MODEL-BASED SPARSE COMPONENT ANALYSIS FOR MULTIPARTY DISTANT SPEECH RECOGNITION

THIS IS A TEMPORARY TITLE PAGE
It will be replaced for the final print by a version
provided by the service academique.

Thèse n. 5723 (2013)
présentée le 21 March 2013
à la Faculté Sciences et Techniques de l'Ingénieur
laboratoire de LIDIAP
programme doctoral en Génie Électrique
École Polytechnique Fédérale de Lausanne

pour l'obtention du grade de Docteur ès Sciences
par

Afsaneh Asaei

acceptée sur proposition du jury :

Prof. Jean-Philippe Thiran, président du jury
Prof. Hervé Bourlard, directeur de thèse
Prof. Christian Jutten, rapporteur
Prof. Dietrich Klakow, rapporteur
Prof. Pierre Vandergheynst, rapporteur

Lausanne, EPFL, 2013

Abstract

This thesis takes place in the context of multi-microphone distant speech recognition in multiparty meetings. It addresses the fundamental problem of overlapping speech recognition in reverberant rooms. Motivated from the excellent human hearing performance on such problem, possibly resulting of sparsity of the auditory representation, our work aims at exploiting sparse component analysis in speech recognition front-end to extract the components of the desired speaker from the competing interferences (other speakers) prior to recognition. More specifically, the speech recovery and recognition are achieved by sparse reconstruction of the (high-dimensional) spatio-spectral information embedded in the acoustic scene from (low-dimensional) compressive recordings provided by a few microphones. This approach exploits the natural parsimonious structure of the data pertained to the geometry of the problem as well as the information representation space.

Our contributions are articulated around four blocks. The structured sparse spatio-temporal representation of the concurrent sources is constituted along with the characterization of the compressive acoustic measurements. A framework to simultaneously identify the location of the sources and their spectral components is derived exploiting the model-based sparse recovery approach and, finally, the acoustic multipath and sparsity models are incorporated for effective multichannel signal acquisition relying on beamforming. This work is evaluated on real data recordings. The results provide compelling evidence of the effectiveness of structured sparsity models for multi-party speech recognition. It establishes a new perspective to the analysis of multichannel recordings as compressive acquisition and recovery of the information embedded in the acoustic scene.

Keywords Distant Multiparty Speech Recognition, Overlapping Speech, Model-based Sparse Component Analysis, Compressive Acoustic Measurements, Reverberant Enclosure, Structured Sparse Coding, Image Method, Sparse Microphone Array, Multipath Sparse Beamforming

Zusammenfassung

Diese Dissertation befasst sich mit der Erkennung entfernter Sprache mittels mehrerer Mikrofone in einer Besprechung mit mehreren Teilnehmern. Sie behandelt das fundamentale Problem der Erkennung überlagerter Sprache in nachhallenden Räumen. Motiviert durch die exzellente Erkennungsleistung von Menschen unter solchen Umständen, möglicherweise basierend auf "spärlicher" auditiver Repräsentation, ist diese Arbeit bestrebt, spärliche Komponentenanalyse (sparse component analysis) in der Vorverarbeitung von Spracherkennungssytemen auszunutzen um die Komponenten des erwünschten Sprechers vor der Spracherkennung von konkurrierenden Überlagerungen (anderer Sprecher) zu separieren. Genauer gesagt wird die Sprache durch spärliche Rekonstruktion (sparse recovery) der (hoch-dimensionalen) räumlich-zeitlichen Informationen, enthalten in der akustischen Szene, aus der (niedrig-dimensionalen) komprimierten, mit wenigen Mikrofonen aufgezeichneten Aufnahme erkannt. Dieser Ansatz nutzt die natürliche, dünnbesetzte Datenstruktur die zur Geometrie des Problems und der räumlichen Informationsdarstellung gehört.

Der Beitrag dieser Arbeit kann in vier Bereiche gegliedert werden. Die Formulierung der strukturierten, spärlichen, räumlich-zeitlichen Repräsentation der konkurrierenden Schallquellen geht einher mit der Charakterisierung der komprimierten akustischen Messdaten. Ein Verfahren um gleichzeitig Ort und spektrale Komponenten der Schallquellen zu bestimmen wird unter Ausnutzung der modellgestützten spärlichen Rekonstruktion hergeleitet und schlussendlich fliessen die akustischen Merhweg- und Sparsity-Modelle in effektive mehrkanalige Signalakquisitionen beruhend auf Beamforming ein. Diese Arbeit wird mit realen Datenaufzeichnungen evaluiert, resultierend in überzeugenden Belegen für die Effektivität der strukturierten Sparsity-Modelle für Spracherkennung mehrerer Sprecher. Dies eröffnet neue Perspektiven für die Analyse von mehrkanaligen Aufnahmen als komprimierte Akquisition und Rekonstruktion der in der akustischen Szene enthaltenen Informationen.

Schagwörter Erkennung entfernter Sprache mehrerer Sprecher, Überlagerte Sprache, Modellgestützte spärliche Komponentenanalyse, Komprimierte akustische Messungen, Nachhallender Raum, Strukturierte spärliche Kodierung, Hallsimulation, Spärliche Mikrofonarrays, Mehrwegspärliches Beamforming

—Swiss German version translated by David Imseng as per the English version

Résumé

Cette thèse s'inscrit dans le contexte de la reconnaissance automatique de la parole distante pour réunions de groupes à l'aide de microphones multiples. Elle traite du problème fondamental de la reconnaissance de parole superposée en environnement réverbérant. L'excellente performance de l'audition humaine dans cette tâche, qui découle peut-être de la parcimonie de la représentation auditive, a motivé notre approche visant à exploiter l'analyse en composantes parcimonieuses dans le traitement acoustique d'un système de reconnaissance de la parole afin d'extraire les composantes du locuteur désiré parmi les interférences (autres locuteurs) avant l'étape de reconnaissance. Plus précisément, la reconstruction et la reconnaissance sont effectuées à partir de la reconstruction parcimonieuse de l'information spatio-spectrale (de haute dimension) contenue dans la scène acoustique à partir d'enregistrements comprimés (à basse dimension) obtenus avec quelques microphones. Cette approche exploite la structure naturellement parcimonieuse des données par rapport à la géométrie du problème ainsi qu'au domaine de représentation de l'information.

Nos contributions s'articulent autour de quatre blocs. La représentation parcimonieuse spatio-temporelle structurée des sources concurrentes est établie avec la caractérisation des mesures acoustiques comprimées. Un système permettant simultanément la localisation des sources et la détermination de leur composantes spectrales est élaboré en utilisant l'approche de reconstruction parcimonieuse basée sur un modèle, et, finalement, la propagation acoustique par trajets multiples et les modèles de parcimonie sont appliqués à l'acquisition efficace de signaux multicanaux basée sur la technique de formation de faisceau. Les données utilisées pour l'évaluation de ce travail proviennent d'enregistrements réels. Les résultats obtenus démontrent de manière convaincante l'efficacité des modèles parcimonieux pour la reconnaissance de la parole multilocuteur. Ce travail porte un nouveau regard sur l'analyse d'enregistrements multicanaux en tant que problème d'acquisition comprimée et de reconstruction de l'information contenue dans la scène acoustique.

Mots-clés Reconnaissance de parole distante multilocuteur, parole superposée, analyse en composantes parcimonieuses basée sur des modèles, mesures acoustiques comprimées, enceinte réverbérante, codage parcimonieux structuré, méthode de l'image, réseau de microphones parcimonieux, formation parcimonieuse de faisceau avec propagation par trajets multiples

— Swiss French version translated by Raphael Ullmann as per the English version

Acknowledgements

If I have seen further it is by standing on the shoulders of giants.
— Issac Newton

I have been incredibly fortunate and honored to be student of Hervé Bourlard, the best supervisor that I could ever have. Hervé guided me to a fundamental research problem which involved all my potentials and scientific passions. I would like to express my gratitude for his consistent support, encouragement and advices which made my doctoral studies a rewarding and fruitful accomplishment. Hervé guided me to achieve a Marie Curie training fellowship within the framework of SCALE [1] network which was perfectly aligned with my scientific longings. The goal of this program was to train scientists to work across traditional boundaries in multidisciplinary themes governed by multiple institutes. I learned enormously from the training that I received from the network of inspirational speech scientists and the futuristic development and educational plans. My PhD studentship has been truly a privilege for me and I wish to sincerely thank you Hervé for all the opportunities.

I have been extremely fortunate and honored to be co-advised by Volkan Cevher, the best co-supervisor that I could ever have. Even though this assignment did not happen officially, Volkan guided, encouraged and supported me through out my research efforts and career opportunities. I wish to express my sincere appreciation for his fundamental insights and ideas motivated and explained with clear objectives. I would like to extend my deep acknowledgment to Bhiksha Raj for accepting me at MLSP and supporting my research and career opportunities. Bhiksha helped me through numerous inspirational discussions. I obtained invaluable professional and personal gifts which I would always benefit from along my academic career. I also would like to thank Rita Singh and Richard Stern for various discussions and the warm welcome and guidance while I was at CMU. I would like to acknowledge David Nahamoo from IBM for the very inspiring discussions and his kind advice and support. I wish to extend my sincere acknowledgment to Mike Davies for accepting me at IDCoM and for the very fruitful discussions and analysis of my research during my PhD secondment. I also would like to thank Mehrdad Yaghoobi at IDCoM for the helpful discussions. I would like to acknowledge Steve Renals for his advice and kind welcome at CSTR during my SCALE secondment. I enjoyed an amazing and influential research environment at Edinburgh. I wish to extend my gratitude to Christian Jutten for the very precise and constructive remarks on my thesis manuscript which helped me to improve the quality of my dissertation. I extend my acknowledgment to the jury committee member Pierre Vandergheynst and the president Jean-Philippe Thiran. I would like to acknowledge John Hansen and Alex Acero for several discussions following up my IEEE spoken language processing award.

1. Speech Communication with Adaptive LEarning, FP7 grant agreement number 213850.

Acknowledgements

Special thanks to Dietrich Klakow my second advisor in SCALE project for the meetings to guide my research and his kind support. I thank his kind welcome and all the interesting aspects about distant speech recognition that we practiced at Saarland university under the guidance of Friedrich Faubel and John McDonough. I thank all of them very much. I extend my acknowledgment to the SCALE team, Louis ten Bosch, Lou Boves and Bert Cranen from Radboud University, Thomas Hain, Phil Green and Roger K. Moore from university of Sheffield, Simon King from CSTR and university of Edinburgh and Ralf Schlüter and Hermann Ney from RWTH Aachen University, Isabel Trancoso from INESC, Kate Knill from university of Cambridge, Sadaoki Furui from TTIC and Cornelia Köck from Saarland University. It was a remarkable opportunity to be in SCALE project. I would like to acknowledge the senior scientists at Idiap, Petr Motlicek, Phil Garner, Mathew Magimai Doss, Fabio Valente and Francois Fleuret for the kind advice. I wish to extend my gratitude to the amazing administrative team of Idiap who enable the institute to be extremely well organized and pleasant to work at. In particular, I would like to thank Nadine Rousseau and Sylvie Millius for all their support. I am very grateful for their kind guidance in all the problems that I encountered in sorting out the official complications in Switzerland. I extend my gratitude to other support staff at Idiap, Yann Rodriguez, Frank Formaz, Norbert Crettol, Cédric Dufour, Ed Gregg, Louis-Marie Plumel, Vincent Spano, Christophe Ecoeur, Bastian Crettol and Sandra Micheloud. I would like to extend my gratitude to Francois Foglia and Jean-Albert Ferrez. In addition, I would like to thank Particia Emonet for her kind help and information about life in Martigny.

I have been fortunate to have the companionship of my dear colleagues. In particular, I would like to acknowledge Lakshmi Saheer and Ramya Rasipouran. I extend my acknowledgment to Raphael Ullmann, Laurent El-shafey, David Imseng, Rahil Mahdiyan, Marc Ferras Font, Sree Harsha Yella, Hamid Reza Abutalebi, Anindya Roy, Samira Sheikhi, Hui Liang, Serena Soldo, Maryam Habibi, Hassan Ghasemzadeh, Gelareh Mohammadi, Katayoun Farrahi, Samuel Kim, Marzieh Razavi, Gwenole Lecorve, Elham Taghizadeh, Majid Yazdani, Ivana Chingovska, Dairazalia Sanchez-Cortes, Dinesh Babu Jayagopi, Marco Fornoni, Milos Cernak, Deepu Vijayasenan, Nesli Erdogmus, André Anjos, Leonidas Lefakis, Benjamin Picart, Daniel Gatica-Perez, Oya Aran, Hayley Hung, Elie Khoury, Kenneth Funes, Sriram Ganapathy, G.S.V.S. Sivaram, Samuel Thomas, Ashtosh Sapru, Manuel Günther, Murali Mohan Chakka, Niklas Johansson, Hari Parthasarathi, Giulia Garau, Alfred Dielmann, Yang Sun, Mahaboob Ali Basha Shaik, Mauro Nicolao, Cassia Valentini-Botinhao, Youssef Oualil, Joan Isaac Biel Tres and Riwal Lefort, Barbara Caputo, Francesco Orabona, Tatiana Tommasi for the memorable coincidence.

Special thanks to my very dear friends Leibny Paola Garcia, Atieh Ghazi, Tara Abdehagh, Narges Razavian, Miad Faezipour, Takayuki and Akiko Arakawa, Soudeh Khoubrouy, Zohreh Ayatollahi, Arsham Farshad and Shirin Ghanbari for their invaluable friendship during my life in USA, Switzerland and UK. I have been supported constantly with love and prayers of my family and my warmest acknowledgment goes to them. And above all, I would like to express my heartfelt gratitude to my husband for his consistent love and desire for me to flourish. He is standing beside me along with all the milestones of my life for more than a decade and giving me strength and confidence to seek for higher and push further while showing me the path through his brilliant mind and wisdom. This dissertation is dedicated to the unlimited gift of love and courage...

Martigny, 21 March 2013 Afsaneh Asaei

Contents

List of Figures

List of Tables

Glossary

Technical acronyms in this thesis are exhaustively listed below in alphabetical order.

ASR, **DSR**	automatic speech recognition, distant speech recognition
CS	compressive sensing
SCA	sparse component analysis
CASA	computational auditory scene analysis
MVDR	minimum variance distortionless response
MMSE	minimum mean square estimator
DOA	direction of arrival
ICA	independent component analysis
SIMO, **MIMO**	single (multi)-input multi-output
MINT	multiple-input-output inverse filtering theorem
RIR	room impulse response
STFT, **DCT**	short-time Fourier transform, discrete cosine transform
BSS-MSR	blind source separation via model-based sparse recovery
DUET	degenerate unmixing estimation technique
LOST	line orientation separation technique
AR	auto-regressive
RMSE	root-mean-square error
IHT	iterative hard thresholding
OMP	orthogonal matching pursuit
SBL	sparse Bayesian learning
SNR	signal to noise ratio

SIR	source to interference ratio
PESQ	perceptual evaluation of speech quality
WRR	word recognition rate
HMM	hidden Markov model
OLA	overlap-add
MFCC	mel-frequency cepstral coefficient
EM	expectation-maximization
MAP	maximum a posterior
CMN	cepstral mean normalization
MLLR	maximum likelihood linear regression
MONC	multichannel overlapping numbers corpus
MC-WSJ	multichannel wall street journal

Notation

Technical notations in this thesis are listed below.

s_n, S_n	signal of the n^{th} source in time and frequency domain
x_m, X_m	signal of the m^{th} microphone in time and frequency domain
h_{mn}, H_{mn}	acoustic channel between the m^{th} microphone and n^{th} source in time and frequency domain respectively
Φ	microphone array manifold matrix; it characterizes the acoustic projections associated to the acquisition of source signals inside a reverberant enclosure
$f; F$	each frequency bin; total number of Fourier coefficients
$\tau; \mathcal{T}$	each frame of speech; total number of speech frames
N	number of sources
M	number of microphones
G	number of cells on a grid
c	speed of sound assumed to be constant
R	order of reflections in a reverberant room
D	number of reflective surfaces within the enclosure
$\circledast$	convolution operation
$.^T$	transpose operation
$.^H$	conjugate transpose operation
$.^\dagger$	pseudo-inverse operation

1 Introduction

There are missing principles in what are so far explored and envisioned for machine listening paradigm. The specific focus of this thesis is on speech recognition task. The current systems can beat human listeners in clean training condition but the performance is very poor in noisy unexpected conditions [Kollmeier et al., 2008b]. This thesis addresses an open problem of research which is recognition in the presence of interfering talkers. The current technology breaks down at this scenario hence, the multiparty condition is where we expect some of the missing premises ensuring robustness are ought to be revealed.

1.1 Big Picture

This research provides a novel perspective to the extraction of information from multi-sensor observations of speech signals, by exploiting the naturally parsimonious structure of the data. Parsimony is a fundamental characteristic of nature, a fact recognized by several philosophers and scientists alike. Nature is known to favor parsimony, in form [Hildebrandt and Tromba, 1996], and in action [Zipf, 1949]. Natural scenes, both visual and auditory, have parsimony of composition: of all the elements that can compose a scene, only a small number are present in any given instance. In fact, it is hypothesized that human perception analogously employs parsimony of representation, e.g. the visual system may in fact characterize visual scenes as parsimonious combinations of a large number of visual elements [Olshausen and Field, 1996]. This natural predilection for parsimony was noted very early by William of Ockham in his Lex Parsimoniae, which postulates parsimony of explanation: the most parsimonious explanation invoking the fewest variables is often the best one, and *Occam's razor* continues to be a driving principle in modern mathematical and statistical models for various natural and artificial phenomena [Blu et al., 2008, Olshausen and Field, 1997, Rissanen, 1978, Elad, 2010, Baraniuk et al., 2010, Allen and Berkley, 1979].

The parsimony, as described above, is a generic structure that is broadly applicable and is not dependent on the specific details of any situation or problem. As such, it provides us with a

situation-agnostic prior that may be used as a guiding principle in a variety of problems. In this research, we have attempted to exploit this structure to recover information from audio data. Specifically, we have addressed the problem of deriving information from measurements obtained through multiple sensors. Although the fundamental structure of the underlying scene is not affected by the presence of multiple sensors, the complementarity in the measurements provided by the sensors can be key to effective recovery of information [Elad, 2010, Baraniuk et al., 2010].

In the context of audio, parsimony can be identified in many factors: *Geometry:* Of all the locations where a sound source may be present, typically very few are actually occupied. *Spectrum:* In most information carrying sounds the spectral entropy tends to be low. *Acoustic:* Reflections in reverberant rooms can be characterized through a very small number of parameters, namely its dimensions and the reflectance of the walls. Other similar instances of structure can be identified.

We devise mathematical tools that can utilize mathematical characterizations of the various forms of structure described above, to learn about acoustic scenes. We do this within a unified framework that encompasses modeling and processing of data. The framework draws on concepts obtained from diverse fields including signal processing, compressive sensing, acoustics, and machine learning. This doctoral dissertation demonstrates the validity of this approach, by successfully learning the structure of reverberant acoustic environments and its embedded sources through microphone array recordings under well-defined experimental conditions. The broader goal of this work, is to develop a unified framework for sensing, processing, and modeling of information embedded in unrestricted acoustic scenes, and for comprehensive audio analysis.

1.2 Motivation

Human listening possesses a fundamental functionality, which is the ability to listen to and follow one speaker in the presence of others in a complex acoustic environment. This phenomenon is termed as the *cocktail party* problem in [Cherry, 1957]. Construction of the machine to perform similarly yet remains an open problem in theory and practice.

The discrepancy between human and machine performance has motivated many feature extraction approaches inspired by modeling the human auditory system. Perceptual modeling indicates that sparse representations exist in the auditory cortex and the more accurate the auditory model, the sparser the representation [Kleinschmidt, 2002b]. One example of auditory sparse representation is depicted in Figure 1.1 [Stern and Morgan, 2012]. We can see that the number of high energy components in the spectrogram reconstructed after peripheral auditory processing is very small. The evidences on a key role of sparse coding and structural grouping in auditory scene analysis and in particular multiparty segregation can be traced back to the work of [Brungart, 2001, Bregman, 1990, Wang and Brown, 2006]. The findings highlight the analogous auditory perception governing mechanisms in relation to the visual perception [Olshausen and Field, 1996, 1997, Palmer, 1999]. More recent work demonstrate excellent reconstruction of the sound from a tiny fraction of the neural activities. It reveals that several attributes of the attending talker remain

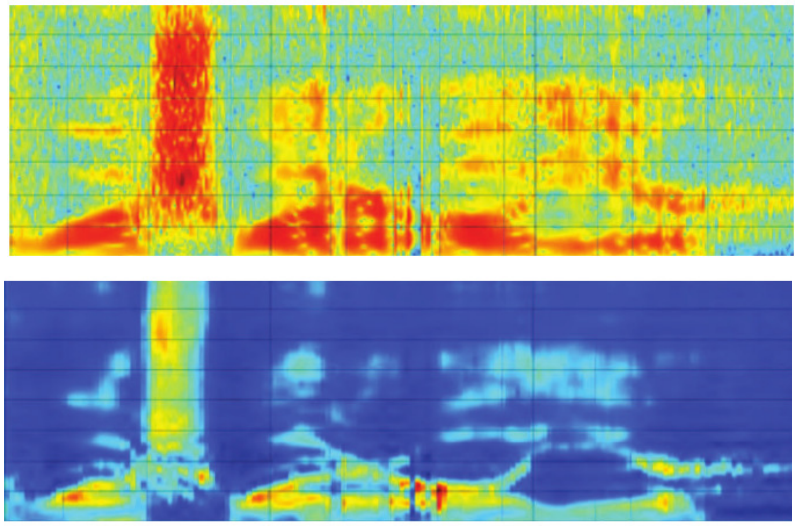

Figure 1.1 – (top) Original spectrogram: MATLAB spectrogram of clean speech. (bottom) Peripheral auditory spectrogram: reconstructed spectrogram of the utterance after peripheral auditory processing based on the model proposed by [Zhang et al., 2001]. This Figure is published in [Stern and Morgan, 2012].

intact in the presence of the interferer which indicates the disjointness of the perceptual cues as a result of sparse representation [Mesgarani and Chang, 2012, Pasley et al., 2012, David et al., 2007]. The machine listening paradigm however, does not incorporate the sparse structures in extraction of information. Hence, this PhD research is motivated to explore and incorporate the sparse coding principles to enable speech recognition systems to achieve robustness in multi-party distant scenarios.

The problem of overlapping speech is one of the major challenges of speech recognition systems in multi-speaker environments and distant-talking applications [Renals et al., 2007]. As identified in [Shriberg et al., 2001], around 10–15% of words or 50% of speech segments in a meeting or telephone conversation contain some degree of overlapping speech. These overlapped speech segments result in an absolute increase in speech recognition word error rate of 15–30%. Therefore, any system designed to recognize speech in multi-speaker environments is required to initially separate the speech from each individual prior to recognition.

Sparse component analysis is a source separation technique exploiting a priori assumption that the sources have a sparse representation in a known basis or frame. These techniques have turned out in the last few years to be a successful tool to estimate the mixing parameters and non-linear recovery of the source components [Gribonval and Lesage, 2006, Araki et al., 2011]. Previous work to evaluate the source separation approaches to perform speech recognition has been largely confined to the algorithmic approaches relying on statistical independence assumption or spatial filtering. There is little literature on evaluating the capability of sparse techniques for speech recognition which motivates the research on applicability of the developed sparse recovery methods to enable multiparty distant speech recognition [Asaei et al., 2010b].

1.3 Thesis Statement

This thesis addresses a fundamental research problem which is recognition in the presence of interfering talkers. The developed theory and practice in the field break down at this scenario hence, we expect that some of the missing principles to ensure multiparty robustness are ought to be identified from this study. We take a new perspective to the analysis of multi-sensor observations and propose a structured sparse coding framework to extract the spatio-spectral information embedded in the acoustic scene. Inspired form the studies on sparse coding of sensory information in biological systems, our goal is to investigate how machine listening paradigm can exploit the sparsity models to process the information. Figure 1.2 illustrates the major building blocks of the research presented in this thesis.

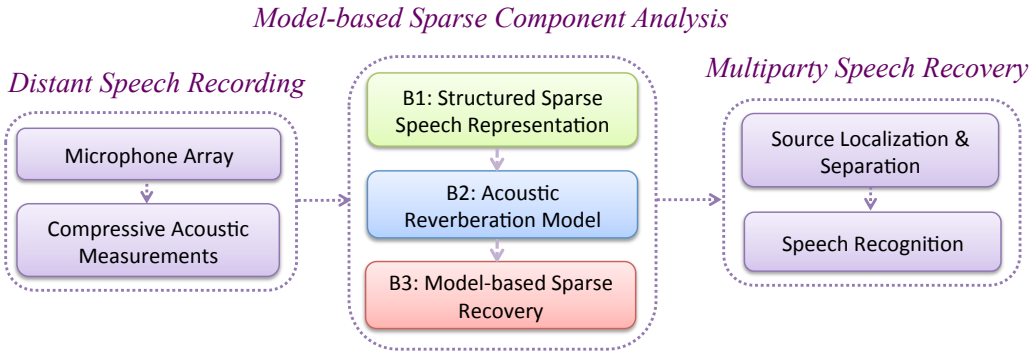

Figure 1.2 – The building blocks of the research presented in this dissertation

The multi-channel microphone recordings provide a compressive sensing of the acoustic field data. We derive a spatio-spectral representation of the concurrent sound sources (B1) and characterize the acoustic reverberation model (B2) to formulate a unified theory for identification of the source locations and the individual spectral components in the presence of acoustic multipath (B3). The proposed framework incorporates the model underlying spectrographic speech representation as well as the acoustic channel for extraction of the information bearing components. We consider various algorithmic approaches to model-based sparse recovery to customize the framework for speech specific models and applications. We further elucidate the links between conventional spatial filtering and sparse recovery and exploit the sparsity and multipath models to derive an acoustic-informed sparse beamforming method for estimation of the desired source information.

The proposed theories are evaluated on real data recordings. The results provide compelling evidence of the effectiveness of model-based sparse component analysis framework for multiparty speech recognition. The evidences support the argument that the salient information for speech recognition are sparse in the spectro-temporal scene and the structured sparsity models inspired from the psychoacoustic of sound perception and propagation enable more efficient signal recovery. The model-based sparse component analysis framework opens up avenues of research on a unified modeling and processing scheme for future generation of multiparty technologies.

1.3.1 Objectives

The key themes involved in this research are the followings

1. Applicability of sparse component analysis for multiparty speech recognition
2. Incorporating structured sparsity for model-based sparse component analysis
3. Characterizing the acoustic projections associated with microphone array recordings
4. Compare and contrast various algorithmic approaches to sparse recovery
5. Analysis of optimum microphone array acquisition and processing objective
6. Performance bounds of the proposed framework in theory and practice

The objective of this research is to investigate how sparse recovery enables speech recognition systems to achieve robustness in multi-party distant scenarios. We validate the hypothesis that information bearing components for speech recognition are sparse in the spectro-temporal domain and they remain disjoint and sparse in the presence of multiple interferences and multipath effect. Hence, sparse component analysis ought to be a potential technique to deal with overlapping speech.

This research takes a step further than sparsity assumptions and shows that the sparse coefficients are aligned with particular structures which play a key role in perception of the sound and acoustic. The goal is thus identification of these structures and incorporate them for signal recovery. A novel framework of model-based sparse component analysis is proposed which leverages the recent algorithmic advances in model-based compressive sensing, and formulates a source separation algorithm for efficient recovery of convolutive speech mixtures in spectro-temporal domain. Compared to the common sparse component analysis techniques, our approach fully exploits structured sparsity models to obtain substantial improvement over the existing state-of-the-art.

Specific attention is paid to characterization of the measurement matrix exploiting the Image model of multipath effect. We show that the geometry of the reflective surfaces can be estimated by localization of the early images of the speakers through sparse approximation of the spatial spectra of the virtual sources in a free-space model. The images are then clustered exploiting the low-rank structure of the spectro-temporal components belonging to each source. This enables us to estimate the room geometry. To further tackle the ambiguity of the absorption ratios, we propose a novel formulation of the reverberation model and estimate the absorption coefficients through a convex optimization exploiting joint sparsity model. The acoustic parameters are then incorporated for separating individual speech signals through either model-based sparse recovery or inverse filtering the acoustic channels.

The ultimate objective is to incorporate the new insights in the framework of optimum spatial filtering. We elucidate the links between sparse reconstruction and spatial filtering and derive the acoustic informed sparse beamforming methods. The proposed theories are evaluated by conducting experiments on real data recordings collected for meeting speech recognition. The deterministic and probabilistic performance bounds are derived for various algorithmic approaches.

1.3.2 Contributions

The research presented in this dissertation features the following contributions:

- ◇ Validating the sparsity prior and thus sparse component analysis as a potential approach to enable speech recognition in multiparty scenarios.
- ◇ Presenting a new perspective to the objective of multichannel processing as sparse recovery of the information embedded in the acoustic field.
- ◇ Characterizing the acoustic projections associated to the microphone array manifold using the Image model of multipath effect from the multiparty recordings.
- ◇ Novel formulation of the reverberation model factorized into virtual sources which enables estimating the absorption factors of the reflective surfaces.
- ◇ Room geometry estimation from recording of multiple unknown sources located at unknown positions exploiting sparse recovery and low-rank clustering techniques.
- ◇ Identifying the sparsity models pertained to sound perception and multipath propagation and analysis of the algorithmic approaches to incorporate those structures.
- ◇ Theoretical analysis of the performance bounds using the generic theory of compressive sensing which suggests a sparse random array topology and empirical validation of the effectiveness of random microphone array in terms of speaker localization and quality of the recovered speech.
- ◇ Unified framework of source and sensor localization and speech recovery relying on sparse coding of spatial signals and drawing probabilistic performance bounds with respect to the deterministic bounds obtained from the compressive sensing theory.
- ◇ Equivalence of the inverse filtering to a two-step procedure of speech separation followed by channel deconvolution to analyze the performance limits of the framework.
- ◇ Elucidating the links between optimum microphone array processing relying on beamforming and sparse recovery to derive a formulation of an acoustic informed sparse beamforming method.

The contributions are communicated through the following scholarly publications:

1. Asaei A., Golbabaee M., Bourlard H., Cevher V., *Structured Sparsity Models for Multiparty Speech Recovery from Reverberant Recordings*, IEEE Transactions on Speech and Audio Processing, (under review) 2012.
2. Asaei A., Bourlard H., Cevher V., *A Method, Apparatus and Computer Program for Determining the Location of a Plurality of Speech Sources*, 2012US-13/654055, US Patent, October 2012.
3. Asaei A., Raj B., Bourlard H., Cevher V., *A Unified Structured Sparse Coding Framework for Spatio-spectral Information Recovery*, IEEE Transactions on Speech and Audio Processing, (under revision) 2013.

4. Asaei A., Raj B., Bourlard H., Cevher V., *A Multipath Sparse Beamforming Method*, Signal Processing with Adaptive Sparse Structured Representations (SPARS), 2013, ***Nominated for Best Paper Award***.
5. Asaei A., Golbabaee M., Bourlard H., Cevher V., *Structured Sparse Acoustic Modeling for Speech Separation*, Signal Processing with Adaptive Sparse Structured Representations (SPARS), 2013.
6. Asaei A., Golbabaee M., Bourlard H., Cevher V., *Room Acoustic Modeling and Speech Dereverberation Exploiting Sparsity and Low-rank Structures*, Journal of Acoustic Society of America (in preparation), 2013.
7. Asaei A., Raj M., Bourlard H., Cevher V., *Structured Sparse Coding for Microphone Array Location Calibration*, The 5^{th} ISCA workshop on Statistical and Perceptual Audition, SAPA-SCALE Conference, 2012.
8. Asaei A., Davies M., Bourlard H., Cevher V., *Computational Methods for structured Sparse Component Analysis of Convolutive Speech Mixtures*, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2012.
9. Asaei A., Taghizadeh M., Bourlard H., Cevher V., *Multi-party Speech Recovery Exploiting Structured Sparsity Models*, The 13^{th} Annual Conference of the International Speech Communication Association (INTERSPEECH), 2011.
10. Asaei A., Bourlard H., Cevher V., *Model-based Compressive Sensing for Distant Multi-party Speech Recognition*, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2011, ***Winner of IEEE Spoken Language Processing Award***.
11. Asaei A., Garner P., Bourlard H., *Sparse Component Analysis for Speech Recognition in Multi-Speaker Environment*, The 12^{th} Annual Conference of the International Speech Communication Association (INTERSPEECH), 2010.
12. Asaei A., Picart B., Bourlard H., *Analysis of Phone Posterior Feature Space Exploiting Class-Specific Sparsity and MLP-Based Similarity*, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2010.

1.4 Thesis Outline

The introduction aims to provide a review of the problem and motivations underlying this study. We outlined the specific objectives along with the contributions that features the findings of this work.

In Chapter 2, the context of this study is reviewed regarding the earlier literature published on the subject. We survey the previous art in two broad categories of linear and nonlinear approaches to speech separation with specific attenuation to the independent component analysis, beamforming, computational auditory scene analysis and sparse techniques. In addition, initial experimental validation confirming that Sparse Component Analysis (SCA) is a potential approach to deal with overlapping problem in multi-party speech recognition are carried out and presented.

In Chapter 3, we extend our initial work on SCA (from Chapter 2) to more practical scenario, where the speech can also be corrupted by reverberation. We leverage the model-based compressive sensing (CS) theory to propose a novel SCA technique built upon natural realization of the CS premises in microphone array recordings. We evaluate our method for separation and recognition of speech in a multi-party scenario. Our studies provide compelling evidence of the effectiveness of sparse recovery formulations in speech recognition.

In Chapter 4, we explain some theory of characterizing sparsity of the acoustic signals and identifying the deterministic models underlying the coefficients. We start with a brief introduction into the theory behind this analysis from the CS perspective. There are two distinct aspects to the structured sparsity models associated to the sound propagation and sound perception which are studied in this chapter.

In Chapter 5, we address specifically the problem of characterizing the acoustic measurements from distant recordings of unknown signals. We describe a procedure to estimate the geometry of the room and estimate the absorption factors of the reflective surfaces to define the Image model of acoustic projections. This model enables us to characterize the microphone array manifold matrix for sensing the information embedded in the acoustic scene.

In Chapter 6, we compare and contrast different algorithmic approaches to model-based sparse recovery namely, convex optimization, hard thresholding and sparse Bayesian learning. We perform some evaluations in terms of source localization accuracy and quality of the recovered signal using sparse approximation of the spatial spectra obtained by different methods.

In Chapter 7, we describe generalized framework of source and sensor localization and speech recovery relying on the principle of structured sparse coding. We elaborate on optimality of inverse filtering to perform speech separation and dereverberation and present some evaluations on a multiparty corpus collected for meeting recognition.

In Chapter 8, we exploit the sparsity and multipath characterization to derive an optimum formulation of multichannel spatial filtering for multiparty recordings. An acoustic informed sparse beamformer is derived for the minimum variance distortionless and minimum mean square error signal estimation criteria. Further experiments are carried out on overlapping speech recognition.

In Chapter 9, a summary of the main findings of the dissertation are listed and the directions of the future research are recommended to address several issues remained yet unsolved within the broad goal of this research.

2 Multiparty Speech Recovery from Multichannel Recordings

This chapter provides an outlook to the state of the art on addressing the problem of speech recovery from the acoustic clutter of interfering voices. We overview the fundamental multichannel speech recovery approaches, and explain the basic idea of sparse component analysis. Inspired from the sparse coding of sensory information for human perception, we overview the general strategy where the computational auditory scene analysis is founded upon to highlight the principles that we can apply in a framework of sparse signal recovery. This survey study puts forward a critical question: *Does Distant Speech Recognition require sparse representation and could it benefit from sparse component analysis?* We recognize the advantages of sparse representation and provide some insights on how a sparse coding framework can lay the foundation of a DSR system robust to overlapping.

2.1 Inverse Problem Statement

The acoustic observation is expressed as a linear convolutive mixing process, stated concisely as:

$$x_m = \sum_{n=1}^{N} h_{mn} \circledast s_n, \quad m = 1, ..., M \tag{2.1}$$

where s_n refers to the n^{th} source signal convolved through the acoustic channel, h_{mn} and acquired at m^{th} microphone signal x_m. N and M denote the number of sources and microphones, respectively. This formulation is stated in time domain. To represent it in a sparse domain, we analyze the discrete Short-Time Fourier Transform (STFT) of speech signals [1]. Following from the convolution-multiplication property of the Fourier transform, the mixtures in frequency

1. also referred to as Gabor expansion; we will see in Section 2.4 that STFT yields approximate sparse representation of speech signals.

domain can be written as

$$X_m(f,\tau) = \sum_{n=1}^{N} H_{mn} S_n(f,\tau), \quad m = 1,...,M \tag{2.2}$$

The objective is to recover and recognize N (unknown) source signals from M recorded mixtures. There is no prior knowledge about N, M or the acoustic mixing channel $H_{\cdot}$. If N is either less than or equal to M, the scenario is referred to as *overdetermined* or *determined*, respectively.

The linear inverse problem dealt with throughout this thesis considers particularly the *underdetermined* scenario where the number of sources N is greater than the number of microphones M. In this scenario, the number of available recordings is less than the number of unknown variables. Hence, the linear system is ill-posed and some prior information is required to identify the solution. We start by providing an overview of the previous art on multichannel techniques addressing the problem of speech separation.

2.2 Linear Speech Recovery

There are two fundamental approaches relying on multichannel linear filtering to recover the individual source signals namely, *independent component analysis* and *beamforming*. In the following, we study the assumptions underlying each approach and the scope of their application.

2.2.1 Independent Component Analysis

The independent component analysis (ICA) approach to blind source separation relies on the assumption of statistically independent sources. The mixture signals are modeled in the standard form as a linear superposition of source signals. The mixing model of a form $x = As$ is assumed where both the mixing matrix A and the source signal vector s are unknown. The ICA separation is thus formulated as estimating the demixing matrix (i.e., inverse of A) such that recovered source signals are statistically independent. Therefore, ICA requires that the number of microphones to be at least equal to the number of sources so that A is invertible. Some efforts have been devoted to relax this requirement but raise further constraints on its application in underdetermined scenarios [Winter et al., 2004, Comon and Jutten, 2010].

To extend the application of ICA for underdetermined separation, further assumptions are incorporated within a hierarchical framework of sparse masking techniques and ICA [Araki et al., 2004a,b, Davies and Mitianoudis, 2004]. A more fundamental limitation is that the mixing matrix A must remain the same (i.e. stationary acoustic assumption) for a period of time to provide a reasonable estimate of a large number of model parameters. This assumption is difficult to fulfill in the realistic scenarios in which speakers turn their heads or move around. In addition, most ICA methods assume the number of sources is given beforehand. Similar to the principle of spatial filtering which will be explained in the next section, the source signals must originate

from different spatial directions to achieve separability. However, the ICA techniques do not require prior information on microphone array configuration or the direction of arrival (DOA) of the source signals. Having estimated a set of linear time-invariant demixing filters, ICA implicitly estimates the source directions by maximizing the independence of the sources, and acts as an adaptive null beamformer that reduces the interfering sources [Sawada et al., 2007].

The method proposed in [Buchner et al., 2007] incorporates characterization of the room acoustics in the separation process. Their approach exploits statistical independence assumption of the sources to perform joint deconvolution and separation of speech signals in overdetermined scenarios. An extension for underdetermined scenario is proposed in [Nesta and Omologo, 2012] where multiple complex valued ICA adaptations jointly estimate the mixing matrix and the temporal activities of multiple sources in each frequency band to exploit the spectral sparsity of speech signals. The method does not explicitly rely on identification of the acoustic channel and recovery of the desired source imposes a permutation problem due to mis-alignment of the individual source components [Nesta and Omologo, 2012, Wang et al., 2011].

2.2.2 Spatial Filtering

The geometric source separation can be performed by steering the beam pattern of the microphone array towards the desired speaker. This process is called beamforming. Given the location of the desired speaker, the directivity pattern steering can spatially filter out interferences from other directions regardless of the signal nature [Taghizadeh et al., 2011, Asaei et al., 2009, Omologo et al., 1998]. As the spatial filtering relies on steering the directivity pattern to capture the signal coming from a specific direction, it can mitigate the effect of reverberation which causes a field of dispersed signals. The limitation of beamforming is that separation is not possible when multiple sounds come from directions that are the same or near to each other [Mccowan et al., 2000, Parra and Alvino, 2002]. Several variants of this approach has been applied for robust speech recognition in multiparty scenarios where the interface of sound capturing is an array of microphone and the principle of spatial directivity are employed [McCowan et al., 2002, Parra and Alvino, 2002, Kumatani et al., 2011, McDonough et al., 2007, Asaei et al., 2008, Taghizadeh et al., 2012].

The minimum-variance distortion-less response (MVDR) beamformer is constrained so that signals from the direction of interest are passed with no distortion, while it suppresses noise and interference. The beamforming weights were calculated using time-domain recursive algorithms [Frost, 1972, Griffiths and Jim, 1982]. Recent work considers a frequency-domain MVDR beamformer which performs sample matrix inversion using statistics estimated from a short sample support [Lockwood et al., 2004]. It outperforms the time-domain recursive algorithms in non-stationary acoustic environments. Unlike the ICA approach, adaptive beamforming requires information about the microphone array configuration and the sources (such as the direction of the desired source). However, adaptive beamforming techniques can attenuate spatially spread and reverberant interferences, and there is no need to determine their number. In general, beamform-

ing can attain excellent separation performance in determined or overdetermined time-invariant mixtures. However, in underdetermined scenarios only partial interference suppression is possible. Recent work considers non-linear mixture of beamformers which incorporate sparsity of the spectro-temporal coefficients to address the underdetermined mixtures [Dmour and Davies, 2011]. The application of this method is however limited to the anechoic mixing and the performance is degraded due to reverberation.

