UNIVERSIDAD
POLITÉCNICA
DE VALENCIA

Multichannel Audio Processing for Speaker Localization, Separation and Enhancement

DOCTORAL THESIS

by
Amparo Martí Guerola

Supervisors:
Dr. Máximo Cobos Serrano
Dr. José Javier López Monfort

Valencia, Spain
July 2013

To Dad

Abstract

This thesis is related to the field of acoustic signal processing and its applications to emerging communication environments. Acoustic signal processing is a very wide research area covering the design of signal processing algorithms involving one or several acoustic signals to perform a given task, such as locating the sound source that originated the acquired signals, improving their signal to noise ratio, separating signals of interest from a set of interfering sources or recognizing the type of source and the content of the message. Among the above tasks, *Sound Source localization* (SSL) and *Automatic Speech Recognition* (ASR) have been specially addressed in this thesis. In fact, the localization of sound sources in a room has received a lot of attention in the last decades. Most real-word microphone array applications require the localization of one or more active sound sources in adverse environments (low signal-to-noise ratio and high reverberation). Some of these applications are teleconferencing systems, video-gaming, autonomous robots, remote surveillance, hands-free speech acquisition, etc. Indeed, performing robust sound source localization under high noise and reverberation is a very challenging task. One of the most well-known algorithms for source localization in noisy and reverberant environments is the *Steered Response Power - Phase Transform* (SRP-PHAT) algorithm, which constitutes the baseline framework for the contributions proposed in this thesis. Another challenge in the design of SSL algorithms is to achieve real-time performance and high localization accuracy with a reasonable number of microphones and limited computational resources. Although the SRP-PHAT algorithm has been shown to be an effective localization algorithm for real-world environments, its practical implementation is usually based on a costly fine grid-search procedure, making the computational cost of the method a real issue. In this context, several modifications and optimizations have been proposed to improve its performance and applicability. An effective strategy that extends the conventional SRP-PHAT functional is presented in this thesis. This approach performs a full exploration of the sampled space rather than computing the SRP at discrete spatial positions, increasing its robustness and allowing for a coarser spatial grid that reduces the computational cost required in a practical implementation with a small hardware cost (reduced number of microphones). This strategy allows to

implement real-time applications based on location information, such as automatic camera steering or the detection of speech/non-speech fragments in advanced videoconferencing systems.

As stated before, besides the contributions related to SSL, this thesis is also related to the field of ASR. This technology allows a computer or electronic device to identify the words spoken by a person so that the message can be stored or processed in a useful way. ASR is used on a day-to-day basis in a number of applications and services such as natural human-machine interfaces, dictation systems, electronic translators and automatic information desks. However, there are still some challenges to be solved. A major problem in ASR is to recognize people speaking in a room by using distant microphones. In distant-speech recognition, the microphone does not only receive the direct path signal, but also delayed replicas as a result of multi-path propagation. Moreover, there are multiple situations in teleconferencing meetings when multiple speakers talk simultaneously. In this context, when multiple speaker signals are present, *Sound Source Separation* (SSS) methods can be successfully employed to improve ASR performance in multi-source scenarios. This is the motivation behind the training method for multiple talk situations proposed in this thesis. This training, which is based on a robust transformed model constructed from separated speech in diverse acoustic environments, makes use of a SSS method as a speech enhancement stage that suppresses the unwanted interferences. The combination of source separation and this specific training has been explored and evaluated under different acoustical conditions, leading to improvements of up to a 35% in ASR performance.

Keywords: Sound source localization, sound source separation, SRP-PHAT, microphone array, speaker detection, automatic speech recognition.

Resumen

Esta tesis se enmarca en el campo del procesado de señales acústicas y sus aplicaciones para entornos de comunicación emergentes. El procesado de señales acústicas es un área de investigación muy amplia que abarca el diseño de algoritmos para el tratamiento de una o varias señales acústicas con el fin de realizar una tarea determinada, como puede ser: la localización de la fuente de sonido que originó las señales acústicas adquiridas, la mejora de la relación señal a ruido de las mismas, la separación de señales de interés a partir de un conjunto de fuentes interferentes o el reconocimiento del tipo de fuente y/o el contenido del mensaje. Entre las tareas anteriores, la localización de fuente de sonidos (SSL, *Sound Source Localization*) y el reconocimiento automático de voz (ASR, *Automatic Speech Recognition*) han sido especialmente tratados en esta tesis. De hecho, la localización de fuentes de sonido en una habitación ha recibido mucha atención por parte de la comunidad científica en las últimas décadas. La mayoría de las aplicaciones reales de arrays de micrófonos necesitan localizar una o más fuentes de sonido activas en condiciones adversas (baja relación señal-ruido y una alta reverberación). Algunas de estas aplicaciones son los sistemas de teleconferencia, videojuegos, robots autónomos, sistemas remotos de vigilancia, la adquisición de señal en modo manos libres, etc. De hecho, la localización robusta de fuentes de sonido bajo condiciones de alto nivel de ruido y reverberación sigue siendo un reto. Uno de los algoritmos más conocidos para la localización de fuentes en entornos ruidosos y reverberantes es el *Steered Response Power - Phase Transform* (SRP-PHAT), que constituye el marco de referencia para las contribuciones que se proponen en esta tesis. Otro desafío en el diseño de algoritmos de SSL es lograr su funcionamiento en tiempo real con una alta precisión en la localización y con un número razonable de micrófonos a un coste computacional reducido. Aunque el algoritmo SRP-PHAT ha demostrado ser un algoritmo de localización efectivo en entornos reales, su aplicación práctica se basa por lo general en un procedimiento costoso de búsqueda por mallado, por lo que el coste computacional de este método supone un problema a considerar. Es por ello que diversas modificaciones y optimizaciones se han propuesto en la literatura para mejorar su rendimiento y aplicabilidad. En esta tesis se propone una nueva estrategia que extiende eficazmente el comportamiento

del algoritmo SRP-PHAT convencional. Este nuevo método realiza una exploración completa del espacio muestreado en lugar de calcular el SRP en posiciones espaciales discretas, aumentando así su robustez y permitiendo un mallado espacial más ancho que reduce el coste computacional requerido en una aplicación práctica, reduciendo también el coste en hardware (menor número de micrófonos). Esta estrategia permite implementar aplicaciones en tiempo real basándose en la información de las posiciones estimadas, como por ejemplo redirigir de forma automática la posición de una cámara o la detección de fragmentos de habla / no habla en sistemas avanzados de videoconferencia.

Como se ha comentado anteriormente, además de las contribuciones relacionadas con SSL, esta tesis está también relacionada con el campo del reconocimiento automático de voz (ASR). Esta tecnología permite a un ordenador o dispositivo electrónico identificar las palabras pronunciadas por una persona para que el mensaje se pueda almacenar y procesar de una forma útil. ASR es utilizado en el día a día en una serie de aplicaciones y servicios, como interfaces hombre-máquina naturales, sistemas de dictado, traductores electrónicos y mostradores de información automática. Sin embargo, aún existen algunos desafíos que hay que resolver. Un problema importante en ASR es reconocer a las personas que están hablando en una habitación mediante el uso de micrófonos a distancia. En el reconocimiento de voz distante, los micrófonos no sólo reciben la señal vía directa de las fuentes de sonido, sino que también reciben réplicas retardadas como resultado de la propagación multitrayecto. Por otra parte, también existen múltiples situaciones en las teleconferencias en las que varios oradores hablan simultáneamente. En este contexto, cuando múltiples señales de voz están presentes simultáneamente, los métodos de separación de fuentes de sonido (SSS, *Sound Source Separation*) pueden emplearse con éxito para mejorar el rendimiento del reconocimiento automático de voz en escenarios con múltiples fuentes. Con el objetivo de mejorar este tipo de situaciones, en esta tesis se ha propuesto un método de entrenamiento diferente. Este entrenamiento, el cual se basa en un modelo robusto construido a partir de voces previamente separadas en diversos entornos acústicos, utiliza las técnicas de separación como una etapa de mejora del habla que suprime las interferencias no deseadas. Se ha estudiado la combinación de la separación de fuentes y el uso de este entrenamiento específico para la mejora del reconocimiento de voz en diferentes condiciones acústicas, dando lugar a mejoras de hasta un 35% en la tasa final de reconocimiento.

Palabras Clave: Localización de fuentes de sonido, separación de fuentes de sonido, SRP-PHAT, array de micrófonos, detección de habla, reconocimiento automático de voz.

Resum

Aquesta tesi s'emmarca en el camp del processament de senyals acústics i les seves aplicacions per a entorns de comunicació emergents. El processat de senyals acústics és una àrea de recerca molt àmplia que abasta el disseny d'algorismes per al tractament d'un o diversos senyals acústics per tal de realitzar una tasca determinada, com pot ser: la localització de la font de so que va originar els senyals acústics aconseguits, la millora de la relació senyal a soroll de les mateixes, la separació de senyals d'interès a partir d'un conjunt de fonts interferents o el reconeixement del tipus de font i / o el contingut del missatge. Entre les tasques anteriors, la localització de font de sons (SSL, *Sound Source Localization*) i el reconeixement automàtic de veu (ASR, *Automatic Speech Recognition*) han estat especialment tractades en aquesta tesi. De fet, la localització de fonts de so en una habitació ha rebut molta atenció per part de la comunitat científica en les últimes dècades. La majoria de les aplicacions reals d'arrays de micròfons necessiten localitzar una o més fonts de so actives en condicions adverses (baixa relació senyal-soroll i una alta reverberació). Algunes d'aquestes aplicacions són els sistemes de teleconferència, videojocs, robots autònoms, sistemes remots de vigilància, l'adquisició de senyal en mode mans lliures, etc. De fet, la localització robusta de fonts de so sota condicions d'alt nivell de soroll i reverberació segueix sent un repte. Un dels algorismes més coneguts per a la localització de fonts en entorns sorollosos i reverberants és el *Steered Response Power - Phase Transform* (SRP-PHAT), que constitueix el marc de referència per a les contribucions que es proposen en aquesta tesi. Un altre desafiament en el disseny d'algorismes de SSL és aconseguir el seu funcionament en temps real amb una alta precisió en la localització i amb un nombre raonable de micròfons a un cost computacional reduït. Encara que el algoritme SRP-PHAT ha demostrat ser un algorisme de localització efectiu en entorns reals, la seva aplicació pràctica es basa en general en un procediment costós de recerca per mallat, pel que el cost computacional d'aquest mètode suposa un problema a considerar. És per això que diverses modificacions i optimitzacions s'han proposat en la literatura per millorar el seu rendiment i aplicabilitat. En aquesta tesi es proposa una nova estratègia que estén eficaçment el comportament de l'algorisme SRP-PHAT convencional. Aquest nou mètode realitza una exploració completa

de l'espai mostrejat en lloc de calcular el SRP en posicions espacials discretes, augmentant així la seva robustesa i permetent un mallat espacial més ample que el cost computacional requerit en una aplicació pràctica, reduint també el cost en hardware (menor nombre de micròfons). Aquesta estratègia permet implementar aplicacions en temps real basant-se en la informació de les posicions estimades, com ara redirigir de forma automàtica la posició d'una càmera o la detecció de fragments de parla / no parla per a sistemes avançats de videoconferència.

Com s'ha comentat anteriorment, a més de les contribucions relacionades amb SSL, aquesta tesi està també relacionada amb el camp del reconeixement automàtic de veu (ASR). Aquesta tecnologia permet a un ordinador o dispositiu electrònic identificar les paraules pronunciades per una persona perquè el missatge es pugui emmagatzemar i processar d'una forma útil. ASR és utilitzat en el dia a dia en una sèrie d'aplicacions i serveis, com a interfaces home-màquina naturals, sistemes de dictat, traductors electrònics i taulells d'informació automàtica. No obstant això, encara hi ha alguns reptes que cal resoldre. Un problema important en ASR és reconèixer a les persones que estan parlant en una habitació mitjançant l'ús de micròfons a distància. En el reconeixement de veu distant, els micròfons no només reben el senyal via directa de les fonts de so, sinó que també reben rèpliques retardades com a resultat de la propagació multitrajecte. D'altra banda, també hi ha múltiples situacions en les teleconferències en què diversos oradors parlen simultàniament. En aquest context, quan múltiples senyals de veu són presents simultàniament, els mètodes de separació de fonts de so (SSS, *Sound Source Separation*) es poden utilitzar amb èxit per millorar el rendiment del reconeixement automàtic de veu en escenaris amb múltiples fonts. Amb l'objectiu de millorar aquest tipus de situacions, en aquesta tesi s'ha proposat un mètode d'entrenament diferent. Aquest entrenament, el qual es basa en un model robust construït a partir de veus prèviament separades en diversos entorns acústics, utilitza les tècniques de separació com una etapa de millora de la parla que suprimeix les interferències no desitjades. S'ha estudiat la combinació de la separació de fonts i l'ús d'aquest entrenament específic per a la millora del reconeixement de veu en diferents condicions acústiques, donant lloc a millores de fins a un 35% en la taxa final de reconeixement.

Paraules Clau: Localització de fonts de so, separació de fonts de so, SRP-PHAT, array de micròfons, detecció de parla, reconeixement automàtic de veu.

Acknowledgements

This thesis would not have been possible without the unconditional support of many people. First and foremost I would like to offer my sincerest gratitude to my supervisors Dr. Máximo Cobos and José Javier López, both have offered me valuable support throughout this thesis. Without their advices and continuous technical support, this thesis would have been not possible for me.

I would like to show my gratitude to all the people at the Universitat Politècnica de València who shared my daily work at the Institute of Telecommunications and Multimedia Applications. In particular, I would like to thank my colleagues: Sandra Roger, Ana Torres, Laura Fuster, Emanuel Aguilera, Fernando Domene, Luis Maciá, Jose Antonio Belloch and Jorge Lorente. I have shared great moments with all these people.

I wish also to thank the Spanish Ministry of Science and Innovation for the received financial support under the FPI program.

Special thanks to my best friend and partner, Raúl, for all his support specially in the bad moments.

Finally, I would like to wish the best to my family. I would like to pay special tribute to my father. I am deeply indebted to him. He was always a hardworking man who taught me to fight for my objectives and never give up. Wherever you are, I hope you're proud of me.

Amparo Martí
July 2013

Contents

List of Figures

Abbreviations and Acronyms

ASR	Automatic Speech Recognition
CFRC	Coarse-to-Fine Region Contraction
DFT	Discrete Fourier Transform
DNN	Deep Neural Networks
DOA	Direction Of Arrival
FDOA	Frequency Difference Of Arrival
FRSCM	Full-Rank Spatial Covariance Model
GCC	Generalized Cross Correlation
HMM	Hidden Markov Models
ICA	Independent Component Analysis
IBM	Ideal Binary Mask
IMTDF	Inter-Microphone Time-Delay Function
ISIP	Institute for Signal and Information Processing
MFCCs	Mel-Frequency Cepstral Coefficients
ML	Maximum Likelihood
MuLeTS	Multi Level Thresholding Separation
MUSIC	Multiple Signal Classification
PCA	Principal Component Analysis
PHAT	Phase Transform
RMSE	Root Mean Square Error
SNR	Signal-to-Noise Ratio
SRC	Stochastic Region Contraction
SRP	Steered Response Power
SSL	Sound Source localization
SSS	Sound Source Separation
STFT	Short Time Fourier Transform
TDOA	Time Delay Of Arrival
TDE	Time Delay Estimation
WDO	W-Disjoint Orthogonality
WER	Word Error Rate
WRR	Word Recognition Rate

Introduction 1

1.1 Background and Motivation

The human ability to distinguish when a sounding object is close or far from us is completely developed when we are just a few months old [1]. In fact, the development of the localization mechanisms used by the human auditory system takes place before being one year old [2]. The localization of sound sources is possible because the human brain analyzes all the signals arriving through our ears, using subtle differences in intensity and other spectral and timing cues to recognize the direction of one or even several sound sources [3; 4].

While localizing sound sources does not require any special effort for a human subject, for machines, sound source localization in a room is a complicated process since not all the sound objects have the same spectral properties, they occur at different time instants and at different spatial positions, and the process is strongly affected by reflections. Acoustic reflections dominate the perception of sound in a room modifying the spatial characteristics of the perceived sources.

Over the last decades, the scientific community has dedicated many efforts to localize sound sources in space by means of microphone array

systems and, today, achieving high localization performance is still a challenge. Microphone arrays have been used in many applications, such as speech recognition [5], teleconferencing systems [6; 7], hands-free speech acquisition [8], digital hearing aids [9], video-gaming [10], autonomous robots [11] and remote surveillance [12]. All these applications require the localization of one or more acoustic sources. In fact many audio processing applications can be improved when the spatial location of the source of interest is known. For this reason, the volume of research works in this area has increased considerably over the last years.

Many current SSL systems assume that the sound sources are distributed on a horizontal plane [13; 14]. This assumption simplifies the problem of SSL in almost all methods. For example, in teleconference applications all talkers are assumed to speak at the same height which is somewhat true, but the talker or other attendees can act as sound blockades between the main talker and the array. Moreover, in most dominant SSL methods, the required computational cost is usually very high, even when the sources are assumed to be on the same plane [15]. Some of these SSL methods have been modified to cover a three dimensional space at a very high computational cost. Thus, the development of 3D localization methods having low complexity is still a very challenging task. There is also another problem to take into account, that is the reflections of the sound signal in the different walls, floor and objects around. These reflections interfere in the system making more difficult the localization. As a result, SSL systems must perform robustly and work in noisy and reverberant environments.

Algorithms for SSL can be broadly divided into indirect and direct approaches [16]. Indirect approaches usually follow a two-step procedure: they first estimate the *Time Difference Of Arrival* (TDOA) [17] between microphone pairs and, afterwards, they estimate the source position based on the geometry of the array and the estimated delays. On the other hand, direct approaches perform TDOA estimation and source localization in one single step by scanning a set of candidate source locations and selecting the most likely position as an estimate of the source location. In addition, information theoretic approaches have also shown to be significantly powerful in source localization tasks [18].

The SRP-PHAT algorithm is a direct approach that has been shown to be very robust under difficult acoustic conditions [19; 20; 13]. The al-

gorithm is commonly interpreted as a beamforming-based approach that searches for the candidate source position that maximizes the output of a steered delay-and-sum beamformer. However, despite its robustness, computational cost is a real issue because the SRP space to be searched has many local extrema [21]. Very interesting modifications and optimizations have already been proposed to deal with this problem, such as those based on *Stochastic Region Contraction* (SRC) [22] and *Coarse-to-Fine Region Contraction* (CFRC) [23], achieving a reduction in computational cost of more than three orders of magnitude.

SSL methods can be a very useful tool for human-machine interaction systems, since the spatial information provided by localization algorithms is essential for mimicking the natural human ability to discriminate the position of a talker. This can be used for setting up a spatial audio reproduction system, detecting active speakers or as a speech enhancement stage aimed at suppressing noise or reverberation from the signal of interest. In this context, another of the contributions of this thesis is the development of a speech/non-speech discrimination technique based on the statistical distribution of location estimates. The automatic detection of speech frames can be very useful to manage audio levels, suppressing noise and improving ASR systems. In fact, human-machine interaction is also closely related, from an acoustic point of view, to ASR, which is another topic covered in this thesis. One major problem in ASR is to recognize people speaking in a room by using a distant microphone [24]. In distant-speech recognition, the microphone does not only receive the direct path signal, but also delayed replicas as a result of multi-path propagation [25]. The existing mismatch between the training and testing conditions limits the performance of ASR systems, thus, robust recognition methods are aimed at reducing this mismatch. In this context, several approaches have been proposed to cope with room reverberation in speech recognition applications. Some methods are based on a speech enhancement stage prior to recognition [26]. In other methods, the recognizer itself is made robust to reverberation by using model compensation or by performing an improved feature extraction process [27]. All these approaches have shown to be useful to improve ASR. In any case, reducing the acoustic mismatch training and testing conditions seems to be a relevant issue for the development of robust ASR systems. Despite the fact that speech recognizers are usually trained on anechoic (or almost anechoic) conditions, the environment where they are usually employed can be rarely considered anechoic. To this end,

the use of different training data matching several acoustic environments has already been suggested [28; 29], yielding a noticeable improvement.

One of the most challenging problems nowadays is to provide a comfortable conversation with a remote partner where one of them or both are in adverse environments. By adverse environment we mean noisy offices, railway stations, airports, shop floors, etc. Similar problems have to be solved when a speech recognition system is used. Under comfortable environment we understand that a speaker does not have to be wired, i.e. to carry or to hold a microphone very close to the mouth. The talker should be able just to talk without caring where the microphone(s) is(are) located. The partner at the remote location or a speech recognition system should just receive the speech signal as clearly as possible.

Many efforts have been made to develop robust ASR systems working in reverberant and noisy conditions, most of them focused on recognizing a single speech source. However, besides noise and reverberation, cocktail-party situations where different speakers are talking at the same time pose a real problem for ASR systems [30; 31]. Source separation algorithms have been described in the literature as a solution for simultaneous speech recognition. However, separated speech signal present an additional mismatch with respect to the training signals used in a conventional ASR system.

1.2 Objectives of the Thesis

Taking into account the above context, the main scope of this thesis is as follows:

To deepen into signal processing algorithms for sound source localization, separation and enhancement with microphone arrays, providing a robust and low-complexity front-end for automatic speech recognition systems.

Some particular aims emerge from the main scope, which are presented as follows:

- To develop robust and high-accuracy SSL algorithms working in adverse acoustic environments.
- To develop low-complexity SSL methods, avoiding the need for high

hardware and computational resources.

- To evaluate the performance of SSL methods in real application environments.
- To integrate other acoustic processing tasks into the proposed SSL frameworks, such as speech/non-speech classification or audio-based camera-steering.
- To apply multichannel source separation techniques to ASR systems, improving the percentage of recognized words in simultaneous speech cases.
- To study the advantages and disadvantages of using source separation methods in the context of ASR.

1.3 Organization of Thesis

This thesis work is carried out in 3 different stages. The advantages of this new algorithm are demonstrated in different acoustic environments, including its application to real videoconferencing systems.

Secondly, a post-processing technique using the information provided by the SSL algorithms is proposed, which is aimed at increasing the signal-to-noise ratio of the captured speech. In this method, the location information is processed to generate a binary time-frequency mask using the advantages provided by SRP-PHAT localization.

Finally, different ways to train an ASR system are proposed to deal with simultaneous speech cases, studying the performance and influence of different kinds of source separation methods.

This thesis work is organized as follows:

- Chapter 2: This chapter is intended to give a comprehensive overview of different acoustical signal processing. The chapter is focused specifically in three different aspects:
  - Sound source localization methods and applications.
  - Automatic speech recognition.

  - Source separation separation.

- Chapter 3: This chapter summarizes the findings of this research work, revisiting each scientific paper associated to this thesis by means of a sort description of the paper and its main contributions.

- Chapter 4-9: These chapters correspond to the scientific papers published throughout this thesis, reformatted to this book style.

- Chapter 10: Finally, conclusions obtained in this dissertation are presented, including some guidelines for future research related to the presented contributions.

State of the Art 2

Although the papers included within this thesis contain a brief review of the state of the art related to each specific contribution, we have considered appropriate to present an extended overview of the state of the art in this chapter. This will make clearer the novelty of the approaches proposed in the rest of the thesis.

The chapter is structured to cover three fundamental topics:

- Sound source localization (SSL) methods and applications.
- Automatic speech recognition (ASR).
- Sound source separation (SSS).

2.1 Sound Source Localization

As commented in the introductory chapter, the localization of sound sources by humans is based on the analysis that the brain performs on the signals arriving to the ears, using subtle differences in intensity and other spectral and timing cues to recognize the direction of the source that emitted the signal [3; 4]. Automatic SSL methods make use of microphone arrays and

complex signal processing techniques to perform the same task, however, undesired effects such as acoustic reflections and noise make this process difficult, being currently a hot research topic in acoustic signal processing. In the next subsections, some well-known localization approaches are explained, which establish the framework for the contributions developed throughout this thesis.

2.1.1 Signal Model

The location of a source can be determined from signals received at several sensors. One of the most effective methods is to use estimates of the TDOA and/or the frequency-difference-of-arrival (FDOA) between pairs of signals received at the sensors [32].

The assumed (anechoic) signal model for the time-delay between receiving signals at sensors, $m_1(t), m_2(t), ..., m_M(t)$, is given by

$$\begin{aligned} m_1(t) &= a_1 s(t - t_1) + w_1(t) \\ m_2(t) &= a_2 s(t - t_2) + w_2(t) \\ &\vdots \\ m_M(t) &= a_M s(t - t_M) + w_M(t) \end{aligned} \tag{2.1}$$

where a_m are the amplitude attenuation factors, t_m are the signal arrival time delays and $w_m(t)$ are additive noise signals. We assume that the noise is stationary white Gaussian noise and uncorrelated with the signal of interest $s(t)$. Given Equation 2.1, $\tau = t_1 - t_2$, would be the TDOA between microphones 1 and 2.

In the frequency domain, the signal model is given by

$$\begin{aligned} M_1(w) &= A_1 S(w) e^{-jwt_1} + W_1(w) \\ M_2(w) &= A_2 S(w) e^{-jwt_2} + W_2(w) \\ &\vdots \\ M_M(w) &= A_M S(w) e^{-jwt_M} + W_M(w), \end{aligned} \tag{2.2}$$

where the signal, noise, and received signal have spectral densities $G_{s,s}(w) = E[S(w)S^*(w)]$, $G_{w_1,w_1}(w) = E[W_1(w)W_1^*(w)]$, and $G_{m_1,m_1}(w) = E[M_1(w)M_1^*(w)]$,

respectively. The amplitudes A_m are dependent on the distance from the source to the microphones, the directivity of the source and the microphone, the properties of the reflective surfaces, and the air absorption. Here the amplitudes are assumed to be equal to unity, i.e., $A_m = 1, \forall m$. This model is assumed for simplicity in the cases studied in this thesis.

2.1.2 Time Difference Of Arrival (TDOA)

Most practical acoustic source localization schemes are based on TDOA estimation because these systems are conceptually simple. They are reasonably effective in moderately reverberant environments and, moreover, their low computational complexity makes them well-suited to real-time implementation with several sensors [33].

In general, an array is composed of M microphones, and each microphone is positioned at a unique spatial location. Hence, the direct-path sound waves propagate along M bearing lines, from the source to each microphone, simultaneously. The orientations of these lines in the global coordinate system define the propagation directions of the wave fronts at each microphone. The propagation vectors for a four-element ($m = 1, \ldots, 4$), linear array are illustrated in Figure 2.1, denoted as $\vec{d}_m$.

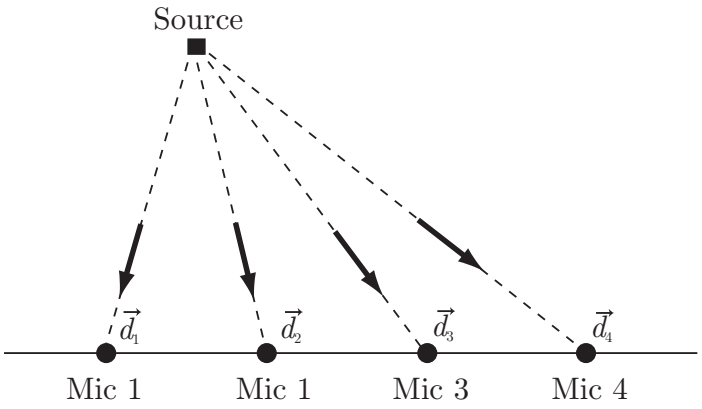


Figure 2.1. Propagation vectors.

Time Delay Estimation (TDE) is concerned with the computation of the relative TDOA between different microphone sensors. It is a funda-

mental technique in microphone array signal processing and the first step in passive TDOA-based acoustic source localization systems. With this kind of localization, a two-step strategy is adopted as shown in Figure 2.2.

The first stage involves the estimation of the TDOA between receivers through the use of TDE techniques [17]. The estimated TDOAs are then transformed into range difference measurements between sensors, resulting in a set of nonlinear hyperbolic range difference equations. The second stage utilizes efficient algorithms to produce an unambiguous solution to these nonlinear hyperbolic equations. The solution produced by these algorithms results in the estimated position location of the source [34]. This data along with knowledge of the microphone positions are then used to generate hyperbolic curves, which are then intersected in some optimal sense to arrive at a source location estimate as shown in Figure 2.3.

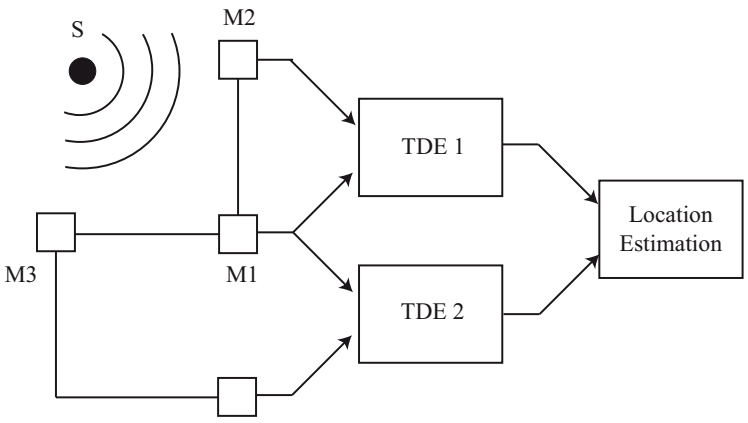


Figure 2.2. A two stage algorithm for sound source localization.

Several variations of this principle have been developed [35]. They differ considerably in the method of derivation, the extent of their applicability (2D versus 3D, near field source versus far field source), and their means of solution.

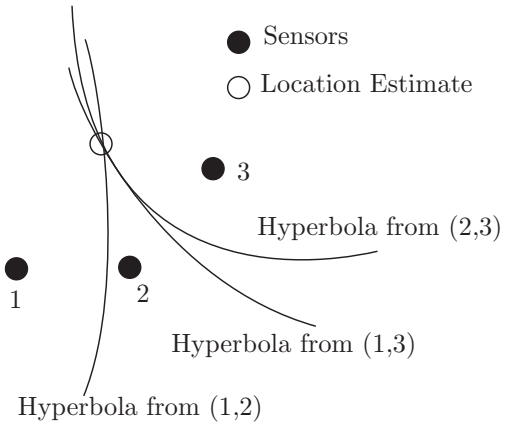


Figure 2.3. Source estimation with three microphones.

2.1.3 Generalized Cross Correlation

The existing strategies of SSL may broadly be divided into two main classes: indirect and direct approaches [16]. Indirect approaches to source localization are usually two-step methods: first, the relative time delays for the various microphone pairs are evaluated and then the source location is found as the intersection of a pair of a set of half-hyperboloids centered around the different microphone pairs. Each half-hyperboloid determines the possible location of a sound source based on the measure of the time difference of arrival between the two microphones. On the other hand, direct approaches generally scan a set of candidate source positions and pick the most likely candidate as an estimate of the sound source location, thus performing the localization in a single step.

For both approaches, techniques such as the *Generalized Cross Correlation* (GCC) method, proposed by Knapp and Carter in 1976, are widely used [36].

Consider the output from microphone l, $m_l(t)$, in an M microphone system. The GCC for a microphone pair (k, l) is computed as

$$R_{m_k m_l}(\tau) = \int_{-\infty}^{\infty} \Phi_{kl}(\omega) M_k(\omega) M_l^*(\omega) e^{-j\omega\tau} d\omega, \tag{2.3}$$

where τ is the time lag, * denotes complex conjugation, $M_l(\omega)$ is the Fourier

transform of the microphone signal $m_l(t)$, and $\Phi_{kl}(\omega) = W_k(\omega)W_l^*(\omega)$ is a combined weighting function in the frequency domain.

The TDE between signals from any pair of microphones can be performed by computing the cross-correlation function of the two signals after applying a suitable weighting step. The lag at which the cross-correlation function has its maximum is taken as the time delay between them.

The type of weighting used with GCC is crucial to localization performance. Among several types of weighting, the phase transform (PHAT) is the most commonly used pre-filter for the GCC because it is more robust against reverberation. The GCC with the phase transform (GCC-PHAT) approach has been shown to perform well in a mild reverberant environment:

$$\Phi_{kl}(\omega) \equiv \frac{1}{|M_k(\omega)M_l^*(\omega)|}. \tag{2.4}$$

Unfortunately, in the presence of even moderate reverberation levels, the algorithm is seriously hampered, due to the presence of spurious peaks. Also reflections of the signal on the walls produce different peaks in the impulse response of the room which can generate peaks in the GCC function that may be strongest than the peak corresponding to the direct path. An example room impulse response is shown in Figure 2.4.

2.1.4 Steered Response Power

Another class of important SSL algorithms is that based on a steered beamformer. When the source location is not known, a beamformer can be used to scan over a predefined spatial region by adjusting its steering parameters. The output of a beamformer is known as the steered response. When the point or direction of scan matches the source location, the SRP will be maximized. However, the localization performance of the conventional steered-beamformer techniques which apply filters to the array signals have been derived to improve its performance. When the phase transform filter is incorporated with the steered-beamformer method, the resulting algorithm (SRP-PHAT) is superior in combating the adverse effects of background noise and reverberation compared to the conventional steered-beamformer method and the pairwise method, GCC-PHAT [36].

Today, the SRP-PHAT algorithm has become a well-known localization

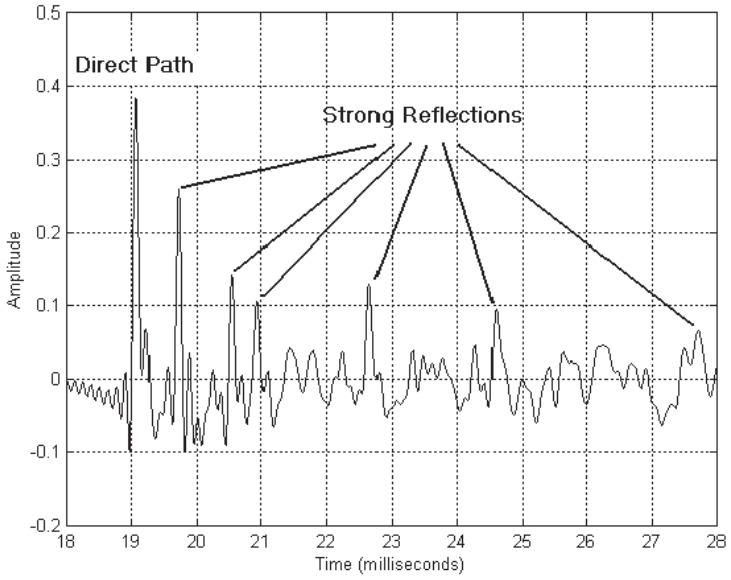


Figure 2.4. Room impulse response from source to one microphone.

method for its robust performance in real environments. However, the computational requirements of the method are large and this makes real-time implementation difficult. Since the SRP-PHAT method was proposed, there have been several attempts to reduce the computational requirements of the intrinsic SRP search process [37; 38].

Array signal processing techniques rely on the ability to focus on signals originating from a particular location or direction in space. Most of these techniques employ some type of beamforming, which generally includes any algorithm that exploits an array's sound-capture ability [39]. Beamforming, in the conventional sense, can be defined by a filter-and-sum process, which applies some temporal filters to the microphone signals before summing them to produce a single, focused signal. These filters are often adapted during the beamforming process to enhance the desired source signal while attenuating others. The simplest filters execute time shifts that have been matched to the source signals propagation delays. This method is referred to as delay-and-sum beamforming; it delays the microphone signals so that

all versions of the source signal are time-aligned before they are summed. The filters of more sophisticated filter-and-sum techniques usually apply this time alignment as well as other signal-enhancing processes.

Beamforming techniques have been applied to both source-signal capture and source localization. If the location of the source is known (and perhaps something about the nature of the source signal is known as well), then a beamformer can be focused on the source, and its output becomes an enhanced version (in some sense) of the inputs from the microphones. If the location of the source is not known, then a beamformer can be used to scan, or steer, over a predefined spatial region by adjusting its steering delays (and possibly its filters). As previously commented, the output of a beamformer, when used in this way, is known as the steered response. The SRP may peak under a variety of circumstances, but with favorable conditions, it is maximized when the steering delays match the propagation delays. By predicting the properties of the propagating waves, these steering delays can be mapped to a location, which should coincide with the location of the source.

For voice capture application, the filters applied by the filter-and-sum technique must not only suppress the background noise and contributions from unwanted sources, they must also do this in way that does not significantly distort the desired signal. The most common of these filters is the phase transform (PHAT), which applies a magnitude-normalizing weighting function to the cross-spectrum of two microphone signals.

We now describe the measurement principle of SRP-PHAT algorithm which is closely related to GCC-PHAT, and then introduce its implementation.

SRP-PHAT algorithm

Consider the output from microphone l, $m_l(t)$, in an M microphone system. Then, the SRP at the spatial point $\mathbf{x} = [x, y, z]$ for a time frame n of length T is defined as

$$P_n(\mathbf{x}) \equiv \int_{nT}^{(n+1)T} \left| \sum_{l=1}^{M} w_l m_l \left(t - \tau(\mathbf{x}, l)\right) \right|^2 dt, \tag{2.5}$$

where w_l is a weight and $\tau(\mathbf{x}, l)$ is the direct time of travel from location $\mathbf{x}$ to microphone l.

Taking into account the symmetries involved in the computation of Eq.(2.5) and removing some fixed energy terms [21], the part of $P_n(\mathbf{x})$ that changes with $\mathbf{x}$ is isolated as

$$P'_n(\mathbf{x}) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} R_{m_k m_l} \left(\tau_{kl}(\mathbf{x})\right), \tag{2.6}$$

where $\tau_{kl}(\mathbf{x})$ is the *Inter-Microphone Time-Delay Function* (IMTDF). This function is very important, since it represents the theoretical direct path delay for the microphone pair (k, l) resulting from a point source located at $\mathbf{x}$. The IMTDF is mathematically expressed as

$$\tau_{kl}(\mathbf{x}) = \frac{\|\mathbf{x} - \mathbf{x}_k\| - \|\mathbf{x} - \mathbf{x}_l\|}{c}, \tag{2.7}$$

where c is the speed of sound, and $\mathbf{x}_k$ and $\mathbf{x}_l$ are the microphone locations.

The SRP-PHAT algorithm consists in evaluating the functional $P'_n(\mathbf{x})$ on a fine grid G with the aim of finding the point-source location $\mathbf{x}_s$ that provides the maximum value:

$$\mathbf{x}_s = \arg\max_{\mathbf{x} \in G} P'_n(\mathbf{x}). \tag{2.8}$$

Implementation

Basically, the SRP-PHAT algorithm is implemented as follows:

1. Define a spatial grid G with a given spatial resolution r. The theoretical delays from each point of the grid to each microphone pair are pre-computed using Eq.(2.7).
2. For each analysis frame, the GCC of each microphone pair is computed as expressed in Eq.(2.3).
3. For each position of the grid $\mathbf{x} \in G$, the contribution of the different cross-correlations are accumulated (using delays pre-computed in 1), as in Eq.(2.6).

4. Finally, the position with the maximum score is selected.

2.1.5 SRP-PHAT Variants

The accuracy of the SRP-PHAT algorithm is limited by the time resolution of the PHAT weighted cross correlation functions [40]. However, despite its robustness, computational cost is a real issue because the SRP space to be searched has many local extrema [13]. Very interesting modifications have already been proposed to improve the SRP-PHAT algorithm. Some of these only affect to the weighting factor [41].

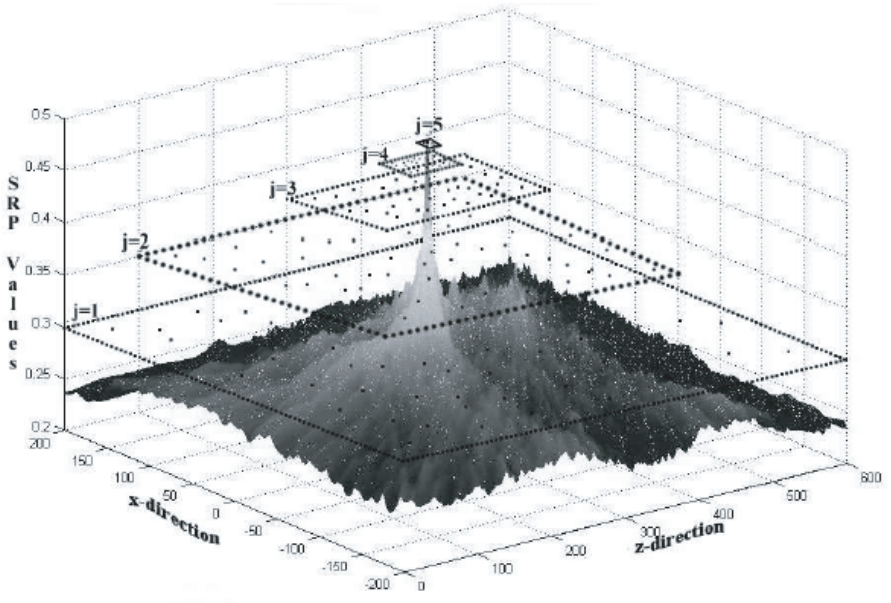


Figure 2.5. 2D example of SRC: j is the iteration index. The rectangular regions show the contracting search regions.

Other modifications of the SRP-PHAT algorithm are focused on reducing the computational cost of that technique, such as those based on *Stochastic Region Contraction* (SRC) [21] and *Coarse-to-Fine Region Contraction* (CFRC) [22]. In SRC, the algorithm starts by considering an initial rectangular search volume containing the desired global optimum and, perhaps, many local maxima or minima. Then, gradually, an iterative process is followed to contract the original volume until a sufficiently small

subvolume is reached, in which the global optimum is trapped (see Figure 2.5 extracted from [22]). A basic stochastic exploration usually controls the contraction operation. In CFRC, which is very similar to SRC, the contraction operation is guided by a sub grid search procedure, reducing also the computational cost of the algorithm in more than three orders of magnitude.

2.2 Automatic Speech Recognition

2.2.1 The Speech Recognition Problem

Communication can be visual, verbal and/or nonverbal. Speech is a verbal method to interact with other people and, for humans, speech is the quickest and most natural form of communication. However, sometimes we need to interact with other people or machines through computers, mobile phones and other interfaces. In this context, ASR systems can be used to identify spoken words received by a microphone and convert them into written text. Although ASR technology is not yet at the point where machines robustly understand speech in any acoustic environment, or by any person, it is used on a day-to-day basis in a number of applications and services such as natural human-machine interfaces, dictation systems, electronic translators and automatic information desks [5].

Speech recognition systems generally assume that the speech signal is a realization of some message encoded as a sequence of one or more symbols [42]. To perform the reverse operation of recognizing the underlying symbol sequence given a spoken utterance, the continuous speech waveform is first converted into a sequence of equally spaced discrete parameter vectors. This sequence of parameter vectors is assumed to be an appropriate representation of the speech waveform for the duration covered by a single vector (typically 10 ms), where speech is approximately stationary. Although not strictly true, it is a reasonable approximation.

The role of the recognizer is to get a mapping between sequences of speech vectors and the wanted underlying symbol sequences. There are three problems which make this task very difficult. Firstly, the mapping from symbols to speech is not one-to-one since different underlying symbols can give rise to similar speech sounds. Furthermore, there are large vari-

ations in the realized speech waveform due to speaker variability, mood, environment, etc. Secondly, the boundaries between symbols cannot be identified explicitly from the speech waveform. Hence, it is not possible to treat the speech waveform as a sequence of concatenated static patterns. Thirdly, a major problem in ASR systems is to recognize people speaking in a room by using a distant microphone [24]. In distant-speech recognition, the microphone does not only receive the direct path signal, but also delayed replicas resulting from multi-path propagation [25]. Consequently, much of the current research in speech processing is directed toward improving robustness to acoustical variability of all types. Two of the major forms of environmental degradation are produced by additive noise of various forms and the effects of linear convolution. There are three different types of distortion [24]:

- *Noise*, also known as *background noise*, which is any sound different to the desired speech, such as that from machines in a factory, air conditioners, or speech from other speakers. As the SNR decreases, it is to be expected that speech recognition will become more difficult. In addition, the impact of noise on speech recognition accuracy depends as much on the type of noise source as on the SNR. Interference by sources such as background music or background speech is especially difficult to handle, as it is both highly transient in nature and easily confused with the desired speech signal.

- *Echo* and *reverberation*, which are reflections of the sound source arriving some time after the signal on the direct path. The temporal structure of speech waveforms is destroyed by the presence of even a small amount of reverberation. This has a very adverse impact on the recognition accuracy that is obtained from distant-speech systems. Today, ASR is more difficult when the effects of common room reverberation are presented than the effects of additive noise, even at fairly low SNRs.

- Other types of distortions are introduced by environmental factors such as *room modes*, the o*rientation of the speakers head*, or the *Lombard effect*.

Mathematical Formulation

The speech recognition problem can be described as a function that defines a mapping from the acoustic evidence to a single or a sequence of words [43]. Let $X = (x_1, x_2, x_3, ..., x_t)$ represent the acoustic evidence that is generated in time (indicated by the index t) from a given speech signal and belong to the complete set of acoustic sequences, χ. Let $W = (w_1, w_2, w_3, ..., w_n)$ denote a sequence of n words, each belonging to a fixed and known set of possible words, w.

In the statistical framework, the recognizer selects the sequence of words that is more likely to be produced given the observed acoustic evidence. Let $P(W|X)$ denote the probability that the words W were spoken given that the acoustic evidence X was observed. The recognizer should select the sequence of words $\widetilde{W}$ satisfying

$$\widetilde{W} = \arg\max_{W \in w} P(W|X). \tag{2.9}$$

However, since $P(W|X)$ is difficult to model directly, Bayes' rule allows us to rewrite such probability as

$$P(W|X) = \frac{P(W)P(X|W)}{P(X)} \tag{2.10}$$

where $P(W)$ is the probability that the sequence of words W will be uttered, $P(X|W)$ is the probability of observing the acoustic evidence X when the speaker utters W, and $P(X)$ is the probability that the acoustic evidence X will be observed. The term $P(X)$ can be dropped because it is a constant under the max operation. Then, the recognizer should select the sequence of words $\widetilde{W}$ that maximizes the product $P(W)P(X|W)$, i.e.,

$$\widetilde{W} = \arg\max_{W \in w} P(X)P(X|W). \tag{2.11}$$

This framework has dominated the development of speech recognition systems since the 1980s.

Evaluating the Performance of ASR

To evaluate the performance of ASR systems, the most common metric is the word error rate (WER). There are different recognition systems. A

simple recognition system consists in recognizing isolated words, so the performance is simply the percentage of misrecognized words. More complex systems consist in continuous speech recognition where such measure is not efficient because of recognized words can contain three types of errors. The first error, known as word substitution, happens when an incorrect word is recognized in place of the correctly spoken word. The second error, known as word deletion, happens when a spoken word is not recognized (i.e., the recognized sentence does not have the spoken word). Finally, the third error, known as word insertion, happens when extra words are estimated by the recognizer (i.e., the recognized sentence contains more words than what actually was spoken). In the following example, the substitutions are bold, insertions are underlined, and deletions are denoted as $*$.

Correct sentence: "Can you bring me a glass of water, please?"

Recognized sentence: "Can you bring $*$ *a glass of* <u>*cold*</u> *water,* ***police****?"*

To estimate the WER, the correct and the recognized sentence must be first aligned. Then the number of substitutions (S), deletions (D), and insertions (I) can be estimated. The WER is defindes as

$$WER = 100\% \times \left(\frac{S + D + I}{|W|} \right) \tag{2.12}$$

where $|W|$ is the number of words in the sequence of word W.

2.2.2 Hidden Markov Models

Acoustic models, $P(X|W)$, are used to compute the probability of observing the acoustic evidence X when the speaker utters W. One of the challenges in speech recognition is to estimate accurately such model. The variability in the speech signal due to factors like environment, pronunciation, phonetic context, physiological characteristics of the speaker make the estimation a very complex task. The most effective acoustic modeling is based on a structure referred to as Hidden Markov Models (HMM).

A HMM is a stochastic finite-state automaton, which generates a sequence of observable symbols. The sequence of states is a Markov chain, i.e., the transitions between states has an associated probability called tran-

sition probability. Each state has an associated probability function to generate an observable symbol. Only the sequence of observations is visible and the sequence of states is not observable and therefore hidden; hence the name hidden Markov model. A HMM, as illustrated in Figure 2.6, can be defined by

- An output observation alphabet $O = o_1, o_2, ..., o_M$, where M is the number of observation symbols. When the observations are continuous, M is infinite.
- A state space $\Omega = 1, 2, ..., N$.
- A probability distribution of transitions between states. Typically, it is assumed that the next state is dependent only upon the current state (first-order Markov assumption). This assumption makes the learning computationally feasible and efficient. Therefore, the transition probability can be defined as the matrix $A = a_{ij}$, where a_{ij} is the probability of a transition from state i to the state j, i.e.,

$$a_{ij} = P(s_t = j \mid s_{t-1} = i), \qquad 1 \leq i, j \leq N \tag{2.13}$$

  where, s_t is denoted as the state at time t.
- An output probability distribution $B = b_i(k)$ associated with each state. Also known as emission probability, $b_i(k)$ is the probability of generating symbol o_k while in state i, defined as

$$b_i(k) = P(v_t = o_k \mid s_t = i) \tag{2.14}$$

  where v_t is the observed symbol at time t. It is assumed that current output (observation) is statistically independent of the previous outputs (output independence assumption).
- A initial state distribution $\pi = \pi_i$, where π_i is the probability that state i is the first state in the state sequence (Markov chain),

$$\pi_i = P(s_0 = i), \qquad 1 \leq i \leq N \tag{2.15}$$

  Since aij, $b_i(k)$, and ϕ_i are all probabilities, the following constraints

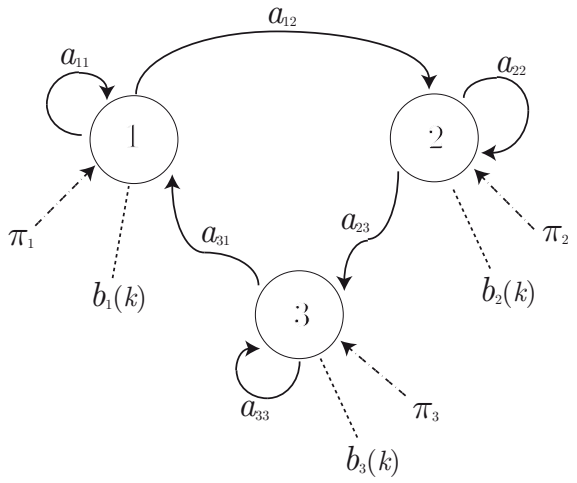


Figure 2.6. A hidden Markov model with three states.

must be satisfied

$$\begin{aligned} a_{ij} &\geq 0, & \Sigma_{j=1}^{N} a_{ij} = 1, \\ b_i(k) &\geq 0, & \Sigma_{k=1}^{M} b_i(k) = 1, \\ \pi_i &\geq 0, & \Sigma_{j=1}^{N} \pi_i = 1, \forall i, j, k. \end{aligned} \tag{2.16}$$

The compact notation $\lambda = (A, B, \pi)$ is used to represent an HMM. The design of an HMM includes choosing the number of states, N, as well as the number of discrete symbols, M, and estimate the three probability densities, A,B, and π.

2.2.3 Speech Recognition Evaluations

Since the beginning of the speech research, several speech recognition systems have been developed for all kinds of purpose. Most of the work was on tasks and speech data elaborated by the developers themselves. The problem is that it is almost impossible to replicate results to perform any type of comparison. Differences in the measurement methodology, task conditions, or testing data can lead to an erroneous comparison between systems.

HTK is a toolkit for building HMMs [42]. HMMs can be used to model any time series and the core of HTK is similarly general-purpose. However, HTK is primarily designed for building HMM-based speech processing tools, in particular recognizers. Thus, much of the infrastructure support in HTK is dedicated to this task. As shown in Figure 2.7, there are two major processing stages involved. Firstly, the HTK training tools are used to estimate the parameters of a set of HMMs using training utterances and their associated transcriptions. Secondly, unknown utterances are transcribed using the HTK recognition tools.

For isolated word recognition; firstly, a HMM is trained for each vocabulary word using a number of examples of that word. Secondly, to recognize some unknown word, the likelihood of each model generating that word is calculated and the most likely model identifies the word.

Speech recognition has significantly improved in the last decade. These improvements are the result of many research efforts in three different areas. Firstly, the use of common speech corpora allows the use of large training sets and makes able to compare results from different ASR systems. Secondly, many developments have been observed in the area of acoustic modeling, such as contributions regarding context-specific HMMs, changes in feature vectors over time or the presence of cross-word effects. Finally, improvements in language modeling and search algorithms have allowed for better recognition of large vocabulary corpora and reduced experimentation cycles, respectively. Unfortunately, most of the above improvements have been developed assuming clean speech. When common ASR systems are used in reverberant and/or noisy environments, the speech signal is degraded and the extracted data vectors differ significantly from the ones expected by the recognizer. In fact, not only are the acoustical conditions responsible for these changes, but also the speaker tends to change his/her voice as a function of the auditory feedback. As a result, to reduce the error-rate in the recognition task, a processing should be included to reduce the differences between training and test environments. This can be done in two ways: by producing changes in the speech model parameters to match the training environment or by transforming the acquired input data to the environment where the models where trained, both methods have been studied.

The existing mismatch between the training and testing conditions limits the performance of ASR systems, thus, robust recognition methods

are aimed at reducing this mismatch. In this context, several approaches have been proposed to cope with room reverberation in speech recognition applications. Increasing the amount of training data generally decreases the WER. However, it is important that the increased training be representative of the types of data in the test. Otherwise, the increased training might not help. Some methods are based on a speech enhancement stage prior to recognition. In other methods, the recognizer itself is made robust to reverberation by using model compensation or by performing an improved feature extraction process. All these approaches have shown to be useful in improving ASR. However, besides noise and reverberation, cocktail-party situations where different speakers are talking at the same time pose a real problem for ASR systems.

Source separation refers to the task of estimating and recovering independent source signals (for example, speech signals) from a set of mixtures in one or several observation channels (microphone signals). Source separation algorithms have been described in the literature as a solution for simultaneous speech recognition, proposing ASR performance as an indicator of the quality achieved by a given source separation algorithm. However, separating speech signals in real acoustic environments is not an easy task and the extracted speech signals are usually corrupted by audible artifacts. Then, separated speech signals present an additional mismatch with respect to the training signals used in a conventional ASR system.

2.3 Source Separation and Enhancement

Source Separation algorithms currently constitute one of the most active research fields in signal processing. Algorithms for source separation have been applied to many areas, ranging from image and video processing to biomedical applications. In the audio context, SSS aims at recovering each source signal from a set of audio mixtures of the original sources, such as those obtained by a microphone array, a binaural recording or an audio CD. Therefore, several applications can emerge from the development of advanced SSS techniques, including music remixing, speech enhancement, automatic music transcription or music information retrieval systems. In the next subsections different scenarios for sound separation techniques, features of the signals and some approaches are presented.

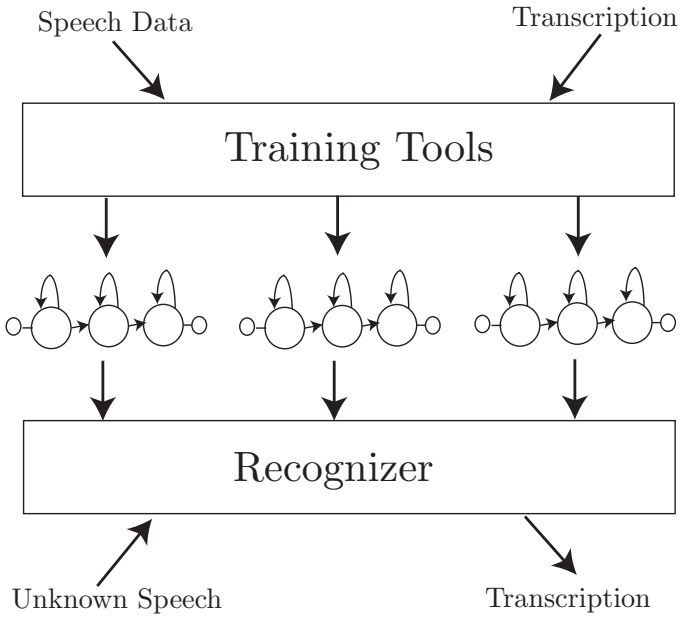


Figure 2.7. The fundamentals of HTK.

2.3.1 Mixing Models

Mixing scenarios are very varied and they determine the nature of the resulting observed mixtures. Generally, the different mixing situations can be mathematically expressed by means of a model that describes how the observations are generated. This is the reason why these models are called *generative models*. Before introducing these models, it is important to clarify the notation used in this section for sampled signals. Although an academic distinction between continuous (t) and discrete $[n]$ time variables is normally used in signal processing works ($s[n] = s(nT_s)$, being T_s the sampling interval), all the signals considered hereafter are assumed to be discrete. Therefore, this distinction is not necessary and the notation (t) has been chosen as the widely adopted in the source separation field, where $t = 1, \ldots, T$ denotes discrete time observations and source signals are indexed by $n = 1, \ldots, N$.

We consider a general setup where M sensors are exposed to N sound sources. As established by the *principle of superposition*, the electrical signal at the m-th channel resulting from this setup can be mathematically

expressed as the scalar addition of the instantaneous amplitudes corresponding to the different *source images*:

$$x_m(t) = \sum_{n=1}^{N} s_{mn}(t), \quad m = 1, \ldots, M, \tag{2.17}$$

where $s_{mn}(t)$ is the image of the n-th source in the m-th microphone at time sample t. These images of the sources represent how the original source signals $s_n(t)$ are recorded at each sensor after being modified by the mixing process (which in the general case can be modeled by a filter $h_{mn}(t)$). Figure 2.8 shows the relations existent between all these signals with an example with two microphones ($M = 2$) and two speakers ($N = 2$).

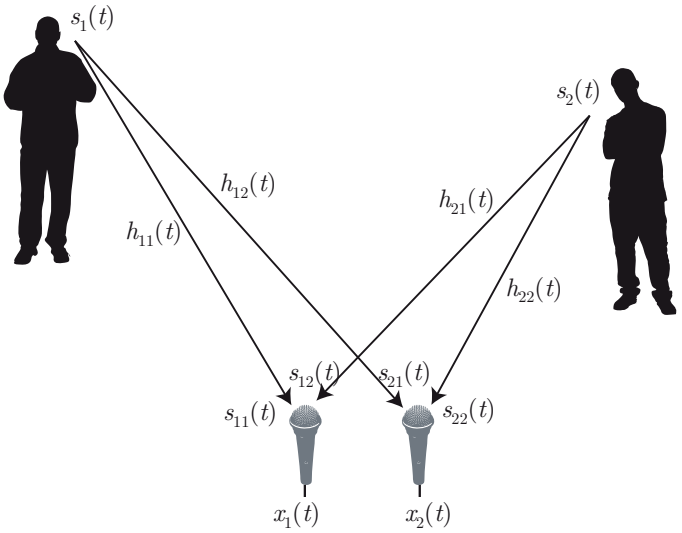


Figure 2.8. Two microphones picking up the signals from two speakers and the signals involved in the mixing process.

In the following subsections the different models in source separation are described. The images of the sources vary depending on the type of mixing considered, which is mathematically represented by a *mixing matrix*. According to the mixing conditions and the nature of this matrix, three mathematical formulations of the mixing process can be defined: the *instantaneous* (or *linear*), the *anechoic* (or *delayed*) and the *convolutive* (or *echoic*) *mixing models*.

Instantaneous Model

The simplest mixing model is the *instantaneous* or *linear* model. In this model, the mixtures are formed by linear combinations of the sources. Therefore, the mixtures are obtained by summing scaled versions of the sources:

$$x_m(t) = \sum_{n=1}^{N} a_{mn} s_n(t), \quad m = 1, \ldots, M, \tag{2.18}$$

where a_{mn} are scalar factors. Thus, the images of the sources are then given by

$$s_{mn}(t) = a_{mn} s_n(t). \tag{2.19}$$

Alternatively, the instantaneous model can be expressed as a system of linear equations in the form

$$\begin{aligned} x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) + \cdots + a_{1N}s_N(t) \\ x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t) + \cdots + a_{2N}s_N(t) \\ &\vdots \\ x_M(t) &= a_{M1}s_1(t) + a_{M2}s_2(t) + \cdots + a_{MN}s_N(t). \end{aligned} \tag{2.20}$$

Taking into account the above system, it is usual to find the mixing models in a compact matrix formulation:

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_M(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{MN} \end{bmatrix} \cdot \begin{bmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_N(t) \end{bmatrix}, \tag{2.21}$$

or equivalently

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{2.22}$$

where $\mathbf{x} = [x_1(t), \ldots, x_M(t)]^{\mathrm{T}}$ is a $M \times 1$ vector of mixtures, $\mathbf{A}$ is the $M \times N$ mixing matrix and $\mathbf{s} = [s_1(t), \ldots, s_N(t)]^{\mathrm{T}}$ is a $N \times 1$ vector of sources. If a collection of individual time samples of the mixture and source signals are considered, the model can be represented as:

$$\mathbf{X} = \mathbf{A}\mathbf{S}, \tag{2.23}$$

where $\mathbf{X}$ is the $M \times T$ matrix corresponding to the sensor data at times $t = 1, \ldots, T$:

$$\mathbf{X} = \begin{bmatrix} x_1(1) & x_1(2) & \cdots & x_1(T) \\ x_2(1) & x_2(2) & \cdots & x_2(T) \\ \vdots & \vdots & \ddots & \vdots \\ x_M(1) & x_M(2) & \cdots & x_M(T) \end{bmatrix}, \tag{2.24}$$

and $\mathbf{S}$ is the $N \times T$ matrix of source signals:

$$\mathbf{S} = \begin{bmatrix} s_1(1) & s_1(2) & \cdots & s_1(T) \\ s_2(1) & s_2(2) & \cdots & s_2(T) \\ \vdots & \vdots & \ddots & \vdots \\ s_N(1) & s_N(2) & \cdots & s_N(T) \end{bmatrix}. \tag{2.25}$$

Note that under this notation, each row of $\mathbf{X}$ and $\mathbf{S}$ corresponds to one of the mixture and source signals, respectively. The notation used in Eq.(2.22) is often referred as the model in *instantaneous notation*, as it represents the generation of the mixtures in a single time sample. On the other hand, the notation of Eq.(2.23) is referred as the model in *explicit notation* and it describes the generation of the mixtures in the whole observation time.

Anechoic Model

The *anechoic* or *delayed model* can be thought as an extension of the instantaneous model where, in addition to different gain factors, different transmission delays between the sources and the sensors are considered. This is equivalent to an anechoic mixing scenario, where only the direct path between each source and sensor has influence on the mixture. The generative model is

$$x_m(t) = \sum_{n=1}^{N} a_{mn} s_n(t - \delta_{mn}), \quad m = 1, \ldots, M, \tag{2.26}$$

where δ_{mn} is the arrival delay between source n and sensor m, and a_{mn} stands for the amplitude factor corresponding to the path between source n and sensor m.

The images of the sources are scaled and delayed versions of the sources:

$$s_{mn}(t) = a_{mn} s_n(t - \delta_{mn}). \tag{2.27}$$

The mixing matrix has the form

$$\mathbf{A} = \begin{bmatrix} a_{11}\delta(t-\delta_{11}) & a_{12}\delta(t-\delta_{12}) & \cdots & a_{1N}\delta(t-\delta_{1N}) \\ a_{21}\delta(t-\delta_{21}) & a_{22}\delta(t-\delta_{21}) & \cdots & a_{2N}\delta(t-\delta_{2N}) \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1}\delta(t-\delta_{M1}) & a_{M2}\delta(t-\delta_{M2}) & \cdots & a_{MN}\delta(t-\delta_{MN}) \end{bmatrix}, \quad (2.28)$$

where $\delta(t)$ are Kronecker[1] deltas. Note that the operator $\delta(t-\delta_{mn})$ is used to denote a delay between source n and sensor m. With this notation, the model can be compactly expressed by

$$\mathbf{x} = \mathbf{A} * \mathbf{s}, \quad (2.29)$$

where $*$ denotes the element-wise convolution operation.

Convolutive Model

In the *convolutive* or *echoic model*, reflections occurring in the mixing environment are considered too. Mathematically, the process can be written as

$$x_m(t) = \sum_{n=1}^{N} \sum_{\tau=1}^{L_{imp}} a_{mn\tau} s_n(t-\delta_{mn\tau}), \quad m = 1, \ldots, M, . \quad (2.30)$$

where L_{imp} is the number of paths the source signal can take to the sensors. Therefore, the images of the sources are filtered versions of the original sources:

$$s_{mn}(t) = \sum_{\tau=1}^{L_{imp}} a_{mn\tau} s_n(t-\delta_{mn\tau}). \quad (2.31)$$

The mixing matrix $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{bmatrix} \sum_{\tau=1}^{L_{imp}} a_{11\tau}\delta(t-\delta_{11\tau}) & \cdots & \sum_{\tau=1}^{L_{imp}} a_{1N\tau}\delta(t-\delta_{1N\tau}) \\ \vdots & \ddots & \vdots \\ \sum_{\tau=1}^{L_{imp}} a_{M1\tau}\delta(t-\delta_{M1\tau}) & \cdots & \sum_{\tau=1}^{L_{imp}} a_{MN\tau}\delta(t-\delta_{MN\tau}) \end{bmatrix}, \quad (2.32)$$

[1]The Kronecker delta is defined in signal processing as $\delta(t) = \begin{cases} 1 & \text{if } t = 0 \\ 0 & \text{if } t \neq 0 \end{cases}$. The alternate notation for Kronecker deltas found in other works, δ_{ij}, must not be here confused with the source-sensor delay δ_{mn}.

thus, a convolutive formulation of the form $\mathbf{x} = \mathbf{A} * \mathbf{s}$ is also used here. Note that the anechoic and instantaneous models can be thought of as particular cases of the convolutive model.

Noisy models

In real life, there is always some kind of noise present in the observations. Noise can come from measuring devices or from any inaccuracies in the model used. Therefore, a noise term is sometimes included in the above models:

$$\mathbf{x} = \mathbf{A} \star \mathbf{s} + \mathbf{n}, \tag{2.33}$$

where $\mathbf{n} = [n_1(t), n_2(t), \ldots, n_M(t)]^{\mathrm{T}}$ is the $M \times 1$ noise vector and $\star$ denotes the model dependent operator (matrix product in instantaneous mixtures and element-wise convolution in the anechoic and convolutive models). Noise is often assumed to be white, Gaussian and uncorrelated, i.e. having diagonal covariance matrix in the form $\sigma^2\mathbf{I}$, where σ^2 is the variance of one of its M components.

The separation methods presented in this thesis are based on noise-free models.

2.3.2 Source Separation Tasks and Approaches

The source separation problem consists in estimating the source spatial images of the sources $s_{mn}(t)$, from the mixture signals $x_m(t)$. Note that the estimation of the single-channel source signals $s_n(t)$ involves undoing the filtering effect of the mixing process (dereverberating), which is an additional problem that will not be considered in this thesis.

The instantaneous mixing model $\mathbf{X} = \mathbf{AS}$, has the form of a conventional system of linear equations. Although it seems that the problem of extracting the sources $\mathbf{S}$ from the mixtures $\mathbf{X}$ can be completely solved by traditional algebraic techniques, this is only possible if the mixing matrix $\mathbf{A}$ is known. However, source separation tries to give solution to this problem in the case were both $\mathbf{S}$ and $\mathbf{A}$ are unknown. Moreover, even if the mixing matrix $\mathbf{A}$ can be accurately estimated, the system is only invertible if $\mathbf{A}$ is square and has full rank, thus, the number of equations must be equal to the number of unknowns ($M = N$). When the problem is overdetermined ($M > N$), dimensionality reduction techniques such as *Principal Component Analysis* (PCA) [44] are usually employed. If the problem is

underdetermined ($M < N$), there is an infinite number of solutions and demixing the sources from the mixtures becomes a very challenging task.

In the next subsections several criteria that are commonly used to classify separation problems are introduced.

Problem Classification

The separation difficulty is mainly related to three different aspects: the relative number of mixture channels and sources, the length of the mixing filters and the time variation of the mixing filters [45]. These three criteria are used to characterize the mixtures in the following way:

- Relative number of mixture channels and sources:
  1. $M > N$ *Overdetermined mixture.*
  2. $M = N$ *Determined mixture.*
  3. $M < N$ *Underdetermined mixture.*
- Mixing filters:
  1. Scalars (zero delay): *Instantaneous mixture.*
  2. Scalars and/or Delays (possibly fractional): *Anechoic mixture.*
  3. Otherwise: *Convolutive mixture.*
- Time variation of the mixing filters:
  1. Static sources or fixed filters: *Time-invariant mixture.*
  2. Moving sources or time-varying filters: *Time-variant mixture.*

Overdetermined and determined situations usually appear in microphone array processing techniques, which are usually used for source localization and tracking [19].

2.3.3 Underdetermined Source Separation

The underdetermined (or *degenerate*) case in *Sound Source Separation* (SSS) is the most challenging one. The challenge resides in the fact that the mixing matrix is not invertible and the traditional method of demixing by estimating the inverse mixing matrix can not be applied in this case.

Unfortunately, most commercial music productions and audio material can be categorized as underdetermined mixtures and perfect separation of instruments or singers from a stereo track is not a solved problem so far.

Sparsity refers to the property by which most of the sample values of a signal are zero or close to zero. This property is the fundamental piece supporting most underdetermined separation algorithms.If the sources are sparse, the mixing directions of a linear instantaneous mixture can be easily observed in the scatter plot. Moreover, in the underdetermined case, higher sparsity is a requirement for good separability of the sources, even in the case when the mixing matrix is known. Thus, an increasingly popular and powerful assumption that has led to many practical algorithms is to assume that the sources have a sparse representation under a given basis. These approaches are motivated by the fact that very often the desired data in the time domain do not represent the required sparsity. Therefore, they have come to be known as *sparse methods*.

The advantage of a sparse signal representation is that the probability of two or more sources being simultaneously active is low. Thus, sparse representations are potentially good for achieving high-quality separation due to the fact that most of the energy in a basis coefficient belongs to a single source. The sparse representation of an audio signal has an interpretation in information theoretic terms: a signal represented by a small number of coefficients corresponds to transmission of information using a code with a small number of bits [46]. Sparse representation of information is a phenomenon that also occurs in the natural world. In the brain, neurons are said to encode data in a sparse way, if their firing pattern is characterized by long periods of inactivity [47].

Throughout this section the general framework for underdetermined source separation will be presented. Most approaches rely on signal transformations that enhance the sparse structure of the sources. Therefore, special attention is paid to the basics of signal decomposition, mainly to time-frequency representations. Sparse distributions and sparsity measures are also introduced, followed by a description of the most common approaches to source estimation.

Sparsity

The reason why high sparsity of source representations is desired in source separation problems is straightforward: the less coefficients are needed to

adequately describe a particular source signal, the less degree of overlapping will occur when mixed with other signals. Sparsity is crucial in most underdetermined situations, specially when very little a prior information is available and the ratio between number of sources and number of mixtures is high.

A sparse representation can be obtained by minimizing a cost function which is the weighted sum of the reconstruction error term $\|\mathbf{X} - \mathbf{ACB}^{\mathrm{T}}\|_F^2$ and the term which incurs a penalty on non-zero elements of $\mathbf{C}$. The variance σ^2 is used to balance between these two.

Measures of sparsity

The sparsity ξ of a signal is usually measured by means of the ℓ_p norm of its coefficient vector $\mathbf{c}$ with the constraint $0 \leq p \leq 1$:

$$\xi = \|\mathbf{c}\|_p = \left(\sum_{i=1}^{C} |c_i|^p \right)^{1/p}, \quad 0 \leq p \leq 1. \tag{2.34}$$

Depending on the value of p, several well-known sparsity measures appear:

- **The ℓ_0 norm**. This measure gives the number of non-zero coefficients in $\mathbf{c}$:

  $$\|\mathbf{c}\| = \#\{i, c_i \neq 0\}, \tag{2.35}$$

  where $\#\{\cdot\}$ denotes the counter operator. This norm is rarely used since it is highly sensible to noise: a slight addition of noise will make a representation completely nonsparse.

- **The ℓ_ϵ norm**. A thresholded version of the ℓ_0 norm in order to be more robust against noise:

  $$\|\mathbf{c}\|_\epsilon = \#\{i, |c_i| \geq \epsilon\}. \tag{2.36}$$

  However, determining a reasonable noise threshold ϵ for unknown signals is a difficult task [48].

- **The ℓ_1 norm**. This measure gives the summation of the modulus of the coefficients:

  $$\|\mathbf{c}\|_1 = \sum_{i=1}^{C} |c_i|. \tag{2.37}$$

The ℓ_1 norm is a popular choice since some algorithms can be implemented with linear programming techniques. The ℓ_2 norm $\|\cdot\|$, for which the order index is usually omitted, corresponds to the traditional Euclidean norm, and to the square root of the energy.

These and others measures of sparsity such as the *normalized kurtosis* were analyzed by Karvanen and Cichoki [48] showing that very different results can be obtained by using different sparsity measures if the distribution does not have a unique mode at zero.

Time-Frequency Masking

Although the sparsity achieved by signal transformations provide a way to deal with underdetermined SSS, the factor that ultimately determine the separation performance is the degree of overlapping that occurs during the mixing process. Sparsity is not the only thing to consider since sparsity alone is useless if there is high overlap among the sources in the mixture. Two sources that are closely positioned in the stereo panoramic will be very hard to separate even if they are sufficiently sparse. Moreover, the correlation properties of the sources also play a role in the degree of overlap of the mixture. The *disjointness* of a mixture can be defined as the degree of non-overlapping of the mixed signals.

Time-frequency masking is another powerful approach for the separation of underdetermined mixtures, especially for the separation of single-channel mixtures. Techniques based on time-frequency masking use a time-frequency representation of the signal, taking profit from the disjointness provided by sparse transformations. Their aim is to identify the dominating source in each time-frequency unit, obtaining a mask that indicates which are the active points of each source in the time-frequency domain.

Formally, the time-frequency source image $\hat{\mathbf{S}}_{mn}(k,r)$ is produced from the m-th mixture $\mathbf{X}_m(k,r)$ by

$$\hat{\mathbf{S}}_{mn}(k,r) = \mathbf{M}_n(k,r) \circ \mathbf{X}_m(k,r), \tag{2.38}$$

where $0 \leq M_n(k,r) \leq 1$, $\forall(k,r)$ and the $\circ$ operator denotes the Hadamard (element-wise) product. Note that this corresponds to filtering the mixture with a set of time-varying frequency responses. The solution to the separation problem consists in deriving the masks from the mixture.

The Ideal Binary Mask

Consider the sum of all signals that interfere with source n in the *Short-Time Fourier Transform* (STFT) domain:

$$\mathbf{U}_n(k,r) = \sum_{n'=1,n'\neq n}^{N} \mathbf{S}_{n'}(k,r). \tag{2.39}$$

The *ideal binary mask* (IBM) for a source $\mathbf{S}_n(k,r)$ is defined as the binary time-frequency mask that is 1 for time-frequency bins where its energy is higher than all the interfering sources:

$$\mathrm{IBM}_n(k,r) = \begin{cases} 1 & \text{if} \quad 20\log\left(\frac{|S_n(k,r)|}{|U_n(k,r)|}\right) \geq 0 \\ 0 & \text{elsewhere} \end{cases}, \quad \forall(k,r). \tag{2.40}$$

This mask has been shown to be optimal when applied to the mixture and this is why the IBM has been suggested as a major computation goal of sound source separation algorithms, since it has proven to be highly effective for robust *automatic speech recognition* and human speech intelligibility in noise [49]. An example of ideal binary mask for a 4 source mixture is shown in Figure 2.9.

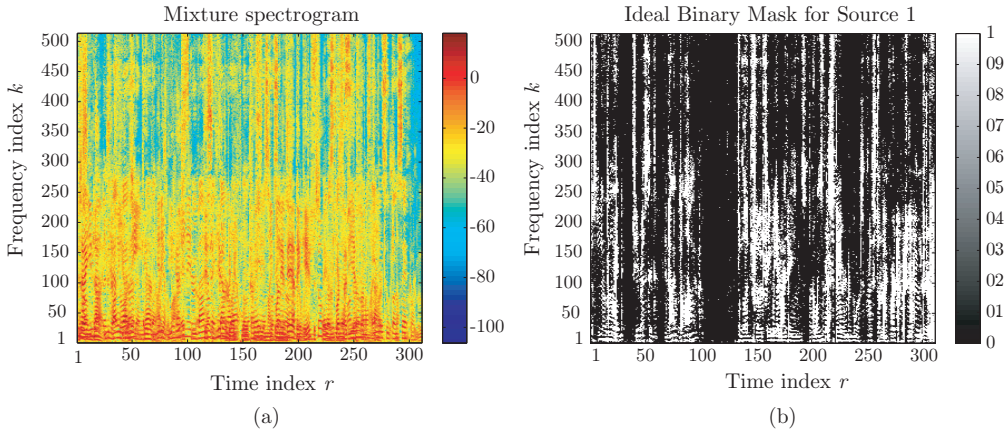


Figure 2.9. Example of ideal binary mask for an instantaneous mixture of 4 sources. (a) Magnitude STFT of one of the mixture channels. (b) Ideal binary mask for one of the sources.

W-Disjoint Orthogonality

Binary time-frequency masking is the special case in which $M_n(k,r)$ can only take the values 0 or 1. It is based on the assumption that every time-frequency point in the mixture with significant energy is dominated by the contribution of one source. This assumption is widely known as the *W-Disjoint Orthogonality* (WDO) assumption. Mathematically, the sources are said to be WDO if

$$S_n(k,r)S_{n'}(k,r) = 0, \quad \forall n \neq n', \forall (k,r), \tag{2.41}$$

where $S_n(k,r)$ and $S_{n'}(k,r)$ are the STFT of any two sources in the mixture. In matrix notation:

$$\mathbf{S}_n(k,r) \circ \mathbf{S}_{n'}(k,r) = \mathbf{0} \quad \forall n \neq n', \tag{2.42}$$

where $\mathbf{0}$ is the zero matrix.

Based on the IBM, a disjointness measure is given by the *average approximate* WDO. Burred provided an excellent analysis of the disjointness properties of speech and music mixtures considering both STFT and frequency-warped representations, showing the advantages of using a non-uniform time-frequency resolution [50; 51].

Yilmaz and Rickard [52] applied binary time-frequency masks to mixtures of several speech sources in the STFT domain considering a two-sensor arrangement. They observed that speech sources are sufficiently disjoint under time-frequency representations and showed that they are approximately WDO in mixtures of up to 10 signals [53]. Note that this aspect is very important, since source sparsity alone is useless if the sources overlap to a high degree.

2.3.4 Time-Frequency Masking Limitations

In practice, perfect separation is very difficult to be achieved and the estimated source spatial images may contain different distortions: musical noise, interference from other sources, timbre distortion and spatial distortion. Musical noise or burbling artifacts appear with time-frequency masking algorithms, being one of the most common distortions in source separation. Musical noise can be reduced by using small STFT hop sizes [54] and non-binary time-frequency masks. To this end, Araki et al. proposed a set

of smoothed masks in [55], showing the tradeoff between source distortion and interference. Nevertheless, the performance achieved by time-frequency masking is sufficient for most practical applications of SSS [56].

Paper Contributions and Discussion 3

In order to provide a compact presentation of the results obtained throughout the course of the degree candidature, this chapter summarizes the most relevant contributions associated to the publications compiled in this thesis from Chapter 4 to Chapter 9. The abstract and a brief discussion of the most relevant contributions are presented for each publication.

Note that the author of this thesis is the primary author of almost all the following publications, and none of them have been previously presented as a part of another thesis. Note also that the papers have been reformatted to this book style in order to avoid infringing copyright issues.

Publications:

- Sound Source Localization

    1. M. Cobos, A. Marti and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," in *IEEE Signal Processing Letters*, vol. 18, no. 1, pp. 7174, 2011.
    2. A. Marti, M. Cobos, J. J. Lopez and J. Escolano, "A low-complexity iterative method for high-accuracy acoustic source

localization," in *Journal of the Acoustical Society of America* (accepted, 2013).

3. A. Marti, M. Cobos and J. J. Lopez, "Real-time speaker localization and detection system for camera steering in multiparticipant videoconferencing environments," in *Proceedings of the 36th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011.

- Automatic Speech Recognition

4. A. Marti, M. Cobos and J. J. Lopez, "Automatic speech recognition in cocktail-party situations: a specific training for separated speech," in *Journal of the Acoustical Society of America*, vol. 131, no. 2, pp. 1529-1535, 2012.
5. A. Marti, M. Cobos and J. J. Lopez, "Evaluating the influence of source separation methods in robust automatic speech recognition with a specific cocktail-party training," in *Proceedings of the 132th AES Convention*, London, UK, 2012.

- Source Separation and Enhancement

6. A. Marti, M. Cobos and J. J. Lopez, "A real-time sound source localization and enhancement system using distributed microphones," in *Proceedings of the 130th AES Convention*, London, UK, 2011.

3.1 A Modified SRP-PHAT Functional for Robust Real-Time Sound Source Localization with Scalable Spatial Sampling

3.1.1 Abstract

SRP-PHAT algorithm has been shown to be one of the most robust sound source localization approaches working in noisy and reverberant environments. However, its practical implementation is usually based on a costly fine grid-search procedure, making the computational cost of the method a real issue. In this paper, we introduce an effective strategy that extends

the conventional SRP-PHAT functional with the aim of considering the volume surrounding the discrete locations of the spatial grid. As a result, the modified functional performs a full exploration of the sampled space rather than computing the SRP at discrete spatial positions, increasing its robustness and allowing for a coarser spatial grid. To this end, the GCC function corresponding to each microphone pair must be properly accumulated according to the defined microphone setup. Experiments carried out under different acoustic conditions confirm the validity of the proposed approach.

3.1.2 Contributions

This paper presented a robust approach to sound source localization based on a modified version of the well-known SRP-PHAT algorithm. The proposed functional is based on the accumulation of GCC values in a range that covers the volume surrounding each point (see Figure 4.3) of the defined spatial grid. The GCC accumulation limits are determined by the gradient of the inter-microphone time delay function corresponding to each microphone pair, thus, taking into account the spatial distribution of possible TDOAs resulting from a given array geometry.

Our results showed that the proposed approach provides similar performance to the conventional SRP-PHAT algorithm in difficult environments with a reduction of five orders of magnitude in the required number of functional evaluations (see Table 4.1). This reduction has been shown to be sufficient for the development of real-time source localization applications.

3.2 A Steered Response Power Iterative Method for High-Accuracy Acoustic Source Localization

3.2.1 Abstract

Source localization using the steered response power (SRP) usually requires a costly grid-search procedure. To address this issue, a modified SRP algorithm was recently introduced, providing improved robustness when using coarser spatial grids. In this letter, an iterative method based on the modified SRP is presented. A coarse spatial grid is initially evaluated with

the modified SRP, selecting the point with the highest accumulated value. Then, its corresponding volume is iteratively decomposed by using a finer spatial grid. Experiments have shown that this method provides almost the same accuracy as the fine-grid search with a substantial reduction of functional evaluations.

3.2.2 Contributions

In this letter, an iterative approach for high-accuracy sound source localization using the modified SRP functional discussed in Chapter 4, was presented. The method starts by performing source localization over a very coarse spatial grid. Then, the grid region having the highest accumulated value is subsequently divided into finer regions until achieving a desired spatial resolution. This iterative process is illustrated in Figure 5.2. A set of experiments have been carried out to evaluate this new approach, comparing its localization accuracy with other well-known approaches in different acoustic conditions, as can be seen in Figure 5.5. The results show that the proposed method has a performance comparable to that of a fine-grid SRP with a reduction of approximately five orders of magnitude in terms of functional evaluations.

The proposed method is aimed at allowing high-accuracy acoustic source localization over systems with limited computational resources.

3.3 A Real-Time Sound Source Localization and Enhancement System Using Distributed Microphones

3.3.1 Abstract

SRP-PHAT algorithm has been shown to be one of the most robust sound source localization approaches operating in noisy and reverberant environments. A recently proposed modified SRP-PHAT algorithm has been shown to provide robust localization performance in indoor environments without the need for having a very fine spatial grid, thus reducing the computational cost required in a practical implementation. Sound source localization methods are commonly employed in many sound processing applications. In our case, we use the modified SRP-PHAT functional for improving noisy speech signals. The estimated position of the speaker is

used to calculate the time-delay for each microphone and then the speech is enhanced by aligning correctly the microphone signals.

3.3.2 Contributions

This paper presented a microphone array system for speech enhancement based on the delay-and-sum beamformer and the modified SRP-PHAT functional developed by the authors. The estimated positions given by the SRP-PHAT algorithm are used to determine the distance from the talker to each microphone and apply the delay-and-sum beamformer. Since the localization step may have a given error due to the limited spatial resolution, an estimation of the time delay error is needed in order to correct the phase alignment of all the microphone signals. Figure 6.2 shows an example of the effect related to the correction of the time-delay error. The estimation of the error is better when the acoustic conditions of the environment are not too adverse, leading to numerically and perceptually better separation results, as proved by the results shown in Figure 6.4.

This paper shows an application of the method developed in Section 3.1. In this case, it is not only used for localizing sound sources, but also to enhance a signal of interest.

3.4 Real-Time Speaker Localization and Detection System for Camera Steering in Multiparticipant Videoconferencing Environments

3.4.1 Abstract

A real time speaker localization and detection system for videoconferencing environments is presented. In this system, a recently proposed modified SRP-PHAT algorithm has been used as the core processing scheme. The new SRP-PHAT functional has been shown to provide robust localization performance in indoor environments without the need for having a very fine spatial grid, thus reducing the computational cost required in a practical implementation. Moreover, it has been demonstrated that the statistical distribution of location estimates when a speaker is active can be successfully used to discriminate between speech and non-speech frames by using

a criterion of peakedness. As a result, talking participants can be detected and located with significant accuracy following a common processing framework.

3.4.2 Contributions

This paper presented a microphone array system for camera-steering to be used in a multiparticipant videoconferencing environment based on the well-known SRP-PHAT algorithm. The distribution of location estimates obtained with a modified SRP-PHAT functional was analyzed, showing that location estimates follow different distributions when speakers are active and allowing to discriminate between speech and non-speech frames under a common localization framework. Figure 7.1 shows the different forms that the histogram of location estimates takes when there is an active speaker or not. The results of experiments conducted in a real room suggest that, using a moderately high number of accumulated location estimates, it is possible to discriminate with significant accuracy between speech and non-speech frames, which is sufficient to correctly detect an active speaker and make the camera point at his/her pre-defined location. The real room where the experiment were carried out is illustrated in Figure 7.2.

This paper shows the great applicability of the method developed in Chapter 4 when combined with video cameras for advanced teleconferencing systems.

3.5 Automatic Speech Recognition in Cocktail-Party Situations: A Specific Training for Separated Speech

3.5.1 Abstract

ASR refers to the task of extracting a transcription of the linguistic content of an acoustical speech signal automatically. Despite several decades of research in this important area of acoustic signal processing, the accuracy of ASR systems is still far behind human performance, especially in adverse acoustic scenarios. In this context, one of the most challenging situations is the one concerning simultaneous speech in cocktail-party environments. Although source separation methods have already been investigated to deal with this problem, the separation process is not perfect and the resulting

artifacts pose an additional problem to ASR performance. In this paper, a specific training to improve the percentage of recognized words in real simultaneous speech cases is proposed. The combination of source separation and this specific training is explored and evaluated under different acoustical conditions, leading to improvements of up to a 35% in ASR performance.

3.5.2 Contributions

Cocktail-party situations where different speakers are talking at the same time pose a real problem for ASR systems. In this paper, a framework for robust ASR in cocktail-party situations was presented. This framework is based on a robust transformed model constructed from separated speech in diverse acoustic environments. Thus, a source separation method is used as a speech enhancement stage that suppresses interferences.

The validity of the method was studied over a meaningful set of experiments, evaluating ASR performance for three different training data-sets: Anechoic training, Reverberant training and Cocktail-party training.

The results showed that both, source separation and specific training, provide a considerable improvement in word recognition rate (WRR) (up to a 35% as shown in Figure 8.5), reducing the existing mismatch between the training and test data.

Moreover, the proposed framework allows to perform both ASR and source separation in real-time, which is a very important feature for practical systems. Future work will be focused on developing efficient double-talk detection methods for real-time ASR model selection.

In this paper, we have combined a working speech recognition system with state-of-the-art signal processing techniques for source separation.

3.6 Evaluating the Influence of Source Separation Methods in Robust Automatic Speech Recognition with a Specific Cocktail-party Training

3.6.1 Abstract

ASR allows a computer to identify the words that a person speaks into a microphone and convert it to written text. One of the most challenging situations for ASR is the cocktail-party environment. Although source separation methods have already been investigated to deal with this problem, the separation process is not perfect and the resulting artifacts pose an additional problem to ASR performance in case of using separation methods based on time-frequency masks. Recently, the authors proposed a specific training method to deal with simultaneous speech situations in practical ASR systems. In this paper, we study how the speech recognition performance is affected by selecting different combinations of separation algorithms both at the training and test stages of the ASR system under different acoustic conditions. The results show that, while different separation methods produce different types of artifacts, the overall performance of the method is always increased when using any cocktail-party training.

3.6.2 Contributions

In the paper corresponding to Chapter 8, the authors proposed a framework for robust ASR in cocktail-party situations. This framework is based on a robust transformed model constructed from separated speech in diverse acoustic environments. In this paper two source separation methods were used as a speech enhancement stage that suppresses interferences, comparing the results obtained by different types of source separation methods (MuLeTS and Full-Rank Spatial Covariance Model). The results showed that both specific trainings provide a considerable improvement in WRR even when the source separation method employed with the test data and training data sets are different.

Since the artifacts introduced by both separation methods are different, the WRR is always higher when the mismatch between training and test data is lower. As expected, the results suggest that the training should be specifically designed for a given source separation method. Nevertheless,

it should be emphasized that this work has only considered two separation methods. Thus, further work is needed to understand better the limitations arising from the mismatch between the separation method used in the training and test ASR stages.

A Modified SRP-PHAT Functional for Robust Real-Time Sound Source Localization with Scalable Spatial Sampling

4

M. Cobos, A. Marti and J. J. Lopez
IEEE Signal Processing Letters, vol. 18, no. 1, pp.71-74, 2011.

Abstract

The Steered Response Power - Phase Transform (SRP-PHAT) algorithm has been shown to be one of the most robust sound source localization approaches operating in noisy and reverberant environments. However, its practical implementation is usually based on a costly fine grid-search procedure, making the computational cost of the method a real issue. In this paper, we introduce an effective strategy that extends the conventional SRP-PHAT functional with the aim of considering the volume surrounding the discrete locations of the spatial grid. As a result, the modified functional performs a full exploration of the sampled space rather than computing the SRP at discrete spatial positions, increasing its robustness and allowing for a coarser spatial grid. To this end, the Generalized Cross-Correlation (GCC) function corresponding to each microphone pair must be properly accumulated according to the defined microphone set-up. Experiments carried out under different acoustic conditions confirm the validity of the

proposed approach.

4.1 Introduction

Sound source localization under high noise and reverberation still remains a very challenging task. To this end, microphone arrays are commonly employed in many sound processing applications such as videoconferencing, hands-free speech acquisition, digital hearing aids, video-gaming, autonomous robots and remote surveillance. Algorithms for sound source localization can be broadly divided into indirect and direct approaches [16]. Indirect approaches usually follow a two-step procedure: they first estimate the *Time Difference Of Arrival* (TDOA) [17] between microphone pairs and, afterwards, they estimate the source position based on the geometry of the array and the estimated delays. On the other hand, direct approaches perform TDOA estimation and source localization in one single step by scanning a set of candidate source locations and selecting the most likely position as an estimate of the source location. In addition, information theoretic approaches have also shown to be significantly powerful in source localization tasks [57].

The *Steered Response Power - Phase Transform* (SRP-PHAT) algorithm is a direct approach that has been shown to be very robust under difficult acoustic conditions [19; 20; 13]. The algorithm is commonly interpreted as a beamforming-based approach that searches for the candidate source position that maximizes the output of a steered delay-and-sum beamformer. However, despite its robustness, computational cost is a real issue because the SRP space to be searched has many local extrema [21]. Very interesting modifications and optimizations have already been proposed to deal with this problem, such as those based on Stochastic Region Contraction (SRC) [22] and coarse-to-fine region contraction [23], achieving a reduction in computational cost of more than three orders of magnitude.

In this paper, we propose a different strategy where, instead of evaluating the SRP functional at discrete positions of a spatial grid, it is accumulated over the *Generalized Cross Correlation* (GCC) lag space corresponding to the volume surrounding each point of the grid. The GCC accumulation limits are determined by the gradient of the inter-microphone time delay function corresponding to each microphone pair, thus, taking

into account the spatial distribution of possible TDOAs resulting from a given array geometry. The benefits of following this approach are twofold. On the one hand, it incorporates additional spatial knowledge at each point for making a better final decision. On the other hand, the proposed modification achieves the same performance as SRP-PHAT with fewer functional evaluations, relaxing the computational demand required for a practical application.

4.2 The SRP-PHAT Algorithm

Consider the output from microphone l, $m_l(t)$, in an M microphone system. Then, the SRP at the spatial point $\mathbf{x} = [x, y, z]$ for a time frame n of length T is defined as

$$P_n(\mathbf{x}) \equiv \int_{nT}^{(n+1)T} \left| \sum_{l=1}^{M} w_l m_l \left(t - \tau(\mathbf{x}, l)\right) \right|^2 dt, \tag{4.1}$$

where w_l is a weight and $\tau(\mathbf{x}, l)$ is the direct time of travel from location $\mathbf{x}$ to microphone l. DiBiase [21] showed that the SRP can be computed by summing the GCCs for all possible pairs of the set of microphones. The GCC for a microphone pair (k, l) is computed as

$$R_{m_k m_l}(\tau) = \int_{-\infty}^{\infty} \Phi_{kl}(\omega) M_k(\omega) M_l^*(\omega) e^{j\omega\tau} d\omega, \tag{4.2}$$

where τ is the time lag, * denotes complex conjugation, $M_l(\omega)$ is the Fourier transform of the microphone signal $m_l(t)$, and $\Phi_{kl}(\omega)$ is a combined weighting function in the frequency domain. The phase transform (PHAT) [36] has been demonstrated to be a very effective GCC weighting for time delay estimation in reverberant environments:

$$\Phi_{kl}(\omega) \equiv \frac{1}{|M_k(\omega) M_l^*(\omega)|}. \tag{4.3}$$

Taking into account the symmetries involved in the computation of Eq.(7.1) and removing some fixed energy terms [21], the part of $P_n(\mathbf{x})$ that changes with $\mathbf{x}$ is isolated as

$$P'_n(\mathbf{x}) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} R_{m_k m_l}\left(\tau_{kl}(\mathbf{x})\right), \tag{4.4}$$

where $\tau_{kl}(\mathbf{x})$ is the *inter-microphone time-delay function* (IMTDF). This function is very important, since it represents the theoretical direct path delay for the microphone pair (k, l) resulting from a point source located at $\mathbf{x}$. The IMTDF is mathematically expressed as

$$\tau_{kl}(\mathbf{x}) = \frac{\|\mathbf{x} - \mathbf{x}_k\| - \|\mathbf{x} - \mathbf{x}_l\|}{c}, \tag{4.5}$$

where c is the speed of sound, and $\mathbf{x}_k$ and $\mathbf{x}_l$ are the microphone locations.

The SRP-PHAT algorithm consists in evaluating the functional $P'_n(\mathbf{x})$ on a fine grid G with the aim of finding the point-source location $\mathbf{x}_s$ that provides the maximum value:

$$\mathbf{x}_s = \arg\max_{\mathbf{x}\in G} P'_n(\mathbf{x}). \tag{4.6}$$

4.3 The Inter-Microphone Time Delay Function

As commented in the previous section, the IMTDF plays a very important role in the source localization task. This function can be interpreted as the spatial distribution of possible TDOAs resulting from a given microphone pair geometry.

The function $\tau_{kl}(\mathbf{x})$ is continuous in $\mathbf{x}$ and changes rapidly at points close to the line connecting both microphone locations. Therefore, a pair of microphones used as a time-delay sensor is maximally sensible to changes produced over this line [58]. An example function is depicted in Figure 4.1(a) for the plane $z = 0$, with $\mathbf{x}_k = [-2, 0, 0]$ and $\mathbf{x}_l = [2, 0, 0]$. The gradient of the function is shown in Figure 4.1(b).

It is useful here to remark that the equation $|\tau_{kl}(\mathbf{x})| = C$, with C being a positive real constant, defines a hyperboloid in space with foci on the microphone locations $\mathbf{x}_k$ and $\mathbf{x}_l$. Moreover, the set of continuous confocal half-hyperboloids $\tau_{kl}(\mathbf{x}) = C$ with $C \in [-C_{\max}, C_{\max}]$, being $C_{\max} = (1/c)\|\mathbf{x}_k - \mathbf{x}_l\|$, spans the whole three-dimensional space.

Theorem: Given a volume V in space, the IMTDF for points inside V, $\tau_{kl}(\mathbf{x} \in V)$, takes only values in the continuous range $[\min(\tau_{kl}(\mathbf{x} \in \partial V)), \max(\tau_{kl}(\mathbf{x} \in \partial V))]$, where ∂V is the boundary surface that encloses V.

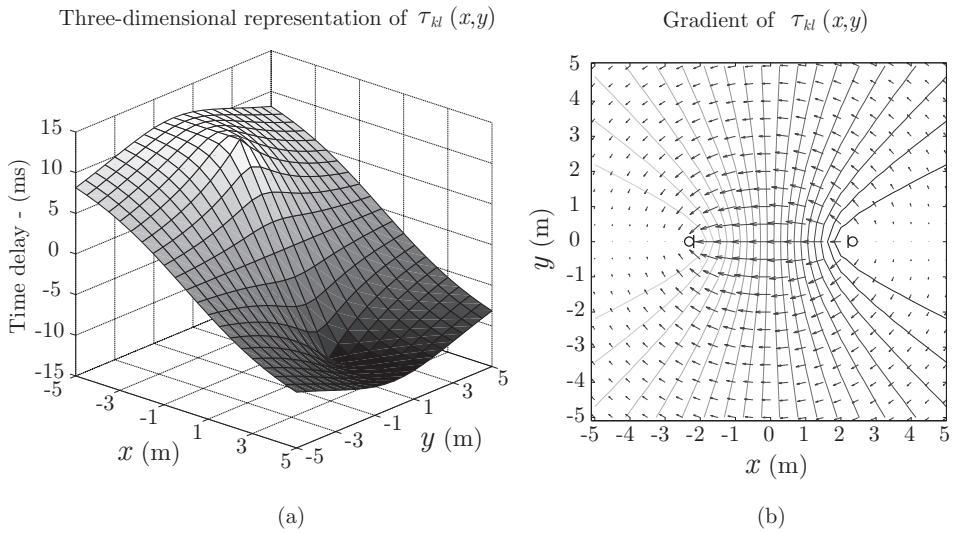


Figure 4.1. Example of IMTDF. (a) Representation for the plane $z = 0$ with microphones located at $[-2, 0, 0]$ and $[2, 0, 0]$. (b) Gradient.

Proof: Let us assume that a point inside V, $\mathbf{x}_0 \in V$, takes the maximum value in the volume, i.e. $\tau_{kl}(\mathbf{x}_0) = \max(\tau_{kl}(\mathbf{x} \in V)) = C_{\max_V}$. Since there is a half-hyperboloid that goes through each point of the space, all the points besides $\mathbf{x}_0$ satisfying $\tau_{kl}(\mathbf{x}) = C_{\max_V}$ will also take the maximum value. Therefore, all the points on the surface resulting from the intersection of the volume and the half-hyperboloid will take this maximum value, including those pertaining to the boundary surface ∂V. The existence of the minimum in ∂V is similarly deduced.

The above property is very useful to understand the advantages of the approach presented in this paper. Note that the SRP-PHAT algorithm is based on accumulating the values of the different GCCs at those time lags coinciding with the theoretical inter-microphone time delays, which are only computed at discrete points of a spatial grid. However, as described before, it is possible to analyze a complete spatial volume by scanning the time-delays contained in a range defined by the maximum and minimum values on its boundary surface. In the next section, we describe how this knowledge can be included in the localization algorithm to increase its robustness.

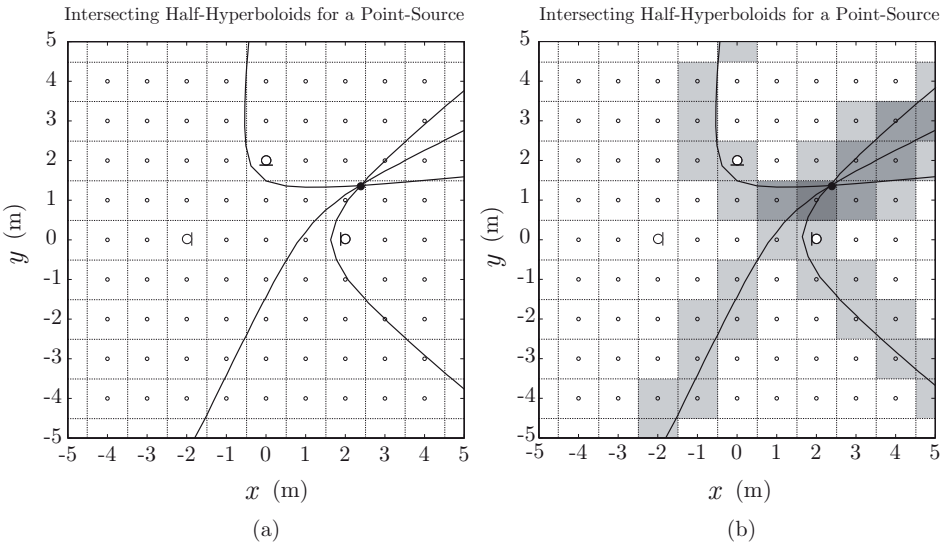


Figure 4.2. Intersecting half-hyperboloids and localization approaches. (a) Conventional SRP-PHAT. (b) Proposed.

4.4 Proposed Approach

Let us begin the description of the proposed approach by analyzing a simple case where we want to estimate the location $\mathbf{x}_s$ of a sound source inside an anechoic space. In this simple case, the GCCs corresponding to each microphone pair are delta functions centered at the corresponding inter-microphone time-delays: $R_{m_k m_l}(\tau) = \delta(\tau - \tau_{kl}(\mathbf{x}_s))$. For example and without loss of generality, let us assume a set-up with $M = 3$ microphones, as depicted in Figure 4.2(a). Then, the source position would be that of the intersection of the three half-hyperboloids $\tau_{kl}(\mathbf{x}) = \tau_{kl}(\mathbf{x}_s)$, with $(k, l) \in \{(1, 2), (1, 3), (2, 3)\}$. Consider now that, to localize the source, a spatial grid with resolution $r = 1$ m is used as shown in Figure 4.2(a). Unfortunately, the intersection does not match any of the sampled positions, leading to an error in the localization task. Obviously, this problem would have been easier to solve with a two step localization approach, but the above example shows the limitations imposed by the selected spatial sampling in SRP-PHAT, even in optimal acoustic conditions. This is not the case of the approach followed to localize the source in Figure 4.2(b) where, using the same spatial grid, the GCCs have been integrated for each

sampled position in a range that covers their volume of influence. A darker gray color indicates a greater accumulated value and, therefore, the darkest area is being correctly identified as the one containing the true sound source location. This new modified functional is expressed as follows

$$P_n''(\mathbf{x}) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} \sum_{\tau=L_{kl1}(\mathbf{x})}^{L_{kl2}(\mathbf{x})} R_{m_k m_l}(\tau). \tag{4.7}$$

The problem is to determine correctly the limits $L_{kl1}(\mathbf{x})$ and $L_{kl2}(\mathbf{x})$, which depend on the specific IMTDF resulting from each microphone pair. The computation of these limits is explained in the next subsection.

4.4.1 Computation of Accumulation Limits

As explained in Section 4.3, the IMTDF inside a volume can only take values in the range defined by its boundary surface. Therefore, for each point of the grid, the problem of finding the GCC accumulation limits of its volume of influence can be simplified to finding the maximum and minimum values on the boundary. To this end, it becomes useful to study

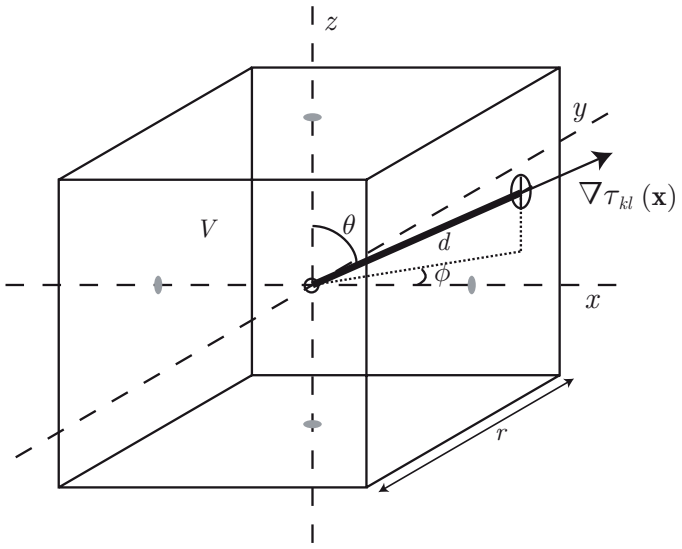


Figure 4.3. Volume of influence of a point in a rectangular grid.

the direction of the greatest rate of increase at each grid point, which is given by the gradient

$$\nabla\tau_{kl}(\mathbf{x}) = [\nabla_x\tau_{kl}(\mathbf{x}), \nabla_y\tau_{kl}(\mathbf{x}), \nabla_z\tau_{kl}(\mathbf{x})]\,, \tag{4.8}$$

where each component of the gradient vector can be calculated with

$$\nabla_\gamma\tau_{kl}(\mathbf{x}) = \frac{\partial\tau_{kl}(\mathbf{x})}{\partial\gamma} = \frac{1}{c}\left(\frac{\gamma-\gamma_k}{\|\mathbf{x}-\mathbf{x}_k\|} - \frac{\gamma-\gamma_l}{\|\mathbf{x}-\mathbf{x}_l\|}\right), \tag{4.9}$$

where γ denotes either x,y or z. The accumulation limits for a symmetric volume surrounding a point of the grid can be calculated by taking the product of the magnitude of the gradient and the distance d that exists from the point to the boundary following the gradient's direction:

$$L_{kl1}(\mathbf{x}) = \tau_{kl}(\mathbf{x}) - \|\nabla\tau_{kl}(\mathbf{x})\| \cdot d, \tag{4.10}$$

$$L_{kl2}(\mathbf{x}) = \tau_{kl}(\mathbf{x}) + \|\nabla\tau_{kl}(\mathbf{x})\| \cdot d, \tag{4.11}$$

Figure 4.3 depicts the geometry for a rectangular grid with spatial resolution r. For this cubic geometry, the distance d can be expressed as

$$d = \frac{r}{2}\min\left(\frac{1}{|\sin\theta\cos\phi|}, \frac{1}{|\sin\theta\sin\phi|}, \frac{1}{|\cos\theta|}\right), \tag{4.12}$$

where

$$\theta = \cos^{-1}\left(\frac{\nabla_z\tau_{kl}(\mathbf{x})}{\|\nabla\tau_{kl}(\mathbf{x})\|}\right), \tag{4.13}$$

$$\phi = \mathrm{atan}_2(\nabla_y\tau_{kl}(\mathbf{x}), \nabla_x\tau_{kl}(\mathbf{x})), \tag{4.14}$$

being $\mathrm{atan}_2(y,x)$ the quadrant-resolving arctangent function.

4.4.2 Computational Cost

Let L be the DFT length of a frame and $Q = M(M-1)/2$ the number of microphone pairs. The computational cost of SRP-PHAT is given by [20]:

$$\begin{aligned}\text{SRP-PHAT}_{cost} \approx\ & [6.125Q^2 + 3.75Q]L\log_2 L \\ & +15LQ(1.5Q-1) + (45Q^2 - 30Q)\nu',\end{aligned} \tag{4.15}$$

where ν' is the average number of functional evaluations required to find the maximum of the SRP space. Since the cost added by the modified functional is negligible and the frequency-domain processing of our approach

remains the same as the conventional SRP-PHAT algorithm, the above formula is valid for both approaches. Moreover, since the accumulation limits can be pre-computed before running the localization algorithm, the associated processing does not involve additional computation effort. However, as it will be shown in the next subsection, the advantage of the proposed method relies on the reduced number of required functional evaluations ν' for detecting the true source location, which results in an improved computational efficiency.

4.5 Experiments

Different experiments with real and synthetic recordings were conducted to compare the performances of the conventional SRP-PHAT algorithm, the SRC algorithm and our proposed method. First, the *Roomsim* Matlab package [59] was used to simulate an array of 6 microphones placed on the walls of a shoe-box-shaped room with dimensions 4 m $\times$ 6 m $\times$ 2 m (Fig. 4.4(a)). The simulations were repeated with two different reverberation times (T_{60} = 0.2 s and T_{60} = 0.7 s), considering 30 random source locations and different Signal-to-Noise Ratio (SNR) conditions. The resultant recordings were processed with 3 different spatial grid resolutions in the case of SRP-PHAT and the proposed method ($r_1 = 0.01$ m, $r_2 = 0.1$ m and $r_3 = 0.5$ m). Note that the number of functional evaluations ν' depends on the selected value of r, having $\nu'_1 = 480 \times 10^5$, $\nu'_2 = 480 \times 10^2$ and $\nu'_3 = 384$. The implementation of SRC was the one made available by Brown University's LEMS at *http://www.lems.brown.edu/array/download.html*, using 3000 initial random points. The processing was carried out using a sampling rate of 44.1 kHz, with time windows of 4096 samples of length and 50% overlap. The simulated sources were male and female speech signals of length 5 s with no pauses. The averaged results in terms of *Root Mean Squared Error* (RMSE) are shown in Figure 4.4(b-d). Since SRC does not depend on the grid size, the SRC curves are the same in all these graphs. As expected, all the tested systems perform considerably better in the case of low reverberation and high SNR. For the finest grid, it can be clearly observed that the performance of SRP-PHAT and the proposed method is almost the same. However, for coarser grids, our proposed method is only slightly degraded, while the performance of SRP-PHAT becomes substantially worse, specially for low SNRs and high reverberation. SRC has

similar performance to SRP-PHAT with $r = 0.01$ m. Therefore, our proposed approach performs robustly with higher grid sizes, which results in a great computational saving in terms of functional evaluations, as depicted in Figure 4.4(e).

Table 4.1. RMSE for the real-data experiment.

r	0.01	0.1	0.5
ν'	$802 \cdot 10^5$	$802 \cdot 10^2$	641
SRP-PHAT	RMSE = 0.29	RMSE = 0.74	RMSE = 1.82
Proposed	RMSE = 0.21	RMSE = 0.29	RMSE = 0.31
SRC	RMSE = 0.34 ($\nu' = 58307$)		

On the other hand, a real set-up quite similar to the simulated one was considered to study the performance of the method in a real scenario. Six omnidirectional microphones were placed at the 4 corners and at the middle of the longest walls of a videoconferencing room with dimensions 5.7 m × 6.7 m × 2.1 m and 12 seats. The measured reverberation time was $T_{60} = 0.28$ s. The processing was the same as with the synthetic recordings, using continuous speech fragments obtained from the 12 seat locations. The results are shown in Table 4.1 and confirm that our proposed method performs robustly using a very coarse grid. Although similar accuracy to SRC is obtained, the number of functional evaluations is significantly reduced.

4.6 Conclusion

This paper presented a robust approach to sound source localization based on a modified version of the well-known SRP-PHAT algorithm. The proposed functional is based on the accumulation of GCC values in a range that covers the volume surrounding each point of the defined spatial grid. The GCC accumulation limits are determined by the gradient of the inter-microphone time delay function corresponding to each microphone pair, thus, taking into account the spatial distribution of possible TDOAs resulting from a given array geometry. Our results showed that the proposed approach provides similar performance to the conventional SRP-PHAT algorithm in difficult environments with a reduction of five orders of magni-

tude in the required number of functional evaluations, with further computational saving than SRC. This reduction has been shown to be sufficient for the development of real-time source localization applications.

4.7 References

The references of this paper have been consolidated in the general bibliography at the end of the book.

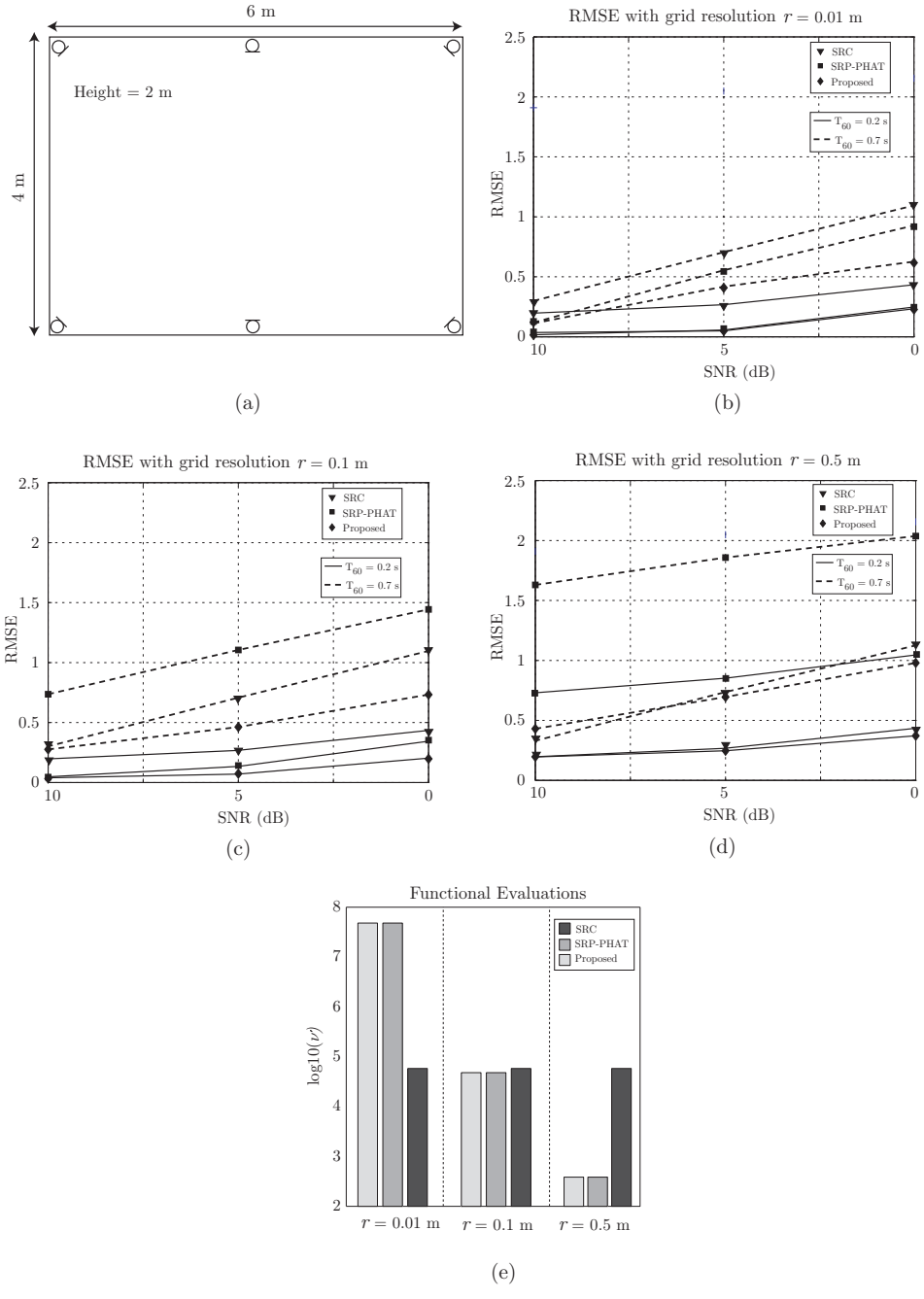


Figure 4.4. Results with simulations. (a) set-up. (b) $r = 0.01$ m. (c) $r = 0.1$ m. (d) $r = 0.5$ m. (e) Functional evaluations.

A Steered Response Power Iterative Method for High-Accuracy Acoustic Source Localization

5

A. Marti, M. Cobos, J. J. Lopez and J. Escolano
Journal of the Acoustical Society of America (accepted, 2013)

Abstract

Source localization using the steered response power (SRP) usually requires a costly grid-search procedure. To address this issue, a modified SRP algorithm was recently introduced, providing improved robustness when using coarser spatial grids. In this letter, an iterative method based on the modified SRP is presented. A coarse spatial grid is initially evaluated with the modified SRP, selecting the point with the highest accumulated value. Then, its corresponding volume is iteratively decomposed by using a finer spatial grid. Experiments have shown that this method provides almost the same accuracy as the fine-grid search with a substantial reduction of functional evaluations.

5.1 Introduction

The localization of sound sources has received a lot of attention in the last decades. Microphone arrays are known for their multiple applications like audio surveillance, teleconferencing, speech enhancement for hearing-aids or camera pointing systems[60]. Most of these applications require location estimators having both high-accuracy and a reasonable computational cost, especially when real-time performance is a real issue[61]. The steered-response power with phase transform (SRP) algorithm has been considered one of the most robust localization algorithms in reverberant environments [21]. However, besides its advantages, the computational cost is considerably high. In this context, several modifications and optimizations have been proposed to improve its performance and applicability. Recently, the authors proposed a modified version of the algorithm to improve the localization performance by accommodating SRP functional evaluations to scalable grid sizes [62]. However, although the computational cost is significantly reduced, the final accuracy is lastly determined by the chosen spatial resolution. In this letter, we propose an extended strategy based on an iterative grid decomposition procedure to improve the modified SRP algorithm. The method is evaluated under different acoustic conditions and compared to other SRP-based algorithms.

5.2 Modified SRP Algorithm

Consider the output from microphone l, $m_l(t)$, in an M-microphone system. The SRP algorithm is based on the calculation of the generalized cross-correlation (GCC) [36] (with phase transform) between microphone pairs (k, l), given by

$$R_{m_k m_l}(\tau) = \int_{-\infty}^{\infty} \frac{M_k(\omega) M_l^*(\omega)}{|M_k(\omega) M_l^*(\omega)|} e^{-j\omega\tau} d\omega, \tag{5.1}$$

where τ is the time lag, * denotes complex conjugation and $M_l(\omega)$ is the Fourier transform of the microphone signal $m_l(t)$. The SRP at spatial point $\mathbf{x} = [x, y, z]^T$ for a time frame n of length T can be expressed as

$$P_n(\mathbf{x}) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} R_{m_k m_l}\left(\tau_{kl}(\mathbf{x})\right), \tag{5.2}$$

where $\tau_{kl}(\mathbf{x})$ is the inter-microphone time-delay function. This function represents the theoretical time delay of arrival for the microphone pair (k, l) resulting from a point source located at $\mathbf{x}$. It is given by

$$\tau_{kl}(\mathbf{x}) = \frac{\|\mathbf{x} - \mathbf{m}_k\| - \|\mathbf{x} - \mathbf{m}_l\|}{c}, \tag{5.3}$$

where c is the speed of sound (340 m/s was used in this work), and $\mathbf{m}_k$ and $\mathbf{m}_l$ are the corresponding microphone locations. To implement the algorithm, the space is usually discretized by taking a spatial grid $\mathcal{G}$ with spatial resolution r, so that Eq.(5.2) takes only into account the GCC value on a discrete space location $\mathbf{x} \in \mathcal{G}$. Note that there are a total of $Q = M(M-1)/2$ microphone pairs that should be processed. The source location at time frame n is assumed to be that maximizing $P_n(\mathbf{x})$. It becomes apparent that, when using a coarse spatial grid, it is more likely to miss the global maximum of the SRP space. To address this issue, the modified SRP is based on accumulating the GCC lag space corresponding to the volume surrounding each point of the spatial grid, resulting in:

$$P'_n(\mathbf{x}) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} \sum_{\tau=L_{kl}^{(1)}(\mathbf{x})}^{L_{kl}^{(2)}(\mathbf{x})} R_{m_k m_l}(\tau). \tag{5.4}$$

The GCC accumulation limits $L_{kl}^{(1)}(\mathbf{x})$ and $L_{kl}^{(2)}(\mathbf{x})$ are determined by the gradient of the inter-microphone time-delay function corresponding to each microphone pair, thus, taking into account the spatial distribution of possible time-differences resulting from a given array geometry. The gradient components $\nabla\tau_{kl}(\mathbf{x}) = [\nabla_x\tau_{kl}(\mathbf{x}), \nabla_y\tau_{kl}(\mathbf{x}), \nabla_z\tau_{kl}(\mathbf{x})]^T$ are given by [62]:

$$\nabla_{x_i}\tau_{kl}(\mathbf{x}) = \frac{1}{c}\left(\frac{x_i - (x_i)_k}{\|\mathbf{x} - \mathbf{m}_k\|} - \frac{x_i - (x_i)_l}{\|\mathbf{x} - \mathbf{m}_l\|}\right), \quad x_i \in \{x, y, z\}. \tag{5.5}$$

The accumulation limits as a function of the gradient components are:

$$L_{kl}^{(1)}(\mathbf{x}) = \tau_{kl}(\mathbf{x}) - \|\nabla\tau_{kl}(\mathbf{x})\| \cdot d, \tag{5.6}$$

$$L_{kl}^{(2)}(\mathbf{x}) = \tau_{kl}(\mathbf{x}) + \|\nabla\tau_{kl}(\mathbf{x})\| \cdot d, \tag{5.7}$$

where, for a cubic spatial grid,

$$d = \frac{r}{2}\min\left(\frac{1}{|\sin\theta\cos\phi|}, \frac{1}{|\sin\theta\sin\phi|}, \frac{1}{|\cos\theta|}\right), \tag{5.8}$$

being $\theta = \cos^{-1}\left(\frac{\nabla_z \tau_{kl}(\mathbf{x})}{\|\nabla \tau_{kl}(\mathbf{x})\|}\right)$ the gradient elevation angle and $\phi = \text{atan}_2(\nabla_y \tau_{kl}(\mathbf{x}), \nabla_x \tau_{kl}(\mathbf{x}))$ the azimuth angle. Finally, the estimated source location $\mathbf{x}_s$ is that maximizing the modified functional over the defined spatial grid, i.e:

$$\mathbf{x}_s = \arg\max_{\mathbf{x}\in\mathcal{G}} \left\{P'_n(\mathbf{x})\right\}. \tag{5.9}$$

5.3 Proposed Approach

5.3.1 Mean-Based Functional

The approach proposed in this paper uses the modified SRP functional with an additional variation. Instead of summing up all the GCC values between the computed limits, we calculate the mean over this interval as follows:

$$\bar{P}'_n(\mathbf{x}) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} \sum_{\tau=L_{kl}^{(1)}(\mathbf{x})}^{L_{kl}^{(2)}(\mathbf{x})} \frac{R_{m_k m_l}(\tau)}{L_{kl}^{(2)}(\mathbf{x}) - L_{kl}^{(1)}(\mathbf{x})}. \tag{5.10}$$

The mean-based functional in Eq.(5.10) is designed to compensate the accumulated GCC, penalizing those values resulting from very large intervals while avoiding large functional values due to accumulated noise. To illustrate this idea, consider the example in Fig.5.1(a), which shows the delay function gradient for a two-microphone set-up over a coarse spatial grid. The lines represent constant-delay half-hyperbolas. Note that the shaded regions **A** and **B** span a different number of constant-delay lines, thus, resulting in different accumulation intervals. Figure 5.1(b) shows a noisy GCC obtained from the same setup and the corresponding summation ranges for **A** and **B** with a sampling frequency of 44.1 kHz. Note that, while interval **A** contains the GCC direct-sound peak, the accumulated value in **B** might be greater due to noise accumulation over a larger interval. The proposed mean-based functional mitigates this undesired effect. It must be emphasized that, unlike narrow-band methods, the effects of spatial aliasing in SRP-based broadband microphone arrays are not so relevant[63] In our case, while no special treatment is applied to avoid spa-

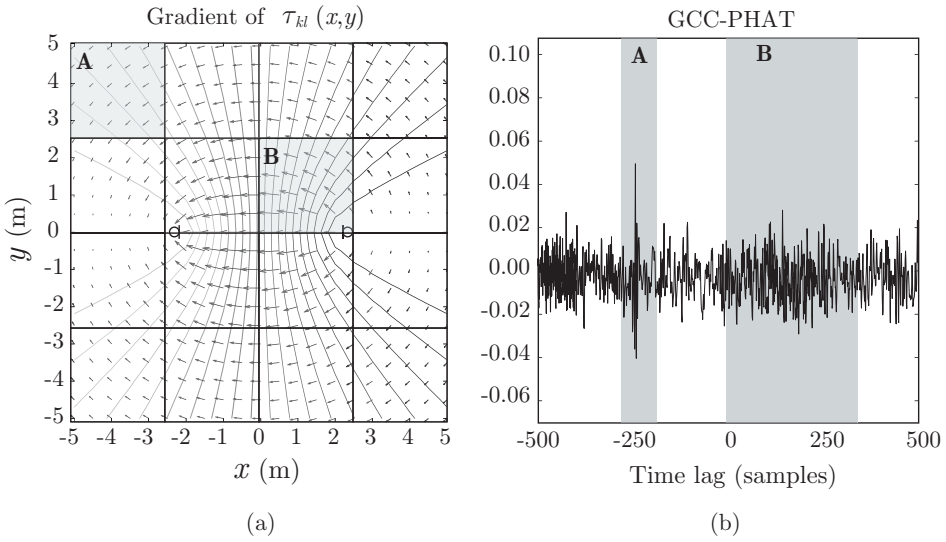


Figure 5.1. **A** and **B** span different accumulation intervals in a coarse grid. (a) Delay function gradient. (b) Noisy GCC.

tial aliasing, the microphone signals are filtered in order to keep only signal components within the speech frequency range.

5.3.2 Iterative Sub-Volume Decomposition

The steps of the algorithm performed at each time frame n are as follows:

1. Compute the GCCs by using the input microphone signals at time frame n. Start with iteration $i = 0$ and initial localization space $V_{-1} = V_{\text{total}}$.

2. Define a spatial resolution r_i and construct a spatial grid $\mathcal{G}_i$ covering the desired localization space V_{i-1}.

3. Apply Eq.(5.10) to all the points in the grid $\mathbf{x} \in \mathcal{G}_i$ and select the point with the greatest value $\mathbf{x}_i = \arg\max_{\mathbf{x}\in\mathcal{G}_i} \left\{\bar{P}'_n\right\}$.

4. The new localization space is that covering all spatial locations closer to $\mathbf{x}_i$ than any other point in the grid, i.e.
$V_i = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{x}_i\| \leq \|\mathbf{x} - \mathbf{x}_{\in\mathcal{G}_i}\|, \forall \mathbf{x}_{\in\mathcal{G}_i} \neq \mathbf{x}_i\}$.

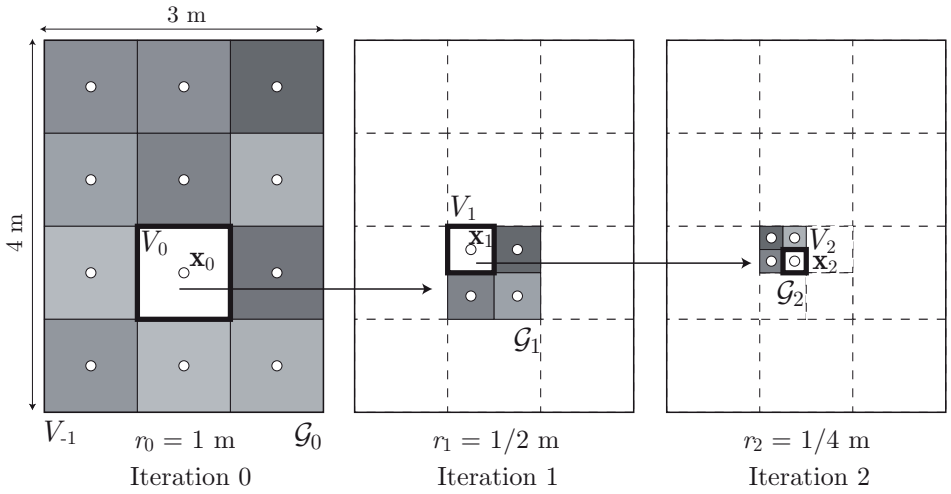


Figure 5.2. Sub-volume division procedure.

5. Go again to Step 2 with iteration $i = i + 1$ and increased spatial resolution $r_i < r_{i-1}$, until reaching a desired final resolution r_f or number of iterations N_T.

Figure 5.2 shows schematically the above process.

5.3.3 Computational Cost

Assuming that the accumulation limits $L_{kl}^{(1,2)}$ are pre-computed in an initialization process, the computational cost of the proposed method only differs from the conventional SRP in terms of the total number of required functional evaluations ν_m. If the resolution at each iteration is defined as a constant scaling of the previous resolution, $r_i = \alpha r_{i-1}$, with $\alpha < 1$, then

$$\nu_m = \frac{V_{\text{total}}}{r_0^3} + (N_T - 1)(1/\alpha)^3, \tag{5.11}$$

resulting in a final resolution $r_f = r_0 \cdot \alpha^{N_T - 1}$. Therefore, the relationship between the number of operations of the conventional SRP having final resolution r_f (denoted as ν_f) and the proposed one is:

$$\frac{\nu_m}{\nu_f} = \frac{r_f^3}{r_0^3} + \frac{N_T - 1}{V_{\text{total}}} \cdot \left(r_f^{(N_T - 2)} r_0 \right)^{\frac{3}{N_T - 1}}. \tag{5.12}$$

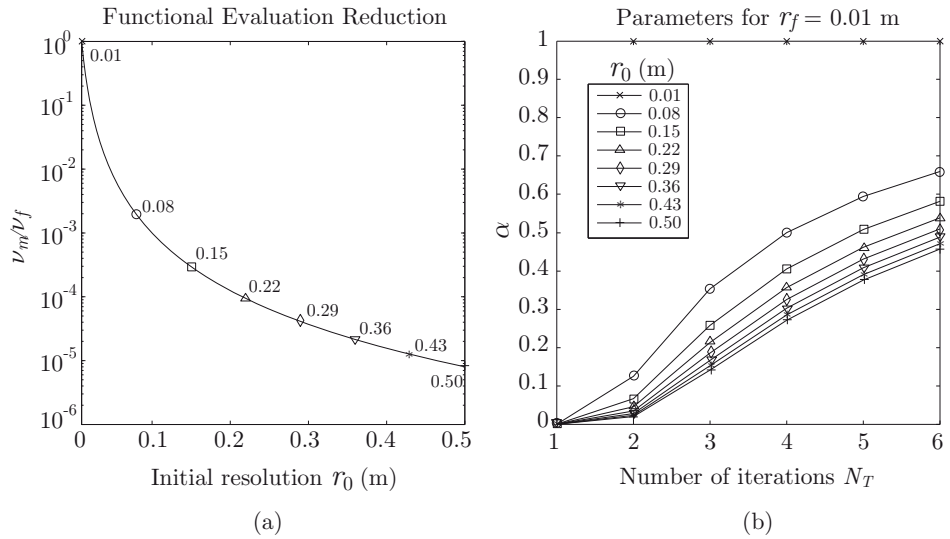


Figure 5.3. (a) Reduction for different initial resolutions r_0 and $r_f = 0.01$ m. (b) Combinations of α and N_T

.

If $N_T \geq 3$, then Eq.(5.12) can be approximated by $(r_f/r_0)^3$. Figure 5.3(a) shows the above relationship for a three-dimensional grid with different initial resolutions r_0 and $r_f = 0.01$ m. In Fig. 5.3(b), possible combinations of α and N_T are shown for the same initial resolutions considered in (a).

5.4 Experiments

In Figure 5.3(a), it is shown that the computational reduction is highly dependent on the chosen initial resolution r_0. The following subsections evaluate the performance of the proposed approach with different initial resolutions and compare it with other SRP-based approaches. To carry out this evaluation, image-source-based acoustic simulations [64] have been performed by considering a rectangular room with dimensions 4 m $\times$ 6 m $\times$ 3 m, with varying wall reflection coefficient ρ and Signal-to-Noise Ratio (SNR). To this end, the *Roomsim* Matlab package [59] was employed. A

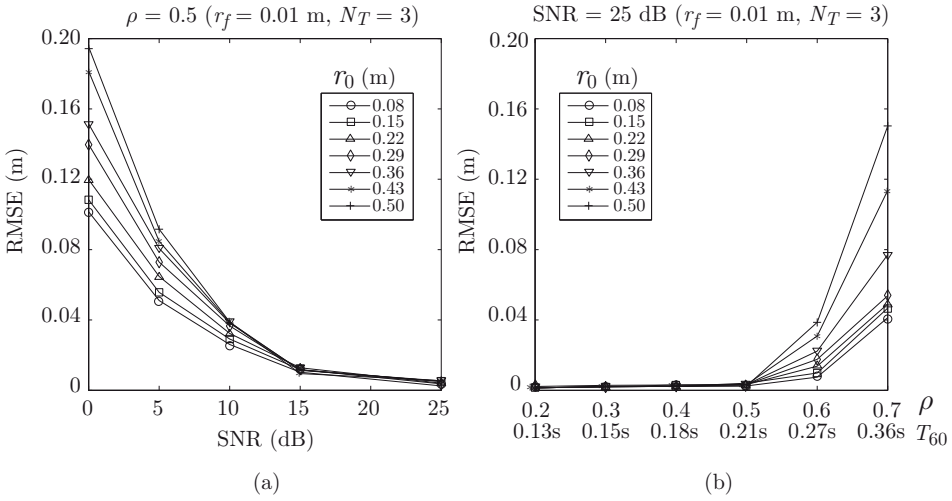


Figure 5.4. (a) RMSE vs. SNR for $\rho = 0.5$. (b) RMSE vs. ρ for SNR = 25 dB.

5 s long speech signal was used as sound source and the results are always presented by averaging 56 different source locations uniformly distributed on a plane. The search volume is restricted to a two-dimensional grid on the same plane. To avoid non-speech frames, the sound source was manually segmented to include only speech frames in the computed results. The localization system consists of $M = 6$ microphones located at the corners of the room and in the middle of the longest wall. The processing was carried out using a sampling rate of 44.1 kHz, with time windows of 4096 samples of length and 50% overlap. The α parameter is always chosen to provide a final resolution $r_f = 0.01$ m after $N_T = 3$ iterations.

5.4.1 Influence of Initial Resolution

Figs.5.4(a) and (b) show the root mean square error (RMSE) for different initial resolutions r_0 as a function of the SNR and the wall reflection coefficient ρ, respectively. The corresponding reverberation time T_{60} (in s) is also provided. In Fig. 5.4(a), ρ is fixed to 0.5, while in Fig. 5.4(b), the SNR is fixed to 25 dB. The initial resolutions tested are the same as the ones shown in Fig. 5.3. Note that, under adequate acoustic conditions (SNR$\geq$ 20 dB

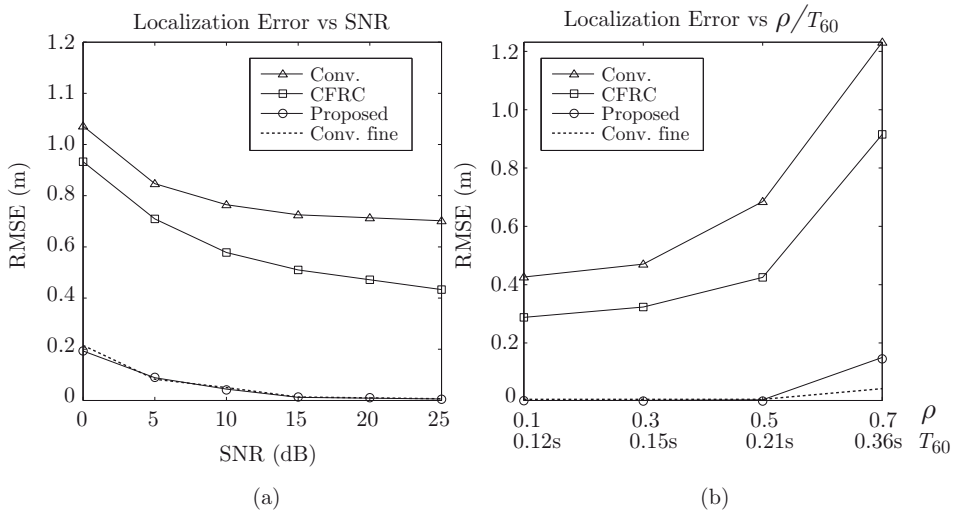


Figure 5.5. Algorithm comparison. (a) RMSE vs. SNR for $\rho = 0.5$. (b) RMSE vs. ρ for SNR = 25 dB.

and $\rho \leq 0.5$), the RMSE is always smaller than r_f (RMSE ≈ 0.0035 m), independently of r_0. The differences among the different initial resolutions are greater when the SNR decreases and/or reverberation increases (higher ρ). This is due to the fact that coarser grids tend to integrate more noise and spurious GCC peaks in adverse conditions, leading to errors in the first iteration. However, the relative performance among the different initial resolutions is not significantly different in moderate acoustic conditions. As a result, a coarse initial spatial grid ($r_0 = 0.5$ m) allows for high-accuracy localization while providing a reduction in functional evaluations of 10^5 when using a three-dimensional search grid.

5.4.2 Algorithm Comparison

This section compares the performance of the proposed method with two other localization algorithms: Coarse-to-Fine Region Contraction (CFRC) [23] and the conventional SRP algorithm (Conv). The proposed method is evaluated for $N_T = 3$, $r_f = 0.01$ m and $r_0 = 0.5$ m. The conventional SRP is evaluated for a grid having a resolution $r = 0.35$ m, while CFRC is evaluated using the suggested parameters [23] ($J = 300$ grid points and

$N = 100$ selected points). Additionally, the performance of the fine-grid SRP algorithm with resolution $r = 0.01$ m (Conv. fine) is also provided as a reference. The experiments were performed in Matlab, using a laptop computer with a 1.7 GHz dual-core processor and 4 GB of RAM. The number of functional evaluations and mean computational time t_c for each case are: 196 for the conventional SRP and the proposed algorithm ($t_c = 44$ ms), 900 for CFRC ($t_c = 46$ ms) and 240.000 for the fine-grid SRP ($t_c = 920$ ms). Figure 5.5(a) and (b) show the results for varying SNR and reflection coefficient, respectively. Note that the proposed method clearly outperforms the rest when using a comparable number of functional evaluations. In fact, the performance of the conventional SRP with $r = 0.01$ m and the proposed method are very similar, having an RMSE that tends to approximate the final resolution when the acoustical conditions are favorable.

5.4.3 Real Setup

A real room, with dimensions $5.7 \times 6.7 \times 2.1$ m and $T_{60} = 0.28$ s, was considered to test the applicability of the method in real-world scenarios. The microphone arrangement was very similar to the one used in the simulations and the algorithm parameters were as in Section 5.4.2. Table 5.1 contains the obtained RMSE for the compared algorithms, showing that the proposed method achieves similar performance to that of the fine-grid SRP.

Table 5.1. RMSE for Real Set-Up

	Conv.	CFRC	Proposed	Conv. fine
RMSE (m)	1.31	0.74	0.30	0.29

5.4.4 Discussion

Note that both the proposed method and CFRC are based on an iterative contraction of the original search volume until a sufficiently small subvolume is reached. However, in CFRC, the functional evaluated over each spatial point corresponds to that of the conventional SRP. As a result, the

volume surrounding each point of the initial grid is not conveniently considered by the algorithm, which makes it more likely to fail in the first step when the number of initial points J is not big enough. Moreover, the contraction operation in CFRC is performed by defining a new region (subvolume) containing the best N points (those with a higher functional value). In our method, it is sufficient to select only the best one, which simplifies considerably the contraction operation and makes the algorithm to converge faster to a desired final resolution.

5.5 Conclusion

In this letter, an iterative approach for high-accuracy sound source localization using the modified SRP functional has been presented. The method starts by performing source localization over a very coarse spatial grid. Then, the grid region having the highest accumulated value is subsequently divided into finer regions until achieving a desired spatial resolution. A set of experiments have been carried out to evaluate this new approach, comparing its localization accuracy with other well-known approaches in different acoustic conditions. The results show that the proposed method has a performance comparable to that of a fine-grid SRP with a reduction of approximately five orders of magnitude in terms of functional evaluations.

5.6 References

The references of this paper have been consolidated in the general bibliography at the end of the book.

A Real-Time Sound Source Localization and Enhancement System Using Distributed Microphones

6

A. Marti, M. Cobos and J. J. Lopez
Proceedings of the 130th AES Convention, London, UK, 2011.

Abstract

The Steered Response Power - Phase Transform (SRP-PHAT) algorithm has been shown to be one of the most robust sound source localization approaches operating in noisy and reverberant environments. A recently proposed modified SRP-PHAT algorithm has been shown to provide robust localization performance in indoor environments without the need for having a very fine spatial grid, thus reducing the computational cost required in a practical implementation. Sound source localization methods are commonly employed in many sound processing applications. In our case, we use the modified SRP-PHAT functional for improving noisy speech signals. The estimated position of the speaker is used to calculate the time-delay for each microphone and then the speech is enhanced by aligning correctly the microphone signals.

6.1 Introduction

Sound Source Localization (SSL) under high noise and reverberation has many applications such as videoconferencing, hands-free speech acquisition, digital hearing aids, video-gaming, autonomous robots and remote surveillance. In this work, we present a microphone array system for speech enhancement based on the well known SRP-PHAT algorithm [19]. The SRP-PHAT method has been shown to be one of the most robust sound source localization approaches operating in noisy and reverberant environments. It is commonly interpreted as a beamforming-based approach that searches for the candidate source positions that maximizes the output of a steered delay-and-sum beamformer. However, the computational requirements of the method are large, making its real-time implementation considerably difficult. Recently, the authors proposed a new strategy based on a modified SRP-PHAT functional that, instead of evaluating the SRP at discrete positions of a spatial grid, it is accumulated over the Generalized Cross Correlation (GCC) lag space corresponding to the volume surrounding each point of the grid [62].

We propose a new application for the SSL systems: speech enhancement. Speech enhancement means the improvement in intelligibility and/or quality of a degraded speech signal by using signal processing tools. Speech enhancement is a very difficult problem for two reasons [65]. First, the nature and characteristics of the noise signals can change dramatically in time and application to application. It is therefore laborious to find versatile algorithms that really work in different practical environments. Second, the performance measure can also be defined differently for each application.

Speech enhancement in the past decades has focused on the suppression of additive background noise. From a signal processing point of view additive noise is easier to deal with than convolutive noise or nonlinear disturbances. Moreover, due to the bursty nature of speech, it is possible to observe the noise by itself during speech pauses, which can be of great value. The experimental setup of the application we propose consist in a limited area (a room) where a set of microphones are located to estimate the position of the speaker. The modified SRP-PHAT functional gives us

the position of the speech source with relatively high accuracy, mostly depending on the selected spatial resolution. Microphone arrays exploit the fact that a speech source is quite stationary and more effectively than any single sensor system. The simplest of all approaches is the delay and sum beamformer that phase aligns incoming wavefronts of the desired source before adding them together. The combination of both systems: SSL and beamforming techniques allow speech improvement in real time processing. This is very useful, since a speech source does not have to be always in the same position for each application and also different sources may exist, for example in a videoconference.

The combination of beamforming-based speech enhancement and SRP-PHAT sound source localization has been recently studied by Levi and Silverman [66]. They proposed a binary masking approach based on a SRP-PHAT discriminator. However, very accurate localization is needed to assign correctly the time-frequency points corresponding to the different sources. In this paper, we propose a method to apply a similar separation framework that takes into account possible errors in the estimation of the sound source locations. Thus, the modified SRP-PHAT using a coarse spatial grid is sufficient to estimate the source separation masks.

The paper is structured as follows. Section 6.2 describes the conventional SRP-PHAT algorithm and our modified functional. Section 6.3 explains how to enhance speech signal using a simple delay-and-sum beamformer and applying a SRP-PHAT based mask to separate different sources in time-frequency domain. Experiments are discussed in Section 6.4. Finally, the conclusions of this work are summarized in Section 6.5.

6.2 SRP-PHAT Sound Source Localization

Consider the output from microphone l, $m_l(t)$, in an M microphone system. Then, the SRP at the spatial point $\mathbf{x} = [x, y, z]$ for a time frame n of length T is defined as

$$P_n(\mathbf{x}) \equiv \int_{nT}^{(n+1)T} \left| \sum_{l=1}^{M} w_l m_l \left(t - \tau(\mathbf{x}, l)\right) \right|^2 dt, \tag{6.1}$$

where w_l is a weight and $\tau(\mathbf{x}, l)$ is the direct time of travel from location $\mathbf{x}$ to microphone l. DiBiase [21] showed that the SRP can be computed by summing the GCCs for all possible pairs of the set of microphones. The GCC-PHAT (GCC using Phase Transform [36]) for a microphone pair (k, l) is computed as

$$R_{m_k m_l}(\tau) = \int_{-\infty}^{\infty} \Psi_{kl}(\omega) e^{j\omega\tau} d\omega, \tag{6.2}$$

where τ is the time lag and

$$\Psi_{kl}(\omega) \equiv \frac{M_k(\omega)}{|M_k(\omega)|} \frac{M_l^*(\omega)}{|M_l(\omega)|}, \tag{6.3}$$

being $M_l^*(\omega)$ the complex conjugated Fourier transform of the microphone signal $m_l(t)$. The term $\Psi_{kl}(\omega)$ is the PHAT filtered cross-spectral power of the kl microphone pair signals.

Taking into account the symmetries involved in the computation of Eq.(7.1) and removing some fixed energy terms [21], the part of $P_n(\mathbf{x})$ that changes with $\mathbf{x}$ is isolated as

$$P'_n(\mathbf{x}) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} R_{m_k m_l}\left(\tau_{kl}(\mathbf{x})\right), \tag{6.4}$$

where $\tau_{kl}(\mathbf{x})$ is the *inter-microphone time-delay function* (IMTDF). This function is very important, since it represents the theoretical direct path delay for the microphone pair (k, l) resulting from a point source located at $\mathbf{x}$. The IMTDF is mathematically expressed as

$$\tau_{kl}(\mathbf{x}) = \frac{\|\mathbf{x} - \mathbf{x}_k\| - \|\mathbf{x} - \mathbf{x}_l\|}{c}, \tag{6.5}$$

where c is the speed of sound, and $\mathbf{x}_k$ and $\mathbf{x}_l$ are the microphone locations.

The SRP-PHAT algorithm consists in evaluating the functional $P'_n(\mathbf{x})$ on a fine grid G with the aim of finding the point-source location $\mathbf{x}_s$ that provides the maximum value:

$$\hat{\mathbf{x}}_s = \arg\max_{\mathbf{x} \in G} P'_n(\mathbf{x}). \tag{6.6}$$

6.2.1 Modified SRP-PHAT

Recently, the authors proposed a new strategy where, instead of evaluating the SRP functional at discrete positions of a spatial grid, it is accumulated over the GCC lag space corresponding to the volume surrounding each point of the grid as follows:

$$P_n''(\mathbf{x}) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} \sum_{\tau=L_{kl1}(\mathbf{x})}^{L_{kl2}(\mathbf{x})} R_{m_k m_l}(\tau). \tag{6.7}$$

The GCC accumulation limits $L_{kl1}(\mathbf{x})$ and $L_{kl2}(\mathbf{x})$ are determined by the gradient of the IMTDF corresponding to each microphone pair, thus, taking into account the spatial distribution of possible TDOAs resulting from a given array geometry, as explained in [62].

6.3 Enhancement

In this section a speech enhancement method is presented. This method, inspired by [66], is based on a simple delay-and-sum beamformer using talkers position estimates given by the modified SRP-PHAT functional to align each microphone signal according to each source position. When more than one talker is active, a SRP-PHAT based mask is applied in order to classify each time-frequency point as belonging to the most probably source.

6.3.1 Delay and Sum Beamformer

One of the simplest approaches for speech enhancement is the delay-and-sum beamformer that phase aligns incoming wavefronts of the desired source before adding them together.

In a M microphone system a talker is situated in a room and his/her position is estimated by the modified SRP-PHAT method introduced in Section 6.2.

Using the information of the microphone positions and the estimated talker location, if we choose the microphone m_1 as reference, then the time

delay between all the microphones and microphone m_1 is expressed by the particularized IMTDF function:

$$\tau_{1l}(\hat{\mathbf{x}}_s) = \frac{\|\hat{\mathbf{x}}_s - \mathbf{x}_1\| - \|\hat{\mathbf{x}}_s - \mathbf{x}_l\|}{c} \quad l = 2 \dots M. \tag{6.8}$$

We steer each microphone to the source location by appropriately delaying the data using the information obtained in Equation (6.8):

$$m_l^d(t) = m_l(t - \tau_{1l}(\hat{\mathbf{x}}_s)), \tag{6.9}$$

where $m_l^d(t)$ is the signal of microphone l delayed accordingly to the estimated talker position $\hat{\mathbf{x}}_s$.

The steered delay-and-sum beamformer is given by

$$m^d(t) = \frac{1}{M} \sum_{l=1}^{M} m_l^d(t). \tag{6.10}$$

In the frequency domain, Equation (6.10) becomes

$$M^d(\omega) = \frac{1}{M} \sum_{l=1}^{M} M_l^d(\omega). \tag{6.11}$$

Note that errors in the estimated $\hat{\mathbf{x}}_s$ will lead to an alignment error. This error will be later discussed in Section 6.3.3.

6.3.2 SRP-PHAT Binary Masking

As previously commented, Equation (6.3) denotes the PHAT filtered cross-spectral power of the microphone signals $m_k(t)$ and $m_l(t)$. After aligning the microphone signals according to the estimated talker location, the delayed cross-spectral density can be expressed as

$$\Psi_{kl}^d(\omega) \equiv \frac{M_k^d(\omega)}{|M_k^d(\omega)|} \frac{M_l^{d*}(\omega)}{|M_l^d(\omega)|}. \tag{6.12}$$

The SRP-PHAT for each frequency will be given by the addition of the normalized cross-spectral densities corresponding to all possible microphone pairs:

$$\Upsilon^d(\omega) = \left| \frac{1}{Q} \sum_{k=1}^{M} \sum_{l=k+1}^{M} \Psi_{kl}^d(\omega) \right|, \tag{6.13}$$

where $Q = M(M-1)/2$ is the number of possible microphone pairs.

In the case we have more than one talker, we get the estimated positions with the modified SRP-PHAT functional and apply the delay-and-sum beamformer explained in Section 6.3.1. To separate the signals corresponding to different simultaneous sources, a binary masking approach is followed using SRP-PHAT discrimination. Therefore, for each analysis frame, a time-frequency point is assigned to talker i if

$$\Upsilon_i^d(\omega) > \Upsilon_j^d(\omega), \quad \forall j \neq i. \tag{6.14}$$

Source separation based on binary masking has already been used by many sound source separation algorithms, such as DUET [52], ADRess [67] or MuLeTS [68]. However, while these approaches are mainly based on inter-channel level and phase differences, the method here proposed uses SRP-PHAT discrimination.

6.3.3 Estimation of Localization Error

The modified SRP-PHAT functional gives us the estimated position of the talker, however, this position has a margin of tolerance defined by the spatial grid resolution used in the algorithm. If the estimated position was perfect and considering only direct-path signals, the angle of $\Psi_{kl}^d(\omega)$ for any microphone pair (k, l) would be equal to zero since the microphone signals are also perfectly aligned. The magnitude of the position error is limited by the spatial grid resolution. This error needs to be estimated and corrected before estimating the separation masks.

Consider that a location estimation error results in erroneous signal alignment, since

$$\tau_{kl}(\hat{\mathbf{x}}_s) = \tau_{kl}(\mathbf{x}_s) + \varepsilon_{kl}. \tag{6.15}$$

Thus, the delayed cross-power spectral density is

$$\Psi_{kl}^{d}(\omega) \equiv \frac{M_k^d(\omega)}{|M_k^d(\omega)|} \frac{M_l^{d*}(\omega)}{|M_l^d(\omega)|} \exp^{j\omega\varepsilon_{kl}} . \tag{6.16}$$

Observing the phase between microphone signals once they have been delayed, we get an estimate of the time delay error for each microphone pair by

$$\hat{\varepsilon}_{kl} = \frac{1}{\omega} \angle \left(\frac{M_k^d(\omega)}{M_l^d(\omega)} \right), \tag{6.17}$$

where $\angle()$ denotes the phase of a complex number.

A better estimation of this error can be obtained by means of an energy-weighted histogram of the different $\hat{\varepsilon}_{kl}$ observed at every time-frequency point. Figure 6.1 shows two example histograms obtained from a mixture of two sources in an anechoic scenario and in a room with short reverberation time. The theoretical error values for a microphone pair are -0.297 ms considering the position of the first source and -0.1109 ms for the second source. The real estimated error values when $\rho = 0$ are -0.3006 ms for the position of the first source and -0.1403 ms for the second source and, when $\rho = 0.5$, estimated error value for source 1 is -0.3106 ms and -0.09018 ms for source 2.

Therefore, once the different errors have been estimated, the corrected cross-spectral densities for each microphone pair are

$$\Psi_{kl}^{dc}(\omega) = \Psi_{kl}^{d}(\omega) \exp^{j\omega\hat{\varepsilon}_{kl}} . \tag{6.18}$$

Finally, the masks for each talker can be applied using the Equation (6.14) with the SRP-PHAT values calculated using the corrected $\Psi_{kl}^{dc}(\omega)$.

Figure 6.2 shows the angle of the cross-spectral density for the pair of microphones of the example mixture in the anechoic case. Figure 6.2(a) shows the phase when the error is present and Figure 6.2(b) shows the phase after correction. Finally, Figure 6.2(c) shows the source 1 mask obtained by comparing the SRP-PHAT for the two sources at every time-frequency point.

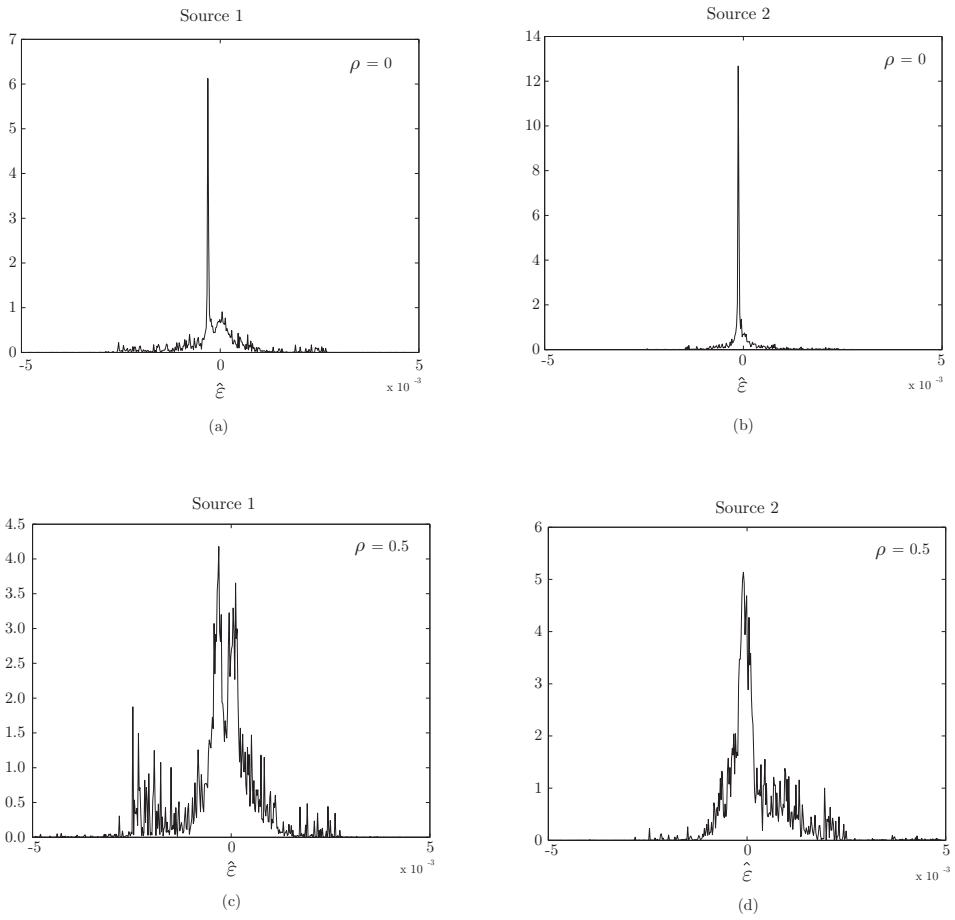


Figure 6.1. Energy-weighted histogram of the different $\hat{\varepsilon}_{kl}$ observed at every time-frequency point. (a) Source 1 with $\rho = 0$. (b) Source with 2 $\rho = 0$. (c) Source 1 with $\rho = 0.5$. (d) Source 2 with $\rho = 0.5$.

6.4 Experiments

To evaluate the performance of the proposed correction method, several simulated recordings have been generated using the *Roomsim* Matlab package [59]. The simulation set-up consisted of six microphones placed on the walls of a shoe-box-shaped room with dimensions 4 m $\times$ 6 m $\times$ 2 m (Fig.

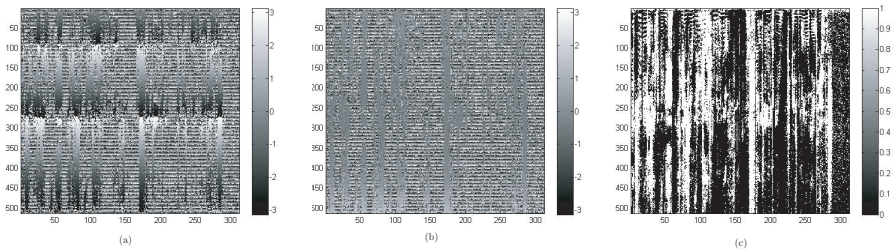


Figure 6.2. Angle of $\Psi_{kl}^{d}(\omega)$. (a) Before error correction. (b) After error correction. (c) Time-frequency mask for isolating source 1

6.3). The simulations were repeated with three different reverberation conditions (using wall reflection factors of $\rho = 0$, $\rho = 0.2$ and $\rho = 0.5$), considering different Signal-to-Noise Ratio (SNR) conditions. The two sources were located at (3,2,0.5) and (1,3,0.5). The modified SRP-PHAT algorithm using a spatial grid resolution of $r = 0.2$ m was used to locate the sources even in the case when both are simultaneously active [69].

The processing was carried out using a sampling rate of 16 kHz, with time windows of 1024 samples of length and 50% overlap. The simulated sources were two male speech signals of length 10 s with no pauses.

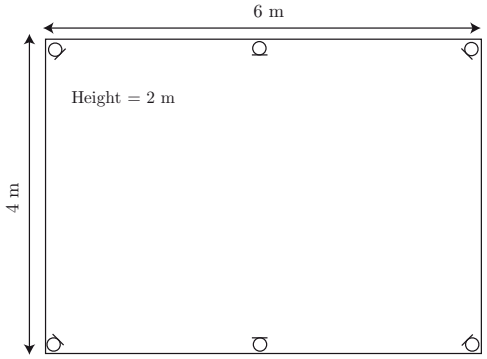


Figure 6.3. Microphone set-up used in the experiments.

6.4.1 Results

Figure 6.4 shows the percentage of time-frequency points which have been correctly assigned in the estimated masks. This percentage has been calculated respect to the ideal binary masks [70].

The results show that, after error correction and in the anechoic case and high SNR, 93% of the time-frequency points (considering only the speech bandwidth from 0 to 5kHz) have been assigned correctly to its real source. This is an indication of very good separation quality. Obviously, with lower SNR and higher reverberation time, the percentage of correctly assigned points decreases. This results in a worse separation quality. Nevertheless, note that in every case the performance after error correction is significantly better.

6.5 Conclusion

This paper presented a microphone array system for speech enhancement based on the delay-and-sum beamformer and the modified SRP-PHAT functional developed by the authors. The estimated positions given by the SRP-PHAT algorithm are used to determine the distance from the talker to each microphone and apply the delay-and-sum beamformer. Since the localization step may have a given error due to the limited spatial resolution, an estimation of the time delay error is needed in order to correct the phase alignment of all the microphone signals. The estimation of the error is better when the acoustic conditions of the environment are not too adverse, leading to numerically and perceptually better separation results.

6.6 References

The references of this paper have been consolidated in the general bibliography at the end of the book.

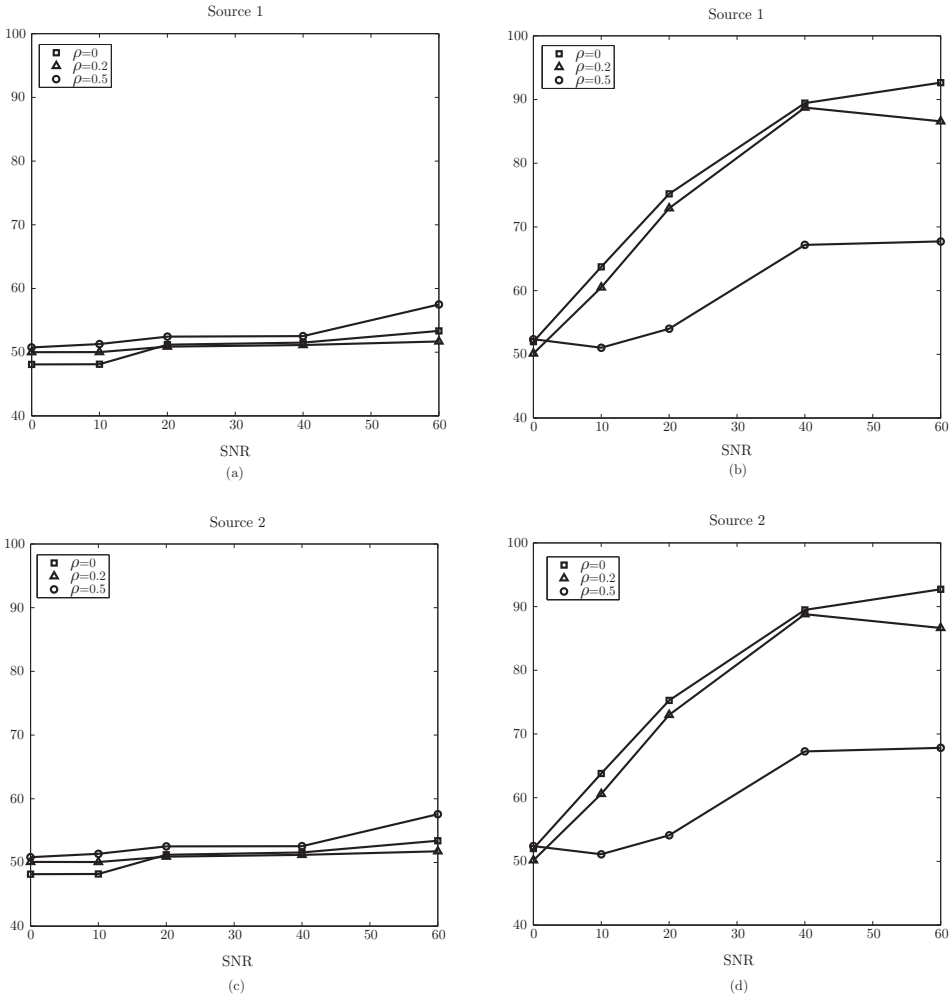


Figure 6.4. Percentage of time-frequency points correctly assigned to its real source for different SNR and wall reflection factor values. (a) Source 1 before error correction. (b) Source 1 after error correction. (c) Source 2 before error correction. (d) Source 2 after error correction.

Real-Time Speaker Localization and Detection System for Camera Steering in Multiparticipant Videoconferencing Environments

7

A. Marti, M. Cobos and J. J. Lopez
Proceedings of the 36th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 2011.

Abstract

A real time speaker localization and detection system for videoconferencing environments is presented. In this system, a recently proposed modified *Steered Response Power - Phase Transform* (SRP-PHAT) algorithm has been used as the core processing scheme. The new SRP-PHAT functional has been shown to provide robust localization performance in indoor environments without the need for having a very fine spatial grid, thus reducing the computational cost required in a practical implementation. Moreover, it has been demonstrated that the statistical distribution of location estimates when a speaker is active can be successfully used to discriminate between speech and non-speech frames by using a criterion of peakedness. As a result, talking participants can be detected and located with signifi-

cant accuracy following a common processing framework.

7.1 Introduction

Many applications, ranging from teleconferencing systems to artificial perception, hands-free speech acquisition, digital hearing aids, video-gaming, autonomous robots and remote surveillance require the localization of one or more acoustic sources. Since the boost of new generation videoconferencing environments, there has been growing interest in the development of automatic camera-steering systems using microphone arrays [6; 7]. In this work, we present a microphone array system for camera-steering to be used in a multiparticipant videoconferencing environment based on the well-known SRP-PHAT algorithm [19]. The SRP-PHAT method has been shown to be one of the most robust sound source localization approaches operating in noisy and reverberant environments. It is commonly interpreted as a beamforming-based approach that searches for the candidate source position that maximizes the output of a steered delay-and-sum beamformer. However, the computational requirements of the method are large, making its real-time implementation considerably difficult. Since the SRP-PHAT method was proposed, there have been several attempts to reduce the computational cost of the method, such as those presented in [22; 23]. Recently, the authors proposed a new strategy based on a modified SRP-PHAT functional that, instead of evaluating the SRP at discrete positions of a spatial grid, it is accumulated over the Generalized Cross Correlation (GCC) lag space corresponding to the volume surrounding each point of the grid [71]. The benefits of following this approach are twofold. On the one hand, it incorporates additional spatial knowledge at each point for making a better final decision. On the other hand, the proposed modification achieves the same performance as SRP-PHAT with fewer functional evaluations, relaxing the computational demand required for a practical application.

In this paper, we analyze the distribution of location estimates obtained with the modified SRP-PHAT functional with the aim of establishing a speaker detection rule to be used in a videoconferencing environment

involving multiple participants. The analysis shows that location estimates follow different distributions when speakers are active, allowing to discriminate between speech and non-speech frames under a common localization framework. Moreover, the distribution of an active speaker remains almost the same for different positions inside the room, which makes easier to select a candidate location following a maximum-likelihood criterion, thus simplifying the camera-steering task.

The paper is structured as follows. Section 7.2 describes the conventional SRP-PHAT algorithm and our modified functional. Section 7.3 explains the proposed localization-based approach to speech/non-speech discrimination and speaker detection. Experiments with real-data are discussed in Section 7.4. Finally, the conclusions of this work are summarized in Section 7.5.

7.2 SRP-Based Source Localization

Consider the output from microphone l, $m_l(t)$, in an M microphone system. Then, the SRP at the spatial point $\mathbf{x} = [x, y, z]$ for a time frame n of length T is defined as

$$P_n(\mathbf{x}) \equiv \int_{nT}^{(n+1)T} \left| \sum_{l=1}^{M} w_l m_l \left(t - \tau(\mathbf{x}, l)\right) \right|^2 dt, \tag{7.1}$$

where w_l is a weight and $\tau(\mathbf{x}, l)$ is the direct time of travel from location $\mathbf{x}$ to microphone l. DiBiase [21] showed that the SRP can be computed by summing the GCCs for all possible pairs of the set of microphones. The GCC for a microphone pair (k, l) is computed as

$$R_{m_k m_l}(\tau) = \int_{-\infty}^{\infty} \Phi_{kl}(\omega) M_k(\omega) M_l^*(\omega) e^{j\omega\tau} d\omega, \tag{7.2}$$

where τ is the time lag, * denotes complex conjugation, $M_l(\omega)$ is the Fourier transform of the microphone signal $m_l(t)$, and $\Phi_{kl}(\omega) = W_k(\omega) W_l^*(\omega)$ is a combined weighting function in the frequency domain. The phase transform (PHAT) [36] has been demonstrated to be a very effective GCC weighting

for time delay estimation in reverberant environments:

$$\Phi_{kl}(\omega) \equiv \frac{1}{|M_k(\omega)M_l^*(\omega)|}. \tag{7.3}$$

Taking into account the symmetries involved in the computation of Eq.(7.1) and removing some fixed energy terms [21], the part of $P_n(\mathbf{x})$ that changes with $\mathbf{x}$ is isolated as

$$P'_n(\mathbf{x}) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} R_{m_k m_l}\left(\tau_{kl}(\mathbf{x})\right), \tag{7.4}$$

where $\tau_{kl}(\mathbf{x})$ is the *inter-microphone time-delay function* (IMTDF). This function is very important, since it represents the theoretical direct path delay for the microphone pair (k, l) resulting from a point source located at $\mathbf{x}$. The IMTDF is mathematically expressed as

$$\tau_{kl}(\mathbf{x}) = \frac{\|\mathbf{x} - \mathbf{x}_k\| - \|\mathbf{x} - \mathbf{x}_l\|}{c}, \tag{7.5}$$

where c is the speed of sound, and $\mathbf{x}_k$ and $\mathbf{x}_l$ are the microphone locations.

The SRP-PHAT algorithm consists in evaluating the functional $P'_n(\mathbf{x})$ on a fine grid G with the aim of finding the point-source location $\mathbf{x}_s$ that provides the maximum value:

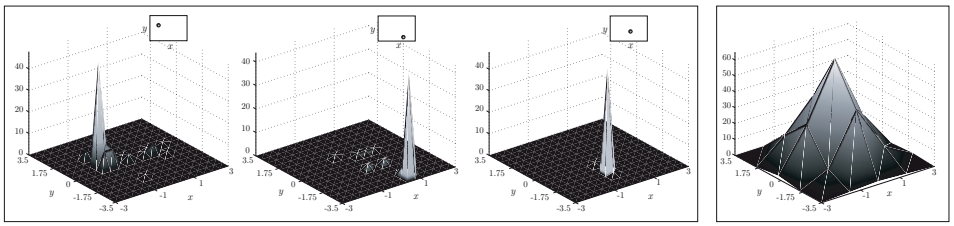
$$\hat{\mathbf{x}}_s = \arg\max_{\mathbf{x}\in G} P'_n(\mathbf{x}). \tag{7.6}$$

7.2.1 Modified SRP-PHAT Functional

Recently, the authors proposed a new strategy where, instead of evaluating the SRP functional at discrete positions of a spatial grid, it is accumulated over the GCC lag space corresponding to the volume surrounding each point of the grid as follows:

$$P''_n(\mathbf{x}) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} \sum_{\tau=L_{kl1}(\mathbf{x})}^{L_{kl2}(\mathbf{x})} R_{m_k m_l}(\tau). \tag{7.7}$$

The GCC accumulation limits $L_{kl1}(\mathbf{x})$ and $L_{kl2}(\mathbf{x})$ are determined by the gradient of the IMTDF corresponding to each microphone pair, thus, taking into account the spatial distribution of possible TDOAs resulting from a given array geometry, as explained in [71].



(a) (b)

Figure 7.1. Two-dimensional histograms showing the distribution of location estimates. (a) Distribution obtained for three different speaker locations. (b) Distribution for non-speech frames.

7.3 Speaker Detection

In the next subsections, we describe how active speakers are detected in our system, which requires a previous discrimination between speech and non-speech frames based on the distribution of location estimates.

7.3.1 Distribution of Location Estimates

Our first step to speaker detection is to analyze the distribution of the location estimates $\hat{\mathbf{x}}_s$ when there is an active speaker talking inside the room from a static position. In this context, six microphones were placed on the walls of the videoconferencing room and a set of 12 recordings from different speaker positions were analyzed to obtain the resulting location estimates. Figure 7.1(a) shows an example of three two-dimensional histograms obtained from different speaker locations. It can be observed that, since the localization algorithm is very robust, the resulting distributions when speakers are active are significantly peaky. Also, notice that the shape of the distribution is very similar in all cases but centered in the actual speaker location. As a result, we model the distribution of estimates as a bivariate Laplacian as follows:

$$p(\hat{\mathbf{x}}_s|H_s(\mathbf{x}_s)) = \frac{1}{2\sigma_x\sigma_y}\exp^{-\sqrt{2}\left(\frac{|x-x_s|}{\sigma_x}+\frac{|y-y_s|}{\sigma_y}\right)}, \tag{7.8}$$

where $p(\hat{\mathbf{x}}_s|H_s(\mathbf{x}_s))$ is the conditional probability density function (pdf) of the location estimates under the hypothesis $H_s(\mathbf{x}_s)$ that there is an active speaker located at $\mathbf{x}_s = [x_s, y_s]$. Note that the variances σ_x^2 and σ_y^2 may depend on the specific microphone set-up and the selected processing parameters. This dependence will be addressed in future works. On the other hand, a similar analysis was performed to study how the distribution changes when there are not active speakers, i.e. only noise frames are being processed. The resulting histogram can be observed in Figure 7.1(b), where it becomes apparent that the peakedness of this distribution is not as significant as the one obtained when there is an active source. Taking this into account, the distribution of non-speech frames is modeled as a bivariate Gaussian:

$$p(\hat{\mathbf{x}}_s|H_n) = \frac{1}{2\pi\sigma_{x_n}\sigma_{y_n}} \exp^{-\left(\frac{x^2}{2\sigma_{x_n}^2}+\frac{y^2}{2\sigma_{y_n}^2}\right)}, \tag{7.9}$$

where $p(\hat{\mathbf{x}}_s|H_n)$ is the conditional pdf of the location estimates under the hypothesis H_n that there are not active speakers, and the variances $\sigma_{x_n}^2$ and $\sigma_{y_n}^2$ are those obtained with noise-only frames.

The suitability of the proposed models has been tested by a fitting procedure based on trust region optimization, having a R-square parameter above 0.95 in both cases.

7.3.2 Speech/Non-Speech Discrimination

In the last subsection, it has been shown that speech frames are characterized by a bivariate Laplacian probability density function. A similar analysis of location estimates when there are not active speakers results in a more Gaussian-like distribution, which is characterized by a shape less peaky than a Laplacian distribution. This property is used in our system to discriminate between speech and non-speech frames by observing the

peakedness of a set of accumulated estimates:

$$\mathbf{C} = \begin{bmatrix} \hat{x}_s(n) & \hat{y}_s(n) \\ \hat{x}_s(n-1) & \hat{y}_s(n-1) \\ \vdots & \vdots \\ \hat{x}_s(n-L-1) & \hat{y}_s(n-L-1) \end{bmatrix} = [\mathbf{c}_x\, \mathbf{c}_y], \tag{7.10}$$

where L is the number of the accumulated estimates in matrix $\mathbf{C}$. A peakedness criterion based on high-order statistics was evaluated. Since the kurtosis of a normal distribution equals 3, we propose the following discrimination rules for active speech frames:

$$\mathrm{Kurt}(\mathbf{c}_x) \begin{cases} \geq 3 & \text{speech} \\ < 3 & \text{non} - \text{speech} \end{cases}, \tag{7.11}$$

$$\mathrm{Kurt}(\mathbf{c}_y) \begin{cases} \geq 3 & \text{speech} \\ < 3 & \text{non} - \text{speech} \end{cases}, \tag{7.12}$$

where a frame is selected as speech if any of the above conditions is fulfilled.

7.3.3 Camera Steering

To provide a suitable camera stability, a set of target positions were pre-defined coinciding with the actual seats in the videoconferencing room. The localization system will be responsible for communicating the camera which of the target positions is currently active. This process involves two main steps. First, it is necessary to discriminate between speech and non-speech frames as explained in Section 7.3.2. If a burst of speech frames is detected, then the estimated target position is forwarded to the camera when it does not match the current target seat. Since all the target positions are assumed to have the same prior probability, a maximum-likelihood criterion is followed:

$$\hat{\mathbf{x}}_t = \arg\max_{\mathbf{x}_t} p\left(\hat{\mathbf{x}}_s | H(\mathbf{x}_t)\right), \quad t = 1 \ldots N_t, \tag{7.13}$$

where $\mathbf{x}_t$ is one of the N_t pre-defined target positions. Given that the likelihoods have the same distribution centered at different locations, the estimated target position $\hat{\mathbf{x}}_t$ is the one which is closest to the estimated location $\hat{\mathbf{x}}_s$.

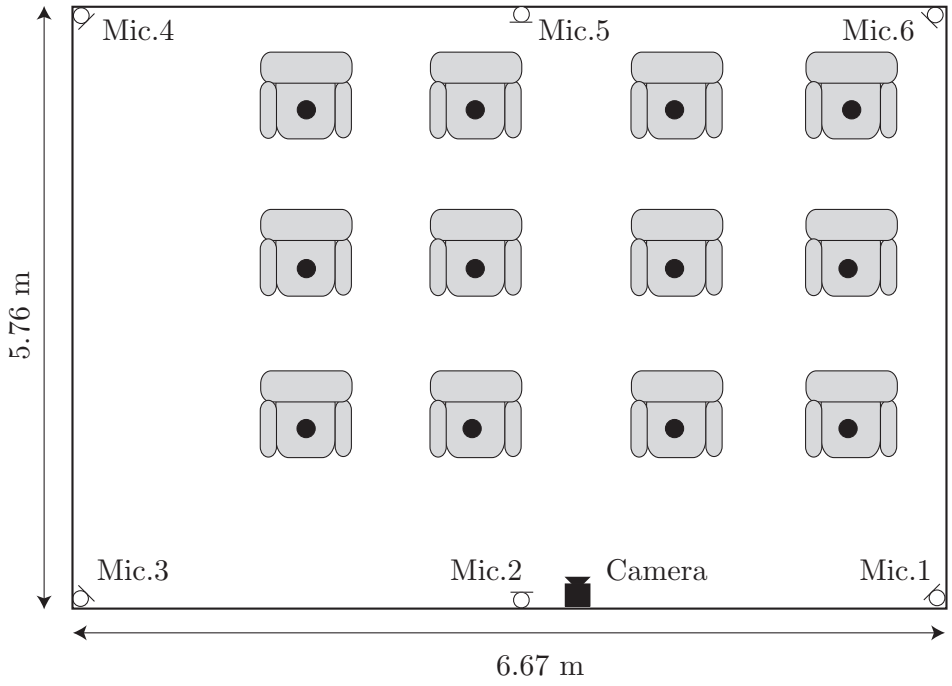


Figure 7.2. Videoconferencing test room and microphones location.

7.4 Experiments

To evaluate the performance of our proposed approach a set of recordings was carried out in a videoconferencing test room with dimensions 6.67 m x 5.76 m x 2.10 m. A set of 6 omnidirectional microphones were placed on the walls of the room. To be precise, 4 of the microphones were situated at the 4 corners of the ceiling of the room and the other two microphones were placed at the same height but in the middle of the longest walls. Figure 9.1 shows the microphone set-up, the camera location and the different seats occupied by the participants. Black dots represent the 12 pre-defined target locations used to select the active speaker seat.

The experiment consisted in recording speakers talking from the different target positions (only one speaker at each time) with the corresponding space of silence between two talking interventions. The recordings were

processed with the aim of evaluating the performance of our system in discriminating speech from non-speech frames and determining the active speaker so that the camera can point at the correct seat. With this aim, the original recordings were manually labeled as speech and non-speech fragments. The processing used a sampling rate of 44.1 kHz, with time windows of 2048 samples and 50% overlap. The location estimates were calculated using the modified SRP-PHAT functional, as explained in Section 7.2. The discrimination between speech and non-speech frames was carried out by calculating the kurtosis of the last L estimated positions, as explained in Section 7.3.2.

7.4.1 Results

Table 1 shows the percentage of correctly detected speech (% SP) and non-speech (% N-SP) frames with different number of accumulated positions $L = 5, 10, 15, 20$. Moreover, the processing was performed considering two different spatial grid sizes (0.3 m and 0.5 m). The percentage of speech frames with correct target positions (% T) is also shown in the table. It can be observed that, generally, the performance increases with a finer grid and with the number of accumulated estimates L. These results were expectable, since the involved statistics are better estimated with a higher number of location samples. Although it may seem that there are a significant number of speech frames that are not correctly discriminated, it should be noticed that this is not a problem for the correct driving of the camera, since most of them are isolated frames inside speech fragments that do not make the camera change its pointing target.

Grid res.	0.5 m				0.3 m			
L	5	10	15	20	5	10	15	20
% SP	52.5	60.4	70.0	74.0	68.9	70.7	83.1	85.4
% N-SP	75.9	64.8	70.9	72.7	81.4	70.9	81.5	82.3
% T	98.2				99.6			

Table 7.1. Performance in Terms of Percentage of Correct Frames

7.5 Conclusion

This paper presented a microphone array system for camera-steering to be used in a multiparticipant videoconferencing environment based on the well-known SRP-PHAT algorithm. The distribution of location estimates obtained with a modified SRP-PHAT functional was analyzed, showing that location estimates follow different distributions when speakers are active and allowing to discriminate between speech and non-speech frames under a common localization framework. The results of experiments conducted in a real room suggest that, using a moderately high number of accumulated location estimates, it is possible to discriminate with significant accuracy between speech and non-speech frames, which is sufficient to correctly detect an active speaker and make the camera point at his/her pre-defined location.

7.6 References

The references of this paper have been consolidated in the general bibliography at the end of the book.

Automatic Speech Recognition in Cocktail-Party Situations: A Specific training for Separated Speech

8

A. Marti, M. Cobos and J. J. Lopez
Journal of the Acoustical Society of America, vol. 131, no. 2, pp.1529–1535, 2012.

Abstract

Automatic Speech Recognition (ASR) refers to the task of extracting a transcription of the linguistic content of an acoustical speech signal automatically. Despite several decades of research in this important area of acoustic signal processing, the accuracy of ASR systems is still far behind human performance, especially in adverse acoustic scenarios. In this context, one of the most challenging situations is the one concerning simultaneous speech in cocktail-party environments. Although source separation methods have already been investigated to deal with this problem, the separation process is not perfect and the resulting artifacts pose an additional problem to ASR performance. In this paper, a specific training to improve the percentage of recognized words in real simultaneous speech cases is

proposed. The combination of source separation and this specific training is explored and evaluated under different acoustical conditions, leading to improvements of up to a 35% in ASR performance.

8.1 Introduction

Automatic Speech Recognition (ASR) can be defined as the independent, computer-driven transcription of spoken language into readable text in real-time [72]. This technology allows a computer or electronic device to identify the words spoken by a person so that the message can be stored or processed in a useful way [73]. ASR is used on a day-to-day basis in a number of applications and services such as natural human-machine interfaces, dictation systems, electronic translators and automatic information desks [5]. However, there are still some challenges to be solved [74]. A major problem in ASR is to recognize people speaking in a room by using a distant microphone [24]. In distant-speech recognition, the microphone does not only receive the direct path signal, but also delayed replicas as a result of multi-path propagation [25]. The existing mismatch between the training and testing conditions limits the performance of ASR systems, thus, robust recognition methods are aimed at reducing this mismatch. In this context, several approaches have been proposed to cope with room reverberation in speech recognition applications. Some methods are based on a speech enhancement stage prior to recognition [26]. In other methods, the recognizer itself is made robust to reverberation by using model compensation or by performing an improved feature extraction process [27]. All these approaches have shown to be useful to improve ASR. In any case, reducing the acoustic mismatch between training and testing conditions seems to be a relevant issue for the development of robust ASR systems. Despite the fact that speech recognizers are usually trained on anechoic (or almost anechoic) conditions, the environment where they are usually employed can be rarely considered anechoic. To this end, the use of different training data matching several acoustic environments has already been suggested [29; 28], yielding a noticeable improvement.

Many efforts have been made to develop robust ASR systems working in reverberant and noisy conditions, most of them focused on recognizing a single speech source. However, besides noise and reverberation, cocktail-party situations where different speakers are talking at the same time pose a real problem for ASR systems [30; 31]. Source separation refers to the task of estimating and recovering independent source signals (for example, speech signals) from a set of mixtures in one or several observation channels (microphone signals) [45]. Source separation algorithms have been described in the literature as a solution for simultaneous speech recognition [56], proposing ASR performance as an indicator of the quality achieved by a given source separation algorithm [75; 76]. However, separating speech signals in real acoustic environments is not an easy task and the extracted speech signals are usually degraded by audible artifacts [77]. As a result, separated speech signals present an additional mismatch with respect to the training signals used in a conventional ASR system.

In this paper, we propose and evaluate the use of a specific training method for ASR in multiple-talk situations. This training includes both the room effects and the particular artifacts resulting from the separation process. The aim is to increase the robustness of ASR systems when a source separation stage is performed before word recognition. Different experiments are conducted by using simulated data in diverse acoustic environments, showing that the proposed training significantly improves ASR performance in multiple simultaneous speech cases.

The paper is structured as follows. Section 8.2 and Section 8.3 present some fundamentals of robust ASR and source separation, respectively. Section 8.4 describes some training approaches to achieve robust ASR in adverse conditions, including cocktail-party situations. Experiments to evaluate ASR performance in multiple acoustic environments are presented in Section 8.5, providing a general discussion in Section 8.6. Finally, the conclusions of this work are summarized in Section 8.7.

8.2 Robust Speech Recognition

Speech recognition has significantly improved in the last decade. Its improvements are the result of many research efforts in four different areas [78]. Firstly, the use of common speech corpora allows the use of large training sets and makes it possible to compare results from different ASR systems. Secondly, many developments have been observed in the area of acoustic modeling, such as contributions regarding context-specific Hidden Markov Models (HMMs), changes in feature vectors over time or the presence of cross-word effects. Finally, improvements in language modeling and search algorithms allow for the better recognition of large vocabulary corpora and reduced experimentation cycles, respectively.

Unfortunately, most of the above improvements have been developed assuming clean speech. When common ASR systems are used in reverberant and/or noisy environments, the speech signal is degraded and the extracted data vectors differ significantly from the ones expected by the recognizer. In fact, not only the acoustical conditions are responsible for these changes, but also the speaker tends to change his/her voice as a function of the auditory feedback [79]. As a result, to reduce the error-rate in the recognition task, a processing should be included to reduce the differences between training and test environments. This can be done in two ways: either by producing changes in the speech model parameters to match the training environment or by transforming the acquired input data to the environment where the models were trained. Three basic approaches summarize the different alternatives [78]:

1. *Use of Robust Features and Similarity Measurements*: the ASR system is assumed to be noise-independent, with the same configuration used for both clean and distorted speech. Thus, these methods are focused on deriving speech features and similarity measures that are robust to environment changes.

2. *Speech Enhancement*: These methods are based on pre-processing the input speech signal by applying a denoising or a dereverberation algorithm. These techniques are not usually designed to improve ASR

performance specifically, but speech quality or intelligibility instead.

3. *Speech Model Compensation*: In these methods, a transformation of the reference speech model is performed to account for specific environment conditions. The usual statistical modeling techniques (HMMs, Neural Networks, etc.) are trained with the model parameters adapted to accommodate distorted speech.

The above techniques usually consider a single speech source scenario. However, as described in the next subsection, cocktail-party scenarios might render an additional challenge to robust ASR.

8.2.1 Speech Recognition in Cocktail-Party Situations

Besides noise and reverberation, ASR systems might deal with multi-talk situations. Humans have the ability to distinguish individual sound sources from complex mixtures of sound. This human ability is usually related to the well-known *cocktail-party effect* discussed by Cherry in 1953 [80], which describes the ability to focus one's listening attention on a single talker among a mixture of conversations and background noises, ignoring other conversations and enabling humans to talk in a noisy place.

Two different situations can be considered regarding ASR with simultaneous speech. In this context, an ASR system might be designed to recognize words or commands from only one source, thus, ignoring the signals coming from other talkers. On the other hand, there might be ASR systems intended to recognize all the words or commands emitted by different simultaneous sources so that different actions are performed consequently. In the first case, there is only one target speech, while the other speech signals are considered to be interferences. In the second case, all the speech signals are target signals and interference signals at the same time.

In both of the above situations, source separation is needed before ASR. However, while in the first case a good separation quality should be achieved only for the target speech signal, in the second case all the extracted speech signals should have sufficient quality for a successful recognition. This fact makes one think of the difficulty to perform ASR in cocktail-party

environments. Moreover, source separation should be performed in real time for a practical ASR system, which makes the problem even harder.

In the next section, we describe the source separation problem in real acoustic environments and some useful approaches aimed at separating speech by means of microphone array processing.

8.3 Source Separation in Real Environments

The solution to the problem of sound separation in real environments is essential for many applications besides ASR, such as hearing-aid systems or hands-free devices. Most of the existing separation algorithms are based on statistical assumptions, mainly, sources statistically independent and non-Gaussian. These assumptions often lead to the *independent component analysis* approach [81]. These algorithms have shown to be successful in the linear complete case, when as many observations as sources are available. In a real situation (mixtures recorded in a room with a set of microphones), the mixing process is said to be convolutive, since each sensor observes the original source signals convolved with the impulse response between each source and sensor [82; 83]. This makes the estimation of the sources even more difficult. Methods based on independent component analysis for convolutive mixtures often apply the separation algorithm separately in each frequency bin using a time-frequency transformation of the observed mixtures. This approach introduces the well-known permutation problem: the different frequency components of the signals are swapped and require an alignment process [84]. Moreover, when there are more sources than observation channels, the problem is underdetermined and other properties of the sources such as sparsity are exploited. When dealing with speech mixtures, it has been shown that they are sparser in the time-frequency domain than in the time domain [53]. Yilmaz et al. [85] assumed that the sources are disjoint in the time-frequency domain, i.e., there exists only one source in a given time-frequency point. This assumption leads to the *time-frequency masking* approach.

Algorithms based on time-frequency masking have shown to provide

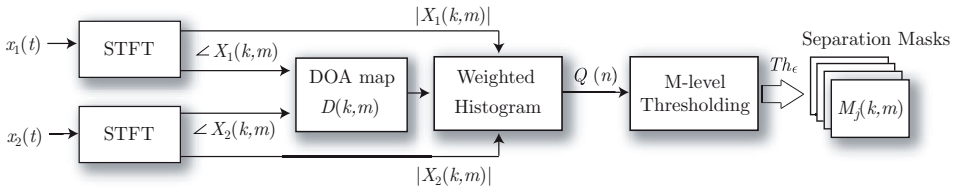


Figure 8.1. Block diagram of the source separation algorithm.

significant results in the separation of realistic mixtures of speech [49]. Recently, the authors developed the multi-level thresholding separation method, which is a time-frequency masking approach based on interclass variance maximization [86]. This method provides good separation quality using only two microphones and can be easily implemented in real-time, which makes it very useful for practical ASR systems. The method, which is next briefly described, is used in this work as a pre-processing stage for cocktail-party ASR.

8.3.1 Speech Separation Based on Interclass Variance Maximization

An overview of the different stages of the multi-level thresholding separation algorithm is depicted in Figure 8.1. Two microphone signals, $x_i(t)$, $i \in \{1, 2\}$, are sufficient to separate N sources, even if the number of sources exceeds the number of microphones. The method can be summarized in the following steps:

1. *Direction-Of-Arrival (DOA) map calculation.* The input microphone signals, $x_i(t)$, are transformed into the *Short-Time Fourier Transform* (STFT) domain, obtaining $X_i(k, m)$. The phase information in each time-frequency point is used to estimate the DOA of a given frequency bin k in the current analysis window m. A DOA map matrix, $D(k, m)$, is formed with the estimations obtained from the analyzed signal.

2. *Coherence test and weighted histogram.* A coherence test is performed with the aim of identifying reliable DOA estimates and increasing the robustness against reverberation. In addition, robustness is further improved by using the STFT magnitude of both input channels to

construct an amplitude-weighted histogram $Q(n)$. This histogram is the input for the multi-level thresholding algorithm.

3. *Multi-level thresholding.* The histogram obtained in the previous step is processed to calculate a set of thresholds, Th_ϵ, that maximizes the interclass variance of the time-frequency points according to their estimated DOA. These thresholds are used to segment the DOA map into the final separation masks, $M_j(k,m)$, with $j = 1,\ldots,N$. The separation masks are directly applied to the input signals $X_i(k,m)$ to estimate the different sources.

Note that a complete description of the separation method is out of the scope of this paper. The interested reader is referred to Cobos, et. al.[86] for a detailed description of each processing stage.

8.4 ASR Training for Reverberant and Simultaneous Speech

The dominant technology in ASR is the HMM [87]. HMMs are based upon a statistical state-sequence known as a Markov chain, consisting of a set of states with transitions between the states characterized by a given probability. HMMs are composed of a non-observable "hidden" Markov chain, and an observation process which links acoustic vectors extracted from the speech signal to the states of the hidden chain. These acoustic vectors are usually based on the calculation of *Mel-Frequency Cepstral Coefficients* (MFCCs) [88]. In this work, several models are constructed by using different training data. Speech signals considering different acoustic environments are artificially simulated to build specific training sets that improve ASR robustness with reverberant speech and simultaneous talking. The use of synthetic data reduces the enormous effort involved in collecting a complete set of training and test data for each environment and has been shown to be very useful in ASR design [29; 28].

8.4.1 Recognizer Specifications and Baseline System

All the models used in this paper are created using a carefully selected subset of the widely used TIDigits database [89]. This subset, provided by the Institute for Signal and Information Processing (ISIP) [90], consists of 941 files used for training and 336 files used for evaluating the system. All the speech signals are normalized so that the average power of each digit is equal across all digit strings. Each sentence has between one and seven digits. Feature vectors are calculated from the speech signal sampled at 8 kHz. The signal is decomposed into frames of length 25 ms with a frame shift of 10 ms. The frames are transformed to the frequency domain using a Hamming window after performing a first order pre-emphasis filtering with a coefficient of 0.97. The analysis filter-bank has 26 channels, from which 12 MFCC coefficients are computed as an output. A 3-state left-to-right HMM has been trained for each of the 11 digits ('0'-'9' and 'oh') and two silence models (short pause and long pause) with one and three states, respectively. The output densities are single Gaussians with diagonal covariance matrix.

8.4.2 Training Data-sets

The baseline system above is modified to take into account different acoustic conditions, resulting in three different training data-sets. Room effects and separation artifacts are both considered to reduce the mismatch between the training and test conditions.

Anechoic training

This training data set is the one corresponding to the baseline system without alterations. Thus, the original TIDigits files are used to build the HMM recognizer. These files were originally collected in an acoustically treated room in a close-talking situation. As opposed to distant-talking ASR, room effects are negligible and the environment can be considered as anechoic [91].

Reverberant training

In this case, the anechoic training data-set is modified to incorporate room reverberation effects. It has been shown that training data-sets using speech

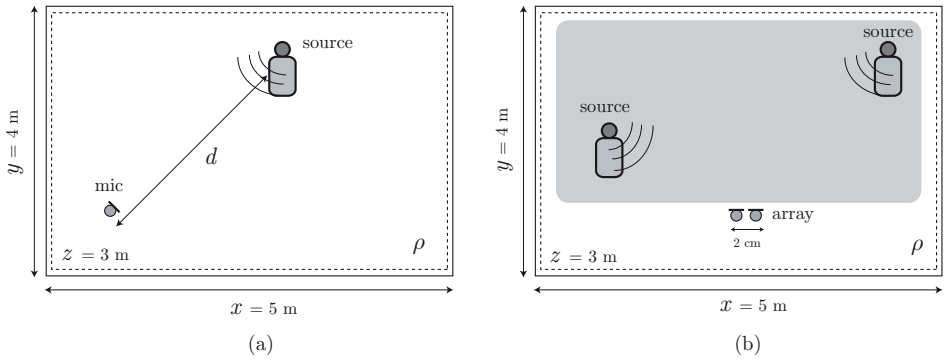


Figure 8.2. Simulation set-up for (a) reverberant training and (b) cocktail-party training. In (a), the training data-set is obtained by means of simulations with only one speaker. In (b), mixtures of two speakers are simulated and separated to form the cocktail-party training data-set. In both training methods, the wall reflection factor ρ and the position of the speakers are randomly changed during the simulations to account for different acoustic conditions. Note that in (b), two microphones are needed to perform the source separation task.

convolved with artificial reverberation helps to improve ASR robustness [92]. Obviously, the optimal training set would be the one matching the test acoustic conditions. However, real-world ASR systems might work under many different environments and having a different training data-set for every possible situation is not very practical. Therefore, a general reverberant training-set is here proposed, where random room conditions are simulated via the image-source model [64]. A shoe-box-shaped room ($x = 5$ m, $y = 4$ m, $z = 3$ m) was considered. The random simulation parameters were the wall reflection coefficient ($\rho \in [0, 1)$) and the source-to-microphone distance ($d \in [0.05, 3]$ m). These parameters affect room reverberation and direct-to-reverberant ratio, which are two important factors affecting ASR performance. The use of the wall reflection factor has been here selected for simulation convenience, however, this factor can be easily related to the room reverberation time by means of Sabine's empirical equation [25]. Figure 8.2(a) shows the simulation set-up to build the reverberant training data-set.

Cocktail-party training

Similarly to the reverberant training previously described, a modified data-set that accounts for source separation artifacts is considered. Although only one separation algorithm is used in this work, time-frequency masking algorithms produce very similar artifacts, which makes the selected approach significantly representative. For this training data-set, pairs of simultaneous speakers are simulated within random acoustic environments. Two-sources were only considered since double-talk situations are the most usual in real application environments. Source separation is applied as explained in Section 8.3.1, performing the STFT analysis with a frame length of 64 ms and 50% overlap. A small two-microphone array is used to separate the two speech signals generated by the sources, which are randomly positioned in the simulation, covering different source-to-microphone distances and DOAs. The inter-microphone distance is 2 cm, which is sufficient to avoid spatial aliasing in the separation process [86]. Figure 8.2(b) shows the simulation set-up for this case. The shaded area reflects possible source positions in the simulated data. The simulated source positions are uniformly distributed within this area. Note that two usual approaches for robust ASR are jointly used here: source separation as an speech enhancement method that suppresses interference and a robust transformed model constructed from separated speech in real environments.

8.5 Experiments

In this section, a set of experiments are carried out to compare the training data-sets previously described in terms of ASR *Word Recognition Rate* (WRR). To this end, the well-known HTK toolkit is employed [42]. Different environments are simulated by means of the image-source model, as already explained in Section 8.4.2. In all cases, a distant talking set-up is considered ($d = 2$ m) to ensure an adverse working condition, while the room reflection coefficient ρ is successively increased to test different degrees of reverberation. The following experiments evaluate ASR performance in three different situations: single speech recognition with and

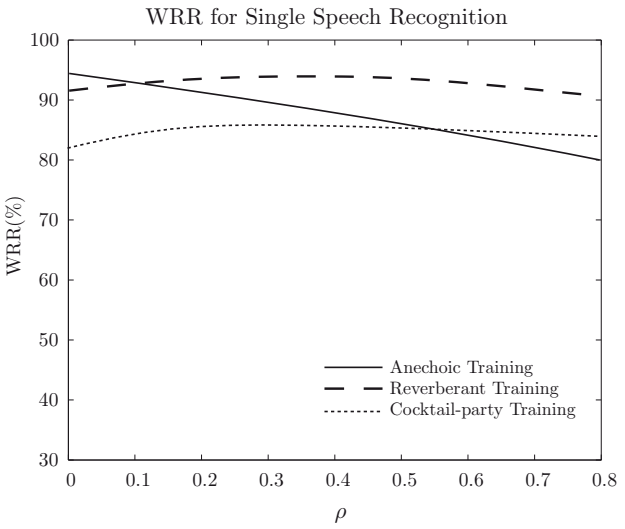


Figure 8.3. Word Recognition Rate (WRR) for single speech recognition as a function of the wall reflection factor (ρ).

without interfering speech signals and simultaneous speech recognition.

8.5.1 Experiment 1: Single Speech Recognition

This experiment evaluates the proposed training data-sets in terms of WRR with different conditions of reverberation. Only one speaker is considered for the test, thus, source separation is not applied here. Figure 8.3 shows the percentage of recognized words for all the training cases exposed in Section 8.4.2 as a function of the room wall reflection factor ρ. As expected, the anechoic training outperforms the other two systems in case of $\rho = 0$, since no reflections occur inside the room. However, as soon as reflections appear ($\rho \neq 0$) the performance is severely degraded, exhibiting a linear decay. On the other hand, the reverberant training data-set performs quite robustly under different reverberation degrees, showing a highly stable performance for all ρ values. A similar behavior is observed for the cocktail-party training, but having a negative offset that reflects the fact that separation artifacts are not present in the test data-set.

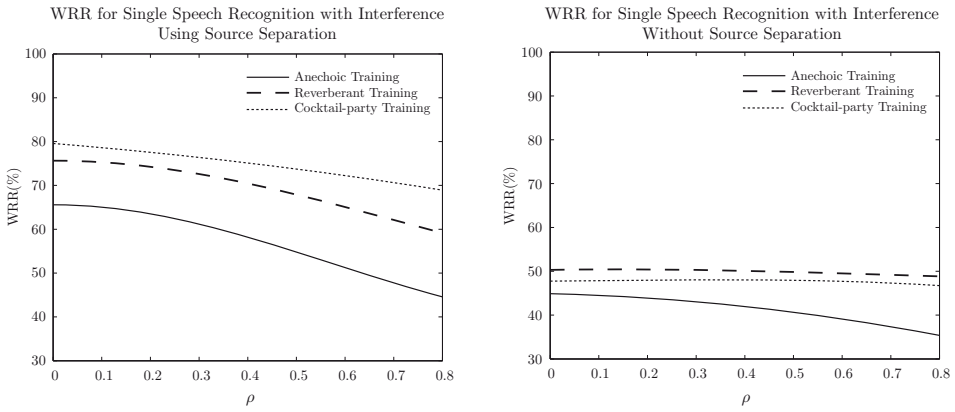


Figure 8.4. Word Recognition Rate (WRR) for speech recognition with interfering speech as a function of the wall reflection factor (ρ). (a) Using source separation. (b) Without source separation.

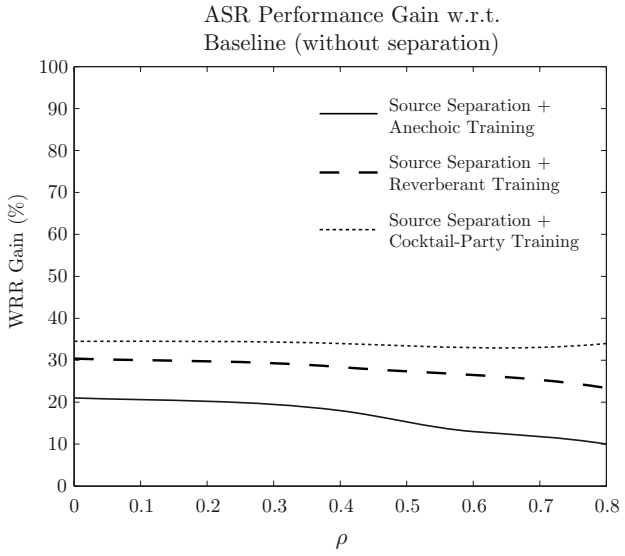


Figure 8.5. Performance Gain in terms of WRR with respect to to the baseline system (without separation).

8.5.2 Experiment 2: Speech Recognition With Interfering Speech

In this case, ASR performance is evaluated when there are two simultaneous speakers but we are only interested in recognizing the words corresponding to one of them. Source separation is applied to segregate the target speech source from the mixture, performing ASR with different training data afterwards. Figure 8.4(a) shows the WRR for the three training data-sets. Obviously, the WRR is lower than in the previous experiment, since the cocktail-party scenario is a more challenging environment. Despite the performance is lower than in the case of Experiment 1 (Figure 8.3), source separation allows for a high ASR improvement. To illustrate this fact, Figure 8.4(b) presents the results over the same test data without using any source separation stage. Note that, for all the training data-sets, the performance is significantly better when using source separation than when source separation is not considered (see Figures 8.4(a)-(b)). If source separation is applied, the best performance is achieved by the cocktail-party training. This is due to the reduced mismatch between the training data and the separated data used in the test. Note also that the difference between the reverberant and cocktail-party training in Figure 8.4(a) becomes greater for higher wall reflection coefficients. This can be explained by the greater amount of artifacts that appear after source separation when the environment is highly reverberant.

Figure 8.5 shows the performance gain in terms of WRR obtained by the proposed training methods in a simultaneous talking situation. The gain is computed with respect to the performance of the conventional baseline system without separation. Thus, it is computed as the difference between Figure 8.4(a) and the anechoic training line in Figure 8.4(b). Note that the proposed cocktail-party training outperforms the other methods, providing a nearly constant gain of 35%.

8.5.3 Experiment 3: Simultaneous Speech Recognition

In this last experiment, the test data is the same as in Experiment 2, but the WRR accounts for errors in both speech signals. This case is interesting to recognize commands from different sources simultaneously. Both signals are targets and interferences at the same time, although the recognition tasks are independently performed for each target. Figure 8.6 shows the

ASR performance for this experiment. The curves are very similar to those obtained in Experiment 2, but having a negative offset due to the errors corresponding to both sources. Nevertheless, the proposed cocktail-party training still performs better than the others.

8.6 Discussion

The experiments conducted have shown that both source separation and specific training provide a considerable improvement in ASR performance. The reduced mismatch between training and test data allows the design of ASR systems with increased robustness. The anechoic training data-set, which represents the conventional approach to ASR design, has shown to be optimum only for an ideal scenario. In real situations with reverberant rooms and a single speaker, a reverberant training seems to be more appropriate, especially under highly adverse conditions. An approximate improvement of 10% in WRR is achieved with respect to the baseline (anechoic) system for $\rho = 0.8$ (see in Figure 8.3). On the other hand, when multiple speech sources are present, source separation has been shown to be strictly necessary, both for simultaneous speech recognition and/or interference rejection. The proposed cocktail-party training accounts for multiple degradations: reverberation, artifacts and residual noise. This training data-set has been shown to achieve a higher recognition rate whenever multiple speech sources are present and ASR follows a source separation stage. As shown in Figure 8.5, the major improvement occurs at higher reverberant conditions, providing a 35% higher WRR with respect to the baseline system without separation. It is interesting here to remark that source separation and ASR must be both performed in real-time during the test stage, which makes our proposed system a practical solution for adverse ASR.

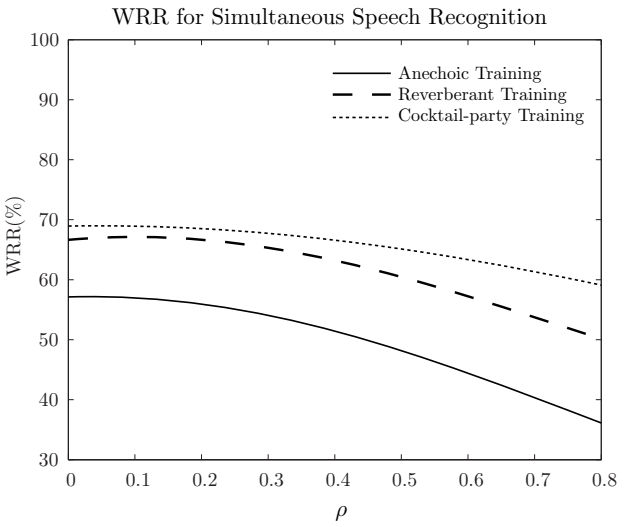


Figure 8.6. Word Recognition Rate (WRR) for simultaneous speech recognition as a function of the wall reflection factor (ρ).

8.7 Conclusion

Cocktail-party situations where different speakers are talking at the same time pose a real problem for ASR systems. In this paper, a framework for robust ASR in cocktail-party situations has been presented. This framework is based on a robust transformed model constructed from separated speech in diverse acoustic environments. Thus, a source separation method has been used as a speech enhancement stage that suppresses interferences. The validity of the method has been studied over a meaningful set of experiments, evaluating ASR performance for three different training data-sets. The results have shown that both source separation and specific training provide a considerable improvement in WRR (up to a 35%), reducing the existing mismatch between the training and test data. Moreover, the proposed framework allows to perform both ASR and source separation in real-time, which is a very important feature for practical systems. Future work will be focused on developing efficient double-talk detection methods for real-time ASR model selection.

8.8 References

The references of this paper have been consolidated in the general bibliography at the end of the book.

Evaluating the Influence of Source Separation Methods in Robust Automatic Speech Recognition with a Specific Cocktail-Party Training

9

A. Marti, M. Cobos and J. J. Lopez
Proceedings of the 132th AES Convention, London, UK, 2012.

Abstract

Automatic Speech Recognition (ASR) allows a computer to identify the words that a person speaks into a microphone and convert it to written text. One of the most challenging situations for ASR is the cocktail-party environment. Although source separation methods have already been investigated to deal with this problem, the separation process is not perfect and the resulting artifacts pose an additional problem to ASR performance in case of using separation methods based on time-frequency masks. Recently, the authors proposed a specific training method to deal with simultaneous speech situations in practical ASR systems. In this paper, we study how the speech recognition performance is affected by selecting different combinations of separation algorithms both at the training and test stages of the ASR system under different acoustic conditions. The results

show that, while different separation methods produce different types of artifacts, the overall performance of the method is always increased when using any cocktail-party training.

9.1 Introduction

Automatic Speech Recognition (ASR) allows a computer or electronic device to identify the words spoken by a person so that the message can be stored or processed in a useful way [73]. Although ASR technology is not yet at the point where machines robustly understand speech in any acoustic environment, or by any person, it is used on a day-to-day basis in a number of applications and services such as natural human-machine interfaces, dictation systems, electronic translators and automatic information desks [5]. A major problem in ASR systems is to recognize people speaking in a room by using a distant microphone [24]. In distant-speech recognition, the microphone does not only receive the direct path signal, but also delayed replicas resulting from multi-path propagation [25]. The existing mismatch between the training and testing conditions limits the performance of ASR systems, thus, robust recognition methods are aimed at reducing this mismatch. In this context, several approaches have been proposed to cope with room reverberation in speech recognition applications. Some methods are based on a speech enhancement stage prior to recognition [26]. In other methods, the recognizer itself is made robust to reverberation by using model compensation or by performing an improved feature extraction process [27]. All these approaches have shown to be useful in improving ASR. In any case, reducing the acoustic mismatch between training and testing conditions seems to be a relevant issue for developing robust ASR systems.

Many efforts have been made to develop robust ASR systems working in reverberant and noisy conditions, most of them focused on recognizing a single speech source. However, besides noise and reverberation, cocktail-party situations where different speakers are talking at the same time pose a real problem for ASR systems [30; 31]. Source separation refers to the

task of estimating and recovering independent source signals (for example, speech signals) from a set of mixtures in one or several observation channels (microphone signals) [45]. Source separation algorithms have been described in the literature as a solution for simultaneous speech recognition [56], proposing ASR performance as an indicator of the quality achieved by a given source separation algorithm [75; 76]. However, separating speech signals in real acoustic environments is not an easy task and the extracted speech signals are usually corrupted by audible artifacts [77].

Recently, the authors proposed a specific training for ASR in multiple-talk situations [93]. This training includes both room effects and the particular artifacts resulting from the separation process. The aim of this specific training was to increase the robustness of ASR systems when a source separation stage is performed prior to word recognition. In that work only a time-frequency masking based separation method was used.

In this paper, we evaluate how the recognition performance is affected by considering the combination of different separation approaches both at the training and test stages of the ASR system. While masking-based source separation methods provide a high isolation of the target source from simultaneous interfering sources, i.e. they have a high Signal to Interference Ratio (SIR), other separation methods present less artifacts, having a higher Signal to Artifact Ratio (SAR) than masking-based approaches. Thus, the aim of this work is to compare two types of separation methods in different acoustic conditions to examine how the speech recognition performance achieved with the proposed training is affected by the choice of the separation method.

The paper is structured as follows. Section 9.2 presents the problems of ASR in cocktail-party situations and describes different proposed training data-sets. Section 9.3 presents some fundamentals of source separation. Experiments are discussed in Section 9.4. Finally, the conclusions of this work are summarized in Section 9.5.

9.2 Speech Recognition in Cocktail-Party Situations

Humans have the ability to focus their listening attention on a single talker among a mixture of conversations and background noises, ignoring other conversations and recognizing a specific voice. This situation is usually referred to as the "cocktail party effect" [80]. Speech recognition technology has significantly improved in the last decade, however cocktail-party situations are still a challenge.

Recently, the authors proposed a specific training which improves considerably the percentage of word recognition rate when different people are talking simultaneously. This cocktail-party training, incorporates room reverberation effects and source separation artifacts. In [92], it was shown that training data-sets using speech convolved with artificial reverberation helps to improve ASR robustness. Obviously, the optimal training set would be the one matching the test acoustic conditions. However, real-world ASR systems might work under many different environments and having a different training data-set for every possible situation is not very practical. Therefore, in the proposed training, random room conditions are simulated via the image-source model and source separation artifacts are considered. Moreover, since the training data consists of blindly separated sources in these simulated environments, source separation artifacts are additionally considered in the training stage. For that specific training only one separation algorithm was used. Thus in this paper, we compare the results of applying a different source separation method, both in the training and test data-sets.

The different training data sets used are [93]:

- **Anechoic training:** This training data-set is composed of clear speech data from the TIDigits database [89] without alterations. These files were originally collected in an acoustically treated room in a close-talking situation. As opposed to distant-talking ASR, room effects are negligible and the environment can be considered as anechoic.
- **Reverberant training:** In this case, the anechoic training data-set

is modified to incorporate room reverberation effects. A shoe-box-shaped room (x = 5 m, y = 4 m, z = 3 m) was considered. The random simulation parameters were the wall reflection coefficient ($\rho \in [0, 1]$) and the source-to-microphone distance (d $\in [0.05, 3]$ m). These parameters affect room reverberation and direct-to-reverberant ratio, which are two important factors affecting ASR performance.

- **Cocktail-party training:** Similarly to the reverberant training, a modified data-set that accounts for source separation artifacts is considered. For this training data-set, pairs of simultaneous speakers are simulated within random acoustic environments. Two sources were only considered since double-talk situations are the most usual in real application environments. In this work two separation algorithms are used and compared. Figure 9.1 shows the simulation set-up for this case. The shaded area reflects possible source positions in the simulated room. The simulated source positions are uniformly distributed within this area.

As we compare two different source separation algorithms for the cocktail-party training we consider two different test data sets: the TIDigits files modified with random reverberation and separation artifacts from both separation techniques, which are briefly described in the next section.

9.3 Sound Source Separation

Speech source separation algorithms are aimed at estimating source speech signals from a set of observed mixtures. Single channel speech separation methods are not able to exploit spatial information about the sources, thus, other features such as harmonicity and spectral modulation are employed [94]. On the other hand, multi-channel methods are usually based on spatial localization cues, which provide very useful information within many separation frameworks [52].

According to the number of sources and the number of available mixture channels, the source separation problem can be classified as overdeter-

mined (more sensors than sources), determined (equal number of sensors and sources) or underdetermined (more sources than sensors). Independent Component Analysis (ICA) [81] methods have been widely used for the separation of determined problems, both with time-domain [95] and frequency-domain approaches [96]. However, ICA methods can not be employed for the separation of underdetermined mixtures, and other methods based on sparsity assumptions are used [45]. Time-frequency (T-F) masking methods [49] make use of this sparseness property by assuming that the energies of the independent source signals rarely overlap in the T-F domain. Moreover, the source separation problem becomes even more difficult in reverberant environments, leading to convolutive source separation approaches [49].

In our previous work [93], the authors studied the effect of source separation in ASR training by using a T-F masking separation method based on the maximization of the inter-class variance found in the distribution of Direction-Of-Arrival (DOA) estimates. This method was shown to provide a good separation performance in real-time, thus, showing that it could be easily integrated within an advanced ASR system. In this paper, we also consider another recent method for source separation in reverberant environments which is based on a full-rank spatial convariance model. In the next subsections, we briefly described both separation algorithms.

9.3.1 Multi-Level Thresholding Separation

The method is able to separate N source signals by using only two microphones, thus it can be applied to underdetermined mixtures. It basically consists of three stages. First, the input signals are transformed to the T-F domain by means of Short-Ti me Fourier Transform (STFT) processing, and a DOA estimate is obtained at each T-F point by analyzing the phase difference existing between the two microphone signals. Second, a coherence test is performed to discard unreliable DOA estimates and the distribution of the selected ones is analyzed by means of an amplitude-weighted histogram. Finally, a number of different classes (sources) is assumed and the histogram is divided by a set of thresholds that are calculated by maximizing the interclass variance. These thresholds define the

binary masks that are used to separate the mixture spectrograms into the separated speech sources.

9.3.2 Separation with Full-Rank Spatial Convariance Models

Duong et. al presented in [82] a separation algorithm for underdetermined reverberant mixtures based on a full-rank spatial covariance model. The STFT coefficients of the source images are modeled as a zero-mean Gaussian random variable with a factored covariance matrix depending on a spatial covariance matrix that encodes their spatial position and spread. Under this model, source separation is performed in two steps. first, the parameters of the model are estimated in the Maximum Likelihood (ML) sense and then, the source images are obtained by means of multichannel Wiener filtering.

The reader is referred to [86] and [82] for further details on both separation methods.

9.4 Experiments

In this section, a set of experiments are carried out to compare different source separation methods in terms of ASR *Word Recognition Rate* (WRR). To this end, the well-known HTK toolkit is employed [42].

The training and test data sets are obtained using a carefully selected subset of the widely used TIDigits database. This subset, provided by the Institute for Signal and Information Processing (ISIP) [90], consists of 941 files used for training and 336 files used for evaluating the system. All the speech signals are normalized so that the average power of each digit is equal across all digit strings. Each sentence has between one and seven digits. Feature vectors are calculated from the speech signal sampled at 8 kHz. The signal is decomposed into frames of length 25 ms with a frame shift of 10 ms. The frames are transformed to the frequency domain using a Hamming window after performing a first order pre-emphasis filtering with a coefficient of 0.97. The analysis filter-bank has 26 channels, from which 12 *Mel-Frequency Cepstral Coefficients* (MFCC) [88] are computed

as an output. A 3-state left-to-right HMM [87] has been trained for each of the 11 digits (0-9 and oh) and two silence models (short pause and long pause) with one and three states, respectively. The output densities are single Gaussians with diagonal covariance matrix.

In this work we proposed two experiments. In each experiment a different sound source separation method has been employed for the test data-set. The experiments evaluate ASR performance when there are simultaneous speakers and the WRR accounts for errors in all speech signals (simultaneous speech recognition). This case is interesting to recognize commands from different sources simultaneously. Therefore, all signals are targets and interferences at the same time, although the recognition tasks are independently performed for each target.

9.4.1 Experiment 1: Test Data-Set Using Multi-Level Thresholding Separation

In this experiment the Multi-Level Thresholding Separation method (MuLeTS), explained in the Subsection 9.3.1, has been used to obtain the test data-set. Figure 9.2 shows the ASR performance for this experiment.

As observed in the case of the MuLeTS test data-set, the WRR for all the training data-sets, the performance is significantly better when using source separation than when source separation is not considered. If source separation is applied, the best performance is achieved by the cocktail-party training. The cocktail-party training which is using MuLeTS as the separation method has a WRR percentage a little bit higher than the other cocktail-party training. This is due to the reduced mismatch between the training data and the separated data used in the test.

9.4.2 Experiment 2: Test Data-Set Using Separation with Full-Rank Spatial Covariance Models

In this experiment the test data-set has been constructed by means of separated sources using the Full-Rank Spatial Covariance Model (FRSCM) separation method described in Subsection 9.3.2. Figure 9.3 shows the result of simultaneous speech recognition using this test data-set.

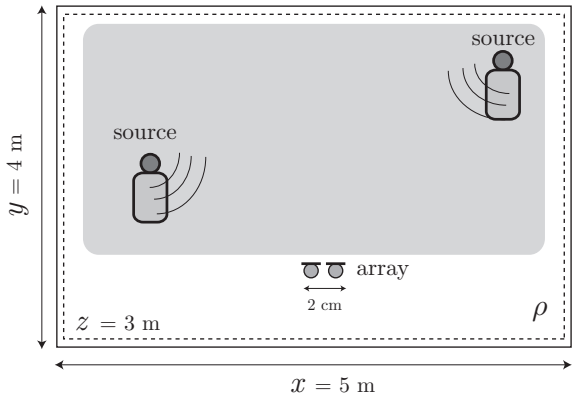


Figure 9.1. Simulation set-up for cocktail-party training. The wall reflection factor ρ and the position of the speakers are randomly changed during the simulations to account for different acoustic conditions. Two microphones are needed to perform the source separation task.

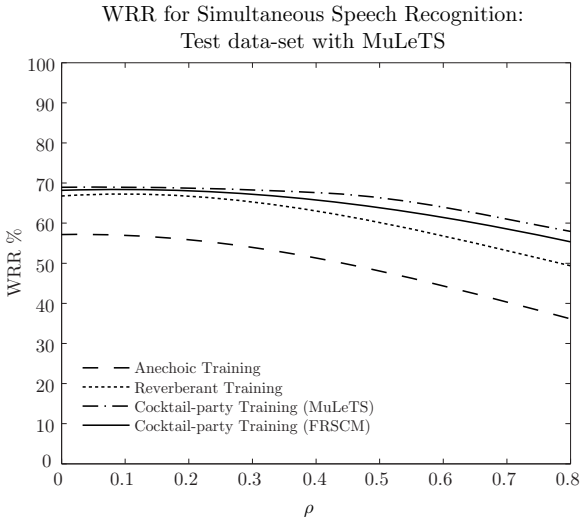


Figure 9.2. WRR for simultaneous speech recognition when using Multi-Level Thresholding Separation for the test data-set.

The curves are very similar to those obtained in Experiment 1. The performance is significantly better when using FRSCM separation which is the same sound source separation method as the test data-set.

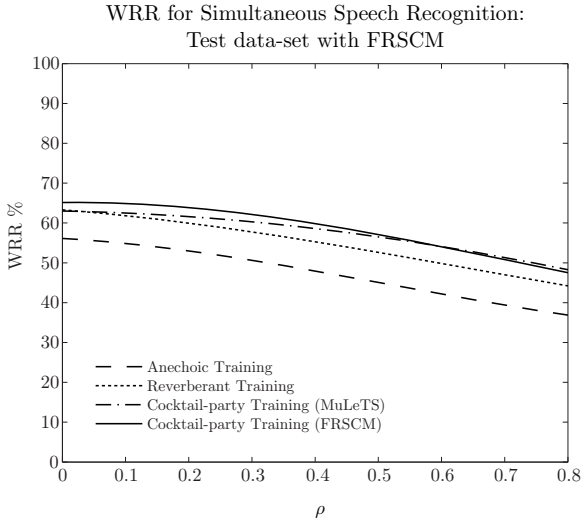


Figure 9.3. Word recognition rate (WRR) for simultaneous speech recognition when using separation with Full-Rank Spatial Covariance Models for the test data-set.

9.5 Conclusion

Cocktail-party situations where different speakers are talking at the same time pose a real problem for ASR systems. Recently, the authors proposed a framework for robust ASR in cocktail-party situations. That framework is based on a robust transformed model constructed from separated speech in diverse acoustic environments. In this paper two source separation methods have been used as a speech enhancement stage that suppresses interferences.

In this work it has been studied the effect of two different sound source separation methods in the specific training for ASR. The results have shown that both specific trainings provide a considerable improvement in WRR even when the source separation method employed at the test data and

training data sets are different. As the artifacts introduced by one type of separation method (Subsection 9.3.1) are different from the artifacts introduced by the other sound source separation method (Subsection 9.3.2) studied at this work, the percentage of WRR is always higher when the mismatch between training and test data is lower. So, as expected, each training has a best performance when the mixtures of the test data-set has been separated with the same method as in the training data-set. However, the results (in terms of WRR) for both trainings with both types of test data are very similar between them. As a result, we can conclude that having a specific training for simultaneous speech ASR improves significantly recognition performance, even if the separation method is not exactly the same. Nevertheless, it should be emphasized that this work has only considered two separation methods. Thus, further work is needed to understand better the limitations arising from the mismatch between the separation method used in the training and test ASR stages.

9.6 References

The references of this paper have been consolidated in the general bibliography at the end of the book.

Conclusions and Future Research 10

This chapter summarizes the findings of this research work, revisiting the research objectives given in the introductory chapter and proposing guidelines for future research lines.

10.1 Summary and Conclusions

This thesis focused on the field of acoustic signal processing and its applications to emerging communication environments. In this context, Sound Source Localization (SSL) and Automatic Speech Recognition (ASR) have been specially addressed in this thesis. Most real-world microphone array applications require the localization of one or more active sound sources in adverse environments. Indeed, performing robust SSL under high noise and reverberation is a very challenging task. To solve this problem, one of the most well-known algorithms for source localization in noisy and reverberant environments is the SRP-PHAT algorithm, which constitutes the baseline framework for many of the contributions developed throughout this thesis.

One of the objectives of this research work was to design accurate SSL

algorithms working in real-time time using a reasonable number of microphones. To address this issue, several modifications have been proposed to the SRP-PHAT algorithm, improving its performance and applicability. Specifically, Chapter 4 presented an effective strategy that extends the conventional SRP-PHAT functional. This approach performs a full exploration of the sampled space rather than computing the SRP at discrete spatial positions, increasing its robustness and allowing for a coarser spatial grid that reduces the computational cost required in a practical implementation with a reduced number of microphones and consequently with an important reduction in hardware cost. The modified SRP-PHAT was further improved in Chapter 5 by proposing an iterative method. To this end, the volume having the highest functional value over an initial coarse spatial grid is iteratively decomposed into smaller sub-volumes. The experiments simulating diverse acoustic conditions have shown that this iterative method provides almost the same accuracy as the fine-grid search with a substantial reduction in the number of required functional evaluations. Additionally, it has been demonstrated that, by using the modified SRP-PHAT, it is possible to implement real-time applications based on location information, such as the presented automatic camera steering or the detection of speech/non-speech fragments in advanced videoconferencing systems (Chapter 6 and Chapter 7). This application has been successfully presented in different conferences. Moreover, a videoconferencing system using this technology has been installed in a demonstration room located in the headquarters of the *Telefonica* company.

As commented before, besides the contributions related to SSL, this thesis is also related to the field of ASR. ASR is used on a day-to-day basis in a number of application and services such as natural human-machine interfaces, dictation systems, electronic translators and mobile phones. However, there are still some challenges to be solved. A major problem in ASR is to recognize people speaking in a room by using distant microphones. Echoes and reverberation of the room and multiple speakers talking simultaneously are very well know problems. In this context, when multiple speaker signals are present, Sound Source Separation (SSS) methods can be successfully employed to improve ASR performance in multi-source scenar-

ios. In Chapter 8 of this thesis, we developed a successful training method for multiple talk situations. This training, which is based on a robust transformed model constructed from separated speech in diverse acoustic environments, makes use of a SSS method as a speech enhancement stage that suppresses the unwanted interferences. The combination of source separation and this specific training has been explored and evaluated under different acoustical conditions, leading to improvements of up to a 35% in ASR performance. In addition, the effect of using different SSS methods in the proposed training was also explored in Chapter 9, showing that the training should match the separation algorithm to account for possible distortion artifacts.

The main **contributions** of this thesis can be highlighted as follows:

- Based on the well known SRP-PHAT SSL method, a modified version that uses a new functional has been developed. The results showed that the proposed approach provides similar performance to the conventional SRP-PHAT algorithm in difficult environments with a reduction of five orders of magnitude in the required number of functional evaluations. Also, an iterative method based on the modified SRP-PHAT algorithm has been developed, achieving the same accuracy as the conventional SRP-PHAT using a fine-grid search.

- All the approaches proposed in this thesis have been evaluated with a 6 microphone array, reducing considerably the computational and economical cost of the experiments. Using only 6 microphones has been shown to be sufficient to develop real-time source localization applications. This is an advantage compared to other approaches found in the literature.

- The SSL methods presented in this thesis have been evaluated in different environments, demonstrating its robustness in real and adverse acoustic situations.

- SSL has been used as a stage prior to speech enhancement, suppressing the time delay between microphones and improving the signal-to-noise ratio. Also, the position estimates give a valuable information

to discriminate between speech and non-speech frames. As a result, it has been possible to develop a videoconferencing system with automatic camera steering. When non-speech frames are detected, the camera pointed to all the audience while, in case of speech detection, the camera pointed to the active source.

- In the field for ASR, a specific training has been introduced which has been demonstrated to achieve better results in terms of WRR. For this specific training, a source separation technique has been employed to let the recognizer perform better in case of double-talk conditions, adding cocktail-party functionality to current ASR systems.
- The specific training developed for ASR has been evaluated with different SSS algorithms, demonstrating that the performance in case of using different SSS algorithms for training and testing, are still better than a baseline ASR system.

10.2 Further Work

From the conclusions of this work, some new and challenging research lines could be proposed, being some of them already open. Future work may follow the lines listed here:

- To evaluate SSL in much bigger rooms as large meeting rooms, conference halls and even concert halls. During the work we have detected that the larger the room the greater the number of microphones are needed to produce accurate and stable results. However, there are not published works related to the estimation of the appropriate number of microphones as a function of the room volume for a good behavior. It is our intention to continue this interesting line.
- To apply the developed SSL algorithms outdoors with the necessary modifications and to design new algorithms for this scenario. Nowadays, outdoor SSL is a hot research topic with many applications, which are mainly related to surveillance, but also to environment protection in cities and high-level acoustic monitoring.

- To combine sound-based localization systems with image-based systems into a multimodal localization system. In some situations, audio information is not sufficient to provide accurate and stable results (very high levels of noise and reverberation with multiple simultaneous sources). In these cases, visual information can be used to improve and complement the information provided by the microphone array, leading to better results. Some applications for ASR apply these techniques by means of lip reading. In the case of localization, the task should be easier, since just detecting lip movement (without reading it) would be enough.

- To try specific ASR training in the context of the new findings related to *Deep Neural Networks* (DNN) and to develop efficient double-talk detection methods for real-time ASR model selection.

Bibliography

[1] E. E. Perris and R. K. Clifton, "Reaching in the dark toward sound as a measure of auditory localization in infants," *Infant Behavior and Development*, vol. 11, no. 4, pp. 473 – 491, 1988. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0163638388900070

[2] M. Kaga, "Development of sound localization," *Acta Paediatr Jpn*, vol. 34, pp. 134–138, 1992.

[3] J. Blauert, *Spatial hearing: The psychophysics of human sound localization.* Cambridge, MA: MIT Press., 1983.

[4] W. Yost, *Fundamentals of hearing: an introduction.* San Diego: Academic Press., 1994.

[5] N. Hataoka, H. Kokubo, A. Lee, T. Kawahara, and K. Shikano, "Network and embedded applications of automatic speech recognition," *ECTI Transactions on Electrical Engineering, Electronics and Communications*, vol. 6, no. 2, pp. 1–8, 2008.

[6] E. Ettinger and Y. Freund, "Coordinate-free calibration of an acoustically driven camera pointing system," in *Second ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC 2008)*, Stanford, CA, 2008.

[7] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '97)*, Washington, DC,, 1997.

[8] A. Y. Nakano, S. Nakagawa, and K. Yamamoto, "Automatic estimation of position and orientation of an acoustic source by a microphone array network," *J. Acoust. Soc.*, vol. 126, pp. 3084–3094, 2009.

[9] J. G. Desloge, W. M. Rabinowitz, and P. M. Zurek, "Microphone-array hearing aids with binaural output - part i: Fixed-processing systems," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 6, pp. 529–542, 1997.

[10] D. Zotkin, R. Duraiswami, L. Davis, and I. Haritaoglu, "An audio-video front-end for multimedia applications," in *Systems, Man, and Cybernetics, 2000 IEEE International Conference on*, vol. 2, 2000, pp. 786–791 vol.2.

[11] J.-M. Valin, F. Michaud, J. Rouat, and D. Letourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Intelligent Robots and Systems, 2003. (IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on*, vol. 2, 2003, pp. 1228–1233 vol.2.

[12] T. Räty, "Survey on contemporary remote surveillance systems for public safety," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 40, no. 5, pp. 493–515, 2010.

[13] P. Aarabi, "The fusion of distributed microphone arrays for sound localization," *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 338–347, 2003.

[14] J. C. Makous and J. C. Middlebrooks, "Two-dimensional sound localization by human listeners," *The Journal of the Acoustical Society of America*, vol. 87, no. 5, pp. 2188–2200, 1990. [Online]. Available: http://link.aip.org/link/?JAS/87/2188/1

[15] H. Schau and A. Robinson, "Passive source localization employing intersecting spherical surfaces from time-of-arrival differences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 35, no. 8, pp. 1223–1225, 1987.

[16] N. Madhu and R. Martin, *Advances in Digital Speech Transmission*. Wiley, 2008, ch. Acoustic source localization with microphone arrays, pp. 135–166.

[17] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–19, 2006.

[18] F. Talantzis and L. C. Constatntinides, A. G. adn Polymenakos, "Estimation of direction of arrival using information theory," *IEEE Signal Processing*, vol. 12, no. 8, pp. 561–564, 2005.

[19] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. Ward, Eds. Springer-Verlag, 2001.

[20] H. F. Silverman, Y. Yu, J. M. Sachar, and W. R. Patterson III, "Performance of real-time source-location estimators for a large-aperture microphone array," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 593–606, 2005.

[21] J. H. DiBiase, "A high accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, Providence, RI, May 2000.

[22] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, 2007.

[23] H. Do and H. F. Silverman, "A fast microphone array SRP-PHAT source location implementation using coarse-to-fine region contraction (CFRC)," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2007)*, 2007.

[24] M. Woelfel and J. McDonough, *Distant Speech Recognition*. Hoboken, New Jersey, USA: John Wiley & Sons, 2009, 594 pages.

[25] H. Kuttruff, *Room acoustics*, S. Press, Ed. Abingdon, Oxford, UK: Taylor & Francis, October 2000, 368 pages.

[26] S. Nakamura and K. Shikano, "Room acoustics and reverberation: Impact on hands-free recognition," in *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH)*, vol. 5, Rhodes, Greece, 1997, pp. 2419–2422.

[27] T. Takiguchi and M. Nishimura, "Acoustic model adaptation using first-order linear prediction for reverberant speech," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, Quebec, Canada, 2004, pp. 869–872.

[28] T. Haderlein, E. Nth, W. Herbordt, W. Kellermann, and H. Niemann, "Using artificially reverberated training data in distant-talking ASR," in *Proceedings of the 8th International Conference on Text, Speech and Dialogue (TSD)*, Karlovy Vary, Czech Republic, 2005, pp. 226–229.

[29] D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Training of HMM with filtered speech material for hands-free recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Phoenix, Arizona, USA, 1999, pp. 449–452.

[30] L. Di Persia, M. Yanagida, H. L. Runer, and D. Milone, "Objective quality evaluation in blind source separation for speech recognition in a real room," *Signal Processing*, vol. 87, no. 8, pp. 1951–1965, 2007.

[31] L. Di Persia, D. Milone, H. L. Runer, and M. Yanagida, "Perceptual evaluation of blind source separation for robust speech recognition," *Signal Processing*, vol. 88, no. 10, pp. 2578–2583, 2008.

[32] M. L. Fowler and X. Hu, "Signal models for tdoa/fdoa estimation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 44, no. 4, pp. 1543–1550, 2008.

[33] A. K. Tellakula, "Acoustic source localization using time delay estimation," Ph.D. dissertation, Indiand Institute od Science, August 2007.

[34] P. Stoica and J. Li, "Source localization from range-difference measurements," *IEEE Signal Processing Magazine*, pp. 63–69, November 2006.

[35] P. Svaizer, M. Matassoni, and M. Omologo, "Acoustic source location in a three-dimensional space cross-power spectrum phase," *Proc. IEEE Int. Conf. Acoust., Speech Signal Processing*, vol. ICASSP-97, pp. 231–234, Munich, Germany, April 1997.

[36] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-24, pp. 320–327, 1976.

[37] B. Mungamuru and P. Aarabi, "Enhanced sound localization," *IEEE Trans Syst, Man, Cybernet Part B: Cybernet 2004;34(3):152640.*

[38] D. Cha Zhang, Florencio, and Z. Zhengyou, "Why does PHAT work well in low noise, reverberant environments." Las Vegas, NV: ICASSP, March 31 2008April 4 2008, pp. 2565–8.

[39] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques.*, N. J. . Englewood Cliffs, Ed. P T R Prentice Hall, 1993.

[40] S. Tervo and T. Lokki, "Interpolation methods for the srp-phat algorithm," *The 11th International Workshop on Acoustic Echo and Noise Control, Seattle, Washington, USA*, vol. IWAENC2008, pp. 14–17, September 2008.

[41] A. Ramamurthy, H. Unnikrishnan, and K. Donohue, "Experimental performance analysis of sound source detection with SRP PHAT-β," in *Southeastcon, 2009. SOUTHEASTCON '09. IEEE*, 2009, pp. 422–427.

[42] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK*

Version 3.2). Cambridge, UK: Cambridge University Engineering Department, 2002, 277 pages.

[43] L. R. Rabiner and B. H. Juang, *Fundamentasl of Speech Recognition*. Prentice-Hall, 1993.

[44] I. T. Jolliffe, *Principal Component Analysis*. New York: Springer, 2002.

[45] P. O'Grady, B. Pearlmutter, and S. Rickard, "Survey of sparse and non-sparse methods in source separation," *International Journal of Imaging Systems and Technology (IJIST)*, vol. 15, no. 1, pp. 18–33, 2005.

[46] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, pp. 978–982, 2006.

[47] M. R. DeWeese, M. Wehr, and A. M. Zador, "Binary spiking in auditory cortex," *Journal of Neuroscience*, vol. 23, pp. 7940–7949, 2003.

[48] J. