by the effect of the room reflections and ambient noise. Humans deal with this real cocktail party effect very efficiently by using only two ears (sensors). These perceptual masking techniques have been already exploited in successful development of MPEG audio coding standard (MP3 players).

Asano et al [1] have proposed the subspace method for reducing the effect of room reflections and ambient noise. Since the subspace method works in the frequency-domain, we must employ frequency-domain ICA (FDICA). The drawback of FDICA is permutation and scaling problem. For the scaling problem, the method proposed by Murata et al [2], in which the separated output is filtered by the inverse of the separation filter, shows good performance.

For the permutation problem, Asano et al [3] proposed a method that utilizes both the coherency of the mixing matrices and the correlation between spectral envelopes at several adjacent frequencies (denoted as inter frequency coherency (IFC)).

The authors [4] previously proposed a perceptually motivated FDICA method for solving the permutation problem. This method uses the simultaneous frequency masking (MPEG psychoacoustic model 1 [5]) for the complete omission of a signal componet at the given frequency.

In this paper, a perceptually motivated FDICA system with subspace approach for solving the permutation problem is proposed. This method utilizes both the simulatneous frequency masking for the complete omission of a signal at the given frequency and thereby using the subspace method for further reduction of room reflections.

This paper is organized as follows. In Section 2, an outline of the proposed perceptually motivated FDICA system with subspace method is presented for solving the permutation ambiguity. In Section 3, simulation results of experiments using both synthetic and real room recording speech data to evaluate the proposed perceptually motivated FDICA system are reported.

2. PERCEPTUAL FDICA SYSTEM

The flow of the proposed perceptual FDICA system is summarized in Fig.1.

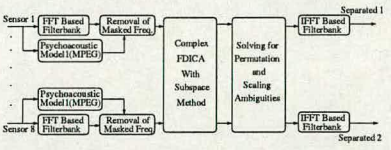


Figure 1: Perceptually Motivated FDICA System

First, the short time Fourier transform (STFT) of the multichannel input signal, $\mathbf{x}(\omega, t)$, is obtained with an appropriate time shift and Hann window function.
Next, psychoacoustic model 1 [5] is used to determine the masking threshold for each segment of speech and thereby obtaining a binary mask for each frequency.
A straightforward means to remove the masked frequency bins would be the multiplication of the complex spectrum of the input speech frame by the binary mask at each frequency bin. Thus, the thresholding in a stereo environment is described by logical AND operation.
The subspace method is then applied to the perceptually relevant spectral components of the input signal. In this stage, room reflections and ambient noise are reduced in advance of the application of FDICA. Next, the FDICA algorithm (complex Infomax [6–9]) is applied to the output of the subspace stage to obtain the separation filter.
After obtaining this filter, permutation and scaling problem is solved by processing the output of separation filter with the permutation and scaling matrices.
Finally, the filter matrices obtained in the above stages are transformed into the time domain and the input speech signal is processed with this time-domain filter network.

2.1. Model of Signal

Let us consider the case when there are N sound sources in the mixing environment with M sensors. By taking STFT of the sensor inputs, we obtain the input vector

$$\mathbf{x}(\omega, t) = [X_1(\omega, t), ..., X_M(\omega, t)]^T \quad (1)$$

Here, $\mathbf{X}_m(\omega, t)$ is STFT of the input signal in the tth time frame at the mth sensor. Further, the input signal is assumed to be modeled as

$$\mathbf{x}(\omega, t) = \mathbf{A}(\omega)\mathbf{s}(\omega, t) + \mathbf{n}(\omega, t) \quad (2)$$

$\mathbf{A}(\omega)$ is the mixing matrix and its (m, n) element, $A_{m,n}(\omega)$, being the transfer function from the nth source to the mth sensor as $A_{m,n}(\omega) = H_{m,n}(\omega)e^{-j\omega\tau_{m,n}}$. $\mathbf{s}(\omega, t)$ consists of the source spectra as $\mathbf{s} = [S_1(\omega, t), ..., S_N(\omega, t)]^T$.

2.2. Psychoacoustic Model 1

The ISO MPEG-1 [5] psychoacoustic model 1 uses a 512 point FFT for high resolution spectral analysis, then selects the perceptually relevant spectral components in each frame of the input speech by means of thresholding. This model assumes that the masking effects are additive.
In perceptual audio coding, thresholding sets the quanization level, here we set a threshold for further processing of the frequencies by ICA according to their psychoacoustic relevance and thereby reducing the computational complexity of solving the permutation problem.
While this thresholding is a nonlinear activity which might at first sight appeared to destroy the linear convolutive properties of the BSS, but it can also be viewed as an irregular sampling rate strategy which is linear. It will however alter the pdf of the signals presented to ICA.
Simultaneous masking refers to a frequency domain phenomenon which has been observed within critical bands. Sharp signal transients create premasking (backward temporal masking) and postmasking (forward temporal masking) in time during which a listener will not perceive signals beneath the elevated audible masking thresholds.
We didn't consider temporal masking based on the fact that our model is principally oriented to the speech signal that is stationary for a period shorter than 50 m sec.

2.3. Modified FDICA Algorithm

Whenever the perceptually masked input speech $\mathbf{x}(\omega, t)$ in one of the channels contains no values, the subspace filter matrix (special case of principal component analysis (PCA) with $M \gg N$, where M and N denote the number of nodes (channels) of the input and the output of PCA, respectively) $\mathbf{W}(\omega)$ is singular, resulting in rank deficiency.
Without loss of generality we have assumed identity matrix of order N for each pair of input nodes as the rank of subspace filter matrix $\mathbf{W}(\omega)$ to avoid this problem while retaining the whitening property of subspace filter.
Then, apply the complex Infomax algorithm for those frequency components of masked input speech, $\mathbf{y}(\omega, t)$, that contains nonzero values in both the channels in order to overcome the rank deficiency of ICA filter, $\mathbf{U}(\omega)$.
Thus the processing of ICA can be avoided whenever the masked input speech in one of the channels is zero.
In the ICA stage, the input signal $\mathbf{y}(\omega, t)$ is processed with the filter matrix $\mathbf{U}(\omega)$ as $\mathbf{z}(\omega, t) = \mathbf{U}(\omega, t)\mathbf{y}(\omega, t)$.
The ICA learning rule is given by

$$\mathbf{U}(\omega, t+1) = \mathbf{U}(\omega, t)+\eta[\mathbf{I}-\varphi(\mathbf{z}(\omega, t))\mathbf{z}^H(\omega, t)]\mathbf{U}(\omega, t) \quad (3)$$

Then, solve the scaling problem of FDICA by filtering individual output of the separation filter, $\mathbf{B}(\omega)$, (product of $\mathbf{W}(\omega)$ and $\mathbf{U}(\omega)$), by its pseudo inverse due to employment of the subspace method [3].
Finally, solve the permutation problem by utilizing both similarity measure among spectral envelopes of the separated output for frequencies that are perceptually relevant and the coherency of perceptually masked mixing matrices in several adjacent frequencies.
Without loss of generality assume zero cross correlation between spectral envelopes of the separated output when one of the channels does not contain any values and thereby avoiding rank deficiency of permutation matrix.

The cost function $F(\mathbf{P})$ is defi ned as

$$F(\mathbf{P}) = \frac{1}{N}\sum_{n=1}^{N}\cos\theta_n \quad (4)$$

Where, the cosine of the angle θ_n between the two location vectors in the adjacent frequencies, $\bar{\mathbf{a}}_n(\omega)$ and $\bar{\mathbf{a}}_n(\omega_0)$, of estimated mixing matrix is defi ned as

$$\cos\theta_n = \frac{\bar{\mathbf{a}}_n^H(\omega)\bar{\mathbf{a}}_n(\omega_0)}{\|\bar{\mathbf{a}}_n(\omega)\|.\|\bar{\mathbf{a}}_n^H(\omega_0\|} \quad (5)$$

In order to get reliable value of the cost function $F(\mathbf{P}, k)$ at $\omega_0 = \omega - k.\Delta\omega$, for $k = 1, \cdots, K$, the confi dence measure defi ned as

$$C(k) = \max_{\mathbf{P}\in\Omega}[F(\mathbf{P}, k)] - \max_{\mathbf{P}\in\Omega'}[F(\mathbf{P}, k)] \quad (6)$$

Here, Ω denotes the set of all possible $\mathbf{P}$ while Ω' denotes Ω without $\hat{\mathbf{P}} = \arg\max_{\mathbf{P}\in\Omega}[F(\mathbf{P}, k)]$. The permutation is then solved at $\omega_0 = \omega - \hat{k}.\Delta\omega \quad (\hat{k} = \max_{\mathbf{P}}[C(k)])$ as

$$\hat{\mathbf{P}} = \arg\max_{\mathbf{P}}[F(\mathbf{P}, \hat{k})] \quad (7)$$

The main contribution of this perceptual auditory masking and subspace method based preprocessor is not only the reduction of frequencies that are processed by ICA algorithm, but also the reduction of frequencies where the similarity to be checked for solving the permutation.

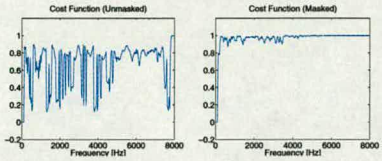
3. SIMULATION RESULTS

3.1. Experiment 1

This experiment was conducted with two speech sources (4 s at 16 kHz) and a circular microphone array (M = 8 and dia = 0.5 m) for simulating the room acoustic environment with reverberation time of 0.4 sec for both the weak and strong early reflection cases [3].

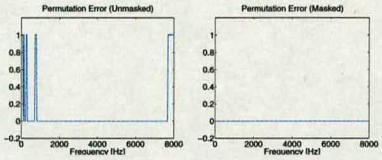
3.1.1. Weak Early Reflection Case

From the Fig.2(a), it can be seen that there are many vertical lines in the measured value of the cost function when unmasked FDICA is considered. These vertical lines show that it is necessary to exchange the output at those frequencies where the permutation problem exists.
From the Fig.2(b), it is clearly evident that the measured value of the cost function is almost unity for all the frequencies except for very low frequencies when the speech is perceptually masked.
Permutation error is defi ned as the case when the result of inter frequency coherency (IFC) differs from that of source output crosscorrelation (SOC) [3]. It is evident



(a) Unmasked (b) Masked

Figure 2: Measured Value of Cost Function for $k = 5$



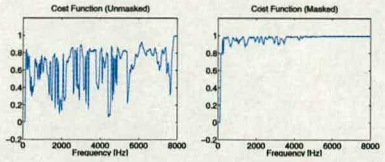
(a) Unmasked (b) Masked

Figure 3: Measured Value of Permutation Error for $k = 5$

from the Fig.3(a) that there are many vertical lines for frequencies below 2 kHz and a very few vertical lines for frequencies above 6 kHz in the measured permutation error when the perceptual masking is not considered.
It is clearly evident from the Fig.3(b) that the measured value of the permutation error is zero for all the frequencies when the speech is perceptually masked.

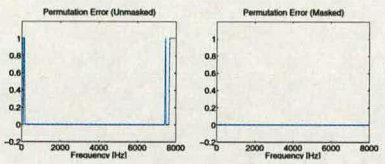
3.1.2. Strong Early Reflection Case

From the Fig.4(a), it can be seen that there are many vertical lines in the measured value of the cost function when unmasked FDICA is considered. These vertical lines show that it is necessary to exchange the output at those frequencies where the permutation problem exists.
From the Fig.4(b), it is clearly evident that the measured value of the cost function is almost unity for all the frequencies except for very low frequencies when the speech is perceptually masked.
It is evident from this Fig.5(a) that there are a very few vertical lines for frequencies below 1 kHz and a few vertical lines for frequencies above 6 kHz in the measured value of the permutation error when the perceptual auditory masking is not at all taken into account.
It is clearly evident from the Fig.5(b) that the measured



(a) Unmasked (b) Masked

Figure 4: Measured Value of Cost Function for $k = 5$



(a) Unmasked (b) Masked

Figure 5: Measured Value of Permutation Error for $k = 5$

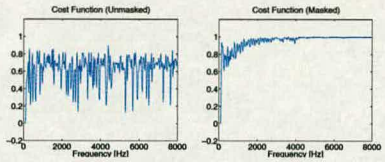
value of the permutation error is zero for all the frequencies when the speech is perceptually masked.

3.2. Experiment 2

The second experiment was chosen to test the algorithm's ability in real room recording condition. To do this, we used real room recorded speech signals (6 s at 16 kHz). The permutation error cannot be computed in this real room recording case as the original sources are unknown. Cost function shown in Fig.6 is similar to that of previous experiment for both unmasked and masked systems.

4. CONCLUSIONS

A perceptually motivated FDICA scheme with subspace method, proposed in the paper, reduces the frequency components that are perceptually irrelevant by exploiting the masking properties of speech.
This system also reduces the computation complexity of similarity measure among spectral envelopes of separated signals for solving the permutation ambiguity.
Further, the crosstalk suppression ratio has been improved by 5 dB when perceptual masking is taken into account.



(a) Unmasked (b) Masked

Figure 6: Measured Value of Cost Function for $k = 5$

The measured permutation error is 7.8% and 6.6% for unmasked FDICA system under both weak and strong early reflection conditions respectively.
On the other hand, the permutation error is zero for perceptually masked FDICA system for both the cases of weak and strong and reflection conditions.

5. REFERENCES

[1] F. Asano et al, "Speech Enhancement Based on the Subsapce Method," *IEEE Trans. Speech, Audio Processing*, Vol. 8, pp. 497-507, Sept. 2000.

[2] N. Murata, S. Ikeda and A. Ziehe, "An Approach to BSS Based on Temporal Structure of Speech Signals," *Neurocomputing*, Vol. 41, pp. 1-24, Oct. 2001.

[3] F. Asano et al, "Combined Approach of Array Processing and ICA for Blind Separation of Acoustic Signals," *Proc. IEEE Trans. Speech and Audio Processing*, Vol. 11, No. 3, pp. 204-215, May 2003.

[4] Rammohana Reddy Guddeti and Bernard Mulgrew, "Perceptually Motivated Blind Source Separation of Convolutive Mixtures," *Proc.of the Int. Conf. IEEE ICASSP2005*, Philadelphia, PA, USA, March 2005.

[5] T. Painter and A. Spanias, "Perceptual Coding of Digital Audio," *Proc. of IEEE*, Vol. 88, No. 4, pp. 451-513, April 2000.

[6] A. Bell and T. Sejnowski, "An Information-Maximization Approach to Blind Separation and Blind Deconvolution," *Proc. Neural Comput.*, Vol. 7, pp. 1129-1159, 1995.

[7] S. Amari, A. Cichocki and A. A. Yang, "A New Learning Algorithm for Blind Signal Separation," *Proc. NIPS'95*, pp. 752-763, 1996.

[8] P. Smaragdis, "Blind Separation of Convolved Mixtures in the Frequency Domain," *Proceedings of Neurocomputing*, Vol. 22, pp. 21-34, 1998.

[9] S. Ikeda and N. Murata, "A Method of ICA in Time-Frequency Domain," *Proc. ICA'99*, pp. 365-371, January 1999.

References

[1] James R Hopgood, *Nonstationary Signal Processing with Application to Reverberation Cancellation in Acoustic Environments*. PhD Thesis, University of Cambridge, UK, November 2000.

[2] B. C. J. Moore, *An introduction to the psychology of hearing*. London: Academic Press, 4th Edition, 1997.

[3] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organisation of Sound*. MIT Press, 2nd Edition, 1999.

[4] A. J. W. van der Kouwe, D. Wang, and G. J. Brown, "A comparision of auditory and blind separation techniques for speech segregation," *IEEE Trans. on Speech and Audio Processing*, vol. 9, pp. 189–195, March 2001.

[5] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Comput. Speech Lang.*, vol. 8, no. 4, pp. 297–336, 1994.

[6] P. Smaragdis, *Redundancy Reduction for Computational Audition, A Unifying Approach*. PhD Thesis, MIT, June 2001.

[7] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. on Antennas and Propagation*, vol. AP-34, pp. 276–280, March 1986.

[8] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Processing Magazine*, vol. 13, pp. 67–94, July 1996.

[9] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," *IEEE Trans. on Speech and Audio Processing*, vol. 8, pp. 497–507, September 2000.

[10] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "Combined approach of array processing and independent component analysis for blind separation of acoustic signals," *IEEE Trans. on Speech and Audio Processing*, vol. 11, pp. 204–215, May 2003.

[11] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley, 2001.

[12] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. John Wiley, 2002.

[13] S. Choi, A. Cichocki, H. M. Park, and S. Y. Lee, "Blind source separation and independent component analysis: A review," *Neural Information Processing - Letters and Reviews*, vol. 6, pp. 1–57, January 2005.

[14] P. Comon, "Independent component analysis, a new concept," *Signal Processing*, vol. 36, pp. 287–314, April 1994.

[15] J. F. Cardoso, "Blind Signal Separation: Statistical Principles," *Proceedings of the IEEE*, vol. 86, pp. 2009–2025, October 1998.

[16] A. Mansour, A. K. Barros, and N. Ohnishi, "Blind Separation of Sources: Methods, Assumptions and Applications," *IEICE Trans. Fundamentals*, vol. E83-A, pp. 1498–1512, August 2000.

[17] K. Torkkola, "Blind separation for audio signals - are we there yet," in *Proc. Workshop on Independent Component Analysis and Blind Signal separation*, (Aussois, France), pp. 1–6, January 1999.

[18] M. Davies, "Audio source separation," *Math. Signal Process.*, vol. V, 2000.

[19] N. Mitianoudis and M. Davies, "Audio Source Separation: Solutions and Problems," *Int. J. Adapt. Control Signal Process.*, vol. 18, pp. 299–314, April 2004.

[20] S. Ikeda and N. Murata, "An approach to blind source separation of speech signals," in *Proc. ICANN'98*, pp. 761–766, September 1998.

[21] A. Papoulis, *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 3rd Edition, 1991.

[22] S. Choi and A. Cichocki, "Adaptive blind separation of speech signals: Cocktail party problem," in *Int. Conf. on Speech Processing (ICSP'97)*, (Seoul, Korea), pp. 617–622, August 1997.

[23] L. Liu and J. He, "On the Use of Orthogonal GMM in Speaker Recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'99)*, vol. 2, pp. 845–848, March 1999.

[24] B. Ans, J. Herault, and C. Jutten, "Adaptive Neural Architectures: Detection of Primitives," in *Proc. of COGNITIVA'85*, (Paris, France), pp. 593–597, 1985.

[25] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind source separation," in *Advances in Neural Information Processing Systems*, vol. 8, pp. 757–763, MIT Press, 1996.

[26] J. F. Cardoso, "Infomax and maximum likelihood for blind source separation," *IEEE Signal Processing Letters*, vol. 4, pp. 112–114, April 1997.

[27] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[28] A. Hyvarinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1483–1492, 1997.

[29] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Networks*, vol. 10, pp. 626–634, May 1999.

[30] A. Cichocki and L. Moszczynski, "A new learning algorithm for blind separation of sources," *Electronic Letters*, vol. 28, no. 21, pp. 1986–1987, 1992.

[31] J. F. Cardoso, "Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'90)*, vol. 5, pp. 2655–2658, April 1990.

[32] J. F. Cardoso and A. Souloumiac, "Blind beamforming for nongaussian signals," *IEE Proceedings - F*, vol. 140, pp. 362–370, December 1993.

[33] J. F. Cardoso and B. H. Laheld, "Equivariant adaptive source separation," *IEEE Trans. on Signal Processing*, vol. 44, no. 12, pp. 3017–3030, 1996.

[34] D. Yellin and E. Weinstein, "Criteria for multichannel signal separation," *IEEE Trans. on Signal Processing*, vol. 42, pp. 2158–2167, August 1994.

[35] D. Yellin and E. Weinstein, "Multichannel Signal Separation: Methods and Analysis," *IEEE Trans. on Signal Processing*, vol. 44, pp. 106–118, January 1996.

[36] A. Belouchrani, K. Abed-Meraim, J. F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Trans. on Signal Processing*, vol. 45, pp. 434–444, Feb. 1997.

[37] U. A. Lindgren and H. Broman, "Source separation using a criterion based on second-order statistics," *IEEE Trans. on Signal Processing*, vol. 46, pp. 1837–1850, July 1998.

[38] T. Gustafsson, U. Lindgren, and H. Sahlin, "Statistical analysis of a signal separation method based on second-order statistics," *IEEE Trans. on Signal Processing*, vol. 49, pp. 441–444, Feb. 2001.

[39] D. T. Pham and J. F. Cardoso, "Blind separation of instantaneous mixtures of nonstationary sources," *IEEE Trans. Signal Processing*, vol. 49, pp. 1837–1848, Sept. 2001.

[40] D. T. Pham, "Exploiting source non-stationary and coloration in blind source separation," in *IEEE Int. Conf. on Digital Signal Processing (DSP'02)*, vol. 1, pp. 458–465, 2002.

[41] M. Kawamoto, K. Matsuoka, and N. Ohnishi, "A method of blind separation for convolved nonstationary signals," *Neurocomputing*, vol. 22, pp. 157–171, 1998.

[42] L. C. Parra and C. D. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. on Speech and Audio Processing*, vol. 8, pp. 320–327, May 2000.

[43] D. T. Pham, C. Serviere, and H. Boumaraf, "Blind separation of speech mixtures based on nonstationarity," in *Proc. 7th Int. Symp. on Sig. Process. and its App. (ISSPA'03)*, vol. 2, pp. 73–76, July 2003.

[44] M. J. T. Alphey, D. I. Laurenson, and A. F. Murray, "The effect of signal non-stationarity on the performance of information-maximisation-based blind separation," in *Proc. Workshop on Neural Networks for Signal Processing*, (Cambridge, U.K), pp. 113–122, September 1998.

[45] M. J. T. Alphey, D. I. Laurenson, and A. F. Murray, "Improvements in the on-line performance of information-maximisation-based blind signal separation," in *Proc. of the 1st Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA'99)*, pp. 49–54, January 1999.

[46] M. J. T. Alphey, *Blind Source Separation: The Effects of Signal Non-Stationarity.* PhD Thesis, The University of Edinburgh, 2002.

[47] H. L. N. Thi and C. Jutten, "Blind source separation for convolutive mixtures," *Signal Processing*, vol. 45, pp. 209–229, 1995.

[48] K. Torkkola, "Blind separation of delayed sources based on information maximization," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'96)*, vol. 6, pp. 3509–3512, May 1996.

[49] K. Torkkola, "Blind separation of convolved sources based on information maximization," in *Proc. IEEE Workshop on Neural Networks for Signal Processing (NNSP'96)*, pp. 423–432, September 1996.

[50] T. W. Lee, A. J. Bell, and R. H. Lambert, "Blind separation of delayed and convolved sources," in *Advances in Neural Information Processing Systems*, vol. 9, pp. 758–764, MIT Press, 1997.

[51] T. W. Lee, A. J. Bell, and R. Orglmeister, "Blind source separation of real world signals," in *Proc. IEEE Int. Conf. on Neural Networks (ICNN-1997)*, vol. 4, pp. 2129–2134, June 1997.

[52] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundemental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. on Speech and Audio Processing*, vol. 11, pp. 109–116, March 2003.

[53] R. H. Lambert, *Multichannel Blind Deconvolution: FIR Matrix Algebra and Separation of Multipath Mixtures.* PhD Thesis, University of Southern California, May 1996.

[54] A. G. Westner, *Object-Based Audio Capture: Separating Acoustic Sounds.* Masters Thesis, Media Lab, MIT, 1998.

[55] A. Westner and V. M. Bove, "Blind separation of real world audio signals using overdetermined mixtures," in *Proc. 1st Int. Workshop on Independent Component Analysis and Signal Separation (ICA1999)*, (Aussois, France), pp. 251–256, January 1999.

[56] S. I. Amari, "Neural Learning in Structured Parameter Spaces-Natural Riemannian Gradient," in *Proc. Advances in Neural Information Processing Systems*, vol. 9, pp. 127–133, MIT Press, 1997.

[57] S. I. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.

[58] S. I. Amari, "Natural Gradient Learning for Over- and Under-complete Bases in ICA," *Neural Computation*, vol. 11, no. 8, pp. 1875–1883, 1999.

[59] S. Haykin, *Adaptive Filter Theory.* Prentice Hall, 4th Edition, 2002.

[60] P. Smaragdis, *Information Theoretic Approaches to Source Separation.* Masters Thesis, Media Lab, MIT, 1997.

[61] P. Samaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[62] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "Blind source separation in reflective sound fields," in *Proc. HSC'01*, (Kyoto, Japan), pp. 51–54, April 2001.

[63] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. on Speech and Audio Processing*, vol. 10, pp. 352–362, September 2002.

[64] S. Amari, T. Chen, and A. Cichocki, "Stability analysis of learning algorithms for blind source separation," *Neural Networks*, vol. 10, no. 8, pp. 1345–1351, 1997.

[65] N. Mitianoudis and M. Davies, "A fixed point solution for convolved audio source separation," in *Proc. IEEE Workshop on Applications of Signal Processing on Audio and Acoustics*, October 2001.

[66] N. Mitianoudis and M. Davies, "New fixed-point solutions for convolved mixtures," in *Proc. 3rd Int. Conf. on Independent Component Analysis and Source Separation*, Dec. 2001.

[67] R. Prasad, H. Saruwatari, and K. Shikano, "Blind Separation of Speech by Fixed-Point ICA with Source Adaptive Negentropy Approximation," *IEICE Trans. Fundamentals*, vol. E88-A, pp. 1683–1692, July 2005.

[68] T. Nishikawa, H. Saruwatari, and K. Shikano, "Blind Source Separation Based on Multistage ICA Using Frequency-Domain ICA and Time-Domain ICA," in *Proc. ICFS-2002*, vol. R-1, pp. 7–12, March 2002.

[69] T. Nishikawa, H. Saruwatari, and K. Shikano, "Comparison of Time Domain ICA, Frequency Domain ICA and Multistage ICA for Blind Source Separation," in *Proc. EUSIPCO-2002*, vol. II, pp. 15–18, September 2002.

[70] T. Nishikawa, H. Saruwatari, and K. Shikano, "BSS of acoustic signals based on Multistage ICA combining Frequency Domain ICA and Time Domain ICA," *IEICE Trans. Fundamentals*, vol. E86-A, pp. 846–858, April 2003.

[71] T. Nishikawa, H. Saruwatari, K. Shikano, S. Araki, and S. Makino, "Multistage ICA for Blind Source Separation of Real Acoustic Convolutive Mixture," in *Proc. ICA2003*, pp. 523–528, April 2003.

[72] T. Nishikawa, H. Abe, H. Saruwatari, K. Shikano, and A. Kaminuma, "Overdetermined Blind Separation for Real Convolutive Mixtures of Speech Based on Multistage ICA Using Subarray Processing," *IEICE Trans. Fundamentals*, vol. E87-A, pp. 1924–1932, August 2004.

[73] N. Mitianoudis and M. Davies, "Audio source separation of convolutive mixtures," *IEEE Trans. on Speech and Audio Processing*, vol. 11, pp. 489–497, September 2003.

[74] L. Tong, R. W. Liu, V. C. Soon, and Y. F. Huang, "Indeterminacy and identifiability of blind identification," *IEEE Trans. on Circuits and Systems*, vol. 38, pp. 499–509, May 1991.

[75] L. Molgedey and H. G. Schuster, "Separation of a mixture of independent signals using time delayed decorrelations," *Phys. Rev. Lett.*, vol. 72, pp. 3634–3636, 1994.

[76] E. Weinstein, M. Feder, and A. V. Oppenheim, "Multi-channel signal separation by decorrelation," *IEEE Trans. on Speech and Audio Processing*, vol. 1, pp. 405–413, October 1993.

[77] L. C. Parra and C. D. Spence, "Convolutive blind source separation based on multiple decorrelation," in *Proc. Int. Workshop Neural Networks Signal Processing*, (Cambridge, U.K), pp. 23–32, September 1998.

[78] H. C. Wu and J. C. Principe, "Simultaneous Diagonalization in the Frequency domain (SDIF) for Source Separation," in *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA'99)*, pp. 245–250, 1999.

[79] D. W. E. Schobben and P. W. Sommen, "A frequency domain blind signal separation method based on decorrelation," *IEEE Trans. on Signal Processing*, vol. 50, pp. 1855–1865, August 2002.

[80] K. Rahbar and J. P. Reilly, "Blind source separation of convolved sources by joint approximate diagonalization of cross-spectral density matrices," in *Proc. IEEE Int. Conf. on Accoustics, Speech and Signal Processing (ICASSP2001)*, vol. 5, pp. 7–11, May 2001.

[81] K. Rahbar and J. P. Reilly, "A new fast-converging method for blind source separation of speech signals in acoustic environments," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 21–24, October 2003.

[82] M. Z. Ikram and D. R. Morgan, "A multiresolution approach to blind separation of speech signals in a reverberant environment," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP2001)*, vol. 5, pp. 2757–2760, May 2001.

[83] W. Wang, J. A. Chambers, and S. Sanei, "Penalty function approach for constrained convolutive blind source separation," in *Proc. 5th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA2004)*, (Granada, Spain), pp. 661–668, September 2004.

[84] W. Wang, S. Sanei, and J. A. Chambers, "Penalty function-based joint diagonalization approach for convolutive blind separation of nonstationary sources," *IEEE Trans. on Signal Processing*, vol. 53, pp. 1654–1669, May 2005.

[85] S. Ikeda and N. Murata, "A Method of ICA in Time-Frequency Domain," in *Proc. ICA1999*, pp. 365–371, January 1999.

[86] J. F. Cardoso, "Multidimensional independent component analysis," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP1998)*, vol. 4, pp. 1941–1944, May 1998.

[87] N. Murata, S. Ikeda, and A. Ziehe, "An approach to BSS based on temporal structure of speech signals," *Neurocomputing*, vol. 41, pp. 1–24, October 2001.

[88] F. Asono and S. Ikeda, "Evaluation and real-time implementation of blind source separation system using time-delayed decorrelation," in *Proc. ICA2000*, pp. 411–415, June 2000.

[89] N. Mitianoudis, *Audio Source Separation using Independent Component Analysis*. PhD Thesis, Dept. of Electronic Engg., Queen Mary, University of London, April 2004.

[90] K. Kamata, X. Hu, and H. Kobatake, "A new approach to the permutation problem in frequency domain blind source separation," in *Proc. 5th Int. Conf. on Independent Component Analysis and Blind Signal separation (ICA2004)*, (Granada, Spain), pp. 849–856, September 2004.

[91] X. Hu and H. Kobatake, "A new method for solving the permutation problem of frequency domain blind source separation," *IEICE Trans. Fundamentals*, vol. E88-A, pp. 1543–1548, June 2005.

[92] M. Z. Ikram and D. R. Morgan, "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP2000)*, vol. 2, pp. 1041–1044, June 2000.

[93] C. Serviere and D. T. Pham, "A novel method for permutation correction in frequency-domain in blind separation of speech mixtures," in *Proc. 5th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA2004)*, (Granada, Spain), pp. 807–815, September 2004.

[94] H. Saruwatari, T. Kawamura, and K. Shikano, "Fast-convergence algorithm for ICA-based blind source separation using array signal processing," in *Proc. IEEE Workshop on Statistical Signal Processing*, pp. 464–467, August 2001.

[95] H. Saruwatari, T. Kawamura, and K. Shikano, "Fast-convergence algorithm for ICA-based blind source separation using array signal processing," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 91–94, October 2001.

[96] H. Saruwatari, T. Kawamura, K. Sawai, A. Kaminuma, and M. Sakata, "Blind source separation based on fast-convergence algorithm using ICA and beamforming for real convolutive mixture," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'02)*, vol. 1, pp. 921–924, 2002.

[97] H. Saruwatari, T. Kawamura, T. Nishikawa, and K. Shikano, "Fast-convergence algorithm for blind source separation based on array signal processing," *IEICE Trans. Fundamentals*, vol. E86-A, pp. 634–639, March 2003.

[98] M. Z. Ikram and D. R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'02)*, vol. 1, pp. 881–884, 2002.

[99] M. Z. Ikram and D. R. Morgan, "Permutation inconsistency in blind speech separation: investigation and solutions," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 1–13, January 2005.

[100] A. Ciaramella, M. Funaro, and R. Tagliaferri, "Some Techniques for the solution of permutation Indeterminacy in Frequency Domain ICA," *Journal of Machine Learning Research*, vol. 1, pp. 1–48, October 2000.

[101] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," in *Proc. 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, (Nara, Japan), pp. 505–510, April 2003.

[102] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech and Audio Processing*, vol. 12, pp. 530–538, September 2004.

[103] S. Makino, H. Sawada, R. Mukai, and S. Araki, "Blind Source Separation of Convolutive Mixtures of Speech in Frequency Domain," *IEICE Trans. Fundamentals*, vol. E88-A, pp. 1640–1655, July 2005.

[104] W. Wang, J. A. Chambers, and S. Sanei, "A novel hybrid approach to the permutation problem of frequency domain blind source separation," in *Proc. 5th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA2004)*, (Granada, Spain), pp. 532–539, September 2004.

[105] T. Rutkowski, A. Cichocki, and A. K. Barros, "Speech extraction from interferences in real environment using bank of filters and blind source separation," in *Proc. IEEE Workshop on Signal Processing Applications*, pp. 14–15, December 2000.

[106] T. Rutkowski, A. Cichocki, and A. K. Barros, "Speech enhancement from interfering sounds using CASA techniques and blind source separation," in *Proc. 3rd Int. Conf. on Independent Component Analysis and Blind Signal separation (ICA2001)*, (SanDiego, California), pp. 728–733, December 2001.

[107] A. K. Barros, H. Kawahara, A. Cichocki, S. Kajita, T. Rutkowski, and N. Ohnishi, "Enhancement of a speech signal embedded in noisy environment using two microphones," in *Proc. ICA2000*, (Helsinki, Finland), pp. 423–428, June 2000.

[108] A. K. Barros and A. Cichocki, "Extraction of specific signals with temporal structure," *Neural Computation*, vol. 13, pp. 1995–2003, September 2001.

[109] A. K. Barros, F. Itakura, T. Rutkowski, A. Mansour, and N. Ohnishi, "Estimation of speech embedded in a reverberant environment with multiple sources of noise," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'01)*, vol. 1, pp. 629–632, May 2001.

[110] A. K. Barros, T. Rutkowski, F. Itakura, and N. Ohnishi, "Estimation of speech embedded in a reverberant and noisy environment by independent component analysis and wavelets," *IEEE Trans. on Neural Networks*, vol. 13, pp. 888–893, July 2002.

[111] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Springer, 2nd Edition, 1999.

[112] W. A. Yost, *Fundamentals of Hearing*. London: Academic Press, 3rd Edition, 1994.

[113] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA: MIT Press, Revised Edition, 1997.

[114] T. W. Parson, *Voice and Speech Processing*. London: McGraw-Hill, 1987.

[115] J. C. Brown, "Calculation of a constant Q spectral transform," *Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[116] C. S. Burrus, R. A. Gopinath, and H. Guo, *Introduction to wavelets and wavelet transforms: a primer*. Prentice Hall, Upper Saddle River, N.J, 1998.

[117] A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing*. Prentice Hall, Englewood Cliffs, N.J, 1989.

[118] R. D. Patterson, M. Allerhand, and C. Giguara, "Time-domain modelling of peripheral auditory processing: a modular architecture and a software flatform," *Journal of the Acoustical Society of America*, vol. 98, pp. 1890–1894, 1995.

[119] M. Slaney, D. Naar, and R. E. Lyon, "Auditory model inversion for sound separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP1994)*, vol. 2, pp. 19–22, April 1994.

[120] D. P. W. Ellis, B. L. Vercoe, and T. F. Quatieri, "A perceptual representation of audio for co-channel source separation," in *Proc. IEEE Workshop on Applications of Signal Processing on Audio and Acoustics*, pp. 73–74, October 1991.

[121] D. P. W. Ellis, "Hierarchic models of hearing for sound separation and reconstruction," in *Proc. IEEE Workshop on Applications of Signal Processing on Audio and Acoustics*, pp. 157–160, October 1993.

[122] D. P. W. Ellis, "A computer implementation of psychoacoustic grouping rules," in *Proc. 12th IAPR Int. Conf. on Signal Processing*, vol. 3, pp. 108–112, October 1994.

[123] D. P. W. Ellis, "Underconstrained stochastic representations for top-down computational auditory scene analysis," in *Proc. IEEE Workshop on Applications of Signal Processing on Audio and Acoustics*, pp. 43–46, October 1995.

[124] D. Ellis, "The weft: a representation for periodic sounds," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97)*, vol. 2, pp. 1307–1310, April 1997.

[125] P. Noll, "Wideband speech and audio coding," *IEEE Communications Magazine*, vol. 31, pp. 34–44, November 1993.

[126] T. Painter and A. Spanias, "A review of algorithms for perceptual coding of digital audio signals," in *Proc. Int. Conf. on Digital Signal Processing (DSP1997)*, vol. 1, pp. 179–208, July 1997.

[127] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, pp. 451–513, April 2000.

[128] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *Journal of the Acoustical Society of America*, vol. 66, no. 6, pp. 1647–1652, 1979.

[129] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, pp. 314–323, Feb. 1988.

[130] [ITU-R BS.1387], *Method for objective measurements of perceived audio quality*. ITU, July 2001.

[131] P. Kabal, *An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality*. TSP Lab. Tech. Report, Dept. of Elect. and Computer Engg., McGill University, December 2003.

[132] D. Pan, "A Tutorial on MPEG/Audio Compression," *IEEE Multimedia*, vol. 2, pp. 60–74, Summer 1995.

[133] Jason Doolittle. http://www.cise.ufl.edu/ jfd/mp3papers/.

[134] M. Lahdekorpi, J. Nurminen, A. Heikkinen, and J. Saarinen, "Perceptual irrelevancy removal in narrowband speech coding," in *Proc. EUROSPEECH 2003*, (Geneva), pp. 1081–1084, September 2003.

[135] M. Lahdekorpi, *Perceptual Irrelvancy Removal in Narrowband Speech Coding*. MS Thesis, Elect. Engg. Dept., Tampere University of Technology, Feb. 2003.

[136] J. F. Cardoso and A. Souloumiac, "Jacobi angles for simultaneous diagonalization," *SIAM J. Math. Anal. Appl.*, vol. 17, no. 1, pp. 161–164, 1996.

[137] Gene H. Golub and Charles F. Van Loan, *Matrix computations*. Oxford: North Oxford Academic, 1983.

[138] Joseph J. K. O'Ruanaidh and W. J. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing*. New York; London, Springer, 1996.

[139] F. Asano, "Circular microphone array (subspace filtering method)." http://asano.media-interaction.jp/English/doc/index.htm.

[140] D. Schobben, K. Torkkola, and P. Smaragdis, "Evaluation of blind signal separation methods," in *Proc. 1st Int. Workshop on Independent Component Analysis and Signal Separation (ICA1999)*, (Aussois, France), pp. 261–266, January 1999.

[141] R. Gribonval, E. Vincent, and C. Fevotte, "Proposals for performance measurement in source separation," in *Proc. 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, (Nara, Japan), pp. 763–768, April 2003.

[142] Wonho Yang, M. Dixon, and R. Yantorno, "A Modified Bark Spectral Distortion Measure Which Uses Noise Masking Threshold," in *Proc. IEEE Workshop On Speech Coding For Telecommunications*, pp. 55–56, Sept. 1997.

[143] Wonho Yang, M. Benbouchta, and R. Yantorno, "Performance of the Modified Bark Spectral Distortion as an Objective Speech Quality Measure," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP1998)*, vol. 1, pp. 541–544, May 1998.

[144] Wonho Yang, *Enhanced Modified Bark Spectral Distortion (EMBSD): An Objective Speech Quality Measure Based on Audible Distortion and Cognition Model*. PhD Thesis, Temple University, May 1999.

[145] L. A. Thorpe and B. R. Shelton, "Subjective test methodology: MOS vs. DMOS in evaluation of speech coding algorithms," in *Proc. IEEE Workshop on Speech Coding for Telecommunications*, pp. 73–74, October 1993.

[146] [ITU-T Recommendation P.861, 1996], *Objective quality measurement of telephone-band (300-3400 Hz) speech codecs*. 1996.

[147] [ITU-T Recommendation P.800, 1996], *Methods for subjective determination of transmission quality*. 1996.

[148] F. Asano, Y. Motomura, and T. Matsui, "Effect of PCA Filter in Blind Source Separation," in *Proc. ICA2000*, (Helsinki, Finland), pp. 57–62, June 2000.

[149] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "A combined approach of array processing and independent component analysis for blind separation of acoustic signals," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP2001)*, vol. 5, pp. 2729–2732, May 2001.

[150] R. R. Guddeti and B. Mulgrew, "Perceptually motivated blind source separation of convolutive mixtures," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP2005)*, vol. 5, (Philadelphia, PA), pp. 273–276, March 2005.

[151] R. R. Guddeti and B. Mulgrew, "Perceptually motivated blind source separation of convolutive audio mixtures with subspace filtering method," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC2005)*, (High Tech Campus, Eindhoven, The Netherlands), September 2005.

[152] T. Verma, S. Bilbao, and T. H. Y. Meng, "The digital prolate spheroidal window," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'96)*, pp. 1351–1354, May 1996.

[153] U. Rass and G. H. Steeger, "Reducing time domain alaising in adaptive overlap-add algorithms," in *Proc. 1999 138th Meeting of the Acoustical Society of America*, (Columbus, Ohio), November 1996.