An alternative method is proposed in [Huang et al., 2005] which relies on characterizing the acoustic channel to achieve speech separation and dereverberation. Their method applies a blind channel identification approach where the mixing procedure is delineated with a multiple-input multiple-output (MIMO) mathematical model. The authors propose to decompose the convolutive source separation problem into sequential procedures to remove spatial interference at the first step followed by deconvolution of the temporal echoes. To separate the speech interferences, the MIMO system of recorded overlapping speech in reverberant environment is converted into the single-input-multi-output (SIMO) system corresponding to the channel associated with each speaker. The SIMO channel responses are then estimated using the blind channel identification through the unconstrained normalized multi-channel frequency-domain least mean square (UNMCFLMS) algorithm [Huang and Benesty, 2003] and dereverberation can be achieved based on the Bezout theorem (also known in the context of room acoustics as the multiple-input/output inverse-filtering theorem (MINT) [Miyoshi and Kaneda, 1988]). A real-time implementation of this approach has been presented in [Rotili et al., 2010], where the optimum inverse filtering is substituted by an iterative technique, which is computationally more efficient and allows the inversion of long room impulse responses in real-time applications [Rotili et al., 2010]. The major drawback of such implementation is that it can only perform channel identification from single talk periods and it requires a high input signal-to-noise ratio.

Another approach to perform joint dereverberation and speech separation extends the maximum likelihood criteria applied in weighted prediction error method for joint dereverberation and separation of individual speech sources from determined and overdetermined mixtures [Yoshioka et al., 2010]. This method does not perform well in estimation of the acoustic channel and assumes that source spectral components are uncorrelated across time frames. It also relies on a single source assumption and thus can not achieve dereverberation when there are multiple sound sources [Nakatani et al., 2011].

2.3 Nonlinear Speech Recovery

There are two fundamental approaches to extraction of descriptions of individual sound sources from the mixture recordings: *Computational Auditory Scene Analysis (CASA)* and *Sparse Component Analysis (SCA)*. Unlike ICA and spatial filtering, these techniques do not estimate a linear demixing filter and the components corresponding to individual sources are identified and recovered usually through nonlinear objectives. In the following, we review the assumptions underlying each approach and the scope of their application.

2.3.1 Computational Auditory Scene Analysis

The goal of CASA systems is to recover the source signals from one or two recordings of an acoustic scene by exploiting the principles of human auditory scene analysis [Wang and Brown, 2006]. The capability of speech segregation is usually attributed to the ability of listeners to "glimpse" the target voice during gaps in the masking interference [Brungart, 2001]. To state it more specifically, the sparse distribution of speech energy in spectro-temporal plane results in gaps in the spectrum of masker during which listeners can obtain an uncorrupted estimate of the target speech signal. The contribution of CASA is thus identifying the spectro-temporal regions that are dominated by a single sound source [Kollmeier et al., 2008a].

The CASA approach may be regarded as a two-stage process. The first stage is decomposition of the acoustic input into a collection of local spectro-temporal regions. The second stage is grouping the segments of the spectro-temporal scene that are likely to have arisen from the same environmental source into a perceptual structure. Bregman recognizes the major primitive grouping principles relying on (1) *Proximity in frequency and time* (2) *Periodicity* (3) *Continuous or smooth transition (forming a continuous trajectory)* (4) *Common onset and offset* (5) *Amplitude and frequency modulation* [Bregman, 1990, Parsons, 1976]. These modeling mechanisms has been the source of various algorithmic approaches to deal with the complex listening situation [Faller and Merimaa, 2004].

Many CASA systems achieve source segregation by computing a mask to weight a spectro-temporal representation of the acoustic input and this has been regarded as the computational goal of CASA [Hu and Wang, 2001, Roman et al., 1991]. The use of binary spectro-temporal mask is motivated by the phenomenon of masking in auditory perception, in which a sound is rendered inaudible by a louder sound within a critical band. In addition, different lines of computational consideration have converged on the use of binary masks. First, [Jourjine et al., 2000, Roweis, 2003] have noted that a speech signal is sparsely distributed in a high-resolution spectro-temporal representation and, as a result, different speech utterances tend not to overlap in individual spectro-temporal units. This observation leads to the property of orthogonality between different speech utterances. In this case, binary masks are sufficient for decomposing sound mixtures into their constituent sources. The orthogonality assumption holds well for mixtures of speech and other sparsely distributed signals (e.g. complex tones), but, is not valid for speech babble or other broadband intrusions. Second, the notion of a spectro-temporal mask is central to the missing data approach to automatic speech recognition (ASR) proposed in [Cooke, 1991]. In this approach, the mask indicates whether each acoustic feature should be regarded as reliable or unreliable evidence of a target speech source, so that it can be treated appropriately during speech recognition. This fundamental approach of CASA leads to the application of missing data techniques for speech recognition where a reliability mask is constructed to identify the uncorrupted regions [Bourlard, 1999, Cooke et al., 2001, Mccowan et al., 2002, Raj et al., 2004, Gemmeke, 2011, D. Kolossa, 2011].

2.3.2 Sparse Component Analysis

Sparse component analysis (SCA) is a relatively young technique that relies upon the assumption that the sources have a sparse representation [Model and Zibulevsky, 2006, Zibulevsky and Pearlmutter, 2001]. Hence, the representation of signal occupies only a small part of a larger space and the mixtures of sparse components are disjoint. In many cases, even if the canonical representation of the sources do not satisfy the sparsity premise, a linear transformation of the signal such as STFT, DCT or wavelet yields sparse representation. The algorithm proposed in [Jourjine et al., 2000, Yilmaz and Rickard, 2004] laid the foundation of exploiting the sparsity and disjoint characteristics of spectro-temporal representation of speech mixture to perform separation. The method is known as Degenerate Unmixing Estimation Technique (DUET). It exploits delay and attenuation differences between the signals captured by two-channel microphones to construct a binary mask and extract the individual signals. The extensions of DUET have been proposed for M-channel microphone array and convolutive mixtures recorded in a reverberant room in [Melia et al., 2005, Melia and Rickard, 2007, Abrard and Deville, 2005]. We see this algorithm in more details in Section 2.4 to provide some empirical insights on speech recognition performance incorporating the sparsity assumptions.

Another algorithmic approach exploits the property that the spectro-temporal scatter plot of the signal of one microphone versus the other forms oriented lines corresponding to individual sources. Hence, the Line Orientation Separation Technique (LOST) is proposed to identify the individual lines and associate the mixture components to each line via masking [O'Grady and Pearlmutter, 2008, 2004a,b]. The line structure is exhibited if the mixing model is instantaneous without reverberation effects.

The sparse approaches share the common strategy of speech separation by masking similar to what is exploited in CASA and the key differences between various algorithmic approaches amounts to the method of clustering the components for mixing matrix estimation and mask construction or sparse recovery for source separation [Bofill and Zibulevsky, 2001, Jafari et al., 2006, Mourad and Reilly, 2010]. Generalized to CASA, in many SCA approaches, a soft mask is applied thus the assumption that each spectro-temporal point belongs to the same source is relaxed and the resulted demixed speech does not suffer from the musical noise and missing entries in their spectrographic representation [O'Grady and Pearlmutter, 2004a, Araki et al., 2005, Kearns et al., 1997].

Previous work to evaluate the source separation approaches to perform speech recognition has been largely confined to the ICA and spatial filtering techniques, and has imposed the constraint that the number of sources must be less than or equal to the number of microphones. When this condition is not satisfied, the problem is under-determined and traditional linear demixing approaches cannot be applied. However, given a sparse representation of the source, it is possible to recover the components belonging to each speaker and obtain the original signal. There is little literature on evaluating the capability of sparse techniques for speech recognition, with [M. Kuhne and Nordholm, 2007, 2008] being of particular note. Previous work incorporated

the sparse components in *missing data* speech recognition [M. Kuhne and Nordholm, 2007]. In this chapter, we show that sparsity of speech in the spectro-temporal domain can be efficiently exploited in more conventional speech recognition systems. The results obtained show significant improvement to the previous missing data speech recognition approach.

2.4 Incorporating Sparsity of Information Bearing Components

The objective of this section is two-fold: Generally, it is about evaluating the capability of sparse techniques to allow speech recognition in overlapping conditions. More specifically, it aims to demonstrate that acknowledging sparsity leads to more robust representation of the speech signal in multi-speaker environments.

This study focuses in particular on DUET to recover the sparse components of speech in the spectro-temporal domain. It is shown that these components in fact preserve the speech information to be recognized by a conventional speech recognition system [Asaei et al., 2010a]. The demixing cues which DUET relies on are shared between SCA and binaural processing inspired from CASA [Faller and Merimaa, 2004]. Exploiting the common principles applied in CASA and SCA was a key motivation to choose DUET for our initial studies.

2.4.1 Degenerate Unmixing Estimation Technique

The underlying principle behind DUET is that each spectro-temporal component corresponds to only one source. This assumption can be stated mathematically as:

$$S_j(f,\tau)S_k(f,\tau) \approx 0 \qquad \forall j \neq k, \tag{2.3}$$

where $S_j(f,\tau)$ is the windowed short-time Fourier transform (STFT) of the source j when the analysis window is centered at time τ, and f indicates the frequency. Given that each spectro-temporal component belongs to only one of the sources, separation of these components can be achieved by applying a function which gives a unique label to the points associated with each source. Assuming that the room is anechoic, thus the mixtures are attenuated and delayed versions of the original signals along a direct path, the mixing model can be approximated as:

$$\begin{bmatrix} X_1(f,\tau) \\ X_2(f,\tau) \end{bmatrix} \approx \begin{bmatrix} 1 & \dots & 1 \\ a_1 e^{-if\delta_1} & \dots & a_N e^{-if\delta_N} \end{bmatrix} \begin{bmatrix} S_1(f,\tau) \\ \dots \\ S_N(f,\tau) \end{bmatrix}, \tag{2.4}$$

where $X_1(f,\tau)$ and $X_2(f,\tau)$ are STFT of the signal captured by distant microphones. a_n and δ_n are the relative attenuation and delay parameters of the source n respectively; their values are proportional to the relative distance of the source to the two microphones. The total number of sources is $N \geqslant 2$; we have only two microphones for recordings. Based on (2.3) and (2.4), an instantaneous estimate of the mixing parameters can be obtained by applying the magnitude and

phase operator onto the complex STFT ratio of the microphone signals:

$$\tilde{a} = |\frac{X_2(f,\tau)}{X_1(f,\tau)}| \quad \text{and} \quad \tilde{\delta} = -\frac{1}{f}\arg\left(\frac{X_1(f,\tau)}{X_2(f,\tau)}\right), \tag{2.5}$$

Given that each source has a unique mixing parameter (spatial signature), the problem is how to identify the actual mixing parameters from these instantaneous estimates and estimate the sources.

Estimation of the Mixing Parameters and Sources

Assuming the contribution of the interfering sources to be independent Gaussian noise and maximizing the likelihood of the mixed signals given the source (S) and mixing parameters (a and δ), a closed-form estimator is obtained [Jourjine et al., 2000, Yilmaz and Rickard, 2004]. We proceed from the result of [Yilmaz and Rickard, 2004] that states that the number of sources and their corresponding mixing parameters can be identified based on the number and location of the peaks in a 2D weighted histogram, where the $(\tilde{\alpha}, \tilde{\delta})$ pairs are used to indicate the indices into the histogram and each point is weighted by

$$|X_1(f,\tau)X_2(f,\tau)|^p f^q, \tag{2.6}$$

where $\tilde{\alpha} = \tilde{a} - 1/\tilde{a}$ is the symmetric attenuation used to obtain Maximum Likelihood (ML) estimate [Yilmaz and Rickard, 2004] and p and q are hyper-parameters chosen for various weighting schemes. In the 2D histogram constructed in this way, clusters of weights will emerge centered on the actual mixing parameter pairs corresponding to the source locations. We found that the following steps are required to achieve high recognition performance.

Hyper-parameter Optimization

The weighted histogram described above is based on a ML estimate for the mixing parameters. It has been shown in [Yilmaz and Rickard, 2004] that it is possible to obtain a maximum likelihood estimation of the hyper-parameters p, q and it has been suggested that $p = 1$ and $q = 0$ is a good default choice. Following a crude grid search, we find that choosing $p = 0.5$ and $q = 1$ applied on subband frequency components gives the best recognition performance.

Subband Weighted Histogram

The histogram based localization approach imposes that the microphones are sufficiently close to avoid the delay estimate from the complex STFT to wrap around. This requires that

$$|f\delta_j| < \pi. \tag{2.7}$$

For the cases where this constraint is not satisfied, we define a safe-delay margin (Δ) based on the maximum high-frequency component and tile a number of histograms constructed from delaying one mixture against the other by products of Δ. The histograms are then appended to obtain a large histogram with a big delay range. To prevent spurious peaks due to phase-wrapping, we propose to consider only the subband frequency components that satisfy (2.7). Figure 2.1 illustrates a subband histogram and its fullband counterpart. As can be seen, the subband histogram contains very localized peaks around the actual mixing parameters whereas the fullband histogram has many spurious peaks that prevent accurate localization of the sources. For a speech signal, there are high energy components below 400 Hz due to pitch and the first formant frequency of the high vowels (e.g., /i/ and /u/). This corresponds to the subband histogram approach being useful for microphone separations up to 40 cm.

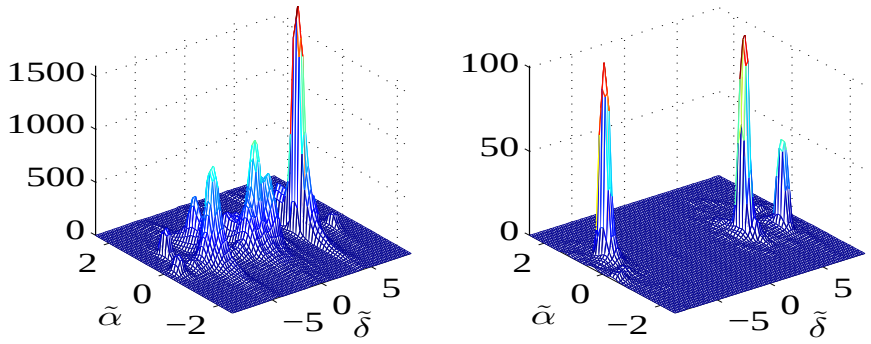

Figure 2.1 – Left: Fullband weighted histogram, Right: Subband weighted histogram ($100 \leqslant f \leqslant 570$). The actual number of sources is 3, $p = 0.5$ and $q = 1$. Distance between microphones is 0.03m.

SNR-based Spectral Smoothing

Having estimated the mixing parameters as the peak centers of the histogram, they can be used to label all spectro-temporal points and construct N disjoint masks to separate the components belonging to each of the speakers. We are interested in evaluating the amount of information recovered based on disjointness assumptions for speech recognition systems. The main difficulty that we observed is that the standard feature extraction approaches for speech recognition are sensitive to the gaps in the spectra resulting from masking. Previous authors [M. Kuhne and Nordholm, 2007] have used missing data techniques to deal with these missing values.

In this study we investigate a two step procedure. First, we set the missing values to zero and use overlap-add (OLA) to reconstruct a time domain signal. Then, we add white Gaussian noise to the signal within a specific signal to noise ratio (SNR). By preventing the effect of zeros on the feature extraction, this will lead to smooth the spectral shape at the discontinuities. OLA is also a convenient mean of changing the DFT size and period.

2.4.2 Empirical Insights on Speech Recognition

We conducted experiments to try to answer the following questions:

1. Are the sparse source separation assumptions valid while incorporated for recognition of overlapping speech? i.e., Does DUET work well with conventional speech recognition systems?
2. What is the limit of performance? i.e., How far we can push the disjointness/sparsity assumptions?
3. Given the spectro-temporal representation of speech signal, is the salient information needed for speech recognition preserved only in a small fraction of the whole components? i.e., How well does it fit a common metric for sparsity?

Overlapping Speech Database

The experiments are all performed in the framework of AURORA2 [Pearce and Hirsch, 2000]. This database is designed to evaluate the performance of speech recognition algorithms in noisy conditions and has become the standard one. A fixed HTK back-end was trained on multi-condition data with different noise types including those of Subway, Babble, Car and Exhibition at 5 SNR levels as well as clean data. Overlapping speech was synthesized by mixing clean AURORA 2 test utterances with interfering sentences from the HTIMIT database. The broad phonetic space of HTIMIT allows the results of our AURORA2 framework to be generalizable for the task of digit recognition in overlapping conditions. For each test sample, interferences are randomly chosen out of this subset to construct two mixtures. The interference files are scaled prior to mixing to achieve the particular baseline and looped to compensate for the difference between the file lengths.

Acoustic Parameters

The planar area of a room with dimension 3×4 is divided into grids with 50 cm spacing (hence 48 grids in total). The sources are assumed to have the same elevations as the sensors (located in the middle with 1.5 m height) and distributed as depicted in Fig. 2.2. The stereo mixtures are recorded from the room center. Room impulse responses are generated with the Image model technique [Allen and Berkley, 1979] using intra-sample interpolation, up to 15^{th} order reflections and omni-directional microphones. The corresponding reflection ratio, β used by the Image model was calculated via Eyring's formula [Eyring, 1930]

$$\beta = \exp(-13.82/[c(L_x^{-1} + L_y^{-1} + L_z^{-1})T]) \tag{2.8}$$

where L_x, L_y and Lz are the room dimensions, c is the speed of sound in the air ($\approx 342 m/s$) and T is the room reverberation time. In our experiments $T = 200$ ms.

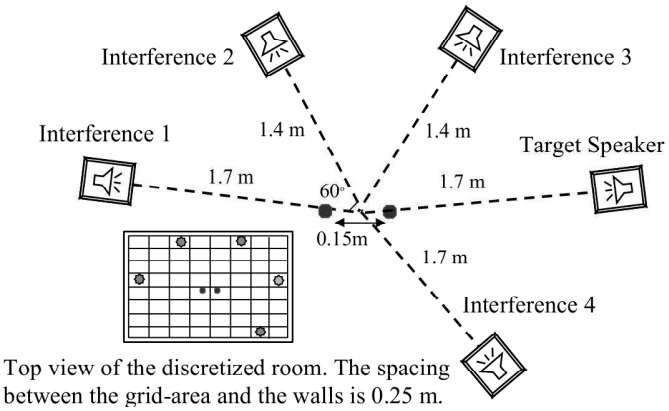

Figure 2.2 – Overhead view of the room set-up. The overlapping scenario is recorded by two microphones located 15 cm apart. The target speaker is almost in the same line as the microphones. There are four competing speakers which are illustrated in the picture.

Analysis Parameters

The subband histogram is constructed for the frequency band of 100–570 Hz. The lower band is chosen based on the lowest human fundamental frequency. Although, HTIMIT and AURORA are both telephony speech, the frequency components below 300 Hz are not completely suppressed. Recall from equation (2.5) that the ratio of the two mixtures is used for instantaneous mixing parameter estimates $(\tilde{\alpha}, \tilde{\delta})$. The upper-band is chosen to satisfy equation (2.7) and prevent phase-wrapping. The size of the bins in the attenuation and delay histograms are 0.06 and 0.14 samples respectively. The histogram attenuation width is $|\tilde{\alpha}| \leqslant 2.5$. For the close-microphone scenario, the histogram delay width is $|\tilde{\delta}| \leqslant 4$. In the far-microphone case, 3 histograms are appended together, each obtained by delaying the second mixture by +4 and -4 samples. Therefore, the delay-width of the big histogram is $|\tilde{\delta}| \leqslant 8$ samples. The target is detected based on the geometric proximity to the position of interest. Notice that the sub-band histogram is only used to estimate the mixing parameters (roughly speaking, source localization) whereas the separation of source components are all performed for the whole frequency band. The analysis and synthesis window for source separation and signal reconstruction is Hann to facilitate OLA. The size of the window is 125 ms. Following a crude grid search, we find that choosing $p = 0.5$ and $q = 1$ and adding white Gaussian noise at 37 dB to the demixed signal gives the best performance.

The separated signal is then presented to the standard HTK-based [Cook et al., 1955b] AURORA2 speech recognition system [2]. The speech signal is processed in blocks of 25 ms with a shift of 10 ms to extract 13 MFCC cepstral coefficients per frame. These coefficients after cepstral mean/variance normalization are appended to their delta and delta-delta derivatives to obtain a 39 dimensional feature vector for every 10 ms of speech.

2. The system was scripted at Idiap Research Institute by Dr. Phil Garner

Speech Recognition Performance

Table 2.1 gives the recognition results of the demixed speech for the clean and multi-condition training. The recognition accuracy shows the significant potential of disjointness assumptions in the spectro-temporal domain to demix the signals while preserving the salient information to perform speech recognition. The best results are obtained for far-microphones.

Table 2.1 – Word accuracy for mixtures and separated components smoothed with OLA and adding white Gaussian noise (WGN). The results are obtained using a speech recognizer.

Mic-dis (cm)	Train Condition	AURORA (%) Baseline	DUET (%)	
			OLA	OLA+WGN
5	Clean	50.09	77.03	79.29
	Multi-Con.	39.62	81.09	91.14
30	Clean	51.35	80.31	86.34
	Multi-Con.	44	79.95	93.35

The histogram peaks for far-microphones are well localized, and the peak regions are clearly distinct, whereas the peak regions in the close-microphone histogram have some degree of overlapping. Experiments on full-band weighted histogram as well as non-frequency weighted histogram ($q = 0$) yielded consistently poor results, more than 10% reduction in recognition rate. For the close-microphone scenario, the proposed sub-band weighting scheme is still the best choice. Furthermore, we observed that adding a negligible amount of Gaussian noise (SNR=37 dB) improves the recognition results up to 14%. The improvement is obtained for both clean and multi-condition training. We can justify this by considering that adding noise specially improves the non-voiced speech which is mainly potential non-disjoint/sparse part of speech signal with an inherent noisy nature. In addition, it prevents the computational issues due to processing zeros in the calculation of features for speech recognition. Figure 2.3 illustrates a sample spectrogram obtained from DUET and the one used for speech (MFCC) feature extraction.

Performance Limit

To approach the question on how far we can push the disjointness/sparsity assumptions, we set up some experiments to do recognition of the separated speech while the number of interferences is increased up to 10. We intend to quantify the sparsity of informative coefficients of speech signal in terms of recognition rate. The set of mixing parameters are chosen to provide a different spatial signature for each speaker. The recognition rate after separation is depicted in Figure 2.4.

We can conclude from this observation that the salient information needed to recognize speech is in fact in a small fraction of the spectro-temporal coefficients and it is quite unlikely for the speech interferences to overlap these sparse structures. Furthermore, we observed that increasing the number of interferences degrades the accuracy of the separation by attenuation and the histogram clusters are separated mainly due to the different delay parameters. This observation emphasizes that the sparse source separation techniques which exploit both the delay and attenuation has

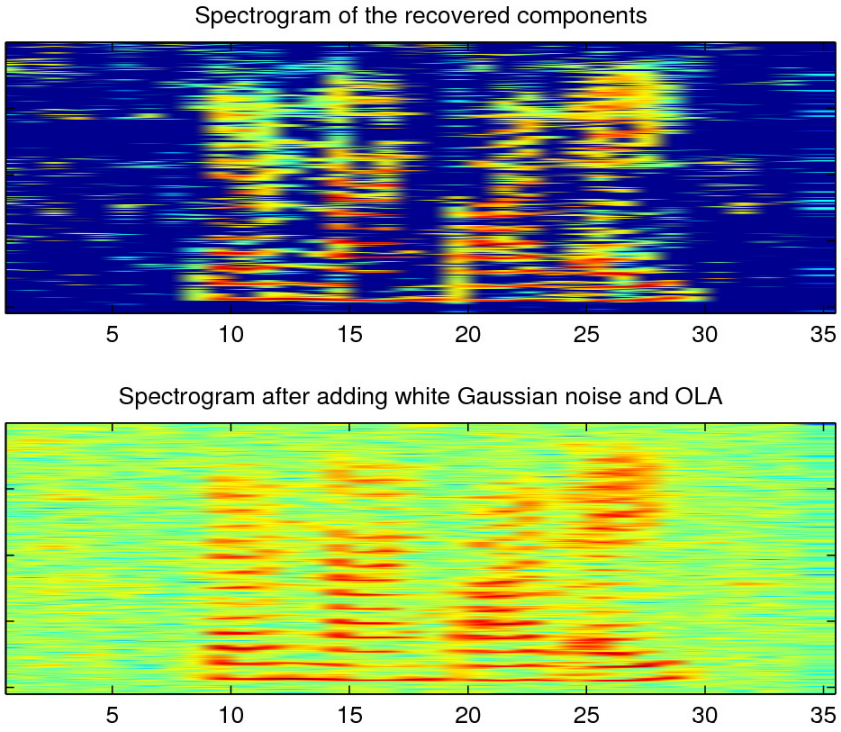

Figure 2.3 – Spectrogram of the recovered speech using DUET and the spectrogram being applied for speech recognition feature extraction

potential to exhibit more robustness.

Sparsity of Speech in Spectro-Temporal Domain

The results of the experiments for the increased number of interferences were very intriguing. To examine how it fits a common metric for sparsity, we investigate the spectro-temporal components recovered by DUET. For the natural signals to be closely approximated as sparse, their representation coefficients S must have a rapid power-law decay when sorted [Baraniuk et al., 2010]

$$|S_{I(i)}| \leqslant \gamma i^{\frac{-1}{r}} \qquad r \leqslant 1, \tag{2.9}$$

where I indexes the coefficients of S when sorted from largest to smallest. Plotting the sorted absolute value of the recovered spectro-temporal components vs. their index is illustrated in Figure (2.5) which satisfies equation (2.9) and exhibits power-law decay, though more than 200 coefficients are needed to get into a regime where the coefficient decay is better than -1. Based on this observation, the speech representation in spectro-temporal space can be approximated to be sparse. This observation has already been shown to be beneficial for speech recognition motivates investigating the sparse features in an integrated framework for source separation and speech recognition. We elaborate more on compressibility and sparse representation in Chapter 4.

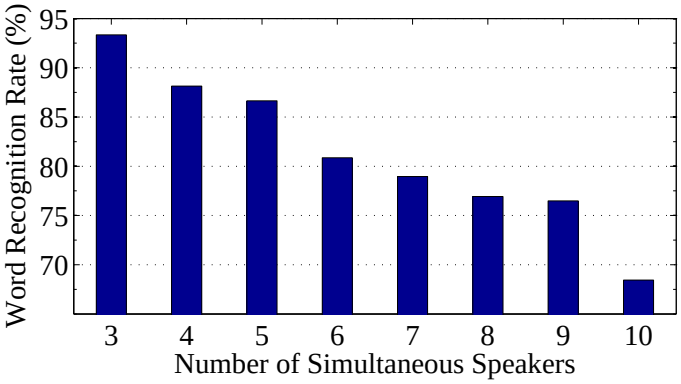

Figure 2.4 – Spectro-temporal disjointness of overlapping speech quantified in terms of ASR performance

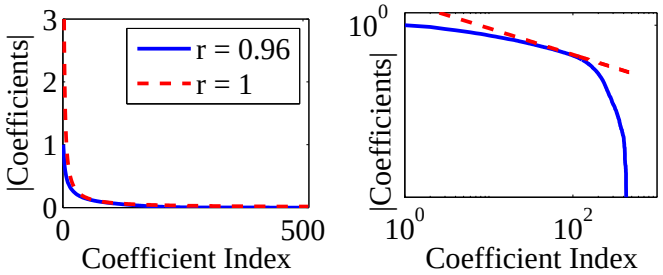

Figure 2.5 – Power-law decay of the recovered spectro-temporal components depicted in linear (left) and logarithmic (right) axes.

2.5 Conclusions

We explored the state-of-the-art techniques in speech separation with a slant towards applications is multiparty speech recognition. We evaluated the sparsity assumptions incorporated in sparse component analysis in the framework of DUET for speech recognition in a multi-speaker environment. Recognition results after demixing show that the salient information needed to recognize speech is in a fraction of disjoint spectro-temporal components. Setting the rest of the coefficients to zero and smoothing the spectral discontinuities by OLA and adding white Gaussian noise, the demixed signal can be recognized using a conventional speech recognition system. These analyses strengthen the benefit of sparse assumptions for recognition of overlapping speech. When the speech signal is represented in a sparse domain, separation of the components becomes straightforward, and these components in fact preserve the information to be recognized well. This motivates more research on identifying a domain where the speech representation is sparse and these coefficients are directly applicable for speech recognition. The framework of DUET has a major limitation which is the underlying assumption that there is no reverberation effect. We propose a new framework which entangles the sparse representation and the model of multipath propagation to address the problem of convolutive speech separation. This subject is discussed in the following Chapter 3.

3 A Compressive Sensing Perspective to Spatio-Spectral Information Recovery

In this chapter, we state the problem of analysis of the multichannel recordings in terms of recovering the high-dimensional signal information from a few microphone measurements. Our formulation relies on a new perspective that acquisition of the signals by microphone array is a natural realization of the Compressive Sensing (CS) framework. We overview the fundamental CS premises and elaborate on realization of the CS components in our formulation.

3.1 Fundamental Premises

Compressed sensing exploits sparsity to acquire high-dimensional signals using very few linear non-adaptive measurements. The theory relies on two fundamental principles: *sparsity*, which pertains to the signals of interest and, *incoherence*, which pertains to the sensing modality.

A signal $\mathscr{S}$ in a G-dimensional space is N-sparse if only $N \ll G$ entries of $\mathscr{S}$ are nonzero. We call the set of indices corresponding to the non-zero entries as the support of $\mathscr{S}$. The CS theory indicates that such a signal can be sampled and reconstructed with only $M = O(N\log(G/N))$ linear measurements [Baraniuk et al., 2010] obtained as

$$\mathscr{X} = \Phi\mathscr{S} \tag{3.1}$$

where $\mathscr{X}$ denotes the microphone array recordings and Φ is an $M \times G$ measurement matrix. A sufficient but not necessary condition on Φ to recover the signal is the Restricted Isometry Property (RIP). An isometry constant δ_N of a matrix Φ is the smallest number such that

$$(1-\delta_N)\|\mathscr{S}\|_2^2 \leqslant \|\Phi\mathscr{S}\|_2^2 \leqslant (1+\delta_N)\|\mathscr{S}\|_2^2; \tag{3.2}$$

where $\|.\|_p$ denotes the ℓ_p-norm defined as $\|\mathscr{S}\|_p := (\sum_i |\mathscr{S}_i|^p)^{1/p}$. The matrix Φ holds RIP property if δ_N is not too close to one. This property implies that all pairwise distances between N-sparse signals must be well preserved in the measurement space or equivalently all subsets of N columns taken from the measurements are in fact nearly orthogonal. The RIP mapping of N-sparse

set into observation space guarantees stable recovery of the high-dimensional data [Blumensath and Davies, 2008].

Relying on (1) sparse representation and (2) incoherent or information preserving measurements, CS guarantees to circumvent the ill-posedness of the recovery problem and reconstruct the N-sparse signal from the compressed measurements by efficient optimization algorithms which search for the sparsest signal that agrees with those measurements.

In practice, signals may not be exactly sparse but, they can be applied in CS framework if the support of the coefficients have a rapid power-law decay when sorted hence, called compressible signals. A new paradigm in CS exploits the inter-dependency structure underlying the support of the sparse coefficients in recovery algorithms to reduce the number of required measurements and to better differentiate true signal information from recovery artifacts, which leads to more robust and efficient recovery [Baraniuk et al., 2010].

3.2 Natural CS Realization

This thesis is built on a new perspective to the objective of multi-channel processing as recovery of high-dimensional data embedded in the acoustic scene from a few observations recorded by an array of microphones. This inspiring analogy between the type of measurements associated with CS and the natural projections manifested by the media Green's function in sensor array recordings was independently recognized and exploited in our work published in [Asaei et al., 2011a].

We later realized that the relationship between compressive sensing and sensor array has been explored in a recent work of Lawrence Carin in the field of antenna array processing. Carin provides rigorous analysis of the observed field as projections of far-field sources located at any arbitrary direction onto the corresponding Green's function and verifies the orthogonality of the projections as long as the antennas are separated by a half wavelength or more [Carin, 2009, Carin et al., 2011]. This natural manifestation of CS through physics of sound propagation holds for general array construction and general linear isotropic media and enables theoretical development and quantitative assessment of sensor array performance using the generic theory of CS [Carin, 2009, Carin et al., 2011].

The following sections explain how the CS principles can be realized within the framework of microphone array processing. In addition to the spectral sparsity exploited in Chapter 2, we incorporate *spatial sparsity* of the concurrent sources to formulate a new objective to recover the spatio-spectral information from compressive acoustic measurements.

3.2.1 Structured Sparse Representation

The first step to formulate a sparse recovery problem requires identification of a domain in which the information has sparse representation. We consider a scenario in which N speech sources are distributed in a planar area spatially discretized into a grid of G cells. We assume to have a sufficiently dense grid so that each speaker is located at one of the cells thus $N \ll G$. The spatial spectra of the sources is defined as a vector with a sparse support indicating the components of the signal corresponding to each grid's cell.

We consider spectro-temporal representation of multi-party speech and entangle the spatial representation of the sources with the spectral representation of the speech signal to form a vector $\mathscr{S} = [S_1^T ... S_G^T]^T \in \mathbb{C}^{GF\times 1}$ where $.^T$ stands for the transpose operator. Each $S_g \in \mathbb{C}^{F\times 1}$ denotes the spectral representation of the g^{th} source signal (located at cell number g) in Fourier domain. If a source is located at cell g, all of its spectral components correspond to that particular cell; hence, there is a *block structure* underlying the support of the coefficients in $\mathscr{S}$. We elaborate more on the structured sparsity models in Chapter 4.

We express the signal ensemble at microphone array as a vector $\mathscr{X} = [X_1^T ... X_M^T]^T$ where each $X_m \in \mathbb{C}^{F\times 1}$ denotes the spectral representation of the recorded signal at microphone m. The sparse vector $\mathscr{S}$ generates the microphone observations as $\mathscr{X} = \Phi\mathscr{S}$ where Φ is the microphone array measurement matrix constituted of the acoustic projections associated to the acquisition of source signals located over the grid.

3.2.2 Compressive Acoustic Measurement

The second step to formulate a sparse recovery problem requires characterization of the measurement mechanism associated to microphone array recordings drawn by the physics of sound propagation. We assume the room to be a rectangular enclosure consisting of finite impedance walls. The point source-to-microphone impulse responses of the room are calculated using the *Image model* technique [Allen and Berkley, 1979]. Taking into account the physics of the signal propagation and multi-path effects, the projections associated with the source located at the cell g where ν_g represents the position of the center of the cell and captured by microphone m located at position μ_m are characterized by the media Green's function and denoted as $\xi_{\nu_g\to\mu_m}$ defined by

$$\xi^f_{\nu_g\to\mu_m} : X(f,\tau) = \sum_{r=1}^{R} \frac{\iota^r}{\|\mu_m - \nu_g^r\|^\alpha} \exp(-jf\frac{\|\mu_m - \nu_g^r\|}{c}) S(f,\tau), \tag{3.3}$$

where $j = \sqrt{-1}$ and ν_g^r indicates the position of r^{th} virtual sources corresponding to the actual source located at cell g with reflective energy ratio of ι^r. The attenuation constant α depends on the nature of the propagation and is considered in our model to equal one which corresponds to the spherical propagation. R denotes the total number of images taken into account. This

formulation assumes that if $s_1(l) = s(l)$ and $s_2(l) = s(l-\rho)$, then $S_2(f,\tau) \approx \exp(-jf\rho)S_1(f,\tau)$; the frame size should be at least the size of the impulse response for this assumption to hold.

Given the source-sensor projection defined in (3.3), we construct matrix $\Xi_{\nu_g \to \mu_m}$ for the measurement of the F consecutive frequencies as

$$\Xi_{\nu_g \to \mu_m} = \begin{bmatrix} \xi^1_{\nu_g \to \mu_m} & 0 & \dots & 0 \\ 0 & \xi^2_{\nu_g \to \mu_m} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \xi^F_{\nu_g \to \mu_m} \end{bmatrix}_{F \times F} \tag{3.4}$$

Acoustic projection (acquisition) of a source located at ν_g by a microphone located at μ_m through $\Xi_{\nu_g \to \mu_m}[S(f_1,\tau)S(f_2,\tau)...S(f_F,\tau)]^T$ yields the signal spectrum as acquired by the distant microphone m. If we consider all possible source locations, the projections associated with the acquisition of the source signals located at any of the grid cells by microphone m is

$$\phi_m = [\Xi_{\nu_1 \to \mu_m} ... \Xi_{\nu_g \to \mu_m} ... \Xi_{\nu_G \to \mu_m}]_{F \times GF} \tag{3.5}$$

Hence, the measurement matrix corresponding to M-channel microphone array recordings is obtained through

$$\Phi = \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_M \end{bmatrix}_{MF \times GF} \tag{3.6}$$

The microphone array manifold matrix as characterized through (3.3)-(3.6) is also referred to as the *forward model* of the acoustic projections. As indicated by (3.3), characterizing the acoustic projections amounts to identifying the location of the *source images* ν^r_g as well as the absorption factors of the reflective surfaces ι^r. We exploit this parametric model to identify the multipath measurement mechanism in Chapter 5.

3.2.3 Model-based Sparse Recovery

Given the spatio-spectral sparse representation stated in Section 3.2.1 and the measurement matrix stated in Section 3.2.2, the sparse vector $\mathscr{S}$ generates the microphone observations as $\mathscr{X} = \Phi\mathscr{S}$. Our goal is to recover $\mathscr{S}$ from a small number of recordings M, where $M < G$. There are infinitely many solutions to this problem; we thus exploit the prior information on sparse properties of $\mathscr{S}$ to circumvent the ill-posedness and cast the underdetermined speech recovery problem as sparse approximation of $\mathscr{S}$. Additionally, we integrate the structures underlying the sparse coefficients in our objective function so the recovery of speech signals amounts to solving

the model-based sparse reconstruction problem stated precisely as:

$$\hat{\mathscr{S}} = \underset{\mathscr{S}\in\mathbb{M}}{\operatorname{argmin}} \|\mathscr{S}\|_0 \quad \text{s.t.} \quad \mathscr{X} = \Phi\mathscr{S} \tag{3.7}$$

where the counting function $\|.\|_0$: $\mathbb{R}^G \to \mathbb{R}$ returns the number of non-zero components in its argument. $\mathbb{M}$ stands for the union of all vectors with a particular support structure (e.g. *block sparsity* as indicated in Section 3.2.1). Incorporating the structured sparsity models limits the degrees of freedom within a smaller subspace. The recovery performance is therefore more efficient from fewer number of measurements [Baraniuk et al., 2010]. The solution stated in (3.7) is NP-hard, hence various techniques are proposed to obtain an estimated $\hat{\mathscr{S}}$ within a computationally feasible framework. Chapter 6 is dedicated to our studies of the computational methods to model-based sparse recovery.

3.2.4 CS Perspective on Performance

Consider a linear sensor array with uniform inter-element spacing Δ; the array is assumed to reside in vacuum. We study the properties of the measurement matrix Φ. Each column of Φ corresponds to the Green's function from a discrete source angle $\theta \in [0, 2\pi]$ to the M uniformly spaced microphones. The components that constitute the i^{th} column of Φ may be expressed (up to a constant) as:

$$[\Phi_{1,i}, \Phi_{2,i}, ..., \Phi_{M,i}]^T = [1, \omega_i, \omega_i^2, ..., \omega_i^{M-1}]^T, \tag{3.8}$$

where $\omega_i = \exp[-j2\pi\frac{\Delta}{\lambda}\cos(\theta_i)]$, and θ_i corresponds to the i^{th} angular bin, with $j = \sqrt{-1}$ and λ denotes the signal wavelength. We observe that each column of Φ corresponds to a sampled Fourier basis function, at angular frequency ω_i, truncated over M samples. By setting $\Delta = \lambda/2$ the desired near-orthogonal projections are achieved [Carin, 2009, Carin et al., 2011].

To extend this framework for analysis of the measurements obtained in a general acoustic environment, we need to consider the coherence of the projections. We characterized the microphone array measurements for a reverberant enclosure in Section 3.2.2 and leveraged model-based sparse recovery algorithms to estimate the individual speech components. In this framework, the theoretical analysis of the performance bounds of our approach is entangled with the performance of the sparse recovery algorithms [Tropp and Wright, 2010]. A fundamental property to guarantee the theoretical performance bounds is the coherence of the measurement matrix defined as

$$\vartheta(\Phi) = \max_{1\leqslant j,k\leqslant G, j\neq k} \frac{|\langle \Phi_{.j}, \Phi_{.k}\rangle|}{\|\Phi_{.j}\|\|\Phi_{.k}\|} \tag{3.9}$$

The coherence quantifies the smallest angle between any pairs of the columns of Φ. The number of recoverable non-zero coefficients (N) using either convexified or greedy sparse recovery is

inversely proportional to the coherence ϑ as

$$\mathrm{N} < \frac{1}{2}(\vartheta^{-1} + 1) \tag{3.10}$$

Therefore, to guarantee the performance of sparse recovery algorithms, it is desired that the coherence is minimized. As the measurement matrix is constructed of the location-dependent projections, this property implies that the contribution of the source to the array's response is small outside the corresponding sensor location or equivalently the resolution of the array is maximized. In general, a large-aperture random design of sensor array (or a sparse microphone array layout) yields the projections to be mutually incoherent, so the projections are spread across all the acoustic scene and each sensor captures the information about all components of $\mathscr{S}$ [Carin, 2009, Carin et al., 2011]. This analysis shows that the performance of our sparse approximation framework is entangled with the microphone array construction design and the algorithmic approaches to sparse recovery. These issues are studied in Chapter 6.

3.3 Empirical Insights on Speech Recognition

In Section 2.4, we showed that sparse component analysis exhibits several advantages when dealing with overlapping problem in speech recognition systems. We achieved excellent word recognition rate (as listed in Table 2.1) with conventional speech recognition under the assumptions that there is no reverberation in the room, hence the spatial cues (direct path delay and attenuation) are reliable to estimate the mixing process and recover the sources. It has been shown in [Yilmaz and Rickard, 2004] that despite the degradation of the spatial cues in reverberant conditions, the components of the overlapping speech in spectro-temporal domain remain disjoint and sparse. This observation was a motivation to formulate the underdetermined source separation as a sparse recovery problem from dimensionality reducing measurements. We thus formulate the sparse representation of multiparty speech as explained in Section 3.2.2 and characterize the multipath projections through the procedure stated in Section 3.2.1. The speech recovery amounts to solving the sparse approximation problem stated in Section 3.2.3.

We apply a model-based CS recovery approach proposed in [Cevher, 2011]. This algorithm is inspired by the development of the first order methods in optimization, most notably based on the algebra used in Nesterov's optimal gradient and smoothing techniques. The sparse recovery is performed by a gradient type of method where the Lipschitz gradient constant is used as the step size to guarantee the fastest convergence speed. To incorporate for the underlying structure of the sparse coefficients, a model approximation is performed along with a gradient calculation at each iteration. Since the sparse coefficients in our model live in at most N blocks, an N-block-sparse signal is approximated by reweighting and thresholding the energy of the blocks. The recovered signal $\mathscr{S}$, contains the contribution of each speaker to the actual sensor observations in the block corresponding to the speaker position. We refer to our method as Blind Source Separation via Model-based Sparse Recovery (BSS-MSR). Contrary to the common SCA practice, our formulation merges the two steps of mixing process estimation and source separation as a joint

localization-separation framework [Cevher et al., 2009]. In the following sections, we evaluate the speech recovery and recognition performance of our method.

3.3.1 Analysis Parameters

The overlapping speech corpus is identical to what is described in Section 2.4.2. The experiments are all performed in the framework of AURORA2 [Pearce and Hirsch, 2000]. Overlapping speech was synthesized by mixing clean AURORA2 test utterances with interfering sentences from the HTIMIT database. For each test sample, interferences are randomly chosen out of this subset to construct two mixtures. The interference files are scaled prior to mixing to achieve the particular baseline and looped to compensate for the difference between the file lengths.

The planar area of a room with dimension 3 m $\times$ 4m is divided into grids with 50 cm spacing (hence 48 grids in total). The sources are assumed to have the same elevations as the sensors (located in the middle with 1.5 m height), and distributed as depicted in Figure 2.2. The stereo mixtures are recorded from the room center. Room impulse responses are generated with the Image model technique [Allen and Berkley, 1979] using intra-sample interpolation, up to 15th order reflections and omni-directional microphones and the room reverberation time is estimated to be about 200 ms.

The speech signals are recorded at 8 kHz sampling frequency and the spectro-temporal representation for source separation is obtained by windowing the signal in 250 ms frames using Hann function with 50% overlapping. The length of the frame is chosen to be greater than the reverberation time to meet the assumptions underlying (3.3). The separated speech is then reconstructed back into time domain. We found this reconstruction to be a convenient mean of changing the FFT size and period.

Speech sources are the separated signals which are presented to the standard AURORA2 speech recognition system. The signals are processed in blocks of 25 ms with a shift of 10 ms to extract 13 MFCC cepstral coefficients per frame. These coefficients after cepstral mean/variance normalization are appended to their delta and delta-delta derivatives to obtain a 39 dimensional feature vector for every 10 ms of speech.

3.3.2 Performance Evaluation

The quality of the separated speech is measured according to the performance criteria proposed in [Vincent et al., 2006]. The idea is motivated by the auditory segregation principles and performs decomposition of the estimated signal into distinct components of interference and noise. The amount of these distortions is computed numerically as the energy ratios expressed in decibels. The objective measures include source-to-interference ratio (SIR) and source-to-noise ratio

(SNR), as defined by

$$\mathrm{SIR} := 10\log_{10}\frac{\|s_{\text{target}}\|_2^2}{\|e_{\text{interf}}\|_2^2}, \qquad \mathrm{SNR} := 10\log_{10}\frac{\|s_{\text{target}}+e_{\text{interf}}\|_2^2}{\|e_{\text{noise}}\|_2^2} \tag{3.11}$$

The value of $s_{\text{target}} + e_{\text{noise}}$ is obtained by least-squares projection of the estimated source image onto the signal subspace spanned by the filtered version of the true source image [Vincent et al., 2006].

In addition to the objective quality measures, we conducted experiments to evaluate the performance of the proposed method for speech recognition systems. Although any number of microphones ($\geqslant 2$) can be accommodated in our framework, we carried out the experiments with stereo mixtures (two microphones) to compare the results with our previous results in Section 2.4.

Due to the limited number of microphone recordings, we first performed a basis reduction using the 20% high-energy coefficients of the observed data. In this procedure, each spectro-temporal coefficient is assigned to one of the cells on the spatial grid by ℓ_1-minimization of $\mathscr{S}$ over the whole grid points. The number of coefficients assigned to each cell denotes the activity of that region. Based on the activity obtained as such some of the cells are selected and the rest are discarded. Therefore, the BSS-MSR algorithm is ran on a smaller subset of the cells to recover the speech sources. The number of selected cells is upper-bounded with 15 (we assumed that the number of concurrent speakers is always less than 15). This activity detection results in reducing the dimensionality of the sparse recovery problem and increases its efficiency.

The source separation is then performed by solving (3.7) using the iterative hard thresholding algorithm proposed in [Cevher, 2011] [1]. The recovered signal $\mathscr{S}$, contains the contribution of each speaker to the actual sensor observations in a block of the spectral components of the source signal and the index of the block corresponds to the cell where the speaker is located. Hence, the proposed framework of BSS-MSR performs joint localization and separation of the concurrent sources. The target speech is selected based on the proximity to the position of interest. The forward model of the room impulse response is not given in the tests and it is approximated by taking an arbitrary value for the reflection coefficients. We repeated the experiments for $\iota = \{0.6, 0.7, 0.9, 1\}$ (considering $\iota \in [0.6, 1]$ for the typical room acoustic) and averaged the results.

We compared our method with the improved (subband) version of Degenerate Unmixing Estimation Technique (DUET) for speech recognition stated in Section 2.4. We also ran the experiments on Line Orientation Separation Technique (LOST) [O'Grady and Pearlmutter, 2004a] since the separation of sparse components in this method is based on ℓ_1-minimization at each spectro-temporal point independently. The quality of the separated target speech is depicted in Figure 3.1. Tables 3.1-3.2 give the recognition results of the demixed speech for the clean and multi-condition training of AURORA2 database. The baseline performance is reported in terms of word recogni-

1. The details of the algorithm are explained in Chapter 6

tion rate using overlapping speech.

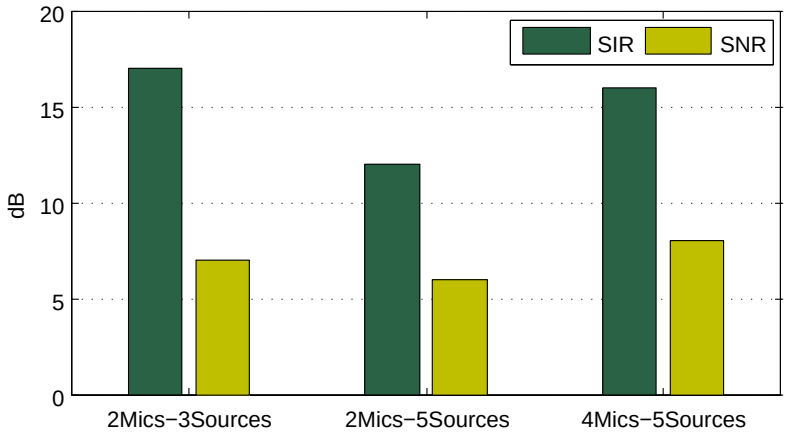

Figure 3.1 – Objective quality measurement of the separated speech in terms of SIR and SNR using either 2 or 4 microphones for separation of 5 sources

Table 3.1 – Word accuracy of the separated speech for stereo echoic mixtures of 3 sources (interferences 1-2 and target speech).

Train Cond. (TC)	Baseline	DUET	LOST	BSS-MSR
Clean (C)	59.3	56.34	61.2	89.3
Multi-Cond. (MC)	61.78	77.25	64.3	92.7

Table 3.2 – Word accuracy of the separated speech for echoic mixtures of 5 sources (interferences 1-4 and target speech). BSS-MSR^1 refers to the stereo recording and BSS-MSR^2 refers to 4-channel circular microphone array recording with the radius = 0.15m.

TC	Baseline	DUET	LOST	BSS-MSR^1	BSS-MSR^2
C	47.3	31.72	49.2	81.7	88.7
MC	58.19	53.32	50.6	91	94

As the results indicate, the proposed method based on model-based sparse recovery can effectively recover the desired speech from the overlapping mixtures. The estimation of the mixing process in reverberant enclosures is highly erroneous within the scheme of DUET and LOST and it results in a poor speech separation performance. However, BSS-MSR incorporates the reverberant mixing model in separation of the speech components.

3.4 Conclusions

We proposed a novel speech separation framework for efficient recovery of convolutive speech mixtures in spectro-temporal domain using model-based compressive sensing theory. Contrary to the common practice in sparse component analysis, our formulation merges the two steps of mixing model estimation and source separation as joint localization-separation based on

spatio-spectral sparsity of overlapping speech signals. The method has been used for separation and recognition of a target speaker in the presence of competing interferences. The results show that the model-based sparse recovery formulation is very effective for distant multiparty speech recognition systems.

The underlying block-sparse structure that we used exists in any signal ensemble recorded by microphone array. It is also exhibited in virtual source images due to the acoustic multi-path effect and can be exploited in further analysis of the room acoustic. The success of the proposed framework motivates incorporation of the sparsity models in sparse component analysis techniques for reverberant conditions. These are the subjects which will be investigated through out Chapters 4-6.

4 Structured Sparse Representation

In the previous chapter, we have seen that there are three premises underlying our proposed framework of model-based sparse component analysis namely, structured sparse representation, compressive measurements and model-based sparse recovery. We now move on to elaborate on the structured sparsity models applicable for speech recovery. The focus of this chapter is on the first building block of our model-based sparse component analysis framework as depicted in Figure 4.2.

This chapter basically investigates theory and practice of characterizing structured sparsity of the acoustic signals in the form of a spatio-spectral scene. We start with a brief introduction into the theory behind this analysis from the compressive sensing perspective. There are two distinct aspects to the structured sparsity models associated either to the perception of sound or propagation which are studied in the following sections.

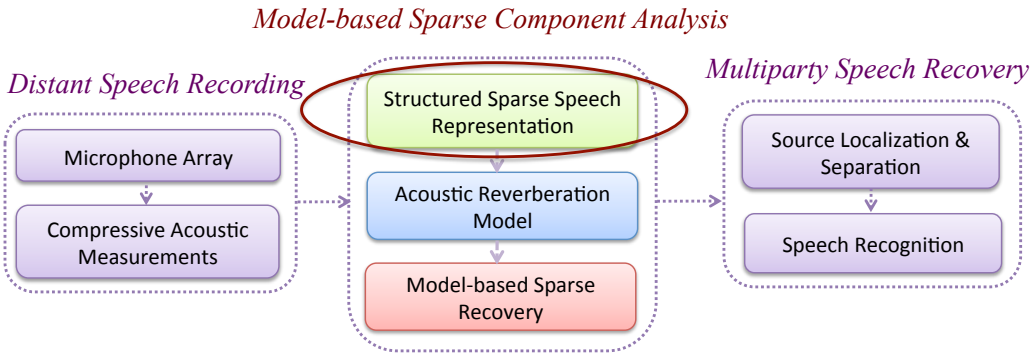

Figure 4.1 – The block diagram of the model-based sparse component analysis framework. The particular focus of this chapter is on *Structured Sparse Representation*.

4.1 Theory of Compressibility

Many natural and man-made signals are not strictly sparse, but they can be closely approximated as sparse if the absolute value of their representation coefficients $\mathscr{S}$ when sorted decay according to the power law hence called compressible signals [1]

$$|\mathscr{S}_{I(i)}| \leqslant \gamma i^{-1/r}, \quad i = 1, ..., N, \quad \text{and} \quad r < 1, \tag{4.1}$$

where I indexes the sorted coefficients in magnitude. Defining the $\mathscr{S}_N$ as the best N-term approximation of $\mathscr{S}$, which is obtained by keeping just the first N terms in $\mathscr{S}_{I(i)}$, the N-sparse approximation error when measured in the ℓ_p norm would have a power law decay exponent $\rho = \frac{1}{r} - \frac{1}{p}$ as N increases:

$$\|\mathscr{S} - \mathscr{S}_N\|_p \leqslant (r\rho)^{-1/p} \gamma N^{-\rho}, \tag{4.2}$$

Relying on the key characteristic of sparse representation, a G-dimensional data can be stably recovered from $M = O(N \log(G/N))$ dimensionality reducing while information preserving measurements through efficient optimization algorithms which search for the sparsest solution [Baraniuk et al., 2010]. The sparse signal recovery algorithms offer provable guarantees for the recovery error of the signals characterized in (4.1) as

$$\|\mathscr{S} - \hat{\mathscr{S}}_N\|_2 \leqslant C_1 \|\mathscr{S} - \mathscr{S}_N\|_2 + C_2 \frac{1}{\sqrt{N}} \|\mathscr{S} - \mathscr{S}_N\|_1; \tag{4.3}$$

with the constants C_1 and C_2 depending on the sparse recovery algorithm.

4.2 Acoustic Sparsity Models of Sound Propagation

There are several structured sparsity patterns exhibited while analyzing the microphone array recordings. In this thesis, we incorporate the sparsity models underlying the perception and propagation of sound. We first consider the acoustic sparsity models associated to the geometry of the problem and multipath propagation. These studies rely on the generative model of room impulse response function known as the *Image model* proposed by [Allen and Berkley, 1979] for modeling the acoustic of a shoe-box room. This model is applicable to general polyhedra assuming that the reflective surfaces are piecewise planar [Borish, 1984]. The Image model states that if there is a sound source on one side of the wall, then the sound field on the same side can be represented as a superposition of the original sound field and the one generated by a mirror image of the source with respect to the wall. Hence we can represent a path involving reflections by a straight line path connecting the listener to a corresponding virtual source.

Assume that the planar area of the room is discretized into a grid of uniform cells such that each of the speakers is expected to be located at one of the cells. To take the effect of the reflective

1. as we already discussed briefly through (2.9)

surfaces into account, we extend the discretization over the room boundaries and consider the actual source and its reflections with respect to surfaces (virtual sources). Hence, we obtain a free-space propagation model of multiple sources equivalent to the single source propagation in a reverberant enclosure. Applying this idea to a rectangular room, a regular lattice of virtual sources is obtained. Figure 4.2 illustrates our free-space model of a reverberant room. This model is applicable to an arbitrary polyhedra considering an irregular lattice to identify the location of the virtual sources. Since the map of actual-virtual sources is defined precisely, we assume that for a potential source located at each cell, the locations of the virtual sources are determined and used for characterization of the acoustic measurements. This structure is called as the acoustic multipath structure exhibited in the spatial representation of the data. We define the Image map as

$$\mathfrak{I} : g \rightarrow \Omega_g, \quad \forall g \in \{1, ..., G\}, \quad \Omega_g \in \{1, ..., \mathscr{G}\} \tag{4.4}$$

where G and $\mathscr{G}$ denote the number of the cells inside the enclosure and the extended grid respectively. In general Ω_g can include any order of reflections. The first order (G1) and the second order (G2) are illustrated in Figure 4.2. The first virtual source generation or G1 echos are the result of the images of the actual source with respect to the surrounding walls. The second virtual source generation or G2 echos are the result of the images of the images of the actual source. We can see that the location of the source images (virtual sources) are sparse in an extended grid. It also has a particular structure identified by the Image model. This mapping characterizes the acoustic multipath structure and determines the time support of the room impulse response function.

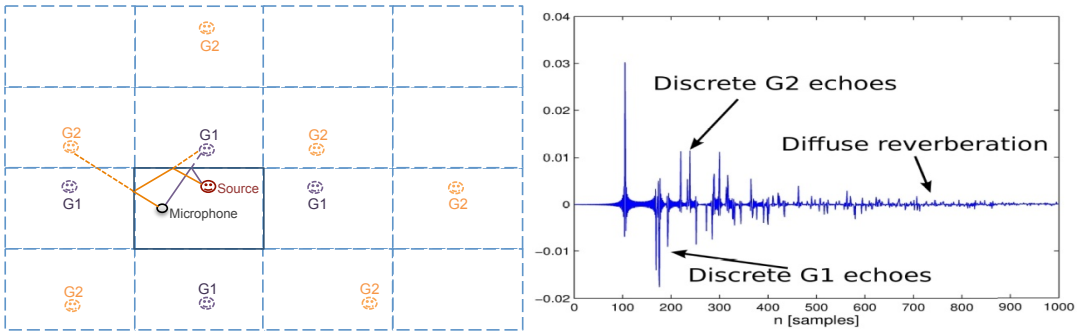

Figure 4.2 – The Image model is illustrated. Left: The actual source is the red face mirrored with respect to the enclosure boundaries and yields virtual sources depicted in dashed faces. The first order virtual sources (G1) are images of actual source with respect to the surrounding walls. The second order virtual sources (G2) are images of G1. Right: A sample room impulse response function is illustrated. The early part is consisted of the G1 and G2 echoes which correspond to the first order and second order generation of the virtual sources respectively. The late part causes the late reverberation which can be modeled as exponentially decaying noise. The right picture is published in [Dokmanic et al., 2011] where they provide the mathematical proof of a unique map between G1 and G2 echoes or correspondingly between the location of the first and second order virtual sources and the geometry of the enclosure.

To state it more precisely, we repeat our linear convolutive model of distant microphone recording as $X_m(f,\tau) = H(f,\mu_m,\nu_g)S_g(f,\tau)$, $m = 1,...,M$, where μ_m and ν_g denote the location of the microphone and source respectively. In case of free-space propagation, $H(f,\mu_m,\nu_g)$ is expressed by the frequency domain Green's function with the particular form of

$$H(f,\mu_m,\nu_g) = \frac{1}{\|\mu_m-\nu_g\|^{\alpha}} \exp(-jf\frac{\|\mu_m-\nu_g\|}{c}), \tag{4.5}$$

The Image model enables us to characterize the room impulse response function for a reverberant enclosure as superposition of free-space propagation of the virtual sources stated as [2]

$$H(f,\mu_m,\nu_g) = \sum_{r=1}^{R} \frac{\iota^r}{\|\mu_m-\nu_g^r\|^{\alpha}} \exp(-jf\frac{\|\mu_m-\nu_g^r\|}{c}), \tag{4.6}$$

where ν_g^r indicates the position of r^{th} virtual sources corresponding to the actual source located at cell g and reflective energy ratio of ι^r. To identify the underlying acoustic structures, we rewrite (2.2) in matrix notation as

$$\mathscr{X} = \mathscr{O}\mathscr{S} \tag{4.7}$$

where $\mathscr{O} \in \mathbb{R}^{M\times\mathscr{G}}$ corresponds to the free-space Green's function [3] for an extended grid with a region of deployment covering R actual and virtual sources. $\mathscr{X} = [X_1^T \cdots X_M^T]^T$ where $.^T$ stands for transpose and $\mathscr{S} = [S_1^T \cdots S_{\mathscr{G}}^T]^T$. The index set of the nonzero coefficients (support) of $\mathscr{S}$ corresponds to the location of the actual source and its images thus holds a particular structure indicated by the Image model as defined by mapping $\mathfrak{I}$ in (4.4). In addition to the structured sparsity model, $\mathscr{S}$ holds a low-rank structure as it is consisted of replications of the original signal impinging on the array from various locations. The rank of $\mathscr{S} \in \mathbb{C}^{\mathscr{G}\times F}$ matrix consisted of F instances of $\mathscr{S}$ is one if there is only one source active and its maximum value is equal to the number of sources N.

The generative model expressed in (4.6) is defined if (1) the geometry of the room is known so that the Image map $\mathfrak{I}$ and accordingly the location of virtual sources can be determined, and (2) the absorption ratios of the reflective surfaces are known. We elaborate on the procedures for estimation of the geometry of the enclosure and the absorption factors exploiting the structured sparsity of acoustic multipath in Chapter 5.

2. The room impulse response function expressed in (4.6) is equal to the acoustic projection defined in (3.3) associated to recording a source by a distant microphone.

3. $\mathscr{O}$ is equal to Φ defined in (3.7) for $R = 0$ and a large extension of the grid beyond the room boundaries. Through out the rest of the thesis, Φ denotes the projection matrix corresponding to reverberant propagation and $\mathscr{O}$ denotes the projection matrix corresponding to direct path propagation exceeding the room reflective boundaries.

4.3 Auditory Sparsity Models of Sound Perception

The sparsity models underlying spectrographic speech representation are identified from the perceptual mechanism that governs auditory organization in human listeners. In this section, we rely on Bregman's framework for auditory scene analysis to characterize these structures [Bregman, 1990, Wang and Brown, 2006]. His theory asserts that the acoustic signal in the form of spectro-temporal scene is decomposed into a collection of segments, which are subsequently grouped to form coherent streams. The mechanism underlying grouping can be either a bottom-up process relying on the intrinsic structure of environmental sound or a top-down process based on prior knowledge on schema patterns (e.g. syllabic). We investigate some of the intrinsic structures that we can exploit to recover the acoustic signal.

Figure 4.3 illustrates the distribution of multiple sources on a spatial grid. Denoting the signal attributed to the source located at cell i as S_i and concatenating the signals corresponding to each cell, the signal vector coming from all over the room can be formed as $\mathscr{S} = [S_1^T, ..., S_G^T]^T$. If we consider one instance of recordings from N speakers recorded by M microphones, $\mathscr{S}$ is a sparse vector with only N non-zero elements. The support of $\mathscr{S}$ corresponds to the N cells where the sources are located. If we consider F instances of recordings and assume that sources are immobile, each instance of the signal of a particular source implies sparsity in exactly the same manner as every other instances as they all correspond to the one particular cell where the source is located. Hence, a block or group sparsity structure is exhibited underlying the support of the sparse coefficients.

This model is in accordance with the proximity principle of structural grouping in auditory system [Wang and Brown, 2006]. According to the rule of proximity, the closer acoustic components are in frequency or time, the greater is the tendency to group them into the same stream. Imposing this proximity constraint on the structure of the elements in $\mathscr{S}$ results in a *block sparsity* model as illustrated in Figure 4.3. Considering a single frequency component of multiple frames, the block structure boils down to a *simultaneous sparsity* which is a special case of block-dependency when the projections (basis vectors) corresponding to a group of sparse components are the same. This structure enables capturing the sequential structure of the components as simultaneous sparsity is exhibited while processing multiple frames.

In addition to the proximity principle, the set of acoustic components that are harmonically related (i.e. have frequencies which are integer multiples of the same fundamental frequency) tend to be grouped together. The harmonic spectral structure is manifested in voiced speech as it typically comprises a small number of spectral peaks at harmonics of a fundamental frequency; at other frequencies the energy is typically low or negligible. We can therefore model the distribution of energy over frequencies as having *harmonic sparsity* model.

Another principle of structural grouping relies on *amplitude and frequency modulation*. Frequency components that exhibit the same modulation tend to be grouped together while analyzing the acoustic scene. This principle applies to both amplitude modulation and frequency modulation.

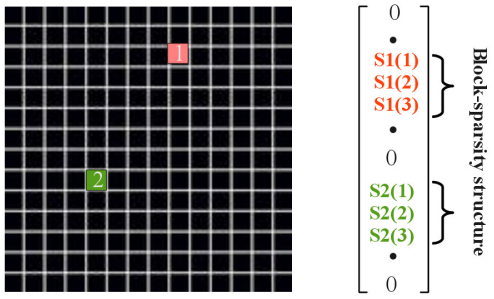

Figure 4.3 – The spatial sparsity of the speakers inside the room is illustrated through discretization of the planar area of the room into a grid of G cells. The sources occupy only two cells marked as 1 and 2. Hence, the spatial representation of the source signals generated inside the room is sparse.
Assuming that the sources are immobile, if we denote an arbitrary F (e.g. $F = 3$) instances of the signal attributed to the speaker at cell g as $S_g(l), l \in \{1, \cdots, F\}$ and concatenate the signals corresponding to each cell, the signal vector of the room can be formed as $\mathscr{S} = [S_1^T, ..., S_G^T]^T \in \mathbb{C}^{GF \times 1}$. We can see that support of $\mathscr{S}$ exhibits the block-sparsity structure as there are only two blocks of non-zero elements corresponding to the two speakers. The size of each block is the number of recording instances. The *proximity* principle of auditory scene analysis indicates that the adjacent components are grouped as belonging to one source.

The dependencies can be captured through linear prediction modeling by fitting an autoregressive (AR) model to find the optimal linear combination of a fixed-length history to predict the next sample. In addition to the AR model, the spectro-temporal modulation can be captured using 2D-Gabor prototypes. This framework offers an auditory-inspired yet easy to tune scheme of feature representation for the automatic speech recognition systems [Kleinschmidt, 2002a,b]. Incorporating the dependencies of the sparse coefficients to extract the information bearing components is the subject of our studies in Chapter 6.

As discussed below, the structured sparse representation introduced in this chapter entangles the *synthesis sparsity* of the geometry with the *analysis sparsity* of the signal processing domain. The canonical statement of the spatial representation exhibits sparsity models, a property referred to as *synthesis sparsity*. Furthermore, the measurements and the unknown acoustic signals are treated in Fourier domain hence, the FFT transformation is applied to yield *analysis sparsity*. Although the theoretical implications of the synthesis and analysis sparsity models are not within the scope of this thesis, we would like to bear in mind that when we talk about analysis sparsity, the transformation always plays a key role [Tan and Fevotte, 2005]. We chose to process the signals in Fourier domain as the phase information is preserved but, the appropriate selection or design of the sparsifying dictionary remains a question demanding further research in particular for the speech recognition task.

4.4 Empirical Insights on Spectrographic Speech Sparsity

To study the approximate sparsity or compressibility of the spectrographic speech, we investigate reconstruction bound of sparse approximation as stated in (4.3). The spectro-temporal representation of 50 minutes speech signal taken from Wall Street Journal database [Lincoln et al., 2005] are obtained by short Time Fourier Transform (STFT) with different window sizes. The speech signals are sampled at 16 kHz. The spectro-temporal coefficients are sorted in absolute value. The largest K coefficients are taken and the rest are forced to zero. Therefore, a K-sparse approximation is obtained. Figure 4.4 shows the decay of the error norms of K-sparse approximation of the clean speech for various frame sizes.

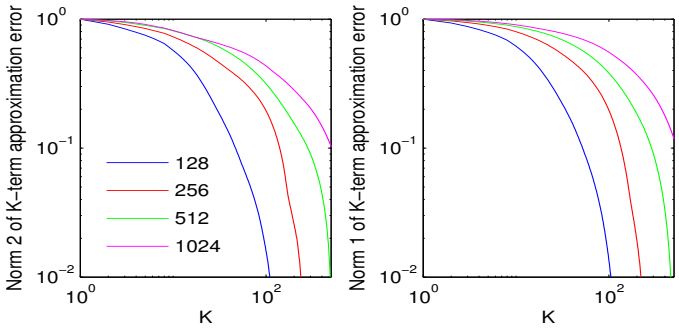

Figure 4.4 – Decay of the norm of error of K-sparse approximation of spectrographic speech for window sizes equal to 128, 256, 512 and 1024 samples.

Table 4.1 summarizes the percentage of the coefficients required for 10 dB reconstruction of spectro-temporal components as explicitly quantified in (4.3), where $C_1 = 0$ and $C_2 = 2$ are set as the theoretical lower-bounds. In addition to clean speech, we have studied reverberant speech in two acoustic conditions, with moderate reverberation and high reverberation with the corresponding 200 ms and 500 ms reverberation time. The concept of room reverberation time defines the time required for the multi-path energy to decay 60 dB from the direct path; hence denoted by RT_{60}.

Table 4.1 – Percentage of coefficients required for 10 dB reconstruction of speech spectro-temporal representation along with the 95% confidence interval.

STFT Window (ms)	64	128	256	512
Clean	14.7±0.54	14±1.48	15±3.84	16±9.9
RT_{60} = 200ms	22.5±0.56	21±1.48	20±3.73	19.8±9.6
RT_{60} = 700ms	27±0.56	25.19±1.52	23.44±3.92	22±10

As the results indicate, the sparsity is preserved in reverberant speech. The maximum sparsity is obtained for 64 ms analysis window for clean as well as moderately reverberant speech while in highly reverberant conditions, larger windows seem to give sparser coefficients. We have already seen in Section 2.4, with less than 30% of the spectro-temporal coefficients, it is possible to perform word recognition with more than 90% accuracy. This observation confirms the

hypothesis that the information bearing components for automatic speech recognition (ASR) are sparse; hence they can be applied in the framework of SCA for multi-party ASR. The discrepancy between speech sparsity in terms of speech reconstruction and recognition motivates ASR-specific sparse representation. These results further motivate exploiting the underlying structure of the sparse coefficients to reduce the number of required measurements and to improve the recovery performance. We study this subject in Chapter 6.

We perform further analysis of the linear dependencies among the spectro-temporal coefficients using the autoregressive model. In this study, the frequency band is divided into blocks of size 16 and accordingly 16 temporal frames are concatenated together; hence, a 16×16 matrix of spectro-temporal coefficients is obtained. The oblique (joint spectro-temporal) dependencies are then modeled by learning the 16-order linear prediction filter coefficients on the diagonal elements of this matrix. Figure 4.5 demonstrates the 2-D average AR coefficients obtained for 10 min analysis of speech utterances. The excess (very small) lines on the demonstrated curve show the variance around the average points. We see that the variances are very small. This experiment highlights a fact that strong correlation exists among *oblique* spectro-temporal components. Similar study can be conducted on spectral dependencies or temporal dependencies of particular frequency bands. We elaborate on this subject in Section 6.4 and incorporate these correlations in recovery of spectro-temporal components.

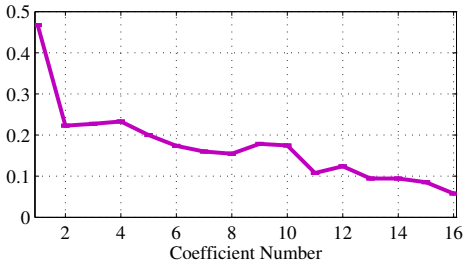

Figure 4.5 – 16-order average spectro-temporal AR coefficients estimated for 10min speech. The variance of the coefficients over the entire frames is shown as the (very small) cross lines emerging from the point estimates

4.5 Conclusions

We studied the structured sparsity models underlying the spectral and spatial representation of speech signal in a reverberant enclosure. The spectral structures are inspired from the auditory principles that govern the perception of sound. These structures are in accordance with the structural grouping mechanism of the auditory scene analysis techniques. We considered the proximity, harmonicity and modulation principles as the particular sparsity models incorporated in our studies. In addition to the spectral structures, spatial structures are specified. The spatial structures are exhibited due to the geometry of the problem and multipath propagation. The acoustic multipath structure is identified using the Image model given the location of the source images. In the following Chapter 5, we elaborate on characterizing the compressive acoustic measurements.

5 Compressive Acoustic Measurements

In Chapter 3, we briefly reviewed how characterizing the acoustic projections amounts to identifying the geometry of the enclosure and the absorption factors of the reflective surfaces. The experiments conducted there assumed that we know the geometry and the absorption factors. This chapter shows how to estimate these parameters.

In Chapter 4, the structured sparsity models underlying multipath propagation were elaborated. This chapter deals with the problem of characterizing the *compressive* acoustic measurements associated to the projection of the high-dimensional acoustic scene data to the low-dimensional manifold of microphone array. Figure 5.1 illustrates the goal of this chapter within the broad picture of our framework. We exploit the structured sparsity models and propose some algorithmic approaches to the problem of estimating the geometry of the room and the absorption coefficients from recordings of unknown concurrent speech sources.

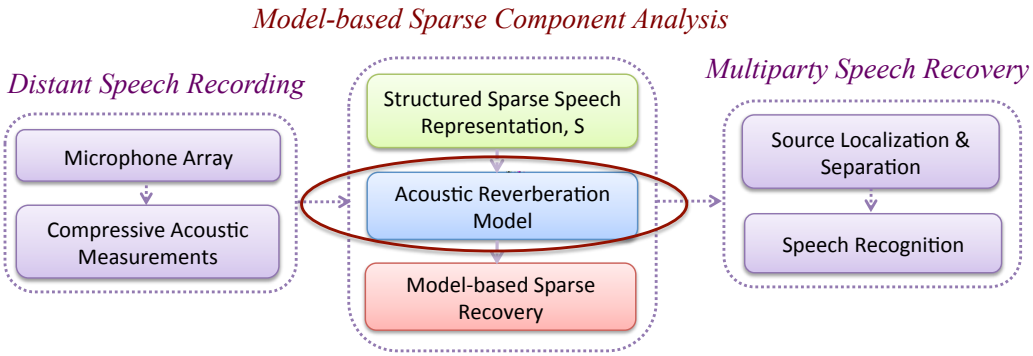

Figure 5.1 – The block diagram of the model-based sparse component analysis framework. The particular focus of this chapter is on characterizing the *acoustic reverberation model*.

5.1 Estimation of the Room Geometry

The projection expressed in (3.3) corresponds to characterization of the forward model of the room acoustic channel as

$$H(f,\mu_m,\nu_g)=\sum_{r=1}^{R}\frac{\iota^r}{\|\mu_m-\nu_g^r\|}\exp(jf\frac{\|\mu_m-\nu_g^r\|}{c}). \tag{5.1}$$

$H(f,\mu_m,\nu_g)$ indicates the room impulse response function between the microphone located at μ_m and a source located at ν_g. Hence, identifying the locations of the R images of the source corresponds to identifying the support of the room impulse response function in time domain. According to the Image model of multipath propagation (as we have already seen in Section 4.3), if the geometry of the enclosure is known, it is possible to identify the source images up to any arbitrary order [Allen and Berkley, 1979].

Recent studies have provided the mathematical proof that the impulse response function is a unique signature of the room and the geometry can be reconstructed given that up to second order of reflections are known [Dokmanic et al., 2011]. Relying on this fact, we propose to localize the source images using the sparse recovery algorithm with a free space measurement model, i.e., $R=0$, while the deployment of the grid captures the location of early reflections. The time support of the acoustic channel, $\{\nu_r|1<r<R\}$ corresponds to the cells where the recovered energy of the signal is maximized. We consider the localized source signals in a close proximity to the microphone array within a pre-specified threshold as the actual sources generating the signals $S_n, n=\{1,...,N\}$. The localized images are sorted up to the order of $D(D+1)/2$ where D indicates the number of reflective surfaces, according to the cosine angle between the estimated signals and the source signal and regarded as the images associated to the n^{th} source. The cosine angle is the appropriate distance measure to cluster the components which are geometrically aligned, i.e., *images* of the same source. The bound of $D(D+1)/2$ guarantees a unique map to the geometry of the enclosure as proved in [Dokmanic et al., 2011]. Given the location of the source images, we estimate the room geometry by crude search to identify the dimensions which generate the least-squares approximation of the location of source images from the location of the actual sources [Asaei et al., 2011b, 2012c]. Algorithm 1 summarizes the procedure for room geometry estimation.

The approach that we presented in this section can estimate the room geometry from recordings of a single source or multiple unknown sources. The experiments on real data recordings are presented in Section 5.3.3. Applying the Image model to a rectangular room, a regular lattice of virtual sources is obtained (as depicted in Figure 4.2). Hence, the Image model of multipath propagation insinuates temporal sparsity of the early part of the room impulse response function with a particular structure identified by the lattice of the virtual sources. We refer to this property as the acoustic structured sparsity and exploit it to address the problem of estimating the absorption coefficients.

Algorithm 1 Room geometry estimation

(i) Run sparse source localization algorithm with a free-space measurement model.

(ii) Run k-means clustering using cosine angle as the distance metric.

- ▷ Select the centroid of the clusters as the nearest (actual) sources to the array center.
- ▷ Measure the cosine angle between components of virtual and actual sources.
- ▷ Keep the closest $D(D+1)/2$ sources as the cluster members.

(iii) Find the room geometry by identifying the dimensions which yield the best approximation of the location of source images in least-squares sense.

5.2 Estimation of the Absorption Coefficients

In Section 5.2.1, we propose an approach for estimating the absorption coefficients if there are only single speaker utterances. In Section 5.2.2, we elaborate on a novel model of the room reverberation which enables accurate estimation of the absorption ratios from a plurality of speech sources.

5.2.1 Single-Source Absorption Coefficient Estimation

We consider the linear convolutive model of the reverberant enclosure and denote the time-domain acoustic channel between the source and microphone i as $h_i(l)$. Hence, the signal of the microphone i, $x_i(l)$ is a filtered version of source signal $s(l)$ as $h_i(l) \circledast s(l)$. It is straightforward to see that

$$x_i(l) \circledast h_j(l) = x_j(l) \circledast h_i(l); \tag{5.2}$$

Considering an $(L+1)$-tap acoustic filter, for $l = L, ..., \mathscr{L}$, where $\mathscr{L}$ is the length of the recorded signal, (5.2) becomes:

$$\begin{aligned} &[\chi_i(L) - \chi_j(L)] \begin{bmatrix} h_j \\ h_i \end{bmatrix} = 0; \quad h := [h(L), ..., h(0)]^T, \\ &\chi_i(L) = \begin{bmatrix} x_i(L) & x_i(L+1) & \dots & x_i(2L) \\ x_i(L+1) & x_i(L+2) & \dots & x_i(2L+1) \\ \vdots & \vdots & \ddots & \vdots \\ x_i(\mathscr{L}-L) & x_i(\mathscr{L}-L+1) & \dots & x_i(\mathscr{L}) \end{bmatrix} \end{aligned} \tag{5.3}$$

Equation (5.3) forms the basic idea for blind channel identification by least squares optimization [Xu et al., 1995]. Relying on the structured sparsity as indicated by the Image model of multipath effect, we propose the optimization algorithm constrained on the structured sparsity to capture the main reflections characterized by the Image model [Asaei et al., 2011b].

Despite the existence of various reflective objects inside the room, in many scenarios the majority

of the reverberant energy is due to the enclosure walls as the multi-path signal energy is a function of the reflective areas and many objects can be considered as acoustically transparent [Ba et al., 2010]. In practice, big flat objects like a meeting table change the boundary conditions and the resulting sound field in the enclosure. We consider this empirical insights for the purpose of the desired experiments described in Section 5.3. In addition, we verified the effectiveness of the structured sparsity constraint for identification of the real acoustic impulse responses from noisy reverberant data generated by the impulse responses available in the Aachen impulse response (AIR) database [AIR, 2010]. In Section 5.3.4, an example is demonstrated.

Given the room geometry and the source location, the support of the highest energy components of Room Impulse Response (RIR) is determined by the Image model and denoted by Ω_d, which refers to the direct path component calculated precisely as η and Ω_g which refers to the support of the reflections. We define

$$\mathscr{H} := [h_j^T \quad h_i^T]^T, \qquad \Pi := [\chi_i(L) \quad -\chi_j(L)]; \tag{5.4}$$

The structured sparse acoustic filter is obtained by the following optimization

$$\begin{aligned} \hat{\mathscr{H}} &= \arg\min \|\mathscr{H}\|_1 \\ &\text{s.t.} \quad \|\Pi\mathscr{H}\|_2 \leqslant \epsilon, \quad \mathscr{H}_{\Omega_d} = \eta, \quad \mathscr{H}_{\Omega_g} > 0 \end{aligned} \tag{5.5}$$

where ϵ is a small threshold tuned according to the noise level and the subscript $\Omega_\cdot$ determines the indices of the vector representation. The inequality condition stated in $\mathscr{H}_{\Omega_g} > 0$ is point-wise inequality. The taps of RIR filter can become zero or negative; this inequality imposes a positive support on the components corresponding to the source images thus enables estimation of the absorption coefficients to identify the parameters of our model stated in (5.1). The absorption coefficients are estimated by least squares fitting and used to characterize the acoustic channels of all cell positions in order to identify the microphone array measurement matrix.

5.2.2 Multi-Source Absorption Coefficient Estimation

This section elaborates on a novel formulation of the reverberant recordings which exploits structured sparsity indicated by the Image model and the spatio-spectral sparsity of multiparty recordings for joint estimation of the absorption coefficients and recovery of the source locations [Asaei et al., 2012c]. The approaches proposed earlier in the literature, rely on estimation of the acoustic channels from the intervals of single speaker [Huang et al., 2005]. Generalizing the algorithm proposed in Section 5.2.1, we can estimate the frequency-dependent absorption factors in a multi-source environment.

Factorized Formulation of the Reverberant Recordings

We formulate the *reverberation model* factorized into permutation (corresponding to the source images) and attenuation (corresponding to the absorption factors) of the sources in an unbounded space. We assume that the G-cells grid of the room containing N sources is expanded into $\mathcal{G}$-cells free space grid where the actual-virtual sources are active. If each of the sources have R images, $N(R+1)$ actual-virtual sources are active [1]. Given the geometry of the room, the Image model maps the position index $g \in \{1,\dots,G\}$ of each source to a group $\Omega_g \subset \{1,\dots,\mathcal{G}\}$ containing the location indices of this source and its images (the corresponding virtual sources) in $\mathcal{G}$-points. Consequently, a free-space propagation model can be considered between $\mathcal{G}$ actual-virtual source locations and the positions of M microphones. Hence, the forward model between sources and the microphone recordings can be concisely stated as follows:

$$X = OPS \tag{5.6}$$

This model holds for each particular independent frequency f of the speech spectrum. We thus discard the frequency dependency in our mathematical formulation for the sake of brevity. Given $X \in \mathbb{C}^{M\times\mathcal{T}}$, the *observation* matrix of $\mathcal{T}$ frames consisted of spectro-temporal representation of M microphones at a particular frequency band, we decompose the microphone recordings into the following terms:

- $S \in \mathbb{C}^{G\times\mathcal{T}}$ is the *source* matrix whose rows contain $\mathcal{T}$ frames of the spectro-temporal representation of the *actual* sources located in G positions inside the room. Given a fine discretization of the room such that each source occupies an exclusive cell, only $N \ll G$ cells are occupied with active sources and contain nonzero elements and the *support* of S representing the position of those N active sources is sparse. In other words, the *spatial sparsity* indicates S to be a row-sparse matrix with a support corresponding to the position of the actual sources.
- $P \in \mathbb{R}_+^{\mathcal{G}\times G}$ is the *permutation* matrix such that its i^{th} column contains the absorption factors of $\mathcal{G}$ points on the grid of actual-virtual sources with respect to the reflection of the i^{th} actual source. Since the Image model characterizes the source groups, each column $P_{.,i}$ is consequently supported only on the corresponding group Ω_i i.e., $\forall i \in \{1\dots,G\}$, $\forall j\ \Omega_i, P_{j,i} = 0$.
- $O \in \mathbb{C}^{M\times\mathcal{G}}$ is the *free-space Green's function* matrix such that each $O_{j,i}$ component indicates the sound propagation coefficients, i.e. the attenuation factors and the phase shift due to the direct path propagation of the sound source located at cell i (on a $\mathcal{G}$-point grid of actual-virtual sources) and recorded at the j^{th} microphone. Given the $\mathcal{G}$-cell discretization, O is computed from the propagation formula stated in (3.3) and it is equal to Φ when $R = 0$.

1. This formulation can be generalized to three-dimensional acoustic modeling by considering the parallel grids at given heights

Source Localization and Absorption Coefficient Estimation

Relying on spatio-spectral sparsity of multiple competing sources, the covariance matrix of the reverberant recordings exhibits structured sparsity determined by the Image model. We exploit this structured sparsity to identify the location of the active sources and their corresponding absorption coefficients consisting the columns of P. Given the model of the microphone recordings stated in (5.6), the covariance matrix of the observations is

$$\begin{aligned} \mathsf{C} = \mathsf{X}\mathsf{X}^{\mathrm{H}} &= \mathsf{O}\Sigma\mathsf{O}^{\mathrm{H}} \\ &= \sum_{i=1}^{G} \mathsf{O}_{.,\Omega_i}\Sigma_{\Omega_i,\Omega_i}\mathsf{O}^{\mathrm{H}}_{.,\Omega_i}, \quad \Sigma = \mathsf{P}\mathsf{S}\mathsf{S}^{\mathrm{H}}\mathsf{P}^{\mathrm{H}} \end{aligned} \tag{5.7}$$

where $.^{\mathrm{H}}$ denotes conjugate transpose and the subscript $.,\Omega_i$ denotes all the rows with particular column indices determined by Ω_i. Note that the spatio-spectral sparsity of concurrent speech sources implies that $\mathsf{S}\mathsf{S}^{\mathrm{H}}$ is a diagonal matrix whose diagonal elements specifies the energy of the individual sources - Section 5.3.2 provides some empirical insights on the properties of the covariance matrix. The second equation follows because of the structure of the permutation-attenuation matrix P which indicates that Σ is supported only on the set $\bigcup_i \Omega_i \times \Omega_i$ i.e.,

$$\begin{aligned} \Sigma_{j,i} = 0 \qquad \forall (j,i) \bigcup_{i=1}^{G} \Omega_i \times \Omega_i, \\ \Sigma_{\Omega_i,\Omega_i} = \|\mathsf{S}_{i,.}\|_2^2 \mathsf{P}_{\Omega_i,.}\mathsf{P}^{\mathrm{H}}_{\Omega_i,.}, \end{aligned} \tag{5.8}$$

where $\mathsf{S}_{i,.}$ denotes the i^{th} row and $\|\mathsf{S}_{i,.}\|_2 = \sqrt[2]{\mathsf{S}_{i,.}\mathsf{S}^{\mathrm{H}}_{i,.}}$. As we can see, recovering the diagonal elements of $\Sigma_{\Omega_i,\Omega_i}$ is sufficient to identify the energy of the corresponding source i and the absorption coefficients $\mathsf{P}_{\Omega_i,.}$. We thus focus on recovering these sub-matrices for all $i \in \{1,\ldots,G\}$ from the observation covariance matrix C. Using the property of the Kronecker product, we can rewrite (5.7) as

$$\mathsf{C}_{vec} = \underbrace{\begin{bmatrix} \mathsf{B}(1) & \mathsf{B}(2) & \ldots & \mathsf{B}(G) \end{bmatrix}}_{\mathscr{B}} \underbrace{\begin{bmatrix} v(1) \\ v(2) \\ \vdots \\ v(G) \end{bmatrix}}_{\mathscr{V}} \tag{5.9}$$

$$\begin{aligned} &\forall i \in \{1\ldots,G\}: \\ &\qquad v(i) \triangleq (\Sigma_{\Omega_i,\Omega_i})_{vec}, \\ &\qquad \mathsf{B}(i) \triangleq \overline{\mathsf{O}}_{.,\Omega_i} \otimes \mathsf{O}_{.,\Omega_i}. \end{aligned}$$

where $\otimes$ denotes the Kronecker product between two matrices and $\overline{\mathsf{O}}_{.,\Omega_i}$ is the *element-wise* conjugate of $\mathsf{O}_{.,\Omega_i}$. In a typical problem setup, very few microphones are used for recording,

i.e. $M \ll \mathscr{G} < \sum_{l=1}^{G} |\Omega_i|$ where operator $|.|$ indicates the cardinality of the set; thus recovering $\Sigma_{\Omega_i,\Omega_i}$ requires solving an underdetermined system of linear equations and therefore, in general (5.7) admits infinitely many solutions and recovery is not feasible.

To circumvent the ill-posedness of the inverse problem, we exploit yet another kind of *block-sparsity* structure that is exhibited in our formulation of the reverberant multi-party recordings. The block sparsity of the actual-virtual sources implies that only $N \ll G$ groups of $\nu(i)$s (or correspondingly $\Sigma_{\Omega_i,\Omega_i}$) contain nonzero elements, and thus, identifying those groups equivalently determines the positions of the active sources S. In addition, by recovering the corresponding elements of $\mathscr{V}$ and then normalizing them by the sources energies, we can identify the absorption coefficients (i.e., the columns of P) which correspond to the attenuation for each source due to the multipath reflections.

We simplify the notation by using $\Sigma^i \triangleq \Sigma_{\Omega_i,\Omega_i} \in \mathbb{R}^{|\Omega_i|\times|\Omega_i|}$. Our block-sparse recovery approach can then be formulated by the following convex minimization problem:

$$\begin{aligned} \underset{\Sigma^1,\ldots,\Sigma^G}{\arg\min} \quad & \sum_{i=1}^{G} \|\Sigma^i_{vec}\|_{L_2} \\ \text{subject to} \quad & \|C_{vec} - \mathscr{B}\mathscr{V}\|_{L_2} \leqslant \varepsilon \\ & (\,\mathscr{V} = \left[(\Sigma^i_{vec})^T,,\ldots,(\Sigma^i_{vec})^T\right]^T) \\ & \Sigma^i = (\Sigma^i)^H \quad \forall i \in \{1,\ldots,G\} \\ & \Sigma^i_{l,j} \geqslant 0 \qquad \forall l,j,i \end{aligned} \tag{5.10}$$

We recall that minimizing the sum of the L_2 norms of a group of vectors induces the block-sparsity structure in the solution so that, only few subsets of vectors in the group (i.e. few Σ^is) contain nonzero elements. Indeed, if Σ^is have the same size (i.e. $|\Omega_1| = |\Omega_2| = \ldots = |\Omega_G|$) the objective function of (5.10) becomes equivalent to the L_1L_2 norm [2] of a matrix whose rows are populated by $(\Sigma^i_{vec})^T$, which as mentioned earlier is a popular convex approach for block (group) sparse approximation. We solve (5.10) by using the iterative proximal splitting algorithm [Combettesand and Pesquet, 2011].

We can see from (5.8) that $\Sigma^i = \|S_{i,.}\|_2^2 P_{\Omega_i,.} P^H_{\Omega_i,.}$ is a rank one matrix so we can replace the objective of (5.10) by $\sum_{i=1}^{G} \|\Sigma^i\|_*$ to perform *low-rank recovery*. The matrix nuclear norm is defines as $\|\Sigma\|_* = \sum_j \sigma_j$, where σ_j denotes the singular values (henceforth assuming singular values sorted in descending order). In our future studies, we will compare the results with joint sparse recovery and explore the advantages of incorporating the low-rank structures for room acoustic modeling and dereverberation [Asaei et al., 2013a, Golbabaee and Vandergheynst].

To summarize, we obtain the location of the sources and their images which also corresponds

2. The $\|.\|_{L_1L_2}$ mixed-norm of a matrix is defined as the sum of the L_2 norms of its rows as defined in (6.18)

to the support of the room impulse response function for multiple sources. The components of $\Sigma_{\Omega_i,\Omega_i}$ normalized by the energy of the sources corresponds to the attenuation factors. We entangle the room geometry with the absorption coefficients to characterize the acoustic projections *for any order of desired* R, as expressed in (5.1).

5.3 Experimental Analysis

The experiments are conducted in the framework of Multichannel Overlapping Numbers Corpus (MONC). Additionally, we present some evaluations on synthetic data to demonstrate the performance bounds of the proposed theory in various scenarios.

5.3.1 Data Recordings Set-up

The MONC database is recorded at Idiap Research Institute [Mccowan, 2003]. The database is acquired by playback of utterances from the Numbers Corpus release 1.0, prepared by the Center for Spoken Language Understanding at the Oregon Graduate Institute. The recordings were made in a 8.2m $\times$ 3.6m $\times$ 2.4m rectangular room containing a centrally located 4.8m $\times$ 1.2m rectangular table. The positioning of loudspeakers was designed to simulate the presence of 3 competing speakers seated around a circular meeting room table of diameter 1.2m. The loudspeakers were placed at 90°$\pm$ spacings at an elevation of 35cm (distance from table surface to center of main speaker element). An eight-element, 20cm diameter, circular microphone array placed in the center of the table recorded the mixtures. The recording scenario is illustrated in Figure 5.2. One hour of speech signals recorded at 8 kHz sampling frequency. The average signal to noise ratio (SNR) of the recordings is 10dB.

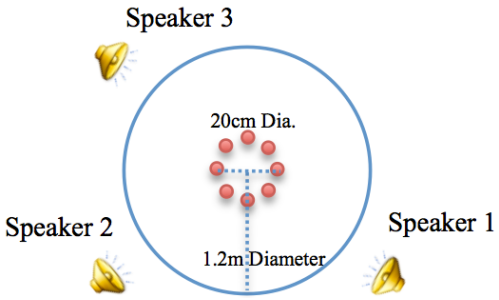

Figure 5.2 – Loudspeaker and microphone placement used for recording MONC corpus [Mccowan, 2003].

5.3.2 Orthogonality of Spectrographic Speech

We carried out experiments to investigate the orthogonality of multiple speech sources in Fourier domain. In this experiment, 3 speech signals of 9s each are analyzed in frames of size 128ms (fft-size = 1024) with 50% overlap; thus we obtain 3 matrices of 512 by 140 corresponding to the

STFT of each source [3]. The orthogonality is measured for each frequency band independently. We construct the matrix $\mathsf{S}_{3\times 140}$ where each row corresponds to each source and has the frequency components of a particular band along 140 frames. In case of perfectly orthogonal sources, $C = \mathsf{S}\mathsf{S}^{\mathrm{H}}$ is identity and the energy of the diagonal of the matrix is equal to the matrix Frobenius norm. Figure 5.3-right-hand-side illustrates the diagonal-L_2-norm/matrix-Frobenius-norm. In addition, we performed some experiments by pointwise multiplication of the STFTs of two utterances and plot the histograms of the resulted values. Figure 5.3-left-hand-side illustrates the obtained histogram.

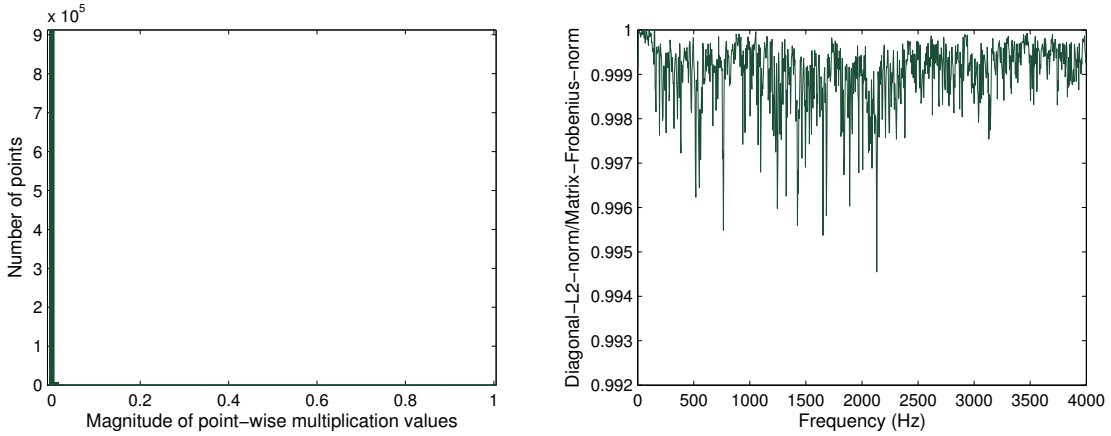

Figure 5.3 – Orthogonality of multiple speech utterances in spectro-temporal domain: Left-hand-side illustrates the energy histogram of the component-wise multiplication of speech utterances and Right-hand-side illustrates the diagonal-L_2-norm/matrix-Frobenius-norm of the covariance matrix constructed per frequency.

We can see that the distribution mass of the energy of the elements of the correlation matrix is localized around 0. This phenomenon indicates that the majority of the high energy components in spectro-temporal domain are non-overlapping or disjoint. It verifies the low-rank clustering algorithm explained in Section 5.2 to identify the images of the individual sources and confirms the hypothesis underlying the algorithm explained in Section 5.2.2 for source localization and absorption coefficient estimation.

5.3.3 Room Geometry Estimation

The first step to characterize the room acoustic is to estimate the room geometry. We accomplish this step through localization of the images of multiple sources on a large extended grid using the sparse recovery framework with a free space forward model [4]. The location of the source images corresponds to the time support of the room impulse response function. The energies of the recovered signals are sorted and truncated to the order of $\mathrm{D}(\mathrm{D}+1)/2$ where D denotes the

3. The total number of samples is $9*8000 = 72000$ and the number of samples per frame is $128*8/2 = 512$; the division by two follows from the 50% overlapping. Hence the total number of frames is $72000/512 \cong 140$.

4. It corresponds to $\mathrm{R} = 0$ in (5.1)

number of reflective surfaces and it is equal to 4 in our study to cover the support of the early reflections of the walls and guarantee the uniqueness [5] of the solution.

The estimated support of the room impulse response function is then used for estimation of the room rectangular geometry by generating the room impulse responses for various room geometries and identifying the best fit to the estimated support in least-squares sense by crude search. The crude search has a computational cost depending on the number of reflective surfaces D. To overcome the computational burden, some heuristic approaches can be employed. We start from an initial guess about the boundaries as the half-way wall between the source and its earliest image. The estimates are then enhanced around the initial state through least square regression of all source images. In our synthetic and real data evaluations, the microphone array was located at the center of the room thus the search space is reduced by half [Asaei et al., 2012c].

The planar area of the room is divided into cells with 25cm spacing [6]. The distance threshold to identify the actual sources is selected as 1m. To achieve a higher estimation, we restrict our discretized gird to the orthogonal subspaces corresponding to the orthogonal walls. We can estimate the geometry of the room up to 50cm per wall position error from the recordings of 3 sources in a close proximity to the microphone array as depicted in Figure 5.2.

5.3.4 Room Impulse Response Estimation

The second step to characterize the room acoustic is to estimate the absorption coefficients of the reflective surfaces. We accomplish this step through estimation of the Room Impulse Response (RIR) function by implementing the technique explained in Section 5.2.1. We used the CVX software package [Grant and Boyd] for optimization formulated in (5.5) while ϵ is chosen 0.1. The data was provided by concatenating 20 stationary single speaker speech utterances from MONC. To establish an oracle performance evaluation of the proposed method, a RIR function is synthesized for a scenario similar to MONC. Fig. 5.4 shows the effectiveness of the room impulse response estimation using the structured sparsity constraints and the alternative least-squared optimization proposed in [Xu et al., 1995] as the basic idea behind the method explained in 5.2.1. The rest of the experiments are conducted to real MONC recordings.

The super-resolution source localization is performed based on the energy recovered from each cell using sparse recovery framework while the forward model corresponds to the direct path propagation and the support of the RIR function was determined considering a 6-sided model of an enclosure with the known geometry. We assumed that the reflections of the carpet floor are trapped under the table; hence, the meeting table was considered as the floor in our Image model. The room reverberation time is measured about 100 ms from the energy decay curve of the estimated RIR and the reflection coefficients are estimated as 0.1 for the walls as well as the ceiling and 0.6 for the meeting table. Our estimation matches the empirical Sabin-Franklin's

5. The unique map between the geometry of the room and the first and second order of reflections is mathematically proved in [Dokmanic et al., 2011]

6. There is no algorithmic impediment for a generalized three-dimensional grid construction

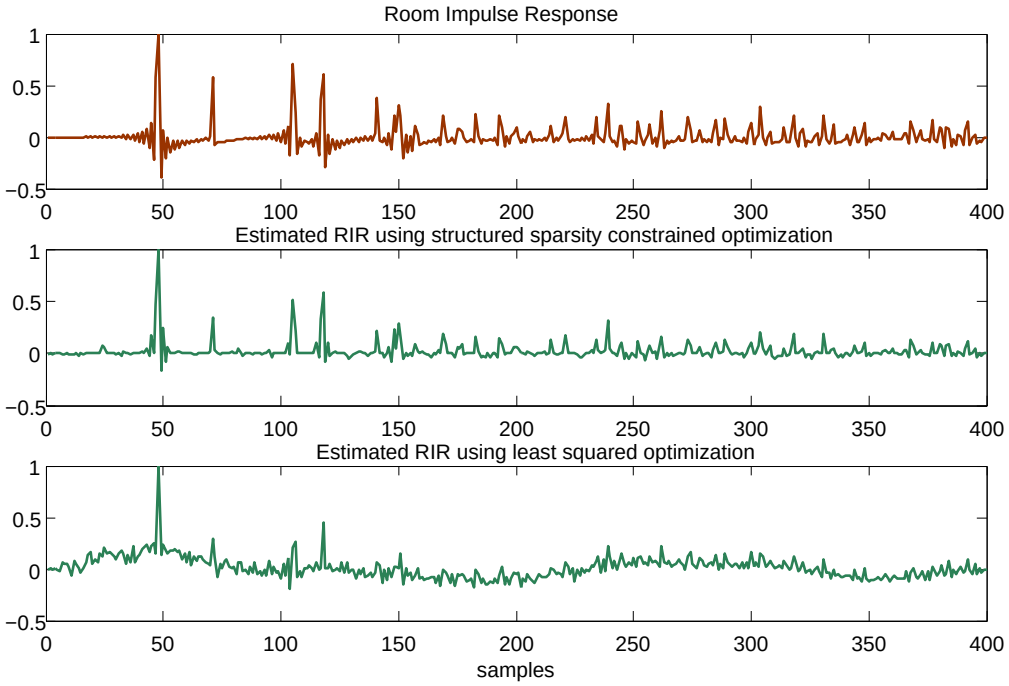

Figure 5.4 – Room impulse response (RIR) estimation from noisy measurements:(Top) simulated RIR, (Middle) estimated RIR from (5.5), and (Bottom) least-squared (no sparsity constraint) estimated RIR (it is based on L_2-minimization of $\mathscr{H}$ stated in (5.5) [Xu et al., 1995]. The normalized distances between the actual RIR and eatimated RIR using structured sparse recovery and least-squared optimization are 0.4 and 0.9 respectively.

formula [Habets, 2007]:

$$\mathrm{RT}_{60} = \frac{24\ln(10)V}{c\sum_{i=1}^{6} W_i(1-\iota_i^2)}, \tag{5.11}$$

where V denotes the enclosure volume in m^3 of the room, ι_i and W_i correspond to the reflection coefficient and area of the i^{th} wall respectively. The MONC recordings were made in a $8.2\mathrm{m}\times3.6\mathrm{m}\times2.4\mathrm{m}$ rectangular room (i.e., $V = 70.848$) containing a centrally located $4.8\mathrm{m}\times1.2\mathrm{m}$ rectangular table. If we calculate (5.11) using these numbers and consider ι_i of all the walls equal to 0.1 and the meeting table (virtual floor) equal to 0.6, RT_{60} is obtained as 128 ms.

Although our method is blind, we verified our estimates of the absorption coefficients by estimating the impulse response and the corresponding reflection coefficients through a least mean squares (LMS) adaptive filtering technique using the clean speech provided from the original Numbers corpus which was played back in MONC scenario to create an overlapping reverberant database.

5.3.5 Multi-party Acoustic Modeling

We perform some initial evaluations of the algorithm explained in Section 5.2.2 using synthetic data in various noisy and reverberant conditions. The results of these experiments demonstrate the empirical performance bounds for absorption coefficient estimation and signal recovery using block sparse recovery algorithm.

We consider the following scenarios: (1) 8-channel circular microphone array positioned in the middle of the room, (2) 12-channel microphone array: two sets of 6-channel circular arrays, each located 1 m far apart with respect to the center of the room, (3) 16-channel microphone array: two sets of 8-channel circular arrays, each located 1 m far apart with respect to the center of the room. We considered about 3 cm displacement of the microphones. The reverberant channel is simulated Habets [2007] for a four-sided $3 \times 4\,\mathrm{m}^2$ enclosure. The area of the room is discretized into a grid of uniform cells of size $0.5 \times 0.5\,\mathrm{m}^2$ adding up to 40 cells inside the room. The reflection coefficients of the walls are selected as 0.4, 0.6, 0.8 and 0.9. Evaluations are carried out using $\mathrm{N} = \{1,2,3\}$ omni-directional sources distributed arbitrarily in the room with the following characteristics (a) Spectrum of orthogonal random broad-band sources at 52 auditory-centered frequencies and (b) Spectrum of independent speech sources at the frequency-bands which contain 80% of the total energy. Fig. 5.4 demonstrates the estimated room acoustic impulse response from recordings of two concurrent speech sources recorded by 8-channel microphone array using our structured sparse acoustic modeling technique. Alternatively, the blind channel impulse response estimation referred to as the Cross-Relation technique Xu et al. [1995] is used to recover the channel from recording of a single source; the normalized distance quantified as $\|\mathrm{H} - \hat{\mathrm{H}}\|_2 / \|\mathrm{H}\|_2$ is calculated as 0.4 and 0.9 respectively. To our knowledge, the state-of-the-art techniques can not recover the acoustic channel from recordings of multiple unknown speech sources. The results of source localization (SL), absorption coefficients estimation (AC) and signal recovery (SR) are illustrated in Figs. (5.5) and (5.6).

The results of Fig. 5.5 demonstrates the performance bound of the algorithm presented in Section 5.2.2. We can see that in noiseless condition, SL is achieved almost 100% correct per frequency band for any number of (one to three) sources. However, estimates of the absorption coefficients are not exact; the root mean square error (RMSE) is proportional to the number of microphones used to collect the data. The best estimate is achieved when 16 microphones are used; increasing the number of concurrent sources results in about 5% error increase in estimation of AC. Similarly, estimations of the source coefficients (SR) is obtained up to 4% error if there is only one source active. Increasing the number of sources reduces the accuracy about 5% per source. Contrasting these results with the bar charts obtained for speech sources does not show any degradation in 16-microphones scenario. In more under-sampled regimes, the degradation is less than 5% in SL and upper bounded by 10% in AC and SR.

If we consider adding white Gaussian noise to the recorded signals, the errors in AC estimation and SR are increased up to 8% and 50% respectively. In a similar way, considering the effect of additive noise and reverberation mismatch (obtained by adding noise to the AC coefficients), the

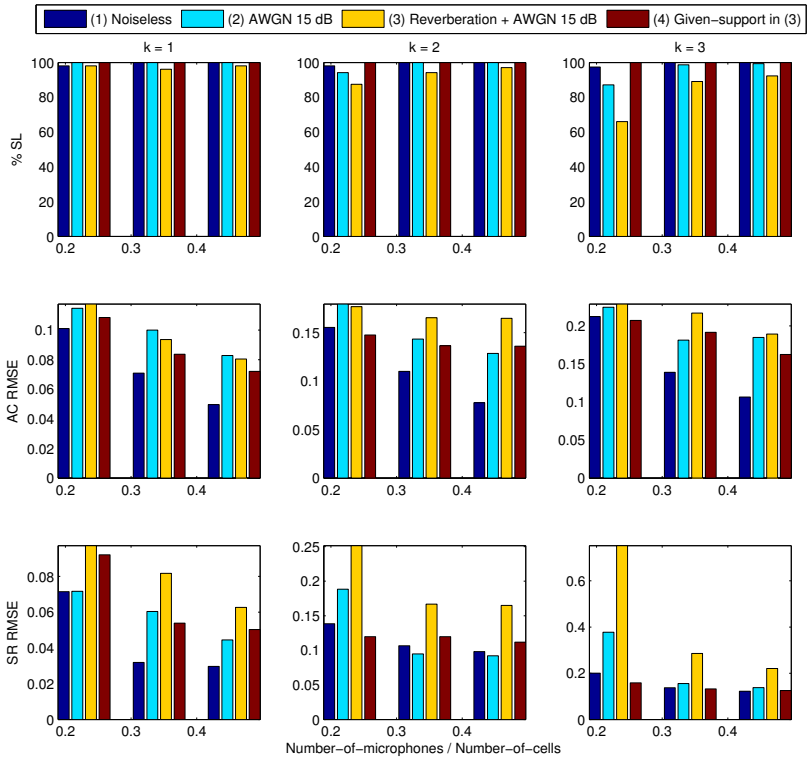

Figure 5.5 – Performance of the algorithm in terms of Source Localization (SL), Root Mean Squared Error (RMSE) of Absorption Coefficients (AC) estimation as well as Signal Recovery (SR). The test data are random orthogonal sources and the measurement matrix is the free-space Green function

distortion of AC estimates is bounded by the noise level whereas the recovered source coefficients (SR) are highly degraded. Contrasting these results with the speech bar charts demonstrates up to 40% SR distortion using only 8 microphones whereas AC estimation is achieved more accurately and degraded less than 5% using the approximately orthogonal speech sources; the average error of AC estimation is expected around 10-20% in noisy and reverberant condition. These results show a good robustness with increasing the number of concurrent sources. In addition, we observe a noticeable reduction in support recovery (SL) or localization of speech sources per frequencies; this effect could be justified as the spectrographic speech is approximately sparse and many of the components have a small energy which are drawn in noise. Hence, exploiting model-based sparse recovery or considering the broadband speech spectrum is crucial to achieve a reasonable localization performance.

Given that support recovery (i.e., SL) is obtained 100% correct by considering the broadband spectrum of speech signal and assuming that the sources are immobile, we can use the identified support for AC estimation and speech recovery. The resulted accuracy is upper bounded by noise level and in particular it enables a great improvement in SR. Hence, we carried out the AC estimation experiments on real data recordings where the support of the sparse coefficients (i.e. location of the active sources) is estimated from the first initial (5) frames and absorption

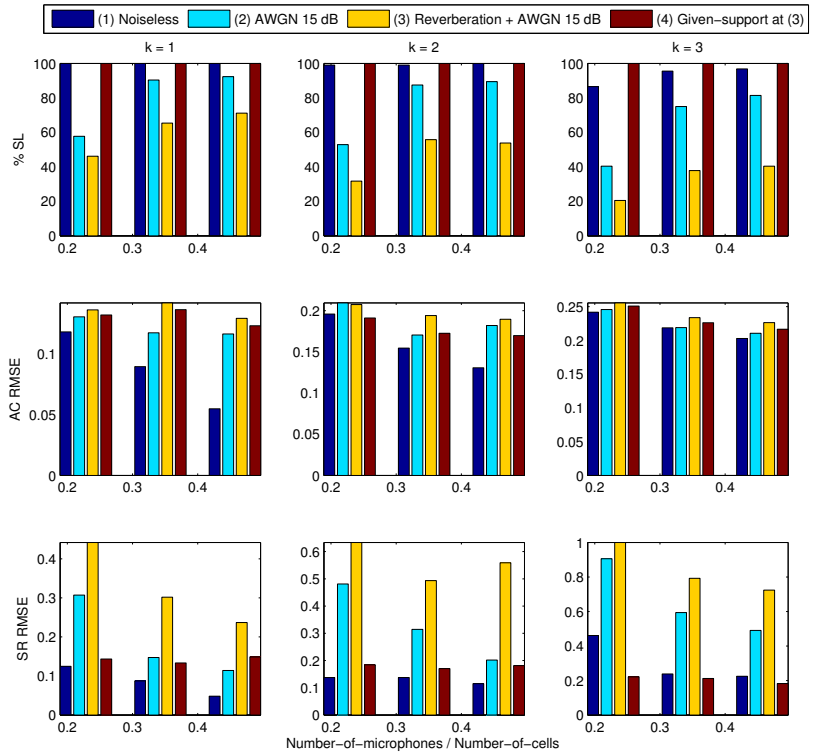

Figure 5.6 – Performance of the algorithm in terms of Source Localization (SL), Root Mean Squared Error (RMSE) of Absorption Coefficients (AC) estimation as well as Signal Recovery (SR). The test data are random speech utterances and the measurement matrix is the free-space Green function

coefficients are recovered given the support. If the number of microphones is more than the number of sources, then support estimation (source localization) enables very accurate results for estimation of the absorption coefficients. Similarly for speech recovery, we can perform inverse filtering to separate the individual sources. This scenario is investigated in Section 7.6.3. We computed the average time per frequency for the absorption coefficient estimation of six-sided walls using 8-channel microphone array as 17.16 seconds. This computational cost grows linearly with the dimension of the sparse vector and the number of microphones. Estimating the support from the first initial frames, enables estimation of the coefficients by pseudo-inversion which decreases the computational cost to a fraction of a second.

Real data evaluations

The scenario of the real data evaluations is explained in Section 5.3.1 which is similar to the first set-up described above. The location of the desired source is fixed through out the whole session (i.e. stationary condition). The estimated absorption coefficients are plotted using the data in the following conditions: (I) single speech utterances, (II) Two simultaneous speech utterances, (III) Three simultaneous speech utterances. The estimates are run over 9000 speech files of

MONC corpus Mccowan [2003]; the absorption coefficients are computed and averaged for each frequency-band independently. The estimated frequency-dependent absorption coefficients (computed at a resolution of 4 Hz) are illustrated in Fig. 5.7. To enable estimation of the three-dimensional acoustic parameters, we considered two parallel grids at given heights corresponding to the first order reflection of the table and ceiling; the reflections of the carpet floor are trapped under the table hence, the meeting table was considered as the floor in our Image Model Ba et al. [2010]. Thereby, the algorithm explained in Section 5.2.2 is run for a six-sided enclosure. To be more illustrative, the absorption coefficients are depicted for four surrounding walls, although we performed three-dimensional acoustic modeling. The absorption coefficients are estimated independently per frame hence, our method is applicable to the dynamic scenarios where the speaker changes the position at a rate slower than the frame-size.

There is no ground truth of the actual acoustic parameters available. The plots show a consistent estimation using recordings of one, two and three concurrent sources. The database is noisy (SNR$\approx$10 dB); the synthetic data evaluations reported in Fig. 5.6 show an expected 10-20% error in absorption coefficient estimation in noisy condition. Similar uncertainty of the coefficients is observed on real data recordings. Nevertheless, we use an average estimate of acoustic parameters for speech recovery tests conducted in Section 7.6.3.

Despite the point source assumption underlying the Image model, our studies on real data recordings demonstrate that in practical scenarios, the model provides a reasonable approximation to enable estimation of the acoustic properties and convolutive speech separation; more experimental analysis will be presented in Section 7.6.3. We can relax the assumption of discretizing the room and perform room acoustic modeling [Ajdler et al., 2003] and sparse recovery in a continuous Euclidean space [Tang et al., 2012]. The future research along these directions are explained in Chapter 9.

5.4 Conclusions

We proposed some algorithmic approaches to exploit structured sparsity models to characterize acoustic measurements obtained from an array of microphones, and to recover the individual speech sources. We estimated the acoustic response of the recording enclosure using the Image Model through a two-step procedure: first, estimating the room geometry and second, estimating the absorption coefficients.

For simple rectangular rooms, the room geometry is estimated by localizing virtual sources associated with discrete reflections of the original signal, followed by low-rank clustering of the subspaces resulting of each actual source. Location of the virtual sources corresponds to the temporal support of the room impulse response; these are used to estimate the geometry of the room via least square regression. The absorption coefficients associated with the reflective surfaces are then estimated via structured sparse recovery of a factorized formulation of a model of the reverberant room response.

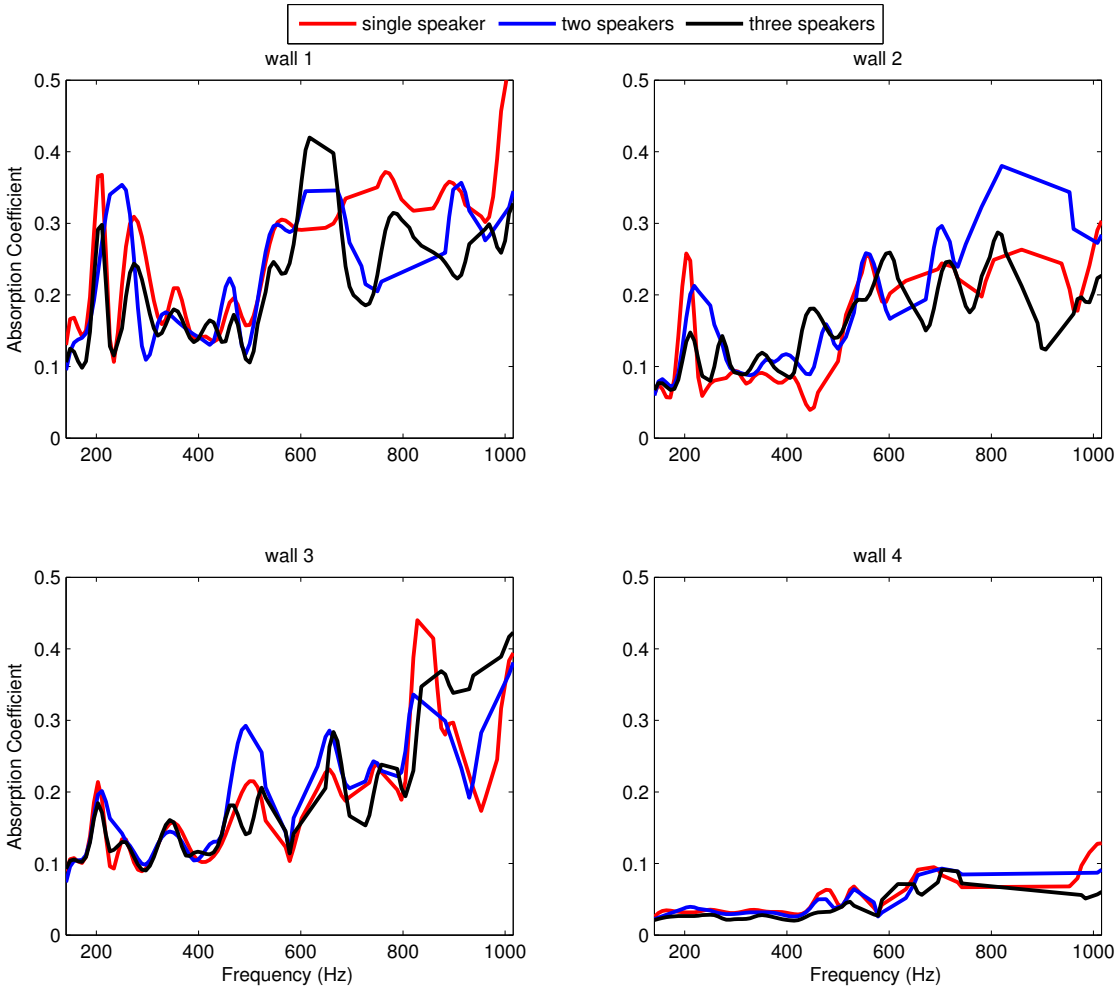

Figure 5.7 – Frequency-dependent absorption coefficients computed for each wall from the utterances of 3 competing speakers for the third speaker.

These algorithms can estimate the room acoustic from recordings of overlapping unknown sources. More studies are required to investigate the cost of acoustic ambiguity due to Image model and discrete space assumption. Although the point-source assumption underlying the Image model could be violated in practice, quantitative assessments on real data recordings verify that estimation of the early support of the room impulse response function enables estimation of the geometry and absorption coefficients, and they can be applied to determine the parameters of the forward model of the reverberant acoustic with sufficient accuracy to enable efficient recovery of speech. Having identified the multipath projections, we can characterize the microphone array measurement mechanism. The speech recovery is then achieved through model-based sparse recovery algorithms, incorporating the structures underlying the sparse coefficients to reconstruct the spatio-spectral representation of the acoustic scene data. This is the subject of the studies in the following Chapter 6.

6 Model-based Sparse Recovery

In Chapter 3, we briefly reviewed the three premises underlying our model-based sparse component analysis framework namely, structured sparse representation, compressive measurements and model-based sparse recovery.

In Chapter 4, we studied the structured sparse representation $\mathscr{S}$ of the acoustic scene information along with the inter-dependency models of the sparse coefficients. In Chapter 5, characterization of the compressive acoustic measurements Φ was elaborated.

This chapter outlines some of the algorithmic approaches to model-based sparse recovery and the performance of each approach is quantified in terms of accuracy of source localization as well as quality of the recovered speech. The focus of this chapter within the broad picture of our framework is depicted in Figure 6.1.

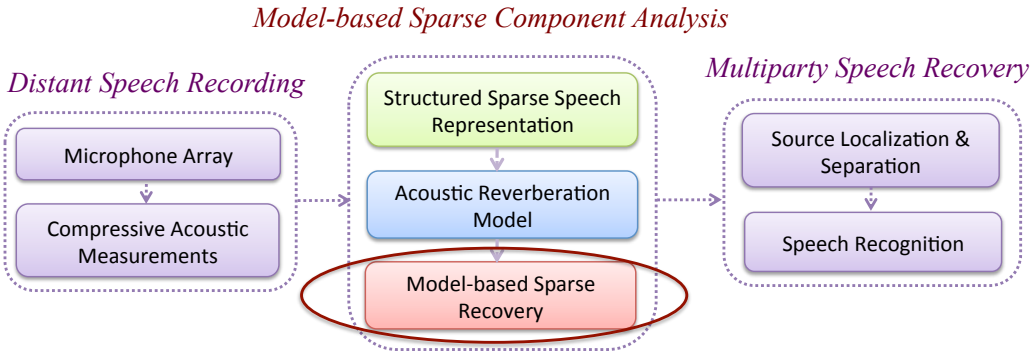

Figure 6.1 – The block diagram of the model-based sparse component analysis framework. The particular focus of this chapter is on *model-based sparse recovery algorithms*.

6.1 Computational Strategies

Defining a set $\mathbb{M}$ as the union of all vectors with a particular support structure, estimation of the structured sparse coefficient vector $\mathscr{S}$ from the microphone recordings $\mathscr{X}$ can be expressed as

$$\hat{\mathscr{S}} = \underset{\mathscr{S}\in\mathbb{M}}{\operatorname{argmin}} \|\mathscr{S}\|_0 \quad \text{s.t.} \quad \mathscr{X} = \Phi\mathscr{S}. \tag{6.1}$$

where the counting function $\|.\|_0$: $\mathbb{R}^{\mathrm{G}} \rightarrow \mathbb{R}$ returns the number of non-zero components in its argument. The major classes of computational techniques for solving sparse approximation problem are *Greedy pursuit*, *Convex optimization* and *Sparse Bayesian learning* [Tropp and Wright, 2010].

- ◇ *Greedy pursuit:* The nonzero components of $\mathscr{S}$ are estimated in an iterative procedure by modifying one or several coefficients chosen to yield a substantial improvement in quality of the estimated signal. We consider in particular extension of iterative hard thresholding and orthogonal matching pursuit to incorporate sparsity structures underlying spectrographic speech [Gribonval and Bacry, 2003, Blumensath and Davies, 2009, Kyrillidis and Cevher, 2011].

- ◇ *Convex optimization:* The counting function in (6.1) is replaced with a sparsity inducing convex norm that exploits the structure underlying $\mathscr{S}$. Therefore, a convex objective is obtained which can be solved using convex optimization. We consider in particular extension of basis pursuit algorithm which relies on ℓ_1 relaxation of the counting objective [Berg and Friedlander, 2008].

- ◇ *Sparse Bayesian learning:* A prior distribution is associated to $\mathscr{S}$ with sparsity inducing hyperparameters and a maximum a posteriori estimation is obtained given the distant microphone measurements, $\mathscr{X}$. We consider in particular the Bayesian framework proposed in [Wipf and Rao, 2007, Zhang and Rao, 2011b, 2012].

6.2 Sparse Recovery Exploiting Temporal Structures

We focus on two types of structures underlying the temporal coefficients, *simultaneous sparsity* and *AR dependency*.

- ◇ *Simultaneous Sparsity:* Suppose that multiple microphone recordings, $\mathscr{X}_1$, $\mathscr{X}_2$,..., have been collected and characterized by different sparse vectors, $\mathscr{S}_1, \mathscr{S}_2, ...$, but with an equivalent measurement matrix Φ. The sparse components simultaneously share a similar sparsity profile. We assume that the sources are stationary hence, the indices of nonzero components of $\mathscr{S}$, or the sparsity profile is the same although the amplitudes may be changing. This scenario indicates that a common subset of basis vectors are relevant in generating each response so we can merge the information contained in all measurements to uncover the underlying sparsity profile. Given L models structurally equivalent to $\mathscr{X} = \Phi\mathscr{S}$, the

simultaneous sparsity model is expressed as

$$\mathsf{X} = \Phi\mathsf{S} \tag{6.2}$$

where $\mathsf{X} = [\mathscr{X}_1, ..., \mathscr{X}_L]$, and $\mathsf{S} = [\mathscr{S}_1, ..., \mathscr{S}_L]$. The simultaneous sparsity as explained here indicates that several rows of S are zero, i.e. row sparse matrix. In the statistics literature, (6.2) is referred to as multiple output model [Hastie et al., 2001] or multiple measurement model (MMV) [Wipf and Rao, 2007].

◇ *AR Dependency*: Similar to what we explained above, we make the assumption that the multiple measurement vectors have the same, but unknown, sparsity profile. An additional consideration is the correlation among the entries in a non-zero row corresponding to each source which we model using an Auto Regressive (AR) process [Zhang and Rao, 2012]. We adopt the notion that $\mathsf{S}_{.l}$ represents the l^{th} column of S while $\mathsf{S}_{r.}$ represents the r^{th} row of S. Likewise, $\mathsf{S}_{r,l}$ refers to the r^{th} element in the l^{th} column of S. The sources (i.e. rows of S) are mutually independent, but each source satisfies a first-order AR model given by

$$\mathsf{S}_{r,l+1} = \beta_r \mathsf{S}_{r,l}, \quad r \in \{1, ..., G\},\ l \in \{1, ..., L\}. \tag{6.3}$$

where $\beta_r \in (-1, 1)$ is the AR coefficient [1]. We assume that $\mathsf{S}_{r,l} \sim \mathscr{N}(0, \alpha_r)$ so the source coefficients can be modeled by Gaussian distribution parameterized with hyper-parameter α_r in accordance with *Sparse Bayesian Learning* (SBL) formulations [Wipf and Rao, 2007, Tipping, 2001]. Having $\alpha_r = 0$ indicates that the r^{th} row of S is zero. In Section 6.4.2, we go through some studies on AR model of temporal and spectral sequences which verifies applicability of first-order AR for general speech recovery.

The first-order AR modeling assumption indicates that the linear combination of the univariate Gaussian would be Gaussian hence, the joint distribution of $\mathsf{S}_{r.} = [\mathsf{S}_{r1}, ..., \mathsf{S}_{rL}]$ is a multivariate Gaussian, expressed by

$$p(S_{r.}; \alpha_r, \beta_r) \sim \mathscr{N}(0, \alpha_r \beta_r^{-1}) \tag{6.4}$$

where B_r is a Toeplitz matrix defined by

$$B_r \equiv \begin{bmatrix} 1 & \beta_r & \dots & \beta_r^{L-1} \\ \beta_r & 1 & \dots & \beta_r^{L-2} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_r^{L-1} & \beta_r^{L-2} & \dots & 1 \end{bmatrix}^{-1} \tag{6.5}$$

The model-based sparse recovery algorithms have been proposed to incorporate the underlying structure of the sparse coefficients in recovering the unknown sparse components. Due to the

1. Incorporating higher AR models in this framework is straightforward.

common sparsity profile along the temporal sequences, these models are appropriate when the set-up is stationary. [2] We consider three model-based sparse recovery algorithms, Multiple measurement FOCal Underdetermined System Solver *MFOCUSS* [Cotter et al., 2005], Multiple measurement Sparse Bayesian Learning *MSBL* [Wipf and Rao, 2007] and Temporally correlated MSBL *TMSBL* [Zhang and Rao, 2011b].

◇ *MFOCUSS*: This algorithm employs an ℓ_p-norm like diversity measure to provide an estimate of the sparse coefficients, i.e.:

$$\hat{\mathsf{S}} = \arg\min_{\mathsf{S}} \|\mathsf{X} - \Phi\mathsf{S}\|_{\mathscr{F}}^2 + \lambda \sum_{r=1}^{G} (\|\mathsf{S}_{r.}\|)^p, \tag{6.6}$$

where $p \in [0,1]$ is a user-defined parameter [3], to prevent models with many nonzero rows. The factored-gradient approach is used for developing an algorithm to minimize the objective measure stated in (6.6). The update rules are summarized as follows:

$$\begin{aligned}
&\mathscr{W}^{i+1} = \text{diag}\,((c_{l.}^{i})^{1-p/2}),\\
&\qquad \text{where } c_{l.}^{i} = \|\mathsf{X}_{l.}^{i}\| = \left(\sum_{l=1}^{L} (\mathsf{X}_{l.}^{i})^2\right)^{1/2},\ p \in [0,1]\\
&Q^{i+1} = (\Phi\mathscr{W}^{i+1})^{\dagger}\mathsf{X}\\
&\mathsf{S}^{i+1} = \mathscr{W}^{i+1}Q^{i+1}
\end{aligned} \tag{6.7}$$

The update rules stated in (6.7) are guaranteed to converge monotonically to a local minimum of the objective. The MFOCUSS objective function can be derived using a generalized Gaussian prior on the row norms of spectro-temporal components. Thereby, it admits the Bayesian analysis on maximum a posteriori (MAP) estimation of the sparse coefficients given the observation $\mathscr{X}$ and linear measurement matrix Φ [Saab et al., 2007]. The difficulty with this procedure is twofold: either the prior is not sufficiently sparsity-inducing [Cevher, 2009] and the MAP estimates sometimes fail to be sparse enough, or a combinatorial number of suboptimal local solutions must be dealt with if a highly sparse prior is chosen. To circumvent these issues, an alternative rather empirical Bayesian strategy is proposed and explained as follows.

◇ *MSBL*: This approach takes an explicit probabilistic standpoint to the problem and explore a Bayesian model relying on the concept of automatic relevance determination (ARD) to encourage sparsity [Neal, 1996]. In the algorithm proposed in [Wipf and Rao, 2007], they postulate $p(\mathsf{X}|\mathsf{S})$ to be Gaussian with an unknown noise variance σ^2. Hence, for each $\mathsf{X}_{.l}, \mathsf{S}_{.l}$ pair we have

$$p(\mathsf{X}_{.l}|\mathsf{S}_{.l}) = (2\pi\sigma^2)^{-M/2}\exp\left(-\frac{1}{2\sigma^2}\|\mathsf{X}_{.l} - \Phi\mathsf{S}_{.l}\|_2^2\right) \tag{6.8}$$

2. The dynamicity can be handled by choosing a suitable sliding time-window in which the data can be approximately modeled as a MMV model [Zhang and Rao, 2011a].

3. A typical value for speech-specific task can be selected as 0.6 [Saab et al., 2007] or 0.8 as suggested by the authors of the MFOCUSS algorithm [Cotter et al., 2005]

Applying the ARD principle, they posit a prior distribution modulated by a vector of hyper-parameters controlling the prior variance of the rows of S. An L-dimensional Gaussian prior with an unknown variance parameter γ_r is assigned to the r^{th} row of S as $p(\mathsf{X}_{.l};\mathsf{S}_{.l}) \triangleq \mathcal{N}(0,\gamma_r I)$. Combining likelihood and prior, the posterior density of the l^{th} column of S becomes

$$p(\mathsf{S}_{.l}|\mathsf{X}_{.l};\gamma_r) \triangleq \mathcal{N}(\mu_{.l},\Sigma) \tag{6.9}$$

with covariance and mean for $\forall\, l = 1,...,L$ given by

$$\begin{aligned}
&\Sigma \triangleq \mathrm{Cov}[\mathsf{S}_{.l}|\mathsf{X}_{.l};\gamma_r] = \Gamma - \Gamma\Phi^T\Sigma_x^{-1}\Phi\Gamma \\
&\mathcal{M} = [\mu_{.1},...,\mu_{.L}] \triangleq \mathrm{E}[\mathsf{S}|\mathsf{X};\gamma] = \Gamma\Phi^T\Sigma_x^{-1}\mathsf{X} \\
&\Gamma \triangleq \mathrm{diag}(\gamma), \quad \Sigma_x \triangleq \sigma^2 I + \Phi\Gamma\Phi^T.
\end{aligned} \tag{6.10}$$

Given this formulation, the point estimate for $\hat{\mathsf{S}}$ would be the posterior mean, $\mathcal{M}$. Row sparsity is naturally achieved whenever a γ_r is equal to zero. Therefore, estimating the sparsity profile of $\hat{\mathsf{S}}$ amounts to estimating a hyper-parameter vector. The ARD framework applies maximization of the marginal likelihood to determine an appropriate γ, leading to the cost function as follows

$$\begin{aligned}
&\mathscr{L}(\gamma) \triangleq -2\log\int p(\mathsf{X}|\mathsf{S})p(\mathscr{S};\gamma)d\mathsf{S} \\
&-2\log p(\mathsf{X};\gamma) \equiv L\log|\Sigma_x| + \sum_{l=1}^{L}\mathsf{X}_{.l}^T\Sigma_x^{-1}\mathsf{X}_{.l}.
\end{aligned} \tag{6.11}$$

To minimize $\mathscr{L}(\gamma)$ with respect to γ, the unknown weights S are treated as hidden data and EM algorithm is used. The E step requires computation of the posterior moments using (6.10) or alternatively in a noise-free sparse representation when $\sigma^2 \to 0$ using the modified moments as

$$\begin{aligned}
&\Sigma = [I - \Gamma^{1/2}(\Phi\Gamma^{1/2})^\dagger\Phi]\Gamma \\
&\mathcal{M} = \Gamma^{1/2}(\Phi\Gamma^{1/2})^\dagger\mathsf{X};
\end{aligned} \tag{6.12}$$

while the M step can be achieved by taking the derivative of (6.11) with respect to γ, equating to zero, and forming an update rule as

$$\gamma_r^{new} = \frac{\frac{1}{L}\|\mu_{r.}\|_2^2}{1-\gamma_r^{-1}\Sigma_{rr}}, \quad \forall r = 1,...,G. \tag{6.13}$$

This particular approximation enforces a common sparsity profile and consistently places its prominent posterior mass on the appropriate region of space necessary for sparse recovery. The resultant algorithm is called MSBL because it can be posed as a multiple response extension of the standard sparse Bayesian learning (SBL) framework [Tipping, 2001]. Note that in this approach the common sparsity pattern is exploited without considering any

correlation among the coefficients. We will see the details of the *TMSBL* algorithm to incorporate the AR dependencies in the next Section 6.3.

6.3 Sparse Recovery Exploiting Spectral Structures

We focus on two types of structures underlying the spectral coefficients: *block structure* and *harmonic structure*.

◇ *Block structure* is exhibited if some interconnections between the adjacent frequencies exist. In case of the vector $\mathscr{S}$, the block dependency model indicates that the spatial sparsity structure is the same at all neighboring discrete frequencies. In other words, a block of b consecutive frequencies corresponds to the same cell so the signal of the individual sources is recovered with a structure of independent blocks defined as

$$\mathscr{F}_B = \{[f_1, ..., f_b], [f_{b+1}, ..., f_{2b}], [f_{F-b+1}, ..., f_F]\}. \tag{6.14}$$

◇ *Harmonic structure* is exhibited if there are some interconnections between frequencies which are the harmonics of a fundamental frequency. In voiced speech, most of the energy in the speech signal occurs at harmonics of a fundamental frequency. The harmonicity model captures this structure as indicates that at any cell of the grid, energy is present in all frequencies that can be expressed as harmonics of a fundamental frequency. To state it more precisely, the support of vector $\mathscr{S}$ has the following $\mathscr{F}_H$ structure defined as

$$\mathscr{F}_H = \{kf_0 | 1 < k < K\}, \tag{6.15}$$

where f_0 is the fundamental frequency and K is the number of harmonics.

We consider different model-based sparse recovery algorithms to incorporate the aforementioned underlying structure of the sparse coefficients in recovering the unknown sparse vector. The algorithms are Iterative hard thresholding *IHT* [Blumensath and Davies, 2008], Orthogonal Matching Pursuit *OMP* [Gribonval and Bacry, 2003], L_1L_2 [Berg and Friedlander, 2008] and Block Sparse Bayesian Learning *BSBL* [Zhang and Rao, 2012].

◇ *IHT*: Iterative hard thresholding (IHT) offers a simple yet effective approach to estimate the sparse vectors [Blumensath and Davies, 2008]. It seeks an N-sparse representation $\hat{\mathscr{S}}$ of the observation $\mathscr{X}$ iteratively to minimize the residual error. We use the algorithm proposed in [Kyrillidis and Cevher, 2011] which is an accelerated scheme for hard thresholding methods with the following recursion

$$\begin{aligned} &\hat{\mathscr{S}}_0 = 0 \\ &\mathscr{R}_i = \mathscr{X} - \Phi\hat{\mathscr{S}}_i \\ &\hat{\mathscr{S}}_{i+1} = \mathscr{M}^{\mathscr{F}.}\left(\hat{\mathscr{S}}_i + \kappa\Phi^T\mathscr{R}_i\right) \end{aligned} \tag{6.16}$$

where the step-size κ is the Lipschitz gradient constant to guarantee the fastest convergence speed [Nesterov, 1983]. To incorporate for the underlying structure of the sparse

coefficients, the model approximation $\mathscr{M}^{\mathscr{F}_\cdot}$ is defined as reweighting and thresholding the energy of the components of $\hat{\mathscr{S}}$ with either $\mathscr{F}_{\mathrm{B}}$ or $\mathscr{F}_{\mathrm{H}}$ structures.

◇ *OMP*: The Orthogonal Matching Pursuit (OMP) is a greedy pursuit algorithm which iteratively refines a sparse solution by successively identifying one or more components that yield the greatest improvement in quality [Tropp and Gilbert, 2007, Blumensath and Davies, 2008, Gribonval and Bacry, 2003]. To describe our model-based OMP in mathematical formulation, we consider an index set Λ which selects a subset of columns from Φ. Denoting the set difference operator as $\setminus$, the columns of $\Phi_{\setminus\Lambda}$ corresponding to $\mathscr{F}_{\mathrm{B}}$ or $\mathscr{F}_{\mathrm{H}}$ (denoted by $\mathscr{F}_\cdot$) structures are searched per iteration and Λ is expanded so as the mean-squared error of the signal approximation is minimized. The signal estimation algorithm would thus have the following recursion

$$\begin{aligned}
&\Lambda_0^{\mathscr{F}_\cdot} = 0 \\
&\lambda_i = \underset{\lambda \in \Phi_{\setminus\Lambda_{i-1}^{\mathscr{F}_\cdot}}}{\operatorname{argmin}} \ \|\mathscr{X} - \Phi_{\Lambda_{i-1}^{\mathscr{F}_\cdot}\cup\lambda}\Phi^{\dagger}_{\Lambda_{i-1}^{\mathscr{F}_\cdot}\cup\lambda}\mathscr{X}\|_2 \\
&\Lambda_i^{\mathscr{F}_\cdot} = \Lambda_{i-1}^{\mathscr{F}_\cdot} \cup \lambda_i \\
&\hat{\mathscr{S}}_i = \Phi^{\dagger}_{\Lambda_i}\mathscr{X}
\end{aligned} \tag{6.17}$$

◇ L_1L_2: Another fundamental approach to sparse approximation replaces the combinatorial counting function in the mathematical formulation stated in Equation (3.7) with the L_1 norm, yielding convex optimization problem that admits a tractable algorithm referred to as basis pursuit [Berg and Friedlander, 2008]. We use a multiple-measurement version of basis pursuit algorithm by re-arranging the components of $\mathscr{S}$ as a row-sparse matrix S with the $n^{\mathscr{F}_\cdot}$ columns corresponding to the common sparsity structure $\mathscr{F}_\cdot$ referring to either $\mathscr{F}_{\mathrm{B}}$ or $\mathscr{F}_{\mathrm{H}}$. Hence, the optimization problem to recover the block or harmonic sparse coefficients would be the following

$$\begin{aligned}
&\hat{\mathsf{S}} = \arg\min \|\mathsf{S}\|_{L_1,L_2} \quad \text{s.t.} \quad \mathsf{X} = \Phi\mathsf{S}, \\
&\|\mathsf{S}\|_{L_1,L_2} = \left(\sum_{r=1}^{G} \left[\sum_{l=1}^{n^{\mathscr{F}_\cdot}} \mathscr{S}_{r,l}^2 \right]^{1/2} \right)
\end{aligned} \tag{6.18}$$

◇ *BSBL*: The correlation among the coefficients modeled as AR dependencies is incorporated by [Zhang and Rao, 2012] in the framework of SBL proposed in [Wipf and Rao, 2007]. The sources $\mathscr{S}_r$ are assumed to be mutually independent, and the density of each $\mathscr{S}_r$ is Gaussian, given by

$$p(\mathscr{S}_r; \gamma_r, B_r) \sim \mathscr{N}(0, \gamma_r B_r), \quad r = 1, ..., G \tag{6.19}$$

where γ_r is a non-negative hyper-parameter controlling the row sparsity of $\mathscr{S}$ and $B_r \in \mathbb{R}^{b\times b}$ is a positive definite matrix that captures the correlation structure of $\mathscr{S}_r$. Assuming elements in the noise vector has a Gaussian distribution with variance σ^2. For the block

model (6.28), the Gaussian likelihood is

$$p(\mathscr{X}|\mathscr{S};\sigma^2) \sim \mathscr{N}(\Phi\mathscr{S},\sigma^2 I) \tag{6.20}$$

The prior for $\mathscr{S}$ is given by

$$p(\mathscr{S};\gamma_r,B_r,\forall r) \sim \mathscr{N}(0,\Sigma_0), \tag{6.21}$$

where Σ_0 is

$$\Sigma_0 \equiv \begin{bmatrix} \gamma_1 B_1 & & \\ & \ddots & \\ & & \gamma_G B_G \end{bmatrix} \tag{6.22}$$

Applying the Bayes rule we obtain the posterior density of $\mathscr{S}$, which is also Gaussian,

$$p(\mathscr{S}|\mathscr{X};\sigma^2,\gamma_r,B_r,\forall r) = \mathscr{N}(\mu_s,\Sigma_s) \tag{6.23}$$

with the mean and the covariance matrix as

$$\begin{aligned} \mu_s &= \frac{1}{\sigma^2}\Sigma_s\Phi^T\mathscr{X} \\ \Sigma_s &= (\Sigma_0^{-1}+\frac{1}{\sigma^2}\Phi^T S)^{-1}\mathscr{X} \\ &= \Sigma_0-\Sigma_0\Phi^T(\sigma^2 I+\Phi\Sigma_0\Phi^T)^{-1}\mathscr{X} \end{aligned} \tag{6.24}$$

Having all the hyper-parameters σ^2, γ_r, B_r, $\forall r$, the MAP estimate of $\mathscr{S}$ is given by

$$\begin{aligned} \hat{\mathscr{S}} \triangleq \mu_s &= (\sigma^2\Sigma_0^{-1}+\Phi^T\Phi)^{-1}\Phi^T X \\ &= \Sigma_0\Phi^T(\sigma^2 I+\Phi\Sigma_0\Phi^T)^{-1}X \end{aligned} \tag{6.25}$$

where the last equation follows from the matrix identity $(I+AB)^{-1}A \equiv A(I+BA)^{-1}$, and Σ_0 is the block diagonal matrix given by (6.22) with many diagonal block matrices being zeros. Clearly, the block sparsity of $\hat{\mathscr{S}}$ is controlled by the γ_r's in Σ_0. More specifically, during the learning procedure, when $\gamma_r = 0$, the associated r^{th} block in $\hat{\mathscr{S}}$ becomes zeros. The framework proposed in [Zhang and Rao, 2011b], derives the EM-based learning rule to learn the noise variance, σ^2, the prior dependency density matrix (i.e., AR parameters) B_r as well as γ. We will see in Section 6.4.2 that the dependency matrix can be estimated offline for an speech-specific task. Estimation of γ_r can be obtained by minimizing the modified cost function of

$$\mathscr{L}(\gamma) \triangleq \log|\sigma^2 I+\Phi\Sigma_0\Phi^T|+\mathscr{X}^T(\sigma^2 I+\Phi\Sigma_0\Phi^T)^{-1}\mathscr{X}, \tag{6.26}$$

The EM algorithm to find the optimal parameters requires computation of the posterior

moments as expressed in (6.24) and the following updating rule for $\gamma_r (r = 1, ..., G)$ by equating to zero the derivative of the cost function with respect to γ_r:

$$\begin{aligned} &\gamma_r \leftarrow \frac{\mathrm{Tr}[B_r^{-1}(\Sigma_s^r + \mu_s^r(\mu_s^r)^T)]}{L}, \quad r = 1, ..., G \\ &\mu_s^r \triangleq \mu_s((r-1)L+1 : rL) \\ &\Sigma_s^r \triangleq \Sigma_s((r-1)L+1) : rL, (r-1)L+1 : rL \end{aligned} \tag{6.27}$$

◇ *TMSBL*: [4] The procedure of BSBL is applicable to exploit the temporal correlation modeled as AR dependencies of coefficients with a common sparsity profile by transforming the multiple measurement model stated in (6.2) to the single measurement model by letting

$$\begin{aligned} &\mathscr{X} = \mathrm{vec}(X) \in \mathbb{C}^{ML \times 1}, \\ &\Psi = \Phi \otimes I_L, \\ &\mathscr{S} = \mathrm{vec}(S) \in \mathbb{C}^{GL \times 1}, \end{aligned}$$

Thereby, the MMV model stated in (6.2) is transformed to

$$\mathscr{X} = \Psi \mathscr{S}, \tag{6.28}$$

with a block structure obtained as

$$\mathscr{X} = [\phi_1 \otimes I_L, ..., \phi_G \otimes I_L][S_1^T, ..., S_G^T]^T = \sum_{r=1}^{G} (\phi_r \otimes I_L) S_r,$$

where $S_r \in \mathbb{C}^{L \times 1}$ is the r^{th} block in $\mathscr{S}$ and $S_r = S_{r\cdot}^T$. Having N nonzero rows in S means N nonzero blocks in $\mathscr{S}$. Thus, $\mathscr{S}$ is block-sparse. Given the vectorized formulation stated in (6.28), the rest of the procedure is similar to BSBL algorithm. This algorithm is referred to as TMSBL which is an extension of MSBL framework incorporating the temporal correlations.

6.4 Experimental Analysis

The experiments are conducted to evaluate the performance of the algorithm in terms of *speaker localization* and *speech recovery*. More specifically, the analyses aim to address the following aspects of the model-based sparse component analysis framework:

1. Evaluation of various algorithmic approaches to model-based sparse recovery considering temporal and spectral structures for source localization and speech recovery.
2. Analysis of the appropriate topology of microphone array used for distant recordings.
3. Sensitivity of the sparse recovery algorithms to noise and mismatch in acoustic parameters as well as the coherence of the measurements.
4. Computational cost of different approaches.

4. This algorithm belongs to Section 6.2; the procedure is explained here as it relies on *BSBL* framework.

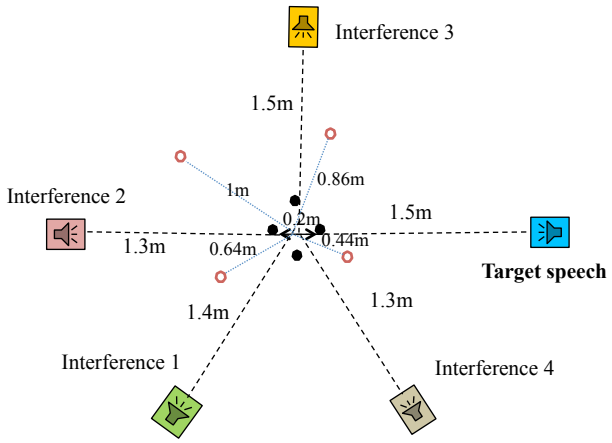

Figure 6.2 – Overhead view of the room set-up for uniform (black dots) and random microphone array (white dots)

6.4.1 Overlapping Speech Database

The speech utterances are taken from the Wall Street Journal (WSJ) corpus [Lincoln et al., 2005]. The WSJ corpus is a 20000-word corpus consisting of read Wall Street Journal sentences. The sentences are read by a range of speakers (34 in total) with varying accents (including a number of non-native English speakers). The broad phonetic space of WSJ allows the results to be generalizable for the task of speech separation in overlapping conditions. The overlapping speech was synthesized by mixing sentences from the test utterances with interfering sentences from development set. The interference files are normalized and scaled to yield -20 dB input SNR.

The audio is recorded using four channels microphone array. The planar area of a room with dimension $3\text{m}\times3\text{m}\times3\text{m}$ is divided into cells with 50 cm spacing. The data collection setup is depicted in Figure 6.2. The scenarios include *random* and *compact* topologies of microphone array in clean as well as reverberant and noisy conditions. Room impulse responses are generated with the Image model technique [Allen and Berkley, 1979] using intra-sample interpolation, up to 15^{th} order reflections and omni-directional microphones for a room reverberation time equal to 180 ms. The speech signals are recorded at 16 kHz sampling frequency and the spectro-temporal representation for source separation is obtained by windowing the signal in 250 ms frames using Hann function with 50% overlapping.

6.4.2 Speaker Localization Performance

The probabilistic performance bounds of multi-speaker localization are obtained by averaging the results over an exhaustive set of configurations [5]. The results are evaluated over all configurations consisted of $\text{N} \in \{5-10\}$ sources. The probabilistic evaluations are necessary to form a realistic expectation of our sparse recovery framework as the deterministic performance bounds explained

5. Similar topologies are exclusively considered.

in 3.2.4 are derived for the worst case scenario which is not likely to occur [Boufounos et al., 2011, Asaei et al., 2012b,a].

To guarantee the performance of the sparse recovery algorithms, the first issue that we are particularly capable of handling in speech applications is the *coherence of the measurement matrix* as expressed in (3.9). Although the acoustic measurements are forced by the natural projections associated to the media Green's function, they are frequency dependent and the broad spectrum of speech signal enables us to have some control over selection of the appropriate frequencies. To analyze the measurement matrix for the broadband speech spectrum, we compute the condition number of Φ for different frequency bands. The results are illustrated in Figure 6.3. The results suggest subband processing in order to increase the efficiency of sparse recovery algorithm.

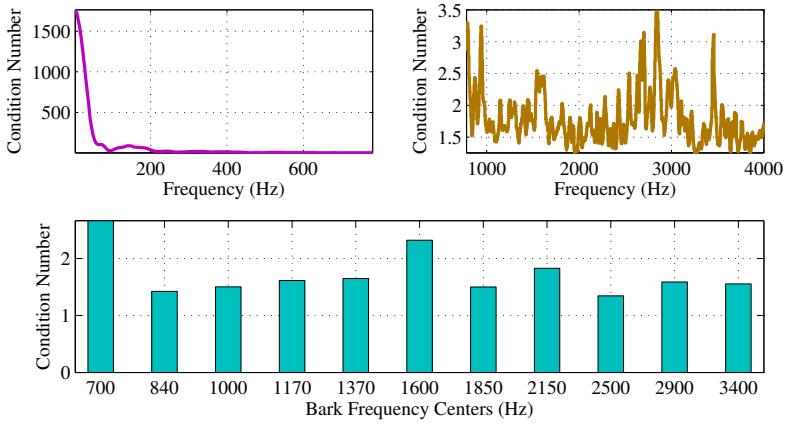

Figure 6.3 – Condition number computed per frequency band for the measurement matrix

The first set of evaluations consider temporal structures. We process L frames of speech jointly and investigate the sensitivity to different selections of L. All algorithms perform better with $L > 1$. We can empirically observe that $L \in \{10, ..., 15\}$ is a good choice for the performance of different algorithms. The results of multi-speaker localization are illustrated in Figure 6.4 for $L = 13$.

The TMSBL algorithm can learn the AR parameters during the optimization [Zhang and Rao, 2011b]. However, the procedure is very expensive in terms of computational cost and we also observe that the algorithm performs similar to MSBL (which does not incorporate the AR correlations). Hence, we carry out some studies on an average AR model for speech signal which can be exploited for source localization. Figure 6.6 illustrates the average AR model learned for 10 min of speech signal and averaged across all frequencies. The first-order AR coefficient is estimated as 0.3. We can further learn an average AR model per frequency but, we observe that the frequency-dependent AR model does not improve the performance. Hence, a first-order average AR model is a good approximation to capture the temporal structure across multiple frames and it offers a significant speed up in the algorithm. The initial algorithm assumes that all

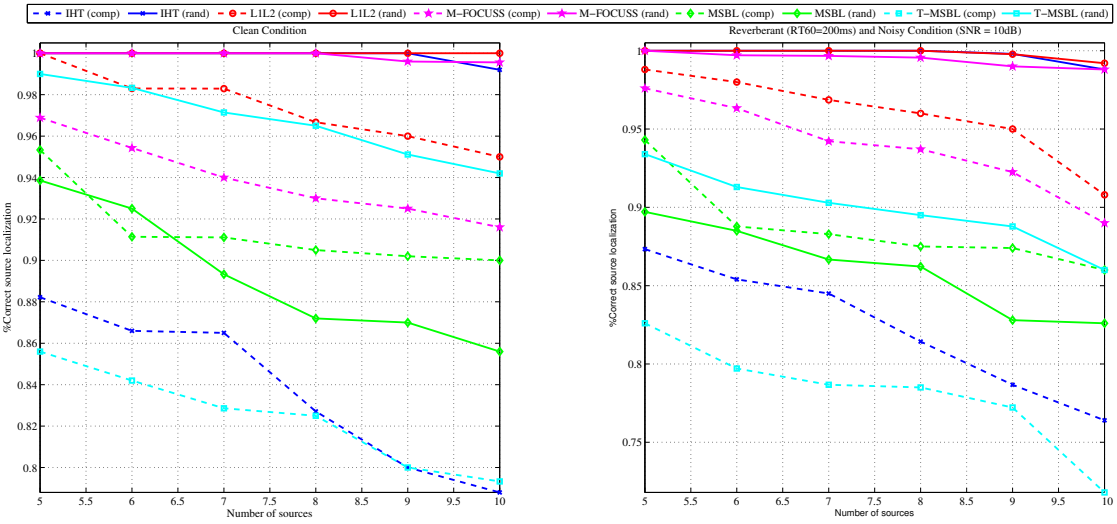

Figure 6.4 – Speaker localization performance evaluated for 5-10 sources exploiting *temporal* structured sparsity models

sources have similar correlation structure. The experimental analysis on speech-specific average AR model verifies this assumption.

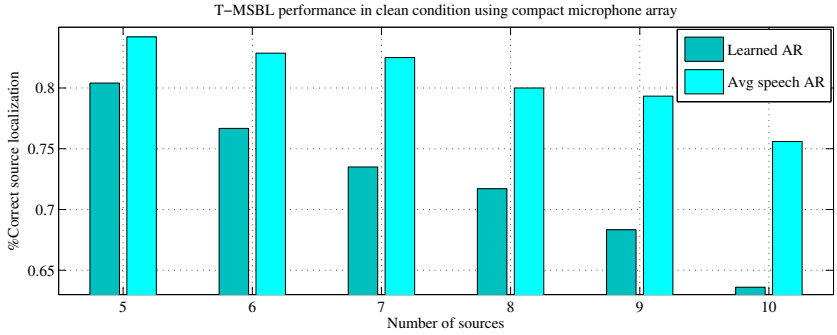

Figure 6.5 – It is beneficial to learn the average AR coefficients for speech source localization

The performance of the TMSBL algorithm improves using the average AR model. The difference is more noticeable when the algorithm does not perform very well (e.g. TMSBL in clean compact scenario). Figure 6.5 shows the improvement obtained from an average AR model. We also observe that increasing L is particularly useful when the number of sources is small. The average AR coefficients are estimated from voiced speech segments using an energy-based voice activity detector.

In addition to the temporal structures, we investigate the spectral sparsity models. Similar to the temporal dependencies, we verify that modeling the blocks as a first-order AR process is sufficient to incorporate the intra-block correlation. To estimate the AR coefficients, the frequency band (nfft = 2048*4) is split into blocks of size 16 and processed independently. Figure 6.7 demonstrates the frequency domain average AR model for 10 min speech signal. The first-order

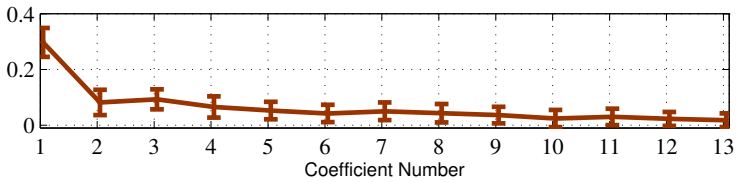

Figure 6.6 – 13-order average temporal-AR coefficients estimated for 10min speech. The cross lines show the variance of the estimates.

coefficient is estimated as 0.45. The results of multi-speaker localization exploiting spectral structure are illustrated in Figure 6.8 for B = 16. All the algorithms are run for the stopping threshold fixed to 1e-2 and the maximum iteration of 150. The value of p is selected as 0.8 suggested by the authors of MFOCUSS [Cotter et al., 2005]. We evaluated other values as suggested in [Saab et al., 2007] but, no difference was obtained.

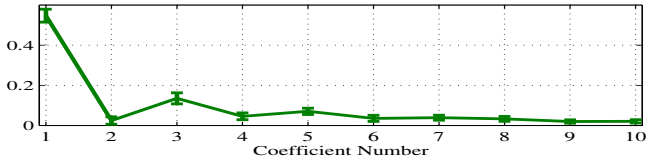

Figure 6.7 – 10-order AR coefficients estimated for 10min speech signal. The cross lines illustrate the variance of estimates

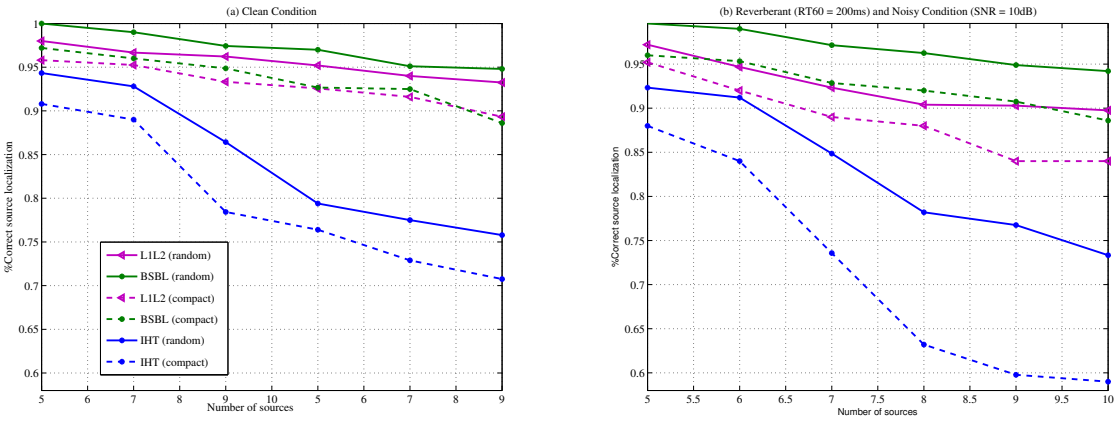

Figure 6.8 – Speaker localization performance evaluated for 5-10 sources exploiting *spectral* structured sparsity models

We can see that exploiting the frequency structures yield very strong results. The number of microphones is only 4 whereas we can localize up to 9 sources with 95% accuracy. These results are beyond the deterministic performance quantified in Section 3.2.4. The orthogonality or disjointness of spectrographic speech signals (as explained in Section 2.4 and 5.3) is a key property to achieve this bound of performance. Another important observation is that the ad-hoc layout of microphone array improves the results for all sparse recovery algorithms.

The algorithms work very well for speaker localization per frame exploiting the spectral model;

hence, we can drop the temporal stationarity assumptions [Asaei et al., 2012a]. In addition to the temporal and spectral dependencies, we can also consider oblique structures attributed to the spectro-temporal correlations. We could verify the correlation (as illustrated in Figure 4.5) in a similar manner as we studied temporal and spectral dependencies. However, we did not achieve reasonable localization results. The results of harmonic sparse recovery were comparable to the block-sparse recovery demonstrated in Figure 6.8 hence, they are not further elaborated here. We observe that considering large block sizes has a significant impact on localization accuracy but, in terms of signal recovery, it results in some artifacts in signal reconstruction. This subject is investigated in the next Section 6.4.3.

6.4.3 Speech Recovery Performance

There are 5 concurrent speech sources, which are separated using the recordings of 4 microphones. The simulations are done in MATLAB 7.14 on 4 Core(TM) i7 CPU @ 2.8-GHz, 11.8-GiB RAM PC and the absolute elapsed times (in seconds) are measured for each algorithm. The quality evaluation results exploiting the temporal structures in clean and noisy conditions are summarized in Figures 6.9 and 6.10. The noisy condition includes both the effect of additive noise as well as mismatch in acoustic condition. The SNR of noisy scenario is 10 dB by adding white Gaussian noise and absorption coefficients of the forward model stated in (5.1) are 25% deviated from the actual parameters. The number of frames, L is equal to 2. The quality of the recovered speech is evaluated in terms of source to interference ratio (SIR) and source to noise ratio (SNR) as defined in (3.11). In addition, we evaluated the perceptual speech quality in terms of PESQ [ITU-T, 2001] and computed the weighted spectral slope distance measure (WSS) [Persia et al., 2008]. In the following, we briefly describe PESQ and WSS distortion measures.

PESQ uses several processing steps in order to predict the perceived quality that would be given to a processed speech signal in a subjective listening test [ITU-T, 2001]. It does so by computing an internal representation of human sound perception for both the original and processed signals, using a transformation to a short-time spectrum with perceptual frequency (Bark) and intensity (loudness) scales. In particular, the calculation of the short-time loudness spectrum considers the effect of simultaneous masking in the human auditory system. Before transforming the input signal to the internal representation, PESQ performs several pre-processing steps with the goal of compensating differences between both signals. These include (possibly time-varying) differences in the delay, frequency response and gain between both signals, which are perceptually insignificant as long as they are within some limits. This is a direct consequence of the subjective listening test scores that PESQ aims to predict, in which subjects rate the quality of the processed signal without knowledge of the original.

After the compensation and transformation steps, differences in the internal short-time representations between both signals are computed; the (signed) difference over time is called disturbance density. Two disturbance vectors, one for positive and one for negative differences are computed, owing to the more objectionable nature of added distortions compared to missing frequency

components in the processed speech signal [ITU-T, 2001]. Finally, the values of both disturbance vectors are aggregated over various time spans with different norms and combined to a raw PESQ score in the range [-0.5, 4.5]. Following the standardization of PESQ as Recommendation P.862, the standardization sector of the International Telecommunications Union (ITU-T) created a simple mapping function to allow comparisons between the raw PESQ score and the traditional MOS (Mean Opinion Score) in the range [1.0, 5.0]. This function is specified in Recommendation P.862.1 [ITU-T, 2003] and maps scores in the range [-0.5, 4.5] to the range [1.0, 4.5]. The choice of 4.5 as maximum score instead of 5.0 is due to the absolute (i.e., without knowledge of the original signal) quality rating that subjects perform in a listening test, where some subjects will not assign the highest score even for signals without any distortions.

WSS distortion: This spectral distance measure is based on comparison of smoothed spectra from the clean and distorted speech samples. The smoothed spectra can be obtained from either linear prediction analysis, or filter-bank analysis. One implementation of WSS can be defined as follows,

$$\mathrm{d}_{WSS} = \frac{1}{\mathscr{T}+1}\sum_{\tau=0}^{\mathscr{T}} \frac{\sum_{\mathrm{j}=1}^{\mathrm{B}} W(\mathrm{j},\tau)(\mathrm{S}(\mathrm{j},\tau)-\hat{\mathrm{S}}(\mathrm{j},\tau))^2}{\sum_{\mathrm{j}=1}^{\mathrm{B}} W(\mathrm{j},\tau)} \tag{6.29}$$

where B is the number of bands, $\mathscr{T}$ is the total number of frames, and $\mathrm{S}(\mathrm{j},\tau)$ and $\hat{\mathrm{S}}(\mathrm{j},\tau)$ are the spectral slopes (typically the spectral differences between neighboring bands) of the j^{th} band in the τ^{th} frame for clean and recovered speech, respectively. $W(\mathrm{j},\tau)$ are weights, which can be calculated described in [Klatt, 1982]. The WSS of clean speech signal is 0.

The next experiment measures the performance of the algorithms exploiting spectral structures. The block-size, B is equal to 2 as it yields the best results. In the harmonic model, we consider that $\mathrm{f}_0 \in [150-400]$ Hz. Those frequencies that are not harmonics of f_0 are recovered independently in H-IHT and H-$\mathrm{L}_1\mathrm{L}_2$. We also considered that the harmonic structures are non-overlapping and k spans the full frequency band. The harmonic sparse recovery approach does not require estimation of f_0. We start from $\mathrm{f}_0 = 50$ and consider all of its harmonics within the frequency band (i.e., $\mathrm{f} \leqslant 4000$); hence, a block of size $\mathrm{K} = 80$ of harmonics of $\mathrm{f}_0 = 50$ are recovered jointly. Then we move to $\mathrm{f}_0 = 51$ and proceed up to $\mathrm{f}_0 = 400$. Therefore, the size of the blocks are variable. To prevent overlapping, the priority is given to the first seen frequency components. In other words, if a particular frequency is first included in the harmonics of $\mathrm{f}_0 = 50$, it is excluded from the harmonics of $\mathrm{f}_0 = 100$. The remaining frequency components are recovered independently. For H-OMP, the harmonic subspaces are used to select the bases while projection is performed for the full frequency band. This procedure is applied on all of the frames regardless of the voiced/unvoiced characteristics. Therefore, we expect the model to be more effective if the ratio of the voiced segments is greater than the unvoiced segments; a combination of *block* and *harmonic* model could be considered for effective model-based speech recovery. The results are summarized in Figures 6.11 and 6.12.

We can see that the harmonic model is strong in terms of interference suppression as quantified by

		BSBL	MSBL	MFOCUSS	L_1L_2	IHT	OMP
Random	SIR	12.15	24.16	**26.4**	14.52	14.24	12.47
	SNR	3.53	26.35	**28.3**	4.99	4.66	4.54
	PESQ	1.8	1.72	**3.09**	1.8	1.69	1.8
	WSS	69.23	40.39	**34.25**	62.59	93.88	86.5
Compact	SIR	8.24	14.57	**24.12**	14.09	14.16	12.72
	SNR	2.67	15.7	**23.47**	4.75	4.23	3.62
	PESQ	1.79	2.59	**2.96**	1.86	1.79	1.85
	WSS	67.55	49.32	**35.3**	66.46	93.15	87.92
	Time	17.6	2.96	2.55	132.46	4.62	**0.75**

Figure 6.9 – Performance of speech recovery exploiting *temporal structure* in clean acoustic condition

		BSBL	MSBL	MFOCUSS	L_1L_2	IHT	OMP
Random	SIR	11.13	**15.37**	14.71	10.39	12.45	10.63
	SNR	3.14	15.86	**18.35**	4.61	4.32	4.32
	PESQ	1.8	2.09	**2.19**	1.73	1.73	1.74
	WSS	68.67	69.89	**63.7**	70.62	74.94	74.13
Compact	SIR	6.47	4.2	4.08	8.26	**9.78**	9.07
	SNR	2.22	3.99	3.22	**4.29**	4	3.28
	PESQ	**1.82**	1.05	1.34	1.66	1.72	1.79
	WSS	**69.2**	87.78	71.46	75.02	82.73	86.8
	Time	18.3	3.6	2.88	136.3	7.17	**0.75**

Figure 6.10 – Performance of speech recovery exploiting *temporal structure* in noisy scenario. SNR = 10 dB by adding white Gaussian noise and acoustic parameters are 25% deviated from the actual parameters

SIR however, the level of artifacts is higher. Similar experiments on Numbers corpus [Mccowan, 2003] yields better results using the harmonic model [Asaei et al., 2012c]. The difference can be justified as the harmonicity of numbers (pronunciation of $0-20$) are generally higher than the average phonetically rich speech utterances provided in MC-WSJ corpus. Comparing the results of random array with the conventional uniform-array, we observe that the ad-hoc setting of microphones array improves the quality of the separated speech.

To compare and contrast various algorithmic approaches exploiting temporal or spectral sparsity structures and the topology of the microphone array, the graphs are demonstrated in Figure 6.13. The bar charts illustrate the amount of separation (SIR) [Vincent et al., 2006], perceptual quality (PESQ) [ITU-T, 2001] as well as weighted spectral slope distance measure (WSS) [Persia et al., 2008] obtained using different structures in signal acquisition set-up and speech recovery.

The results of incorporating the spectral structures are superior in terms of SIR and PESQ to the temporal-based sparse recovery which confirms that stronger dependencies exist between the spectral coefficients. As the results indicate, we observe that the highest perceptual quality are obtained using the sparse Bayesian learning framework (BSBL) and convex optimization (L_1L_2).

		Clean				Noisy			
		L_1L_2	IHT	OMP	BSBL	L_1L_2	IHT	OMP	BSBL
Random	SIR	20	**22.69**	16.59	15.71	16.45	**19.32**	14.31	12.63
	SNR	**25.52**	13.54	11.22	11.33	**16.14**	13.36	8.98	8.037
	PESQ	3.11	2.45	**2.46**	2.58	2.31	2.32	**2.27**	2.45
	WSS	**31.01**	104.58	59.71	50.46	52.38	66.74	59.49	50.75
Compact	SIR	**15.85**	15.25	15.19	9.3	11.72	**14.67**	9.92	7.25
	SNR	**20.58**	14.21	8.61	7.026	**13.65**	10.32	9.27	4.95
	PESQ	**2.98**	2.38	2.39	2.49	2.04	2.26	2.07	**2.36**
	WSS	**33.57**	87.91	75.74	53.62	62.85	73.74	67.53	**57.55**
	Time	148.29	4.25	**0.973**	5.85	139.21	4.79	**0.846**	7.2

Figure 6.11 – Performance of speech recovery exploiting *block spectral structure* in clean and noisy scenario. The noisy scenario has SNR = 10 dB by adding white Gaussian noise and acoustic parameters are 25% deviated from the actual parameters

		Clean				Noisy			
		L_1L_2	IHT	OMP	BSBL	L_1L_2	IHT	OMP	BSBL
Random	SIR	19.75	14.40	9.99	**20.85**	**16.69**	15.46	9.35	15.60
	SNR	5.68	4.68	**6.02**	5.38	5.46	4.57	**6.05**	5.35
	PESQ	**2.44**	1.99	1.94	2.16	**1.45**	1.42	1.27	1.37
	WSS	**32.83**	50.91	43.55	46.49	52.29	**51.86**	58.04	62.33
Compact	SIR	15.55	10.32	6.49	**21.43**	**14.08**	11.13	5.89	3.84
	SNR	5.37	4.40	**5.97**	5.16	5.14	4.31	**5.95**	3.89
	PESQ	**2.36**	1.97	1.86	2.11	**1.45**	**1.45**	1.34	0.73
	WSS	**35.09**	54.00	50.38	42.16	**56.72**	56.77	70.24	81.12

Figure 6.12 – Performance of speech recovery exploiting *harmonic spectral structure* in clean and noisy scenario. The noisy scenario has SNR = 10 dB by adding white Gaussian noise and acoustic parameters are 25% deviated from the actual parameters

This can be due to the zero-forcing spirit of greedy approaches. This deficiency is particularly exhibited for speech-like signals which do not possess high compressibility (as discussed earlier in Chapter 4). However, in some applications such as speech recognition, where the reconstruction of the signal is not desired, we can exploit the sparsity of the information bearing components in greedy sparse recovery approaches which offer a noticeable computational speed in efficient implementations [Kyrillidis and Cevher, 2011, Blumensath and Davies, 2008] and a reasonable performance (as already listed in Table 3.1 and 3.2).

We also observe that increasing the size of the blocks degrades the quality of the separated speech although in some cases it increases the signal to interference ratio (SIR) which is in accordance to the perceptual grouping principles relying on proximity in spectro-temporal space [6] [Wang and Brown, 2006]. The block structure underlies the phonetic information [Yang and Hermansky, 2000] which motivates further consideration for speech recognition applications. Considering that the speech signal is consisted of voiced and unvoiced segments, the block-interdependency mostly corresponds to the unvoiced speech while the harmonicity is exhibited in the voiced

6. In Section 4.3 we mentioned that the frequencies which are close together are grouped in our auditory analysis as being attributed to the same source

segments; hence we expect that a combination of both of the structures is beneficial for structured sparse recovery of speech signal.

6.5 Conclusions

In this chapter, we outlined some of the algorithmic approaches to model-based sparse recovery and evaluated each approach exploiting the speech-specific structures in terms of source localization accuracy and speech recovery performance. The numerical assessments confirm that the performance of sparse signal recovery is improved if both sparsity and correlation structure of the signals are exploited. For speaker localization, we can learn an average AR model and pre-specify it to the algorithm. This approach can yield better results while it offers computational speed. For speech recovery however, we can explore class-specific parameters learned during the optimization procedure. The experimental analysis on speech recovery demonstrates better performance of incorporating spectral structures than the temporal dependencies. The sparse Bayesian learning framework and convex optimization yield the highest perceptual quality and the iterative hard thresholding is very effective in terms of interference suppression.

In addition, we considered the impact of construction layout of the microphone array in the performance of sparse recovery algorithms. Initial studies presented in Section 3.2.4 motivates an ad-hoc design of microphone array to acquire less coherent measurements. We performed empirical evaluations on the aforementioned theory which demonstrated that considering the design specifications acknowledged by the generic theory of compressive sensing and sparse signal recovery leads to significant improvement in speech separation performance hence, the uniform microphone array set-up is not an optimal design. More studies are required to analyze the residual error to incorporate an effective speech enhancement scheme after sparse recovery. The generalization and optimality of our model-based sparse component analysis framework will be studied in the following Chapter 7.

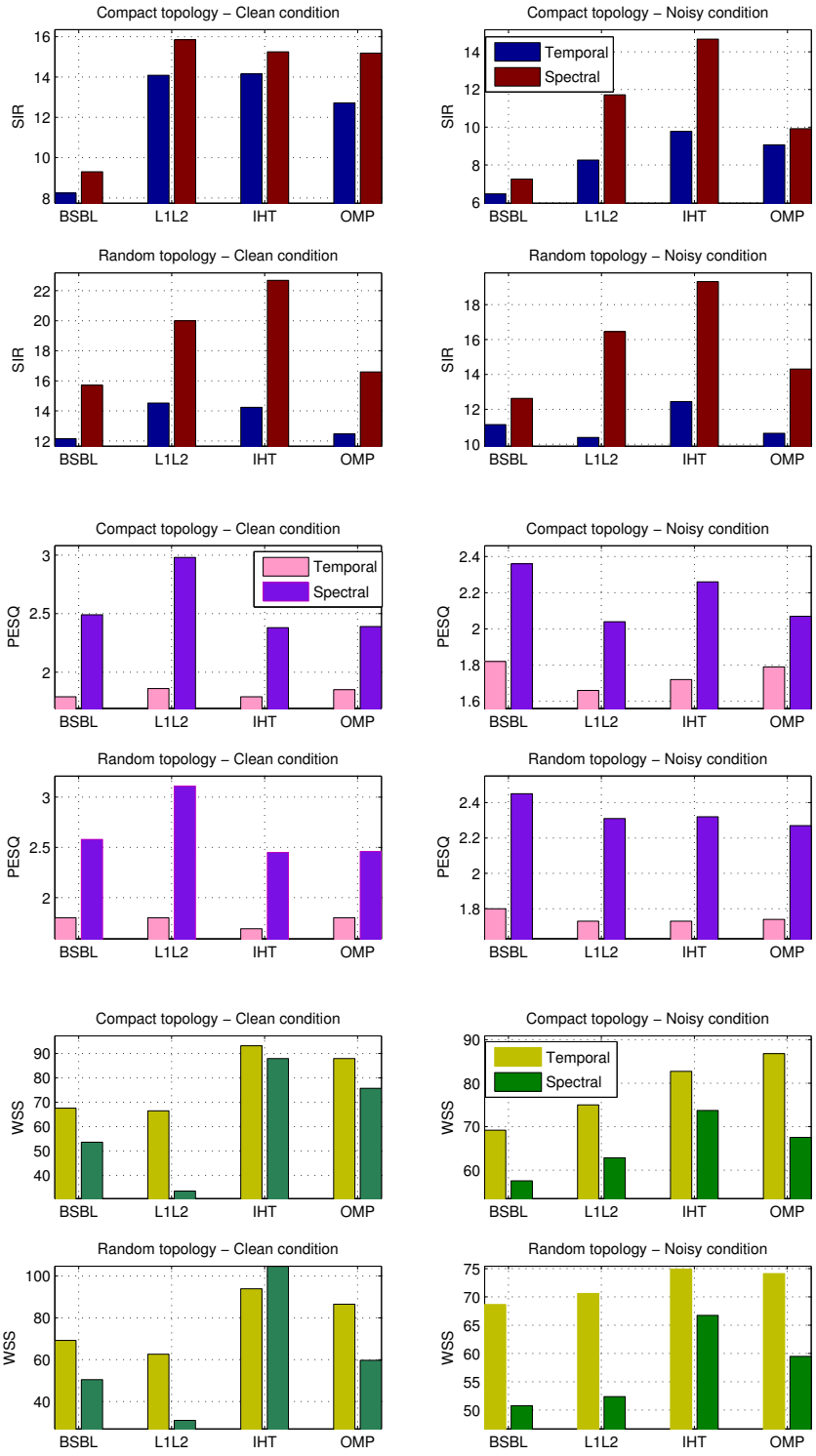

Figure 6.13 – Speech recovery performance in terms of source to interference ratio (SIR), perceptual evaluation of speech quality (PESQ) and weighted spectral slope (WSS). SIR measures the amount of interference suppression. PESQ and WSS are more perceptually motivated metrics which show high correlation with speech recognition performance [Persia et al., 2008]

7 Optimum Structured Sparse Coding

The model-based sparse component analysis framework was established in Chapter 3 along with the three fundamental components. The first component is structured sparse representation which was elaborated in Chapter 4. The second component is compressive acoustic measurements which was characterized in Chapter 5 and the third component is model-based sparse recovery algorithm that we have studied in the previous Chapter 6. This framework assumed that the geometrical set-up of microphone array is known in advance. The recent studies presented in Section 6.4.3 demonstrate that the conventional microphone arrays are not an optimal design and the sparse recovery techniques yield higher performance using ad-hoc microphone topology. Hence, in this chapter we draw a generalization to our framework by formulating a unified structured sparse coding scheme for source-sensor localization and speech recovery. Having the source and sensors being localized, we elaborate on optimality of inverse filtering to perform speech separation and dereverberation.

7.1 Generalized Formulation

We consider a scenario in which an unknown sound signal $S_g(f)$ at frequency f emanates from an unknown location ν_g in an enclosure and impinges on an array of M microphones located at $\mathscr{L} = \{\mu_1, \mu_2, \cdots, \mu_M\}$ on a 2-D plane. The room response $H(f, \nu_g, \mu_m)$ from the source location ν_g to the location μ_m is known for each of the M microphone locations [1]. The signal captured by a microphone located at μ_m would therefore be $X_m(f) = H(f, \nu_g, \mu_m)S_g(f)$ and representing $\mathscr{X}(f) = [X_1^T(f) \cdots X_M^T(f)]^\top$ and $\phi_{\nu_g,\mathscr{L}}(f) = [H(f, \nu_g, \mu_1) \cdots H(f, \nu_g, \mu_M)]^\top$, we can write

$$\mathscr{X}(f) = \phi_{\nu_g,\mathscr{L}}(f)S_g(f) \tag{7.1}$$

$\phi_{\nu_g,\mathscr{L}}(f)$ is also known as the *array manifold* vector which is specific to the source location ν_g and the locations of the microphones at $\mathscr{L} = \{\mu_1, \mu_2, \cdots, \mu_M\}$. Our objective is to estimate the location of the source and sensors and recover the speech signal.

1. Please refer to Chapter 5 for an intricate discussion on characterization of the room acoustic channel.

We can obtain an estimate of the source as $\hat{S}_g(f) = \phi^{\dagger}_{\nu_g,\mathscr{L}}(f)\mathscr{X}(f)$, where $\phi^{\dagger}_{\nu_g,\mathscr{L}}(f)$ represents the pseudo-inverse of $\phi_{\nu_g,\mathscr{L}}(f)$. Given the estimated $\hat{S}_g(f)$ obtained using $\phi_{\nu_g,\mathscr{L}}(f)$, then

$$\hat{\mathscr{X}}(f) = \phi_{\nu_g,\mathscr{L}}(f)\hat{S}_g(f) = \phi_{\nu_g,\mathscr{L}}(f)\phi^{\dagger}_{\nu_g,\mathscr{L}}(f)\mathscr{X}(f); \tag{7.2}$$

this now gives us an effective handle to estimate $\nu_g, \mathscr{L}$ as

$$\nu_g, \mathscr{L} = \underset{\nu_g,\mu_1,\mu_2,\cdots,\mu_M}{\arg\min} \|\mathscr{X}(f) - \hat{\mathscr{X}}(f)\|_2^2, \tag{7.3}$$

where $\hat{\mathscr{X}}(f) = \phi_{\nu_g,\mathscr{L}}(f)\phi^{\dagger}_{\nu_g,\mathscr{L}}(f)\mathscr{X}(f)$ is the projection of $\mathscr{X}(f)$ onto the array manifold vector $\phi_{\nu_g,\mathscr{L}}(f)$. Having estimated $\nu_g, \mathscr{L}$ by solving (7.3), the source signal is recovered as

$$\hat{S}(f) = \phi^{\dagger}_{\nu_g,\mathscr{L}}(f)\mathscr{X}(f) = (\phi^{T}_{\nu_g,\mathscr{L}}\phi_{\nu_g,\mathscr{L}}(f))^{\dagger}\phi^{T}_{\nu_g,\mathscr{L}}$$

The discerning reader may note that the objective function of (7.3) is merely

$$\|(I - \phi_{\nu_g,\mathscr{L}}(f)\phi^{\dagger}_{\nu_g,\mathscr{L}}(f))\mathscr{X}(f)\|_2^2$$

which is minimized if $\phi_{\nu_g,\mathscr{L}}(f)$ is chosen such that the solitary non-unity singular value of $I - \phi_{\nu_g,\mathscr{L}}(f)\phi^{\dagger}_{\nu_g,\mathscr{L}}(f)$ goes to zero. This may appear to be independent of $\mathscr{X}(f)$; however this is not so – the corresponding eigenvector must also be maximally aligned to $\mathscr{X}(f)$ for the objective to be minimized. Nevertheless, the formulation expressed in (7.3) introduces greater dependence on data.

The above modification can be succinctly stated in matrix form as follows. Let $F = \{f_1, f_2, \cdots, f_B\}$ represent a set of B adjacent frequencies. We define an array manifold *matrix* for M sensors in locations $\mathscr{L} = \{\mu_1, \mu_2, \cdots, \mu_M\}$ as the $MB \times B$ matrix $\phi_{\nu_g,\mathscr{L}}(F)$ obtained by stacking a set of diagonal matrices obtained from $H(f, \nu_g, \mu_1)$ to $H(f, \nu_g, \mu_M)$. Let

$$\begin{aligned} H^{\text{diag}}(F, \nu_g, \mu_m) &= \text{diag}([H(f_1, \nu_g, \mu_m)\ H(f_2, \nu_g, \mu_m) \cdots H(f_B, \nu_g, \mu_m)]), \\ \phi_{\nu_g,\mathscr{L}}(F) &= [H^{\text{diag}}(F, \nu_g, \mu_1) \cdots H^{\text{diag}}(F, \nu_g, \mu_M)]^{\top} \end{aligned} \tag{7.4}$$

We define

$$\begin{aligned} X_m(F) &= [X_m(f_1)\ X_m(f_2) \cdots X_m(f_B)]^{\top}, \\ \mathscr{X}(F) &= [X_1^T(F) \cdots X_M^T(F)]^{\top}, \\ S_g(F) &= [S_g(f_1)\ S_g(f_2) \cdots S_g(f_B)]^{\top}. \end{aligned}$$

The extended equivalent of (7.1) is given by

$$\mathscr{X}(F) = \phi_{\nu_g,\mathscr{L}}(F)S_g(F). \tag{7.5}$$

The source-sensor locations and the speech signal can be estimated as

$$\begin{aligned} \nu_g, \mathscr{L} &= \underset{\nu_g, \mu_1, \mu_2, \cdots, \mu_M}{\arg\min} \|\mathscr{X}(F) - \phi_{\nu_g, \mathscr{L}}(F)\phi^{\dagger}_{\nu_g, \mathscr{L}}(F)\mathscr{X}(F)\|_2^2 \\ \hat{S}_g(F) &= \phi^{\dagger}_{\nu_g, \mathscr{L}}(F)\mathscr{X}(F). \end{aligned} \tag{7.6}$$

This formulation indicates a parametric approach to source-sensor localization problem where $\mathscr{L}$ is estimated directly by minimizing the objective function stated in (7.6). It defines the locations as continuous random vectors in a 2-D plane and results in a non-linear objective which is difficult to optimize. Hence, we resort to a non-parametric method and formulate the source-sensor localization problem as structured sparse coding where we leverage the greedy sparse recovery algorithm to find the solution. This idea is described in the following Sections 7.2 and 7.3.

7.2 Codebook of Spatial Signals for Sparse Representation

We consider a scenario in which M microphones are distributed on a discrete grid of G cells sufficiently dense so that each microphone can be assumed to lie at one of the cells and $M \ll G$. Accordingly, we also assume that a source is located at an unknown cell. We then define a selector vector P as

$$P = [p_i],\ \forall i \in \{1, 2, ..., \eth\}, \quad p_i \in \{0, 1\} \tag{7.7}$$

where $\eth$ denotes the size of the codebook. If a component of $p_i = 1$, it indicates that a microphone exists at a particular location inferred from the codebook construction. With this notation, note that the number of microphones M is equal to $\|P\|_0$ [2]. Thereby, the source-sensor localization problem can be converted into a linear regression and the solution is formulated as

$$\begin{aligned} \hat{P} = \arg\min\{\|\mathscr{X}(F) - \mathscr{C}(F)P\|_2^2\} \quad \text{s.t.} \quad \|P\|_0 = M, \\ \mathscr{C}(F) = \phi_{\nu_g, \mathscr{L}}(F)\phi^{\dagger}_{\nu_g, \mathscr{L}}(F)\mathscr{X}(F) \end{aligned} \tag{7.8}$$

The possible number of microphone positions is $\binom{G}{M}$. Assuming that a source could co-locate with the microphones, the total combinations of source-sensor positions adds up to $G\binom{G}{M}$ [3]. Corresponding to each of these $G\binom{G}{M}$ arrangements is an array manifold matrix and any one of these can represent the *true* manifold matrix for the microphone array. The complexity of this problem is very high so we take an iterative greedy sparse recovery approach.

The basic idea of a greedy pursuit algorithm to solve the problem stated in (7.8) is to find the location of a small subset of K-microphones at each iteration while the location of the source can be determined at the first iteration. If the location of $M - K$ of the sensors is known *a priori* and

2. The ℓ_0(pseudo)-norm of P is defined as the number of non-zero elements in the vector

3. In case of having the source being co-located with the microphone, a closed form exact solution to microphone localization problem exists which is explained in Section 7.4

only K sensor locations are unknown, then the choice of possible manifold matrices reduces to $\binom{G}{K}$. In the discussion below we assume $K = 1$ for simplicity, but the argument is easily extended to higher values of K. Given the multi-channel signal recording $\mathscr{X} \in \mathbb{C}^{MB\times 1}$ and assuming that the position of $M-1$ of the microphones are known, we construct a codebook denoted by $\mathscr{C} \in \mathbb{C}^{MB\times GB}$, composed from projections of $\mathscr{X}$ onto G array manifold matrices as given by (7.2). The g^{th} manifold matrix corresponds to a microphone array with $M-1$ microphones at known positions and the M^{th} microphone at cell g. Since the support of P corresponds to the location of the microphone on the grid, it is a 1-*sparse* vector.

Given the array observations and the codebook of the signals projected onto the manifold matrices corresponding to G cells, the source-sensor localization problem amounts to sparse approximation of P. The solution to (7.8) finds the location of the microphones one by one by taking into account the already localized microphones; however, it generalizes trivially to the case of K unknown microphone locations. In the following Section 7.3, we elaborate on the codebook construction and the greedy sparse recovery approach for source-sensor localization.

7.3 Greedy Algorithm for Source-Sensor Localization

The design of the code book $\mathscr{C}$ is based on the reconstruction of the acoustic field from multi-channel recordings. Consider a source signal S_g at an unknown location, which is recorded by each of M microphones. Let the location of m^{th} microphone be μ_m. The signal X_m captured by the m^{th} microphone is obtained by passing S_g through the acoustic channel of the room from the source location ν_g to μ_m, $H(f,\nu_g,\mu_m)$. Hence, we have a linear model of the M microphone observations in spectral domain stated as

$$\begin{bmatrix} X_1(f) \\ \vdots \\ X_M(f) \end{bmatrix} = \begin{bmatrix} H(f,\nu_g,\mu_1) \\ \vdots \\ H(f,\nu_g,\mu_M) \end{bmatrix} S_g(f), \tag{7.9}$$

or, more succinctly, representing

$$\mathscr{X}(f) = [X_1^T(f)\cdots X_M^T(f)]^\top \qquad \phi_{\nu_g,\mathscr{L}}(f) = [H(f,\nu_g,\mu_1)\cdots H(f,\nu_g,\mu_M)]^\top$$

and as earlier, $\mathscr{X}(f) = \phi_{\nu_g,\mathscr{L}}(f)S_g(f)$. Like throughout the earlier chapters, we refer to this equation as the *forward model*. In order to characterize the forward model, we consider the recording environment to be a rectangular enclosure consisting of finite-impedance walls. The point source-to-microphone impulse responses $H(f,\nu_g,\mu_m)$ are calculated using the Image method. The details of characterizing the room acoustic channel are elaborated in Chapter 5. We rely on the model presented in this thesis and consider the room acoustic channel being estimated.

Assuming we know the locations of the source and $M-1$ microphones and only the M^{th} microphone must be located, there are G possible valid array configurations to consider in the construction of the codebook $\mathscr{C}$ in (7.8). We compose the corresponding set of array manifold

matrices $\phi_{\nu_g,\mu_1}(F), \phi_{\nu_g,\mu_2}(F), \cdots, \phi_{\nu_g,\mu_G}(F)$, where $\phi_{\nu_g,\mu_g}(F)$ represents the manifold matrix for the array configuration where the first $M-1$ microphones are in their known locations, and the g^{th} microphone is at μ_g, $\forall g \in \{1,\cdots,G\}$. We now write the codebook as

$$\mathscr{C} = [\phi_{\nu_g,\mu_1}\phi^{\dagger}_{\nu_g,\mu_1}(F)\mathscr{X}(F), \ldots \phi_{\nu_g,\mu_G}\phi^{\dagger}_{\nu_g,\mu_G}(F)\mathscr{X}(F)]. \tag{7.10}$$

The $\mathscr{X}$ in localization model of (7.8) must correspondingly be taken to actually represent $\mathscr{X}(F)$. P is now a $GB \times 1$ vector, with the property that it is B-*sparse* with a block structure: at most B consecutive entries beginning at index Bg can be non-zero. Figure 7.1 illustrates an example of the codebook of the spatial signals.

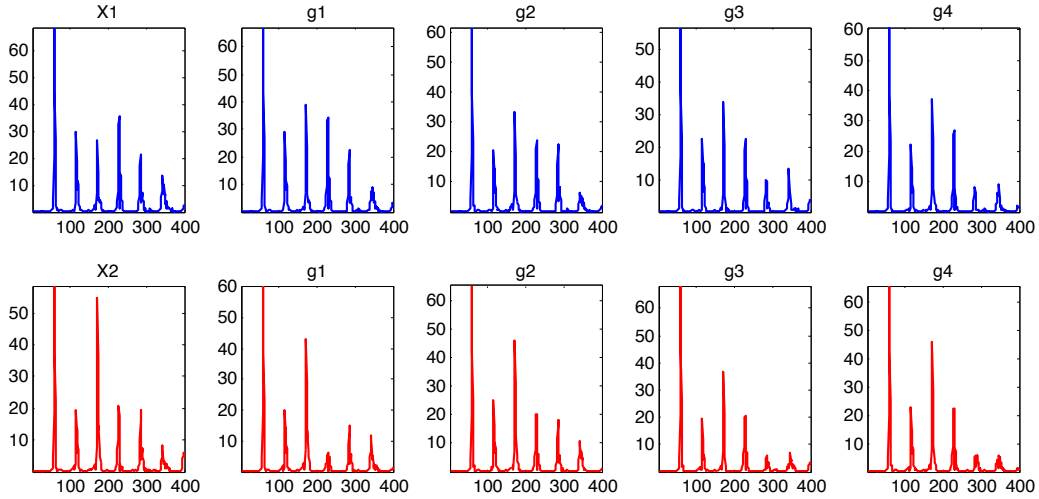

Figure 7.1 – An example of four signals in the codebook for a two-channel recording $[X_1\ X_2]$, projected onto four grid points g1, g2, g3 and g4. The x-axis is the frequency index with the resolution of 3.9Hz per band. The y-axis is the magnitude of the speech spectrum.

The summary of the localization procedure is given in Algorithm 2. The sparse coding model expressed in (7.8) indicates that once the codebook is constructed of all the spatial projections of the multi-channel signals, finding the unknown locations amounts to sparse approximation of the encoding vector P which selects the projections corresponding to the right manifold matrix. Since the codebook is constructed of F adjacent frequencies, the non-zero components of P has a block structure corresponding to the common support/cell where the unknown microphone is located. To incorporate the underlying structure of the sparse coefficients, we use the block sparse recovery algorithm proposed in [Kyrillidis and Cevher, 2011] which is an accelerated scheme for hard thresholding methods with the following recursion

$$P_{i+1} = \mathscr{M}\left(P_i + \kappa\mathscr{C}^{\top}(\mathscr{X} - \mathscr{C}P_i)\right), \tag{7.11}$$

where the step-size κ is the Lipschitz gradient constant to guarantee the fastest convergence speed. To incorporate for the underlying block structure, the model projection operator $\mathscr{M}$

thresholds and retains only the one (or more generally K) B-block with the highest energy, with subsequent renormalization [Kyrillidis and Cevher, 2011]. The support of estimated $\hat{P}$ determines the microphone location. To estimate the *1-sparse* solution, it is also possible to find the combinatorial solution of (7.8) through a linear search. We performed the combinatorial optimization during the experimental analysis and the results were similar to what we obtained by hard thresholding expressed in (7.11).

Algorithm 2 A Greedy pursuit algorithm for source-sensor localization

1. Initialize $K = 2$ as the number of microphones used for codebook construction.
2. Construct the codebook of spatial signal projections for K microphones and one source.
3. Find the *1-sparse* solution to (7.8) for $M = K$ and fix the source location.
4. Choose another microphone at an unknown position and construct the codebook and set $K = K + 1$.
5. Find the *1-sparse* solution to (7.8) for $M = K$.
6. Repeat 4 and 5 until all the microphones are localized.

Once the locations are estimated, we can obtain the source estimate as $\hat{S}(f) = \phi_{\nu_g,\mathscr{L}}(f)^{\dagger} X(f)$. Localization of multiple sources is possible when the full network of microphones are localized [4]; the codebook may be constructed as 7.10 assuming a source being present at any of the cells [Asaei et al., 2012a]. Alternatively, the model-based sparse component analysis framework as formulated in Chapter 3 can be applied for localization of the speakers (as studied in Section 6.4.2). Given the location of the sources and assuming that the number of sources is smaller than the number of microphones, we can estimate the signals by inverse filtering the acoustic channel. In Section 7.5, we provide rigorous analysis of the optimality of multiparty speech recovery by showing the equivalence of inverse filtering to speech separation followed by channel deconvolution.

7.4 Exact Closed-form Solution

If the source signal is known, for instance the source is co-located with one of the microphones, we can formulate the microphone array localization as follows

$$X = C\ \phi_{\nu_g,\mathscr{L}}\ S \tag{7.12}$$

where $X = [X_1 \cdots X_M]^T$, $\phi_{\nu_g,\mathscr{L}} \in \mathbb{C}^{G \times F}$ and $S \in \mathbb{C}^{F \times F}$. Matrix $C \in \{0,1\}^{M \times G}$ is a matrix consisted of binary values with a single 1 at each row and all the other components equal to 0 where the column indices are exclusive; hence, $\|C\|_0 = M$ and $C_{i,j} = 1$ selects one of the G cells where a microphone is located. For example, if we have two microphones located at cells 10 and 20, only $C_{1,10}$ and $C_{2,20}$ are 1. The formulation stated in (7.12) considers a rearrangement of the spectral coefficients for the frequency components constituting the rows of the matrices (as

4. There is also no algorithmic impediment to generalize the formulation presented in Sections 7.2-7.4 for multiple sources.

opposed to vector concatenation in (7.5)).

Matrix $\phi_{\nu_g,\mathscr{L}}$ is consisted of the room acoustic channel corresponding to the G microphone cells and the source location ν_g. We assume to process F frequencies of the source hence, columns of $\phi_{\nu_g,\mathscr{L}}$ are obtained from frequency-dependent Green's function propagation formula as defined in (7.4) and S is the diagonal matrix consisted of F frequency components of the source. Given this formulation of the microphone localization problem, we can find the exact solution for C as follows

$$\begin{aligned} X^T &= S^T \phi_{\nu_g,\mathscr{L}}^T C^T \\ C^T &= (S^T \phi_{\nu_g,\mathscr{L}}^T)^{-1} X^T \\ C &= X (S^T \phi_{\nu_g,\mathscr{L}}^T)^{-T} \end{aligned} \tag{7.13}$$

Therefore, matrix C is exactly calculated which encodes the location of the microphones. Localization of the microphones is also referred to as *microphone calibration* [Asaei et al., 2012d].

7.5 Equivalence to Speech Separation and Dereverberation

Given the source-sensor locations, speech recovery can be achieved by inverse filtering the acoustic channel. This section shows that this operation is equivalent to a two stage procedure: *speech separation* followed by *speech dereverberation*. For the sake of brevity, the proof is explained for two sources recorded by three sensors. The generalized proof is provided in Appendix A.

Consider the multiple input multiple output (MIMO) finite impulse response acoustic system which is represented in matrix notation as

$$\underbrace{\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}}_{\mathscr{X}} = \underbrace{\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \\ H_{31} & H_{32} \end{bmatrix}}_{\mathscr{H}} \underbrace{\begin{bmatrix} S_1 \\ S_2 \end{bmatrix}}_{\mathscr{S}} \tag{7.14}$$

Hence, $\mathscr{S} = \mathscr{H}^\dagger \mathscr{X}$ where

$$\mathscr{H}^\dagger = \frac{1}{|\mathscr{H}^T \mathscr{H}|} \overbrace{\begin{bmatrix} \sum_{i=1}^3 H_{i2}^2 & \sum_{i=1}^3 H_{i1} H_{i2} \\ \sum_{i=1}^3 H_{i1} H_{i2} & \sum_{i=1}^3 H_{i1}^2 \end{bmatrix} \begin{bmatrix} H_{11} & H_{21} & H_{31} \\ H_{12} & H_{22} & H_{32} \end{bmatrix}}^{\mathscr{A}}$$

$$\mathscr{A} = \begin{bmatrix} H_{11}\sum_{i=1}^3 H_{i2}^2 - H_{12}\sum_{i=1}^3 H_{i1}H_{i2} & H_{21}\sum_{i=1}^3 H_{i2}^2 - H_{22}\sum_{i=1}^3 H_{i1}H_{i2} & H_{31}\sum_{i=1}^3 H_{i2}^2 - H_{32}\sum_{i=1}^3 H_{i1}H_{i2} \\ -H_{11}\sum_{i=1}^3 H_{i1}H_{i2} + H_{12}\sum_{i=1}^3 H_{i1}^2 & -H_{21}\sum_{i=1}^3 H_{i1}H_{i2} + H_{22}\sum_{i=1}^3 H_{i1}^2 & -H_{31}\sum_{i=1}^2 H_{i1}H_{i2} + H_{32}\sum_{i=1}^3 H_{i1}^2 \end{bmatrix}$$

Thereby, S_1 is obtained by

$$\begin{aligned} S_1 &= \frac{1}{|\mathscr{H}^\top \mathscr{H}|} \sum_{j=1}^{3} \left(H_{j1} \sum_{i=1}^{3} H_{i2}^2 - H_{j2} \sum_{i=1}^{3} H_{i1} H_{i2} \right) X_j \\ &= \frac{1}{|\mathscr{H}^\top \mathscr{H}|} \big[H_{22}(H_{11}H_{22} - H_{12}H_{21}) + H_{32}(H_{11}H_{32} - H_{12}H_{31}) \big] X_1 + \\ &\qquad \big[H_{12}(H_{21}H_{12} - H_{22}H_{11}) + H_{32}(H_{21}H_{32} - H_{22}H_{31}) \big] X_2 + \\ &\qquad \big[H_{12}(H_{31}H_{12} - H_{32}H_{11}) + H_{22}(H_{31}H_{22} - H_{32}H_{21}) \big] X_3. \end{aligned} \tag{7.15}$$

We show that the equation on the right-hand side can be factorized such that

$$S_1 = F^{-1} G \mathscr{X}, \quad F = \begin{bmatrix} H_{22}H_{11} - H12H_{21} \\ H_{32}H_{11} - H12H_{31} \\ H_{32}H_{21} - H_{22}H_{31} \end{bmatrix}, \quad G = \begin{bmatrix} H_{22} & -H_{12} & 0 \\ H_{32} & 0 & -H_{12} \\ 0 & H_{32} & -H_{22} \end{bmatrix} \tag{7.16}$$

The factorized system is shown in Figure 7.2. Defining $\mathscr{Y} = G\mathscr{X}$, we obtain

$$Y_{S_1,1} = H_{22}X_1 - H_{12}X_2, \quad Y_{S_1,2} = H_{32}X_1 - H_{12}X_3, \quad Y_{S_1,3} = H_{32}X_2 - H_{22}X_3$$

Thereby, S_2 is separated from the three channel outputs Y_1, Y_2, Y_3 by the first linear operation $G\mathscr{X}$. The second operation performs channel deconvolution hence,

$$S_1 = \frac{1}{|F|} \big[(H_{22}H_{11} - H_{12}H_{21}) Y_{S_1,1} + (H_{32}H_{11} - H_{12}H_{31}) Y_{S_1,2} + (H_{32}H_{21} - H_{22}H_{31}) Y_{S_1,3} \big]$$

Substituting Y_1, Y_2, Y_3 with their values, we obtain

$$\begin{aligned} S_1 = \frac{1}{|F|} &\big[H_{22}(H_{11}H_{22} - H12H_{21}) + H_{32}(H_{11}H_{32} - H_{12}H_{31}) \big] X_1 + \\ &\big[H_{12}(H_{21}H_{12} - H_{22}H_{11}) + H_{32}(H_{32}H_{21} - H_{22}H_{31}) \big] X_2 + \\ &\big[H_{12}(H_{31}H_{12} - H_{32}H_{11}) + H_{22}(H_{31}H_{22} - H_{32}H_{21}) \big] X_3 \end{aligned} \tag{7.17}$$

We see that the nominator is equivalent to the nominator in (7.15). We can further see that the determinant of F is

$$|F| = (H_{22}H_{11} - H_{12}H_{21})^2 + (H_{32}H_{11} - H_{12}H_{31})^2 + (H_{32}H_{21} - H_{22}H_{31})^2$$

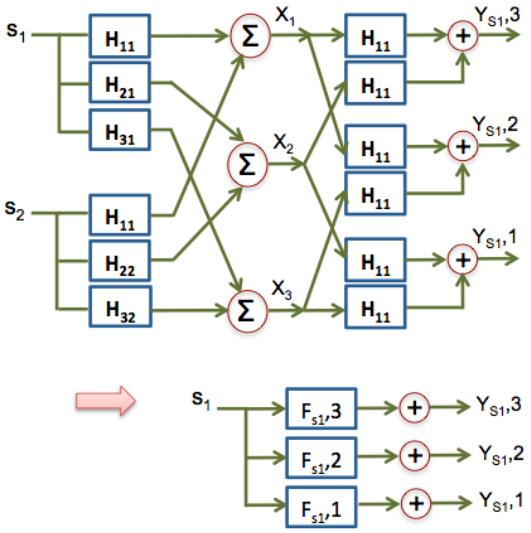

Figure 7.2 – Source separation filters for a MIMO system: $F_{S_1,1}$, $F_{S_1,2}$ and $F_{S_1,3}$ are calculated from the acoustic channel corresponding to multiple sources as stated in (7.16)

which is equivalent to determinant of $H^T H$ as

$$\begin{aligned}|H^T H| &= (H_{11}^2 + H_{21}^2 + H_{31}^2)(H_{12}^2 + H_{22}^2 + H_{32}^2) - (H_{11}H_{12} + H_{21}H_{22} + H_{31}H_{32})^2 \\ &= \underbrace{H_{11}^2 H_{22}^2 + H_{21}^2 H_{12}^2 - 2H_{11}H_{12}H_{21}H_{22}}_{(H_{22}H_{11} - H_{12}H_{21})^2} \\ &+ \underbrace{H_{11}^2 H_{32}^2 + H_{31}^2 H_{12}^2 - 2H_{11}H_{12}H_{31}H_{32}}_{(H_{32}H_{11} - H_{12}H_{31})^2} \\ &+ \underbrace{H_{21}^2 H_{32}^2 + H_{31}^2 H_{22}^2 - 2H_{11}H_{22}H_{31}H_{32}}_{(H_{32}H_{21} - H_{22}H_{31})^2} = |F|^2\end{aligned}$$

Similar to what was shown for S_1, the same equality holds for S_2. The equivalence of inverse filtering to the two stage procedure (1) separation of interference signals and (2) deconvolution of the signal from the early room impulse response function enables analysis of the stability and performance of the system [Huang et al., 2005]. Having the interferences separated as $\mathscr{Y} = G\mathscr{X}$, multiple interference-free speech signals are obtained $Y_{S_1,1}$, $Y_{S_1,2}$, $Y_{S_1,3}$ which are corresponding to the following acoustic paths:

$$\begin{aligned}F_{S_1,1} &= H_{22}H_{11} - H_{12}H_{21} \\ F_{S_1,2} &= H_{32}H_{11} - H_{12}H_{31} \\ F_{S_1,3} &= H_{32}H_{21} - H_{22}H_{31}\end{aligned} \tag{7.18}$$

Hence, the separated signals can sound more reverberant due to the prolonged impulse response of the equivalent channels. To perform exact deconvolution of the signals, Bezout theorem for

dereverberation of an acoustic system is exploited in [Miyoshi and Kaneda, 1988] referred to as multiple-input/output inverse-filtering (MINT) theory of acoustic channel deconvolution. The Bezout theorem is mathematically expressed as follows:

$$\begin{aligned} &\gcd[F_{S_m,1}, F_{S_m,2}, ..., F_{S_m,P}] = 1 \\ &\quad \Leftrightarrow \exists U_{S_m,1} U_{S_m,2} ..., U_{S_m,P} : \sum_{p=1}^{P} F_{S_m,p} U_{S_m,p} = 1. \end{aligned} \tag{7.19}$$

where gcd[.] denotes the greatest common divisor of the involved polynomials. According to the Bezout theorem, if the polynomials $F_{S_m,p}\,(p = 1,2,..,P)$ have no common divisor or equivalently acoustic channel responses H_{nm}, $n = 1,2,...,N$ do not share any common zeros, it is possible to equalize each of the M channels. Therefore, having multiple channel responses eliminates the requirement on minimum-phase impulse response for perfect dereverberation. The considerations raise when channel responses have common zeros. This can happen if $C_{S_m} = \gcd[H_{12}, H_{22}, H_{32}] = \gcd[F_1, F_2, F_3] \neq 1$; having common zeros, the theorem can only partially dereverberate the speech signal up to the polynomial C_{S_m}.

Assuming that the channel responses do not have common zeros, we can deconvolve the signal from the early room impulse response function through inverse filtering [Miyoshi and Kaneda, 1988, Asaei et al., 2013b]. The late reverberation can be statistically modeled as an exponentially decaying white Gaussian noise [Habets, 2010] which also possess the diffuse characteristics [Cook et al., 1955a, McCowan and Bourlard, 2003]. To reduce the effect of late reverberation and enhance the signal, we can apply the post-processing techniques. Among several post-filtering methods proposed in the literature [Wolff and Buck, 2010, McCowan and Bourlard, 2003], the Zelinski post-filtering is a practical implementation of the optimal Wiener filter; while a precise realization of the later requires knowledge about the spectrum of the desired signal, the Zelinski post-filtering method uses the auto- and cross-power spectra of the multi-channel input signals to estimate the target signal and noise power spectra under the assumption of *zero cross-correlation* between noise on different sensors. We implemented the Zelinski post-filter for the experiments conducted in Section 6.4.3. The dereverberation of the early impulse response achieved by inverse filtering the acoustic channels enables a more efficient post-filtering as studied in Section 7.6.3.

7.6 Experimental Analysis

The experiments are carried out using the real data overlapping corpus collected for MONC [Mccowan, 2003]. We evaluate the source-sensor localization performance as well as quality and recognition rate of the recovered speech. In addition to the real data evaluations, some experiments are carried out on synthesized room impulse responses in a controlled set-up to derive an empirical performance bound for the greedy source-sensor localization procedure. Further analysis are performed on the importance of the structured sparsity models.

7.6.1 Overlapping Speech Database

The database was collected by outputting utterances from the Numbers corpus (telephone quality speech, 30-word vocabulary) on one or more loudspeakers, and recording the resulting sound field using a microphone array and various lapel and table-top microphones. The energy levels of all utterances in the Numbers corpus were normalized to ensure a relatively constant desired speech level across all recordings. The corpus was then divided into 6049-utterances as the training set, 2026-utterances as the cross validation set, and 2061-utterances as the test set. *Competing-speakers* of the cross-validation and test sets were also produced by rearranging the order of their respective utterances. The word loop grammar is used and the task is speaker independent. The acoustic of the enclosure is mildly reverberant and the average SNR is 10 dB for single speaker recordings. The speech signals are recorded at 8 kHz sampling frequency. The geometrical set-up of the recordings is described earlier in Section 5.3.1.

7.6.2 Source-Sensor Localization Performance

The single speaker utterances are used for sensor localization evaluations. The spectro-temporal representation required for speech recovery is obtained by windowing the signal in 256 ms frames using a Hann function with 25% overlap. To perform localization of 8-channel microphones, the greedy pursuit algorithm summarized in Table 2 is used. First, the position of two broadside microphones and the source are estimated. Thereafter, the location of the other microphones are estimated one per iteration. To increase the resolution of the estimates while keeping the dimensionally of the sparse vector bounded, a coarse-to-fine strategy is taken [Malioutov et al., 2005] . More specifically, the area is discretized into 5cm cells. The localized microphones are then re-located in 1cm accuracy using a finer discretization. The average norm of calibration error for the relative geometry is 8.9mm. Given that the complexity of the combinatorial optimization increases as $\mathcal{O}(\mathrm{G}^{\mathrm{M}})$ whereas the greedy sparse recovery has a complexity of $\mathcal{O}(\mathrm{GM})$, it is crucial to employ the sparse recovery algorithms to enable microphone array calibration in our set-up.

In addition, the method proposed in [McCowan et al., 2008] is implemented and used for calibration of the circular array. This method relies on diffuse noise model to find the topology of the array. The results obtained for localization of circular microphones is about 1.2cm using about 10s recording of diffuse noise field. In practice however, the level and length of the available diffuse noise might be challenging to employ the technique proposed in [McCowan et al., 2008]. Hence, our proposed approach which requires only a few speech frames (less than 1s) provides a higher applicability and accuracy.

Empirical Performance Bound

To derive an empirical performance bound, some experiments are conducted on synthetic data recordings using ad-hoc microphones distributed in a 0.4m$\times$0.4m area as illustrated in Figure 4.3. The reference point speaker is located at either 0.5m or 1.5m distance to the center of the

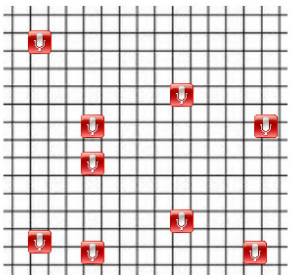

Figure 7.3 – Set-up for microphone placement

grid. We considered a 3m×3m×3m room and synthesized the room impulse responses with the Image model [Allen and Berkley, 1979] with reflective factors of 0.8 for the six walls, which corresponds to 180ms reverberation time according to Eyring's formula [Eyring, 1930] [5]

$$\beta = \exp(-13.82/[c(L_x^{-1} + L_y^{-1} + L_z^{-1})T]), \tag{7.20}$$

where L_x, L_y and Lz are the room dimensions, τ is the speed of sound in the air ($\approx 342 m/s$) and T is the room reverberation time.

We consider the scenario in which the recording condition is perfectly known. To evaluate the sensitivity of the approach to the uncertainties in estimation of the forward model parameters (i.e. errors in estimation of room geometry and absorption coefficients), we consider two mismatched test conditions. In the first scenario (Mismatched1), the observations were generated with a forward model where the codebook is constructed of the spatial projections using a model with up to 25% error in the absorption coefficients corresponding to each of the walls. In the second scenario (Mismatched2), we assume that the room geometry is estimated with an error of 10cm and the absorption coefficients are estimated with 25% error on each of the six reflective walls. The performance of the microphone calibration in terms of root mean squared error (RMSE) is listed in the Table 7.1. The parameter δ indicates the resolution of the grid which is in our case equal to 5cm. We considered all pairs of combinations (denoted by Pairs) to quantify an average expected error to localize the first two-microphones. Then the third microphone is selected for localization and so on until all the microphones are localized. Alternatively, we considered all triples of combinations (denoted by Triples) to quantify an average expected error to localize the first two-microphones. Then the forth microphone is selected for localization and so on until all the microphones are localized.

Importance of Structured Sparsity

Relying on the formulation of sparse coding framework expressed in (7.8), the theoretical analysis of the performance of our approach amounts to the greedy recovery guarantees and it is tied to the properties of the codebook matrix $\mathscr{C}$ [Tropp and Wright, 2010]. A fundamental property of $\mathscr{C}$

5. as stated earlier in Section 2.4.

Table 7.1 – RMSE (cm) of microphone array calibration. Two combinations are considered: Pairs and Triples. δ indicates the resolution of the grid and is equal to 5cm in our experiments

Acoustic condition	Combination	Source-dist = 0.5	Source-dist = 1.5
Match. & Mismatch.1	Pairs	δ	δ
	Triples	δ	δ
Mismatched2	Pairs	14.7	12.3
	Triples	14	6

is the *coherence* between the columns defined in Section 3.2.4. To guarantee the sparse recovery performance, it is desired that the coherence is minimized. Since the codebook is constructed of locations and frequency dependent Green's function projections, this property implies that the contribution of the source to array's response is small outside the corresponding sensor location or equivalently the resolution of the array is maximized. The studies described in Chapter 6 showed that the ad-hoc microphone arrays distributed randomly yield significant improvements in the sparse signal reconstruction performance. In a similar trend, the performance of our sparse recovery framework is entangled with the design of the cells for codebook construction as well as the frequency of the signal. In the discussion below, we investigate the frequency dependency. The grid construction considering non-uniform cells to minimize the coherence will be considered in our future research.

To analyze the codebook for the broadband speech spectrum, the coherence is computed for different frequency bands through (3.9). The results are illustrated in Figure 7.4.

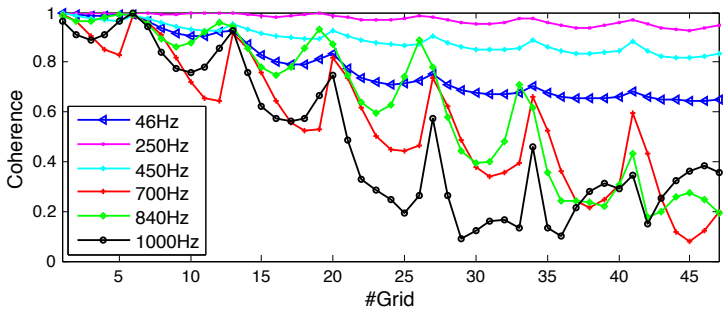

Figure 7.4 – Coherence of the codebook for different frequencies of speech spectrum

This study shows that the coherence of the codebook is smaller for higher frequencies and suggests sub-band processing of the speech signal. Alternatively, joint sparsity models enable us to reduce the ambiguity while exploiting the synergy of the broadband components. This issue is investigated in Figure 7.5.

The results indicate that processing the frequencies independently is more ambiguous due to the high coherence of the codebook over some components. In contrast, the block-sparsity model enables very sharp estimates as a function of the block size. Hence, incorporating joint sparsity models such as the block-dependency structure improves the recovery performance in sparse

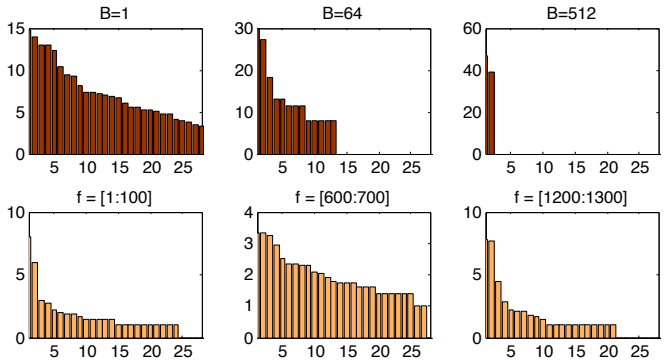

Figure 7.5 – Partial support of $\hat{\mathrm{P}}$; the effect of increasing the block-size B and frequencies to enable less ambiguous estimates of the block sparse vector, P.

modeling framework.

7.6.3 Speech Recovery Performance

Given the location of the sources and the characterized room acoustic channel, the desired signal can be recovered by inverse filtering to perform speech recognition [6]. The automatic speech recognition (ASR) scenario was designed to broadly mirror that of Moore and McCowan [Moore and Mccowan, 2003]. A typical front-end was constructed using the HTK toolkit [Cook et al., 1955b] with 25ms frames at a rate of 10ms. This produced 12 mel-cepstra plus the zeroth coefficient and the first and second time derivatives; 39 features in total. Cepstral Mean Normalization (CMN) is applied to the feature vectors, resulting in speech recognition performance improvement of about 15% relative. The back-end consists of 80 tied-state triphone HMM's with 3 emitting states per triphone and 12 mixtures per state. The ASR accuracy on the clean speech data is about 95%. We perform MAP adaptation by training directly on recovered data. The Zelinsky post-filtering is applied on the recovered speech prior to recognition [McCowan and Bourlard, 2003].

In addition to the speech recognition, we evaluate the quality of the recovered speech using SIR [Vincent et al., 2006] as well as PESQ [ITU-T, 2001]. As the speech recovery approach relies on the principles of spatial diversity, we perform comparison with beamforming which possesses similar essence. We use the super-resolution speaker localization based on sparse recovery to perform near-field beamforming. Room acoustic modeling is achieved through our method explained in Section 5.2.2. As the results indicate, the proposed approach relying on inverse filtering the Room Acoustic Model followed by Post-Filtering (RAM-PF) yields the maximum interference suppression and highest perceptual quality of the recovered speech in multi-party scenarios as quantified in terms of SIR and PESQ. It also outperforms other techniques in terms of

6. The experiments in this Section particularly address the performance of our framework if the number of microphones is equal to or more than the number of competing sources, i.e., $\mathrm{M} \geqslant \mathrm{N}$. If this is not the case, the speech signals are recovered through model-based sparse recovery algorithms; the subject studied considering various algorithmic approaches in Chapter 6.

Table 7.2 – Quality evaluation of the recovered speech in terms of Source to Interference Ratio (SIR), Perceptual Evaluation of Speech Quality (PESQ) and Word Recognition Rate (WRR) using Super Directive (SD) beamforming, vs. inverse filtering using Room Acoustic Modeling (RAM) as formulated in Section 5.2.2. The Zelinski Post-Filtering (PF) is applied after speech recovery. N denotes the number of concurrent sources.

N	Meas.	Baseline	Lapel	SD	SD-PF	RAM	RAM-PF
1	SIR	12.3	19.19	18.5	18.52	16.5	16.1
	PESQ	2.7	3	3.3	3.3	2.91	2.97
	WRR%	89.61	93.21	95	95	93.67	93.3
2	SIR	2.6	18.29	11.8	11.33	12.5	17.5
	PESQ	2	2.35	2.6	2.69	2.6	2.8
	WRR%	55.19	74.53	70.19	68.16	83.37	87.93
3	SIR	-0.7	18.35	10.2	10	10.1	14.2
	PESQ	1.6	2.27	2.4	2.48	2.4	2.62
	WRR%	39.92	68.13	63	61.45	70.88	79.21

word recognition rate. The Zelinski post-processing is derived to reduce the effect of uncorrelated noise. We can observe that the improvement in performance obtained after deconvolution of the room acoustic channel is higher than what we can achieve after standard beamforming.

7.7 Conclusions

This chapter presented a generalized formulation of source-sensor localization relying on structured sparse coding framework. The problem of localization of the microphones as the dual of source localization is particularly addressed. A greedy strategy is suggested to localize a source and a large network of microphones using greedy pursuit sparse reconstruction. Furthermore, a closed-form exact solution is derived if the source is co-located with one of the microphones thus the original signal is known.

The structured sparse coding framework exploits the knowledge about the room acoustic channel and speech recovery is possible through inverse filtering. The equivalence of inverse filtering to speech separation and deconvolution was theoretically proved along with some discussions about the optimality of the solution if the room impulse responses do not have common zeros. The experimental analysis verified the applicability of the proposed method and importance of structured sparse recovery using real data recordings.

There are two fundamental aspects to the success of our framework namely, incorporating the acoustic multipath model and sparsity structures. This information can be integrated with the optimum spatial filtering objective to derive an effective formulation of high quality signal acquisition relying on beamforming. We investigate this subject in the following Chapter 8.

8 Optimum Spatial Filtering

The model-based sparse component analysis framework incorporates the prior information on *structured sparsity models* (Chapter 4) and characterization of the *acoustic multipath projections* (Chapter 5) to obtain the best estimate of the spatio-spectral components matching the microphone array observations. Therefore, the model-based sparse recovery algorithms perform optimization in the observation space. Alternative to this objective, we can optimize the prediction error of the signal which has been the fundamental concept of spatial filtering techniques. Hence, the goal of this chapter is to incorporate the prior information on *structured sparsity models* and *multipath projections* to yield an optimum beamforming formulation.

8.1 A Multipath Sparse Beamforming Method

The model-based sparse component analysis framework stated in Chapter 3 assumes a discrete Euclidean geometry for source locations. We consider similar set-up to derive the formulation of beamforming for spatio-spectral information recovery. The objective is to characterize an optimum beamformer which takes into account the acoustic multipath. Recall from Chapter 3 that the microphones observation vector can be expressed as $\mathscr{X} = \Phi\mathscr{S}$ while $\Phi \in \mathbb{C}^{\mathrm{M}\times\mathrm{G}}$ denotes the array manifold matrix obtained from the generative forward model of the acoustic multipath. For the sake of brevity, we have omitted the frequency index. We will adopt this convention whenever no confusion can arise. The sparse recovery approach is a non-linear framework for joint localization and speech recovery. In this section, we elucidate how spatial filtering can be formulated to achieve the similar objective while taking into account the multipath acoustic. There are several alternatives for designing the beamformers. We focus our attention on minimum variance unbiased estimate of the signal using MVDR as well as the minimum mean square error beamformer. The former serves as a key component in various beamforming structures whereas the later formulates the optimum signal estimation. We assume that the noise is a sample function of a random process with known second-order statistics and it has similar characteristics at all sensors. In the subsequent sections, we derive the formulation to recover a single desired source signal [Asaei et al., 2013c]. This framework can be easily generalized to multiple source signals.

8.1.1 Minimum Variance Distortionless Response Beamformer

The conventional MVDR beamformer performs weighted combination of the microphone recordings to enable separation of the signals and interference on the basis of their spatial characteristics. The spatial filtering operation generates the output as $Y = W^{\mathrm{H}}\mathscr{X}$ where $.^{\mathrm{H}}$ stands for conjugate transpose; the unknown filter weights $W \in \mathbb{C}^{M\times 1}$ are optimized in order to minimize the overall noise and interference power. To guarantee that the signal coming from the desired direction is received without distortion, the optimization is performed subject to the distortionless constraint as follows

$$W^{\mathrm{H}}_{\mathrm{MVDR}} = \underset{W^{\mathrm{H}}}{\arg\min}\, \mathrm{E}\{W^{\mathrm{H}}\mathscr{X}\}, \quad \text{s.t.} \quad W^{\mathrm{H}}\mathrm{d} = 1 \tag{8.1}$$

where E denotes the expectation operation and d is the steering vector for a plane wave coming from the desired direction. The solution to this criterion using a Lagrange multiplier is given by [Trees, 2002]

$$W^{\mathrm{H}}_{\mathrm{MVDR}} = \frac{\mathrm{d}^{\mathrm{H}}\mathrm{R}^{-1}_{\mathscr{X}}}{\mathrm{d}^{\mathrm{H}}\mathrm{R}^{-1}_{\mathscr{X}}\mathrm{d}} \tag{8.2}$$

where $\mathrm{R}_{\mathscr{X}}$ denotes $\mathrm{E}\mathscr{X}\mathscr{X}^{\mathrm{H}}$. The criterion is also known as minimum power distortionless (MPDR) beamformer. In the original MVDR formulation, R_{n} is replaced with $\mathrm{R}_{\mathscr{X}}$, assuming that there are no model-errors, both criteria coincide so we refer to this criteria as MVDR. If R_{n} is available, the MVDR beamformer offers significant robustness to errors in steering vectors (direction of the desired source) estimation [Ehrenberg et al., 2010]. The MVDR beamformer yields a Maximum Likelihood (ML) solution assuming that the frequencies are independent and the signal wavenumber (or direction of arrival (DOA)) is known. However, the MVDR filter scanned in wavenumber space in most cases does not provide the ML estimate of the signal's DOA [Trees, 2002].

Acoustic-Informed Sparse MVDR Beamformer

The conventional beamformer assumes that direction of the desired source is known. To incorporate the multipath effect, we assume that the acoustic channel corresponding to the desired source [1] is known and denoted by ϕ_{s}. Hence, a Multipath MVDR (M-MVDR) beamformer is expressed as

$$W^{\mathrm{H}}_{\text{M-MVDR}} = \underset{W^{\mathrm{H}}}{\arg\min}\{W^{\mathrm{H}}\mathrm{R}_{\mathscr{X}}W\} \quad \text{s.t.} \quad W^{\mathrm{H}}\phi_{\mathrm{s}} = 1, \tag{8.3}$$

1. Room impulse response functions between the source and microphone array

The weights of M-MVDR are optimized such that acquisition of the signal with respect to the desired channel is distortionless; using a Lagrange multiplier, the optimal weights are given by

$$W_{\text{M-MVDR}}^{\text{H}} = \frac{\phi_{\text{s}}^{\text{H}} \text{R}_{\mathscr{X}}^{-1}}{\phi_{\text{s}}^{\text{H}} \text{R}_{\mathscr{X}}^{-1} \phi_{\text{s}}} \tag{8.4}$$

Given the forward model of the reverberant acoustic characterized as Φ, an additional constraint on the weights of an acoustic-informed beamformer is suggested as

$$W^{\text{H}}(\Phi \text{B}) = 1 \tag{8.5}$$

where B is a binary sparse vector whose support corresponds to the source location and selects the acoustic projection affecting the desired source. Therefore, estimation of W^{H} amounts to estimating the vector B. If we ignore the amplitude distortion due to multipath propagation, estimation of the filter weights is achieved by the following optimization

$$\hat{\text{B}} = \arg\min_{\text{B}} \{\text{B}^{\text{H}} \Phi^{\text{H}} \text{R}_{\mathscr{X}} \Phi \text{B}\} \quad \text{s.t.} \quad \text{B}^{\text{H}} \Phi^{\text{H}} \phi_{\text{s}} = 1 \tag{8.6}$$

where ϕ_{s} denotes the multi-path channel corresponding to the desired source. We now solve the problem by imposing the constraint using a Lagrange multiplier. The function that we minimize is

$$\text{F} \triangleq \text{B}^{\text{H}} \Phi^{\text{H}} \text{R}_{\mathscr{X}} \Phi \text{B} + \lambda [\text{B}^{\text{H}} \Phi^{\text{H}} \phi_{\text{s}} - 1] + \lambda^{\text{H}} [\phi_{\text{s}}^{\text{H}} W - 1] \tag{8.7}$$

Taking the complex gradient with respect to B^{H} and solving (8.7) gives

$$\text{B} = -\lambda \frac{\Phi^{\text{H}} \phi_{\text{s}}}{\Phi^{\text{H}} \text{R}_{\mathscr{X}} \Phi} \tag{8.8}$$

To evaluate λ, we use the distortionless constraint and obtain

$$\lambda = -\frac{\Phi^{\text{H}} \text{R}_{\mathscr{X}} \Phi}{\phi_{\text{s}}^{\text{H}} \Phi \Phi^{\text{H}} \phi_{\text{s}}} \tag{8.9}$$

Therefor, the multiparty MVDR (M-MVDR) solution is obtained as

$$\hat{\text{B}} = \frac{\Phi^{\text{H}} \phi_{\text{s}}}{\phi_{\text{s}}^{\text{H}} \Phi \Phi^{\text{H}} \phi_{\text{s}}}, \quad \hat{W}_{\text{M-MVDR}} = \frac{\Phi \Phi^{\text{H}} \phi_{\text{s}}}{\phi_{\text{s}}^{\text{H}} \Phi \Phi^{\text{H}} \phi_{\text{s}}} \tag{8.10}$$

Moreover, we can incorporate the prior information on sparse structure of B and regularize the optimization stated in (8.6) by the ℓ_1-norm of B expressed as follows

$$\hat{\text{B}} = \arg\min_{\text{B}} \{ \|\text{B}^{\text{H}} \Phi^{\text{H}} \mathscr{X}\|_2 + \lambda \|\text{B}\|_1 \quad \text{s.t.} \quad \text{B}^{\text{H}} \Phi^{\text{H}} \phi_{\text{s}} = 1 \} \tag{8.11}$$

The solution can be obtained by sparse recovery algorithm. The formulation of (8.11) enables better null steering so it offers a more robust solution when there are point interferences while

requiring fewer parameters to describe the solution than the conventional beamforming techniques. This formulation can be extended for diffuse noise field to obtain an acoustic-informed sparse superdirective beamformer which is an effective methodology for speech applications [Wolfel and McDonough, 2009].

8.1.2 Minimum Mean-Square Error Estimator

Alternative to MVDR, the linear array processor can be derived to estimate the signal using MMSE criterion. The weights of an MMSE beamformer are adapted to the reference signal S and obtained to minimize the average power in signal recovery error stated precisely as

$$\begin{aligned} W^{\mathrm{H}}_{\mathrm{MMSE}} &= \arg\min_{W^{\mathrm{H}}} \mathrm{E}\left\{|\mathrm{S} - W^{\mathrm{H}}\mathscr{X}|^2\right\}, \\ &= \arg\min_{W^{\mathrm{H}}} \mathrm{E}\left\{\mathrm{S}\mathrm{S}^{\mathrm{H}} - W^{\mathrm{H}}\mathscr{X}\mathrm{S}^{\mathrm{H}} - \mathrm{S}\mathscr{X}^{\mathrm{H}}W + W^{\mathrm{H}}\mathscr{X}\mathscr{X}^{\mathrm{H}}W\right\}, \end{aligned} \tag{8.12}$$

To find the minimum of this function, we take its derivative with respect to W^{H} and set the results equal to zero, thus we obtain

$$W^{\mathrm{H}}_{\mathrm{MMSE}} = \mathrm{R}_{\mathrm{S}\mathscr{X}^{\mathrm{H}}}\,\mathrm{R}^{-1}_{\mathscr{X}} \tag{8.13}$$

where $\mathrm{R}_{\mathrm{S}\mathscr{X}^{\mathrm{H}}} = \mathrm{E}\{\mathrm{S}\mathscr{X}^{\mathrm{H}}\}$. Given $\mathscr{X} = \mathrm{d}.\mathrm{S} + \mathrm{N}$ where d denotes the steering vector and having the uncorrelated assumption between signal S and noise N, the cross-correlation and auto-correlation matrices are

$$\mathrm{R}_{\mathrm{S}\mathscr{X}^{\mathrm{H}}} = \mathrm{R}_{\mathrm{S}}\,\mathrm{d}^{\mathrm{H}}, \qquad \mathrm{R}_{\mathscr{X}} = \mathrm{R}_{\mathrm{S}}\,\mathrm{d}\mathrm{d}^{\mathrm{H}} + \mathrm{R}_{\mathrm{n}} \tag{8.14}$$

This formulation requires acoustic and configuration stationarity assumption to obtain a reasonable estimate of the covariance of signal and noise. Using the matrix inversion formula to invert $\mathrm{R}_{\mathscr{X}}$, we have

$$\mathrm{R}^{-1}_{\mathscr{X}} = \mathrm{R}^{-1}_{\mathrm{n}} - \mathrm{R}^{-1}_{\mathrm{n}}\mathrm{R}_{\mathrm{S}}\,\mathrm{d}(1 + \mathrm{d}^{\mathrm{H}}\mathrm{R}^{-1}_{\mathrm{n}}\mathrm{R}_{\mathrm{S}}\,\mathrm{d})^{-1}\mathrm{d}^{\mathrm{H}}\mathrm{R}^{-1}_{\mathrm{n}}; \tag{8.15}$$

hence, the following solution is obtained [Trees, 2002]

$$W^{\mathrm{H}}_{\mathrm{MMSE}} = \underbrace{\frac{\mathrm{d}^{\mathrm{H}}\,\mathrm{R}^{-1}_{\mathrm{n}}}{\mathrm{d}^{\mathrm{H}}\,\mathrm{R}^{-1}_{\mathrm{n}}\,\mathrm{d}}}_{W^{\mathrm{H}}_{\mathrm{MVDR}}} \cdot \underbrace{\frac{\mathrm{R}_{\mathrm{S}}}{\mathrm{R}_{\mathrm{S}} + \mathrm{d}^{\mathrm{H}}\,\mathrm{R}^{-1}_{\mathrm{n}}\,\mathrm{d}}}_{\text{Wiener post filter}} \tag{8.16}$$

Hence, the MMSE estimator is a shrinkage of the MVDR beamformer which suppresses the interfering signals and noise without distorting the signal propagating along the desired source direction followed by a single-channel Wiener post-filter to yield the optimal signal estimation given the second-order statistics of the noise [Trees, 2002].

Acoustic-Informed Sparse MMSE Beamformer

Given the forward model characterization of the reverberant acoustic, a multipath beamformer can incorporate an additional constraint expressed as $W^{H}(\Phi B) = 1$ so estimating the beamformer weights can be obtained through the following optimization

$$\begin{aligned}\hat{B}^{H} &= \arg\min_{B^{H}} E\{\| S - B^{H}\Phi^{H}\mathscr{X} \|^{2}\}, \\ &= \arg\min_{B^{H}} \{R_{S} - B^{H}\Phi^{H}R_{\mathscr{X}S^{H}} - R_{S\mathscr{X}^{H}}\Phi B + B^{H}\Phi^{H}R_{\mathscr{X}}\Phi B\},\end{aligned} \tag{8.17}$$

where the solution is obtained by equating the derivative with respect to B^{H} to zero. Given $\mathscr{X} = \phi_{s}.S + N$ and assuming that signal and noise are uncorrelated, the cross-correlation and auto-correlation matrices are

$$R_{S\mathscr{X}^{H}} = R_{S}\phi_{s}^{H}, \qquad R_{\mathscr{X}} = R_{S}\phi_{s}\phi_{s}^{H} + R_{n} \tag{8.18}$$

Hence, the multipath MMSE beamformer is obtained as

$$\hat{B}^{H} = \frac{R_{S\mathscr{X}^{H}}\Phi}{\Phi^{H}R_{\mathscr{X}}\Phi}, \quad W^{H}_{\text{M-MMSE}} = \frac{R_{S}\phi_{s}^{H}\Phi\Phi^{H}}{\Phi^{H}(R_{S}\phi_{s}\phi_{s}^{H} + R_{n})\Phi} \tag{8.19}$$

where ϕ_{s} denotes the acoustic channel corresponding to the desired signal.

Moreover, we can incorporate the sparsity constraint in formulation of a multipath sparse MMSE estimator obtained through the following optimization stated in (8.20); the distortionless constraint prevents the trivial zero solution. The sparse MMSE beamformer would then be

$$\hat{S}, \hat{B} = \arg\min_{B,S} \{\|B^{H}\Phi^{H}\mathscr{X} - S\|_{2} + \lambda\|B\|_{1} \quad \text{s.t.} \quad B^{H}\Phi^{H}\phi_{s} = 1\} \tag{8.20}$$

This formulation can be extended to incorporate the structured sparsity models pertained to the representation of source signal S as described in Chapters 4 and 6 [Asaei et al., 2012c, 2013a].

Acoustic Calibration Beamforming

We can calibrate the model parameters of an acoustic-informed beamformer using a known signal at a known location. Recall from Chapter 5 characterization of the compressive acoustic projections. We proposed a novel formulation of the *reverberation model* factorized as $\Phi = OP$ where $O \in \mathbb{C}^{M\times\mathscr{G}}$ is the *free-space Green's function* matrix such that each $O_{j,i}$ component indicates the sound propagation coefficients, i.e. the attenuation factors and the phase shift due to the direct path propagation of the sound source located at cell i (on a $\mathscr{G}$-point grid of actual-virtual sources) and recorded at the j^{th} microphone. Given the $\mathscr{G}$-cell discretization, O is computed from the propagation formula stated in Equation (3.3) and it is equal to Φ when $R = 0$. The other term, $P \in \mathbb{R}_{+}^{\mathscr{G}\times G}$ is the permutation matrix such that its i^{th} column contains the absorption factors

of $\mathcal{G}$ points on the grid of actual-virtual sources with respect to the reflection of the i^{th} actual source. Since the Image model characterizes the actual-virtual source groups, each column $\mathrm{P}_{.,\mathrm{i}}$ is consequently supported only on the corresponding group Ω_{i} i.e., $\forall \mathrm{i} \in \{1\ldots,\mathrm{G}\}$, $\forall \mathrm{j}\ \ \Omega_{\mathrm{i}}, \mathrm{P}_{\mathrm{j},\mathrm{i}} = 0$. We can calibrate the acoustic model parameters through the optimization stated as

$$\hat{\Upsilon} = \arg\min_{\Upsilon} \{\|\Upsilon^{\mathrm{H}}\mathrm{O}^{\mathrm{H}}\mathscr{X} - \mathrm{S}\|_2 + \lambda\|\Sigma\|_{\mathscr{F}}\}, \tag{8.21}$$

where $\Upsilon = \mathrm{PB}$ and $\Sigma = \Upsilon\mathrm{S}\mathrm{S}^{\mathrm{H}}\Upsilon^{\mathrm{H}}$; $\mathscr{F}$ denotes either the joint sparsity norm, $\|.\|_{\mathrm{L}_1,\mathrm{L}_2}$, or nuclear norm,$\|.\|_*$, for structured sparsity or low-rank encoding [Asaei et al., 2012c, 2013a, Golbabaee and Vandergheynst]. The optimization objective formulated in (8.21), tunes the acoustic model parameters such that the expectation of the signal estimation error is minimized. Alternatively, we can derive a *calibration beamformer* as

$$\hat{W} = \arg\min_{W} \{\|W\mathscr{X} - \Phi\mathscr{S}\|_2 + \lambda\|\mathscr{S}\|_1\}, \tag{8.22}$$

where the calibration filter W is learned to minimize the deviation from the forward model $\Phi\mathrm{B}$.

The discerning reader may note that if the condition stated in (8.5) holds, i.e., $W^{\mathrm{H}}(\Phi\mathrm{B}) = 1$, the sparse MMSE estimator stated in (8.20) is equivalent to sparse recovery formulation expressed in Chapter 3 as

$$\hat{\mathscr{S}} = \arg\min_{\mathscr{S}} \|\mathscr{S}\|_1 \quad \text{s.t.} \quad \mathscr{X} = \Phi\mathscr{S} \tag{8.23}$$

In a general set-up however, the two sparse MMSE estimators, (8.20) and (8.23), have different objectives; the earlier optimizes the prediction error and the later optimizes the observation error. Hence, the two formulations may be regarded as alternative approaches with different solution space geometry to address the processing of multiparty recordings.

8.2 Experimental Analysis

The experimental analysis are carried out to study the performance of the proposed methods with two specific scopes presented in Section (8.2.1) on source localization and Section (8.2.2) on signal estimation.

8.2.1 Source Localization and Spatial Resolution

We start our experimental analysis with a computational study on microphone array beam-pattern and spatial resolution. The generic theory of compressive sensing enables quantitative assessment of the side-lobe level in terms of the number of sources, inter-element spacing, number of sensors and acoustic conditions [Carin, 2009]. The theory of the performance bounds was asserted in Chapter 3. We perform empirical demonstrations to study the performance of sparse recovery

and compare it with beamforming.

We consider an 8-channel microphone array. As stated in (3.10), to recover 2 sources, the coherence of the measurements is required to be less than 0.2. We calculate the coherence of the measurement matrix for a random linear array with inter-element spacing, equal to half of the signal wavelength. We assume that the sources are standing towards the array and start from a coarse angular division and increase the resolution below which the required value of the coherence is exceeded. The finest resolution is delineated as the beam-width [Carin et al., 2011]. This experiment is performed for anechoic as well as moderately reverberant acoustic conditions (room reverberation time is 270 ms). In addition, we compute the half-power beam-width of the beamformer as the angular deviation upon which the output power is reduced by 3 dB [Trees, 2002]. The results are listed in Table 8.1. The theoretical relationship between the MVDR beam-width and acoustic multipath is not explicitly defined in beamforming literature so we measured the reverberant beam-width empirically where the output power is reduced by half with respect to the look direction.

Table 8.1 – Calculated beam-width for sparse recovery and conventional beamforming

Acoustic Condition	Sparse Recovery	Beamforming
Anechoic	7°	12°
Reverberant	11°	30

The results show that the resolution at which the two sources are guaranteed to be distinguished is higher for the sparse reconstruction framework compared to the conventional beamforming. The larger aperture-size and random layout of microphone array render the compressive projections less coherent and better performance guarantees in sparse recovery whereas the conventional beamforming are designed for uniform sampling with a small inter-element spacing less than half of the system wave-length to prevent spatial aliasing.

The resolution of source localization is further investigated in another set of experiments. We assume that the array broadside is discretized uniformly and matrix Φ consists of all steering vectors in the DOA range of $[-90^\circ, 0^\circ) \cup (0^\circ, 90^\circ]$ with a sampling interval of 5°. The signal impinges on a uniform linear array consisted of 8 microphones at inter-element spacing equal to half of the wavelength. Randomly generated white Gaussian noises are added to the microphone measurements at signal to noise ratio (SNR=10dB) for evaluations in noisy condition. The MVDR beamformer stated in (8.10) is implemented. The results are shown in Figures 8.1-8.3. We can observe that the formulation of conventional MVDR has more sensitivity to noise and can not distinguish the sources closely located to each other.

8.2.2 Speech Recovery Performance

The initial speech recovery experiments are carried out on synthetic data to obtain the empirical performance bounds in a controlled well-defined scenario. Furthermore, we perform real data

evaluations on Multi-Channel Wall Street Journal (MC-WSJ) corpus [Lincoln et al., 2005].

Synthetic Data Evaluations

We conducted the experiments on two scenarios, *reverberant* and *far-field* set-up.

- ◇ The *reverberant* scenario is recorded using 8-channel circular microphone array with radius 10 cm and spatial resolution 50 cm within an enclosure of dimension $3 \times 3 \times 3$; $\mathrm{RT}_{60} = 180$ ms. The signals are 1000 speech frames sampled at 8 kHz and analyzed using Han function of length 64 ms and the weights are estimated per 10 frames.
- ◇ The *far-field* scenario is recorded using 8-channel uniform array with inter-element spacing equal to half of the wavelength and the directional resolution is 5°. The signals are 1000 trails of random samples of length 100. However, only 10% of data samples are used to compute the statistics of the conventional beamformer.

The signal to noise ratio is 20 dB. We assume that the signal and noise samples are known so the only uncertainty is attributed to the number of reliable samples available for beamforming. The CVX package is used for sparse beamforming optimization [Grant and Boyd]. Table 8.2 summarizes the results. The sparse recovery approach identifies the support of sparse components (i.e., source locations) accurately and enables optimal recovery by inverse filtering as elaborated in Section 7.5 and the accuracy is bounded by the noise level.

Table 8.2 – RMSE of signal recovery. The numbers in parenthesis show the performance of conventional beamforming formulation without sparsity regularization. The multipath sparse MVDR beamformer expressed in (8.11) is compared with multipath MVDR beamformer expressed in (8.3). Similarly the multipath sparse MMSE defined in (8.20) is compared with multipath MMSE defined in (8.19).

Scenario	Beamf.	1 Source	2 Sources	3 Sources	4 Sources
Reverberant	MVDR	0.03 (0.52)	0.27 (0.54)	0.39 (0.83)	0.49 (0.84)
	MMSE	0.66 (0.83)	0.70 (0.84)	0.73 (0.96)	0.75 (0.96)
Farfield	MVDR	0.10 (0.10)	0.10 (0.10)	0.17 (0.35)	0.16 (0.47)
	MMSE	0.07 (0.54)	0.08 (0.55)	0.14 (0.57)	0.16 (0.59)

The differences between the performance of sparse beamforming and the conventional formulation show that the sparse beamformer requires fewer parameters to estimate the solution than the conventional beamforming techniques hence, it enables accurate estimation when the number of reliable samples are limited and suggests a framework for *missing data beamforming*. Additionally, it enables better null steering and offers a more robust solution in multiparty recordings and a better resolution if the sources are closely located. The results of conventional beamformers using only direction of the desired source in a reverberant acoustic were poor so they are not included here. We can calibrate the model parameters of an acoustic-informed beamformer using a known source signal at a given location as formulated in Section 8.1.2.

Real Data Evaluations

The real data evaluations are based on a subset of Multi-Channel Wall Street Journal Audio-Visual (MC-WSJ-AV) corpus [Lincoln et al., 2005] used for PASCAL Speech Separation Challenge II [Himawan et al., 2005, McDonough et al., 2007]. The subset consists of two concurrent speakers who are simultaneously reading sentences form the Wall Street Journal and being recorded with 8-channel circular microphone array. The speakers are either seating or standing at 1.5×1.5 table and a circular microphone array with diameter 20 cm is located at the center of the table. The total number of utterances is 356 (or 178, respectively given that two sentences are read at a time). The geometrical set-up ground-truth are not provided and in particular the height information are missing hence, calculation of the room acoustic channel to perform inverse-filtering incur inaccuracies.

The speech recognition system used in the experiments is identical to the one in [Mahdian et al., 2012, McDonough et al., 2007], except that three passes are applied (instead of four). The first pass is unadapted speech recognition. In the second pass, unsupervised maximum likelihood linear regression (MLLR) feature space adaptation is applied and the third pass employs full MLLR adaptation. The estimated speaker positions are the same ones used in [Mahdian et al., 2012, McDonough et al., 2007] for superdirective and delay-and-sum beamforming. The room reverberation time is 0.7 s and the corpus is highly noisy.

In addition to the conventional beamforming, we employed sparse recovery to provide the exact location of the sources on a discretized grid [Asaei et al., 2012a]. The speech recognition results are presented in Figure 8.4 [2]. The results show that providing the accurate positioning obtained by sparse recovery (as illustrated through Figure 8.1- 8.3) improves the performance of the conventional beamforming methods. Future experiments will consider the multipath sparse beamforming method to enable more effective signal recovery.

2. I would like to acknowledge Rahil Mahdian at Saarland University for running the speech recognition scripts on my recovered data. The WRR results after the second pass adaptation for ICA, delay-and-sum (DS) beamforming, superdirective (SD) beamforming are 15%, 16.54%, 25.15% respectively. We did not have access to the results of ICA after third pass adaptation. Given the improvement obtained for beamforming, we can expect that ICA would also improve roughly by 14% thus ≈29% is reported in the figure. The implementation of ICA is available at [Murata et al., 1998]

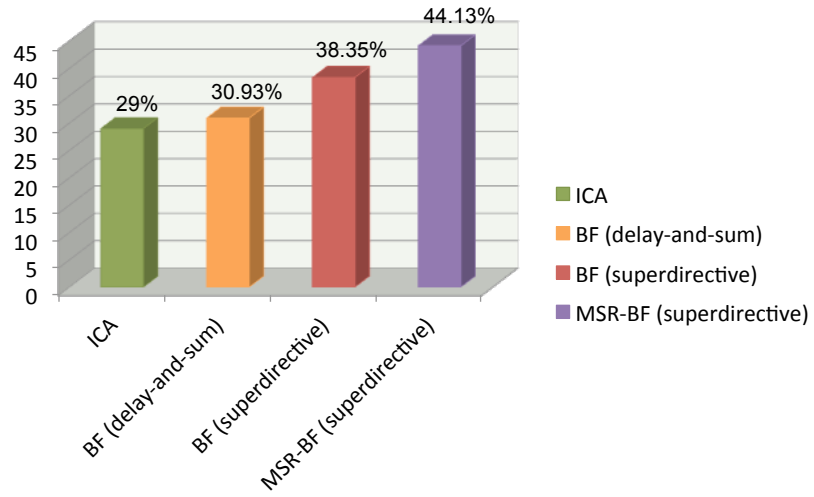

Figure 8.4 – Word Recognition Rate (WRR) using independent component analysis (ICA), delay-and-sum beamformer (DS), superdirective beamformer (SD) and model-based sparse recovery incorporated for superdirective beamformer (MSR-SD). The benchmark WRR for the headset microphone is 76.55% and for the distant microphone is 0%.

8.3 Conclusions

In this chapter, a novel formulation of beamforming is proposed for acquisition of the signals in reverberant acoustic clutter of interferences and noise. We derived the beamforming methods which incorporate the sparsity structure pertained to the acoustic source distribution and multipath propagation. The quantitative assessments demonstrate that sparse beamforming enables effective beampattern steering from far fewer samples than the conventional beamformers. In addition, linear constraint on the desired channel rather than the desired direction improves the signal estimation performance in reverberant enclosures. The experimental analysis in trems of source localization and speech recovery provides strong evidence of the effectiveness of sparsity models to enable high resolution source localization and improving the speech recovery performance.

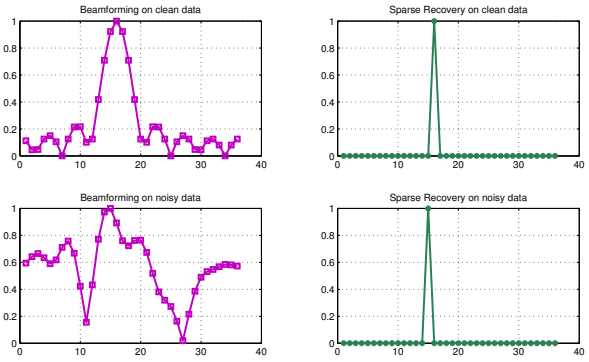

Figure 8.1 – Sparse recovery vs. MVDR beamformer for localization of one source in noisy and clean condition; The SNR of noisy scenario is 10 dB. The x-axis demonstrates the direction bins with a resolution of 5° and y-axis demonstrates the energy of the estimated signal in that corresponding direction.

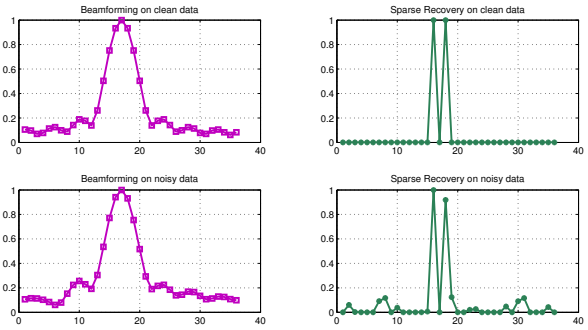

Figure 8.2 – Sparse recovery vs. MVDR beamforming for localization of two sources in clean and noisy condition; the support of the sparse components is exact. The x-axis demonstrates the direction bins with a resolution of 5° and y-axis demonstrates the energy of the estimated signal in that corresponding direction.

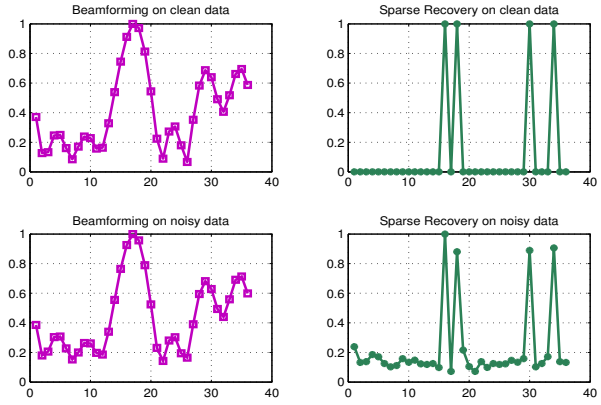

Figure 8.3 – Sparse recovery vs. MVDR beamforming for localization of four sources in noisy condition; the support of the sparse components is exact. The x-axis demonstrates the direction bins with a resolution of 5° and y-axis demonstrates the energy of the estimated signal in that corresponding direction.

9 Conclusion

The present thesis was inspired by two trends (1) auditory sparse coding and (2) recent algorithmic advances in compressive sensing and sparse signal recovery. We have sought to provide the technical detail and experimental justification for a structured sparse coding approach to multiparty reverberant speech recognition, and we have endeavored to offer some understanding of the mechanism through which sparsity models can be utilized. To conclude, we summarize the key messages of our research and recommend future directions along the lines of these findings.

9.1 Summary of Achievements

Our research confirmed the hypothesis that information bearing components for speech recognition are sparse in the spectro-temporal scene and presented a framework of model-based sparse component analysis for extraction of information in a reverberant acoustic. This framework was constructed of three building blocks: structured sparse representation, compressive acoustic measurements and model-based sparse recovery.

The structured sparse representation of concurrent sources was constituted considering the sparsity models pertained to the sound perception (spectral domain) and propagation (spatial domain). The spectral structures were characterized exploiting the harmonic dependency, proximity and modulation of the frequency components. The spatial structures were characterized exploiting the geometry of the source locations and multipath propagation. The spatial representation was appended to the spectral representation to obtain a sparse representation of the spatio-spectral information embedded in the acoustic scene.

The acoustic projections corresponding to acquisition of the sources by microphone array were characterized using the Image model of reverberation effect. It was shown that identification of the acoustic projections were entangled with estimation of the geometry of the enclosure and the absorption properties of the reflective surfaces and effective methodologies were proposed to estimate the parameters using multiparty recordings of unknown speech sources.

A framework was formulated to identify the location of the sources and their spectral components based on model-based sparse recovery algorithms exploiting the structures underlying the speech spectrum as well as the acoustic projections for extraction of the informative patterns. The computational strategies relying on greedy forward selection, iterative hard thresholding, convex optimization and sparse Bayesian learning were compared and contrasted within this framework. The generalization and optimality in terms of speech separation and dereverebartion were elaborated.

Ultimately, the sparsity models were integrated with the formulation of optimum spatial filtering tailored for acquisition of the desired signal in multiparty reverberant scenarios. A multipath sparse beamforming method was derived which can incorporate the model of the acoustic reverberation and sparsity structures for effective signal estimation. This information has not been explicitly exploited in standard spatial filtering techniques.

The evaluations on real data recordings demonstrate the effectiveness of structured sparsity models for multi-party speech recovery and recognition. It establishes a new perspective to the microphone array recordings as a compressive acquisition of the information embedded in the acoustic scene so the spatio-spectral data can be reconstructed using sparse signal recovery techniques. This new formulation of the objective of microphone array processing suggests some theoretical development and quantitative advances which further motivates considerations and alterations in information extraction schemes and microphone array design.

9.2 Future Directions

The future directions could revolve around model-based sparse coding for machine listening in overlapping conditions, characterizing the general acoustic field and incoherent design methodologies for signal acquisition and recovery.

We have shown that information bearing components for speech recognition admit structured sparsity models. However, state-of-the art techniques in feature extraction do not exploit this property. Two directions to extract speech specific features using the model-based sparse coding framework could be investigated. One approach could be to use an auditory-inspired parametric model using Gabor functions [Kleinschmidt, 2002b, Stern and Morgan, 2012]. The Gabor functions are localized sinusoid which can capture all possible modulations so they can be used to design a dictionary for sparse coding with some particular objectives. One objective could be to select a subset of Gabor prototypes optimized for speech recognition while decomposition is designed to yield sparse representation to enable robustness to multiparty overlapping conditions.

Another solution could be exploiting the statistical dependencies to devise a model-based feature extraction scheme. Having pre-trained the desired statistical models of the informative (e.g. phonetic) patterns, this framework ought to be robust to the interferences and noise. We can incorporate the statistical model-based sparse signal recovery using the Boltzmann machine graphical model as proposed in [Peleg et al., 2012].

We characterized the compressive acoustic measurements using the Image model which requires accurate estimation of the geometry of the reflective surfaces and the corresponding absorption coefficients. To enable an acoustic model for an unrestricted set-up, one possibility is exploiting perceptual reconstruction of the plenacoustic function used for spatialization of the sound field [Ajdler et al., 2003]. Plenacoustic function characterizes the room impulse responses from the measurements of impulse responses in a finite number of positions and with this information the total sound field can be recreated up to a corresponding temporal frequency. This information can be incorporated for a flexible characterization of the acoustic measurements.

To address the acoustic ambiguities, the measurement matrix calibration techniques should be considered [Gribonval et al., 2011, Taghizadeh et al., 2013]. Furthermore, an optimal design of sparse ad-hoc layout of microphones could be worked out to provide information-preserving acquisition set-up by incorporating the generic theory of compressive sensing and sparse reconstruction. This procedure could also consider construction of a grid of non-uniform cells to obtain less coherent projections. Finally, we can extend our model-based sparse recovery approach for continuous sparse approximation to enable source localization in a continuous space.

9.3 Concluding Remarks

This thesis was an exploration on the missing principles to enable machine listening in multiparty environment. Motivated from the nature of sound perception and propagation, we formulated a model-based sparse component analysis framework for recovery of the speech-specific information. The present work was a first attempt to coin this formalism which was shown to offer some compelling and advantageous features to be worth of serious consideration. The structured sparsity is a general principle and several possible future directions towards robust machine listening paradigms could be envisioned.

A Equivalence of Inverse Filtering to Source Separation-Deconvolution

In Chapter 7, we provided the proof of equivalence of inverse filtering the room acoustic channel with speech separation and dereverberation considering 3 microphones and 2 sources. In this appendix, we show that inverse filtering of any $M \times N$, $M > N$ MIMO system is equivalent to a two step procedure: *source separation* followed by *channel deconvolution*.

First, we remind the *Cauchy-Binet* theory of matrix determinants.
Cauchy-Binet Formula: Let $\mathscr{H}$ be an $M \times N$ matrix with $M > N$. The determinant of $\mathscr{H}^T\mathscr{H}$ denoted by $|\mathscr{H}^T\mathscr{H}|$ can be computed from the determinant of its submatrices as

$$\begin{aligned} |\mathscr{H}^T\mathscr{H}| &= \sum_{i=1}^{\binom{M}{N}} |\mathscr{H}_i|^2 \\ |\binom{M-1}{N-1}^{-1} \sum_{i=1}^{\binom{M}{N}} \mathscr{H}_i^T\mathscr{H}_i| &= \sum_{i=1}^{\binom{M}{N}} |\mathscr{H}_i|^2 \end{aligned} \tag{A.1}$$

The second follows from the definition of $\mathscr{H}^T\mathscr{H}$.

Denoting the transfer function between microphone m and source n as H_{mn}, a MIMO finite impulse response (FIR) acoustic system can be represented in matrix notation as

$$\underbrace{\begin{bmatrix} X_1 \\ \vdots \\ X_M \end{bmatrix}}_{\mathscr{X}} = \underbrace{\begin{bmatrix} H_{11} & \cdots & H_{1N} \\ \vdots & & \\ H_{M1} & \cdots & H_{MN} \end{bmatrix}}_{\mathscr{H}} \underbrace{\begin{bmatrix} S_1 \\ \vdots \\ S_N \end{bmatrix}}_{\mathscr{S}} \tag{A.2}$$

Hence, the minimum mean square estimation (MMSE) of the source is given by

$$\hat{\mathscr{S}} = \arg\min_{\mathscr{S}} (\mathscr{X} - \mathscr{H}\mathscr{S})^2 = \arg\min_{\mathscr{S}} \frac{\sum_{i=1}^{\binom{M}{N}} (\mathscr{X}_i - \mathscr{H}_i\mathscr{S})^2}{\binom{M-1}{N-1}} \tag{A.3}$$

where submatrix $\mathscr{H}_i$ is an $N \times N$ matrix consisted of N randomly selected rows from $\mathscr{H}$ so we have $\binom{M}{N}$ possible construction for $\mathscr{H}_i$. Corresponding to each $\mathscr{H}_i$, we construct $\mathscr{X}_i$ hence, we obtain

$$\mathscr{X}_i = \mathscr{H}_i \hat{\mathscr{S}}_i \quad \forall i = 1, \cdots, \binom{M}{N} \tag{A.4}$$

where $\hat{\mathscr{S}}_i$ is an estimation of $\mathscr{S}$ using $\mathscr{X}_i$ and $\mathscr{H}_i$. To estimate $\mathscr{S}$, we compute the derivative of (A.3) with respect to $\mathscr{S}$

$$\begin{aligned} &\frac{1}{\binom{M-1}{N-1}} \frac{\partial}{\partial \mathscr{S}} \sum_{i=1}^{\binom{M}{N}} (\mathscr{X}_i - \mathscr{H}_i\mathscr{S})^2 = 0 \\ &\Rightarrow \sum_{i=1}^{\binom{M}{N}} \mathscr{H}_i^T \mathscr{X}_i = \sum_{i=1}^{\binom{M}{N}} \mathscr{H}_i^T \mathscr{H}_i \hat{\mathscr{S}} \\ &\Rightarrow \hat{\mathscr{S}} = \frac{\sum_{i=1}^{\binom{M}{N}} \mathscr{H}_i^T \mathscr{H}_i \hat{\mathscr{S}}_i}{\sum_{i=1}^{\binom{M}{N}} \mathscr{H}_i^T \mathscr{H}_i} \end{aligned} \tag{A.5}$$

We now move onto show the equivalence to a two step procedure of *source separation* and *dereverberation*. Without the loss of generality, we first separate S_1 from S_j, $j = 2, \cdots, N$. From *Cramer's rule* in linear algebra we have

$$|[\mathscr{X}_i \mathscr{H}_i^{/1}]| = |\mathscr{H}_i| \tilde{S}_1^i = y_{S_{1,i}} \tag{A.6}$$

where $\mathscr{H}_i^{/1}$ is $\mathscr{H}_i$ matrix with removing the first column of it. Hence $y_{S_{1,i}}$ is only dependent on S_1 while the components of $S_2, \cdots, S_N$ are separated. By repeating (A.6) for all possible $\mathscr{H}_i$, we get the set of all estimates of S_1 as $y_{S_{1,1}}, \cdots, y_{S_{1,\binom{M}{N}}}$.

In the second step, we perform deconvolution to extract $\hat{S}_1$. We define the matrices

$$\begin{aligned} Y &\triangleq \left[y_{S_{1,1}}, \cdots, y_{S_{1,\binom{M}{N}}} \right] \\ F &\triangleq \left[|\mathscr{H}_1| \cdots |\mathscr{H}_{\binom{M}{N}}| \right] \end{aligned}$$

Hence, deconvolution of the channel from the separated source is achieved via

$$\begin{aligned}\hat{S}_1 &= \frac{Y F^T}{F^T F} = \frac{Y\left[|\mathscr{H}_1|\cdots|\mathscr{H}_{\binom{M}{N}}|\right]^T}{\sum_{i=1}^{\binom{M}{N}}|\mathscr{H}_i|^2} = \frac{\left[|\mathscr{H}_1|\tilde{S}_1^1\cdots|\mathscr{H}_{\binom{M}{N}}|\tilde{S}_1^{\binom{M}{N}}\right]\left[|\mathscr{H}_1|\cdots|\mathscr{H}_{\binom{M}{N}}|\right]^T}{\sum_{i=1}^{\binom{M}{N}}|\mathscr{H}_i|^2} \\ \Rightarrow \hat{S}_1 &= \frac{|\mathscr{H}_1|^2\tilde{S}_1^1+\cdots+|\mathscr{H}_{\binom{M}{N}}|^2\tilde{S}_1^{\binom{M}{N}}}{\sum_{i=1}^{\binom{M}{N}}|\mathscr{H}_i|^2}\end{aligned} \tag{A.7}$$

If we calculate $\hat{S}_2,\cdots\hat{S}_N$ in the same manner, we have

$$\begin{aligned}\hat{S}_2 &= \frac{|\mathscr{H}_1|^2\tilde{S}_2^1+\cdots+|\mathscr{H}_{\binom{M}{N}}|^2\tilde{S}_2^{\binom{M}{N}}}{\sum_{i=1}^{\binom{M}{N}}|\mathscr{H}_i|^2} \\ &\vdots \\ \hat{S}_N &= \frac{|\mathscr{H}_1|^2\tilde{S}_N^1+\cdots+|\mathscr{H}_{\binom{M}{N}}|^2\tilde{S}_N^{\binom{M}{N}}}{\sum_{i=1}^{\binom{M}{N}}|\mathscr{H}_i|^2}\end{aligned}$$

By concatenation we stated in (A.2) , we obtain (A.8) and we believe that the equality to (A.5) can be achieved via *Cauchy-Binet* theory of matrix determinants.

$$\hat{\mathscr{S}} = \frac{|\mathscr{H}_1|^2\hat{\mathscr{S}}_1+\cdots+|\mathscr{H}_{\binom{M}{N}}|^2\hat{\mathscr{S}}_{\binom{M}{N}}}{\sum_{i=1}^{\binom{M}{N}}|\mathscr{H}_i|^2} = \frac{\sum_{i=1}^{\binom{M}{N}}|\mathscr{H}_i|^2\hat{\mathscr{S}}_i}{\sum_{i=1}^{\binom{M}{N}}|\mathscr{H}_i|^2} \tag{A.8}$$

Bibliography

Aachen Impulse Response (AIR) database - version 1.2. Institute of Communication Systems and Data Processing (IND), RWTH Aachen University, available at http://www.ind.rwth-aachen.de/AIR, 2010.

F. Abrard and Y. Deville. A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources. *Elsevier, Signal Processing*, 2005.

T. Ajdler, L. Sbaiz, and M. Vetterli. The plenacoustic function and its sampling. *IEEE Transactions on Signal Processing*, 54(10), 2003.

J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *Journal of Acoustical Society of America*, 60(s1), 1979.

S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada. Underdetermined blind separation for speech in real environments with sparseness and ICA. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004a.

S. Araki, S. Makino, H. Sawada, and R. Mukai. Underdetermined blind separation of convolutive mixtures of speech with directivity pattern based mask and ICA. *Independent Component Analysis and Blind Signal Separation, Lecture Notes in Computer Science*, 3195, 2004b.

S. Araki, H. Sawada, R. Mukai, and S. Makino. A novel blind source separation method with observation vector clustering. In *Proceedings of the 2005 International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2005.

S. Araki, F. Nesta, E. Vincent, Z. Koldovsky, and G. Nolte. The 2011 signal separation evaluation campaign (sisec2011): Audio source separation. *Latent Variable Analysis and Signal Separation, Lecture Notes in Computer Science,*, 7191, 2011.

A. Asaei, M. J. Taghizadeh, and H. Sameti. Far-field continuous speech recognition system based on speaker localization and sub-band beamforming. In *Proceedings of 6th International ACS/IEEE Conference on Computer Systems and Applications,*, 2008.

A. Asaei, M. J. Taghizadeh, M. Bahrololum, and M. Ghanbari. Verified speaker localization utilizing voicing level in split-bands. *Signal Processing*, 89(6), 2009.

A. Asaei, P. N. Garner, and H. Bourlard. Sparse component analysis for speech recognition in mullti-speaker environment. In *Proceeding of INTERSPEECH*, 2010a.

A. Asaei, B. Picart, and H. Bourlard. Analysis of phone posterior feature space exploiting class-specific sparsity and mlp-based similarity. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),*, 2010b.

A. Asaei, H. Bourlard, and V. Cevher. Model-based compressive sensing for distant multi-party speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011a.

A. Asaei, M. J. Taghizadeh, H. Bourlard, and V. Cevher. Multi-party speech recovery exploiting structured sparsity models. In *Proceeding of INTERSPEECH*, 2011b.

A. Asaei, H. Bourlard, and V. Cevher. A method, apparatus and computer program for determining the location of a plurality of speech sources. *2012US-13/654055, US Patent, October*, 2012a.

A. Asaei, M. Davies, H. Bourlard, and V. Cevher. Computational methods for structured sparse component analysis of convolutive speech mixtures. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012b.

A. Asaei, M. Golbabaee, H. Bourlard, and V. Cevher. Structured sparsity models for multiparty speech recovery from reverberant recordings. *submitted to IEEE Transactions on Speech and Audio Processing, available at http://arxiv.org/abs/1210.67666*, 2012c.

A. Asaei, B. Raj, H. Bourlard, and V. Cevher. Structured sparse coding for microphone array position calibration. In *Proceeding of 5th ISCA workshop on Statistical and Perceptual Audition, SAPA-SCALE Conference,*, 2012d.

A. Asaei, M. Golbabaee, H. Bourlard, and V. Cevher. Structured sparse acoustic modelling for speech separation. In *Signal Processing with Adaptive Sparse Structured Representations (SPARS)*, 2013a.

A. Asaei, B. Raj, H. Bourlard, and V. Cevher. A unified structured sparse coding framework for spatio-spectral information recovery. *IEEE Transactions on Speech and Audio Processing (under revisions)*, 2013b.

A. Asaei, B. Raj, H. Bourlard, and V. Cevher. A multipath sparse beamforming method. In *Signal Processing with Adaptive Sparse Structured Representations (SPARS)*, 2013c.

D. Ba, F. Ribeiro, C. Zhang, and D. Florencio. L1 regularized room modeling with compact microphone arrays. In *Proceedings of ICASSP*, 2010.

R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Transactions in Information Theory*, 2010.

E. V. D. Berg and M. P. Friedlander. Probing the pareto frontier for basis pursuit solutions,. *SIAM Journal on Scientific Computing*, 2008. http://www.cs.ubc.ca/labs/scl/spgl1.

T. Blu, P. L. Dragotti, M. Vetterli, P. Marziliano, and L. Coulot. Sparse sampling of signal innovations. *IEEE Signal Processing Magazine*, 25, 2008.

T. Blumensath and M. E. Davies. Gradient pursuits. *IEEE Transactions on Signal Processing*, 56:2370–2382, 2008.

T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

P. Bofill and M. Zibulevsky. Underdetermined blind source separation using sparse representations. *Signal Processing*, 2001.

J. Borish. Extension of the image model to arbitrary polyhedra. *Journal of the Acoustical Society of America*, 75(6), 1984.

P. Boufounos, P. Smaragdis, and B. Raj. Joint sparsity models for wideband array processing. In *Wavelets and Sparsity XIV, SPIE Optics and Photonics*, 2011.

H. Bourlard. Non-stationary multi-channel (multi-stream) processing towards robust and adaptive asr. In *Proceeding of ESCA Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, 1999.

A. S. Bregman. *Auditory Scene Analysis*. MIT Press, Cambride, MA, 1990.

D. S. Brungart. Information and energetic masking effects in the perception of two simultaneous talkers. *Journal of Acoustical Society of America*, 2001.

H. Buchner, R. Aichner, and W. Kellermann. *TRINICON-based blind system identification with application to multiple-source localization and separation*, volume 13. In Blind Speech Separation, S. Makino, T.-W. Lee, and H. Sawada, Eds. New York: Springer, 2007.

L. Carin. On the relationship between compressive sensing and random sensor arrays. *IEEE Antennas and Propagation Magazine*, 51:72–81, 2009.

L. Carin, D. Liu, and B. Guo. Coherence, compressive sensing and random sensor arrays. *IEEE Antennas and Propagation Magazine*, 2011.

V. Cevher. Learning with compressible priors. In *Neural Information Processing Systems (NIPS)*, 2009.

V. Cevher. An ALPS view of sparse recovery. In *Proceedings of ICASSP*, 2011.

V. Cevher, P. Boufounos, R. G. Baraniuk, A. C. Gilbert, and M. J. Strauss. Near-optimal Bayesian localization via incoherence and sparsity. In *Proceedings of IPSN*, 2009.

E. C. Cherry. *On Human Communication*. MIT Press, Cambride, MA, 1957.

P. L. Combettesand and J. C. Pesquet. Proximal splitting methods in signal processing. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer-Verlag*, 49, 2011.

P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, 2010.

R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson. Measurement of correlation coefficients in reverberant sound fields. *Journal of the Acoustical Society of America*, 27(6), 1955a.

R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson. Measurement of correlation coefficients in reverberant sound fields. *Journal of the Acoustical Society of America*, 27(6), 1955b.

M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication, Elsevier*, 34(3), 2001.

M. P. Cooke. Modelling auditory processing and organisation. *PhD thesis, University of Sheffield,*, (also published by Cambridge University Press, Cambridge, UK), 1991.

S. F. Cotter, B. D. Rao, E. Kjersti, and K. Kreutz-Delgado. Sparse solutions to linear inverse problems with multiple measurement vectors. *IEEE Transactions on Signal Processing*, 2005.

R. Haeb-Umbach (Eds.) D. Kolossa. *Robust Speech Recognition of Uncertain or Missing Data Theory and Applications*. Springer, 2011.

S. V. David, N. Mesgarani, and S. A. Shamma. Estimating sparse spectro-temporal receptive fields with natural stimuli. *Network*, 18(3), 2007.

M. Davies and N. Mitianoudis. Simple mixture model for sparse overcomplete ica. *IEE Proceedings on Vision, Image and Signal Processing*, 151(1), 2004.

M. A. Dmour and M. E. Davies. A new framework for underdetermined speech extraction using mixture of beamformers. *IEEE Transactions on Audio, Speech, and Language Processing*, 19: 445–457, 2011.

I. Dokmanic, Y. Lu, and M. Vetterli. Can one hear the shape of a room: The 2-D polygonal case. In *Proceedings of ICASSP*, 2011.

L. Ehrenberg, S. Gannot, A. Leshem, and E. Zehavi. Sensitivity analysis of mvdr and mpdr beamformers. In *IEEE 26th Convention of Electrical and Electronics Engineers in Israel*, 2010.

M. Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, 2010.

C. F. Eyring. Reverberation time in dead rooms. *Journal of the Acoustical Society of America, pp. 217-241*, 1930.

C. Faller and J. Merimaa. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *Journal of the Acoustical Society of America*, 116(5), 2004.

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

J. F. Gemmeke. *Noise robust ASR: Missing data techniques and beyond*. PhD thesis, Radboud Universiteit Nijmegen, The Netherlands, 2011.

M. Golbabaee and P. Vandergheynst. Compressed sensing of simultaneous low-rank and joint-sparse matrices. *submitted to IEEE Transactions on Information Theory, available in*. http://infoscience.epfl.ch/record/181506.

M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21. http://cvxr.com/cvx.

R. Gribonval and E. Bacry. Harmonic decomposition of audio signals with matching pursuit. *IEEE Transactions on Signal Processing*, 51:101–111, 2003.

R. Gribonval and S. Lesage. A survey of sparse component analysis for blind source separation: Principles, perspectives, and new challenges. In *ESANN, 14th European Symposium on Artificial Neural Networks*, 2006.

R. Gribonval, G. Chardon, and L. Daudet. Blind calibration for compressed sensing by convex optimization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

L. Griffiths and C. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. Antennas Propag.*, 30(1), 1982.

E. A. P. Habets. *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement*. PhD thesis, Technische Universiteit Eindhoven, 2007. URL http://alexandria.tue.nl/extra2/200710970.pdf.

E. A.P. Habets. *Speech Dereverberation Using Statistical Reverberation Models*. Speech Dereverberation, Springer, 2010.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. New York: Springer, 2001.

S. Hildebrandt and A. Tromba. *The Parsimonious Universe: Shape and Form in the Natural World*. Springer, 1996.

I. Himawan, I. McCowan, and M. Lincoln. Microphone array beamforming approach to blind speech separation. In *Second International Workshop on Machine Learning for Multimodal Interaction*, 2005.

G. Hu and D. Wang. Speech segregation based on pitch tracking and amplitude modulation. In *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 2001.

Y. Huang and J. Benesty. A class of frequency-domain adaptive approaches to blind multichannel identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 51(1), 2003.

Y. Huang, J. Benesty, and J. Chen. A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment. *IEEE Transactions on Audio, Speech, and Language Processing*, 13(5), 2005.

ITU-T. Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. *International Telecommunications Union, Geneva, Switzerland*, 2001.

ITU-T. Itu-t rec. p.862.1, mapping function for transforming p.862 raw result scores to mos-lqo. *International Telecommunications Union, Geneva, Switzerland*, 2003.

M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies. Sparse coding for convolutive blind audio source separation. In *The 6th International Conference on Independent Component Analysis and Blind Source Separation*, 2006.

A. Jourjine, S. Rickard, and O. Yilmaz. Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2000.

M. Kearns, Y. Mansour, and A. Y. Ng. An information-theoretic analysis of hard and soft assignment methods for clustering. In *The13th conference on Uncertainty in Artificial Intelligence (UAI-97)*. Morgan Kaufmann, 1997.

D. H. Klatt. Prediction of perceived phonetic distance from critical-band spectra: a first step. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1982.

M. Kleinschmidt. Methods for capturing spectro-temporal modulations in automatic speech recognition. *ACUSTICA*, 88(3), 2002a.

M. Kleinschmidt. Robust speech recognition based on spectrotemporal processing. *PhD thesis, Univ. Oldenburg*, 2002b.

B. Kollmeier, T. Brand, and B. Meyer. *Perception of speech and sound*. Springer Handbook of Speech Processing, 2008a.

B. Kollmeier, T. Brand, and B. Meyer. *Perception of speech and sound*, volume Springer, Berlin,. Springer Handbook of Speech Processing (Benesty, Sondhi and Huang Eds.), 2008b.

K. Kumatani, J. W. McDonough, and B. Raj. Maximum kurtosis beamforming with a subspace filter for distant speech recognition. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2011.

A. Kyrillidis and V. Cevher. Recipes on hard thresholding methods. In *Proceedings of CAMSAP*, 2011.

M. Lincoln, I. McCowan, I. Vepa, and H. K. Maganti. The multi-channel wall street journal audio visual corpus (mc-wsj-av): Specification and initial experiments. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2005.

M. Lockwood, D. Jones, R. Bilger, C. Lansing, W. OBrien, W. Wheeler, and A. Feng. Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms. *Journal of Acoustic Society of America*, 115(1), 2004.

R. Togneri M. Kuhne and S. Nordholm. Mel-spectrographic mask estimation for missing data speech recognition using short-time-fourier-transform ratio estimators. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

R. Togneri M. Kuhne and S. Nordholm. Time-frequency masking: Linking blind source separation and robust speech recognition. *Speech Recognition, Technologies and Applications. I-Tech*, 2008.

R. Mahdian, F. Faubel, and D. Klakow. Multi-channel speech separation with soft time-frequency masking. In *SAPA-SCALE Conference*, 2012.

D. Malioutov, M. Cetin, and A. S. Willsky. A sparse signal reconstruction perspective for source localization with sensor arrays. *IEEE Transactions on Signal Processing*, 53(2), 2005.

I. McCowan, D. Moore, and S. Sridharan. Near-field adaptive beamformer for robust speech recognition. *Digital Signal Processing*, 12(1), 2002.

I. McCowan, M. Lincoln, and I. Himawan. Microphone array shape calibration in diffuse noise fields. *IEEE Transactions on Audio,Speech and Language Processing*, 16(3), 2008.

I. A. Mccowan. The Multichannel Overlapping Numbers Corpus. Idiap resources available online: http://www.cslu.ogi.edu/corpora/monc.pdf, 2003.

I. A. McCowan and H. Bourlard. Microphone array post-filter based n noise field coherence. *IEEE Transactions on Audio, Speech, and Language Processing*, 11(6), 2003.

I. A. Mccowan, C. Marro, and L. Mauuary. Robust speech recognition using near-field superdirective beamforming with post-filtering. In *Proceedings of ICASSP*, 2000.

I. A. Mccowan, A. Morris, and H. Bourlard. Improving speech recognition performance of small microphone arrays using missing data techniques. In *IEEE International Conference on Spoken Language Processing*, 2002.

J. McDonough, K. Kumatani, T. Gehrig, E. Stoimenov, U. Mayer, S. Schacht, M. Wolfel, and D. Klakow. To separate speech - a system for recognizing simultaneous speech. In *Fourth International Workshop on Machine Learning for Multimodal Interaction*, 2007.

T. Melia and S. Rickard. Underdetermined blind source separation in echoic environment using desprit. *EURASIP Journal on Advances in Signal Processing*, 2007.

T. Melia, S. Rickard, and C. Fearon. Histogram-based blind source separation of more sources than sensors using duet-esprit technique. In *Proceedings of 13th European Signal Processing Conference (EUSIPCO)*, 2005.

N. Mesgarani and E. F. Chang. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397), 2012.

M. Miyoshi and Y. Kaneda. Inverse filtering of room acoustics. *IEEE Transactions on Audio, Speech, and Language Processing*, 36(2), 1988.

D. Model and M. Zibulevsky. Signal reconstruction in sensor arrays using sparse representations. *Signal Processing*, 86(3), 2006.

D. C. Moore and I. A. Mccowan. Microphone array speech recognition: Experiments on overlapping speech in meetings. In *Proceedings of ICASSP*, 2003.

N. Mourad and J. P. Reilly. Modified hierarchical clustering for sparse component analysis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010.

N. Murata, S. Ikeda, and A. Ziehe. An approach to blind source separation based on temporal structure of speech signals, available at http://www.ism.ac.jp/~shiro/papers/techreport/ism2003-890.pdf. *BSIS Technical Report, code available at http://bsp.teithe.gr/members/downloads/ikedaICA.html*, 1998.

T. Nakatani, T. Yoshioka, and K. Kinoshita. Mathematical analysis of speech dereverberation based on time-varying gaussian source model: Its solution and convergence characteristics. In *Proceedings of IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC),*, 2011.

R. M. Neal. *Bayesian Learning for Neural Networks*. New York: Springer-Verlag, 1996.

F. Nesta and M. Omologo. *Convolutive Underdetermined Source Separation through Weighted Interleaved ICA and Spatio-temporal Source Correlation*, volume 7191. In: Yeredor, A. et al. (eds.) LVA/ICA 2012. LNCS, Springer, Heidelberg, 2012.

Y. Nesterov. A method of solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$, in soviet mathematics doklady. *Soviet Mathematics Doklady*, 27, 1983.

P. O'Grady and B. A. Pearlmutter. Soft-lost: EM on a mixture of oriented lines. *Proceedings of ICA, Lecture Notes in Computer Science, Springer-Verlag*, 2004a.

P. O'Grady and B. A. Pearlmutter. Hard-lost: Modified k-means for oriented lines. In *ISSC Conference*, 2004b.

P. D. O'Grady and B. A. Pearlmutter. The LOST algorithm: Finding lines and separating speech mixtures. *EURASIP Journal on Advances in Signal Processing*, 2008.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381, 1996.

B. A. Olshausen and D. J. Field. Sparse coding with an over-complete basis set: a strategy employed by v1. *Vision Research*, 37, 1997.

M. Omologo, P. Svaizer, and M. Matassoni. Environmental conditions and acoustic transduction in hands-free speech recognition. *Speech Communication*, 1998.

S. E. Palmer. *Vision Science*. MIT Press, Cambride, MA, 1999.

L. C. Parra and C. V. Alvino. Geometric source separation: Merging convolutive source separation with geometric beamforming. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 10(6), 2002.

T. W. Parsons. Separation of speech from interfering speech by means of harmonic selection. *Journal of the Acoustical Society of America*, 60(4), 1976.

B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang. Reconstructing speech from human auditory cortex. *PLoS Biol, Public Library of Science*, 10(1), 2012.

D. Pearce and H. Hirsch. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proceedings of ICSLP*, 2000.

T. Peleg, Y. C. Eldar, and M. Elad. Exploiting statistical dependencies in sparse representations for signal recovery. *IEEE Transactions on Signal Processing*, 60(5), 2012.

L. Di Persia, D. Milone, H. L. Rufiner, and M. Yanagida. Perceptual evaluation of blind source separation for robust speech recognition. *Signal Processing, implementation available at*, 2008. http://www.utdallas.edu/~loizou/speech/software.htm.

B. Raj, M.L. Seltzer, and R. M. Stern. Reconstruction of missing features for robust speech recognition. *Speech Communication, Elsevier*, 43(4), 2004.

S. Renals, T. Hain, and H. Bourlard. Recognition and understanding of meetings the AMI and AMIDA projects. In *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2007.

J. Rissanen. Modeling by shortest data description. *Automatic*, 14, 1978.

N. Roman, D. Wang, and G. J. Brown. Speech segregation based on sound localization. *Journal of the Acoustical Society of America*, 114(4), 1991.

R. Rotili, C. De Simone, A. Perelli, A. Cifani, and S. Squartini. Joint multichannel blind speech separation and dereverberation: A real-time algorithmic implementation. In *Proceedings of 6th International Conference on Intelligent Computing*, 2010.

S T. Roweis. Factorial models and refiltering for speech separation and denoising. In *EUROSPEECH*, 2003.

R. Saab, O. Yilmaz, M. J. Mckeown, and R. Abugharbieh. Underdetermined anechoic blind source separation via ℓ_q-basis-pursuit with q<1. *IEEE Transactions on Signal Processing*, 2007.

H. Sawada, S. Araki, and S. Makino. *Frequency-domain blind source separation*. Speech Enhancement (T.-W. Lee, and H. Sawada, Eds.), Springer, 2007.

E. Shriberg, A. Stolcke A., and D. Baron. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In *Proceedings of Eurospeech*, 2001.

R. M. Stern and N. Morgan. Hearing is believing: Biologically inspired methods for robust automatic speech recognition. *IEEE Signal Processing Magazine*, 29(6), 2012.

M. J. Taghizadeh, P. N. Garner, H. Bourlard, and H. R. Abutalebi. An integrated framework for multi-channel multi-source localization and source activity detection. In *Proceedings of HSCMA*, 2011.

M. J. Taghizadeh, P. N. Garner, and H. Bourlard. Broadband beampattern for multi-channel speech acquisition and distant speech recognition. In *IEEE 7th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2012.

M. J. Taghizadeh, R. Parhizkar, P. N. Garner, and H. Bourlard. Euclidean distance matrix completion for ad-hoc microphone array calibration. In *18th International Conference on Digital Signal Processing (DSP)*, 2013.

V. Y. F. Tan and C. Fevotte. A study of the effect of source sparsity for various transforms on blind audio source separation performance. In *Proceeding of Workshop on Signal Processing with Adaptative Sparse Structured Representations (SPARS)*, 2005.

G. Tang, B. Bhaskar, P. Shah, and B. Recht. Compressed sensing off the grid. *available at arXiv:1207.6053*, 2012.

M. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1, 2001.

Harry L. Van Trees. *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. John Wiley and Sons, Inc., Print ISBN: 9780471093909, 2002.

J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12), 2007.

J. A. Tropp and S. J. Wright. Computational methods for sparse solution of linear inverse problems,. *Proceedings of the IEEE*, 98, 2010.

E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation (code available at http://www.irisa.fr/metiss/sassec07/?show=results). *IEEE transactions on audio, speech, and language processing*, 14, 2006.

D. Wang and G. J. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

L. Wang, H. Ding, and F. Yin. A region-growing permutation alignment approach in frequency-domain blind source separation of speech mixtures. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 19(3), 2011.

S. Winter, H. Sawada, S. Araki, and S. Makino. Overcomplete BSS for convolutive mixtures based on hierarchical clustering. In *Proceeding of Independent Component Analysis (ICA)*, 2004.

D. P. Wipf and B. D. Rao. An empirical bayesian strategy for solving the simultaneous sparse approximation problem. *IEEE Transactions on Signal Processing*, 55(7), 2007.

M. Wolfel and J. McDonough. Distant speech recognition. *New York: John Wiley & Sons*, 2009.

T. Wolff and M. Buck. A generalized view on microphone array postfilters. In *International Workshop on Acoustic Signal Enhancement*, 2010.

G. Xu, H. Liu, L. Tong, and T. Kailath. A least-squares approach to blind channel identification. *IEEE Transactions on Signal Processing*, 1995.

H. H. Yang and H. Hermansky. Search for information bearing components in speech. In *Advances in Neural Information Processing Systems 9*, 2000.

O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52:1830–1847, 2004.

T. Yoshioka, T. Nakatani, M. Miyoshi, and H. Okuno. Blind separation and dereverberation of speech mixtures by joint optimization. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1), 2010.

X. Zhang, M. G. Heinz, I. C. Bruce, and L. H. Carney. A phenomenological model for the response of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression. *Journal of Acoustic Societ of America*, 109(2), 2001.

Z. Zhang and B. D. Rao. Exploiting correlation in sparse signal recovery problems: Multiple measurement vectors, block sparsity, and time-varying sparsity. In *ICML 2011 Workshop on Structured Sparsity: Learning and Inference*, 2011a.

Z. Zhang and B. D. Rao. Sparse signal recovery with temporally correlated source vectors using sparse bayesian learning. *IEEE Journal of Selected Topics in Signal Processing*, 5(5), 2011b.

Z. Zhang and B. D. Rao. Extension of sbl algorithms for the recovery of block sparse signals with intra-block correlation. *IEEE Transactions on Signal Processing*, 2012.

M. Zibulevsky and B. A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation*, 13(4), 2001.

G. K. Zipf. *Human behavior and the principle of least effort*. Addison-Wesley Press, 1949.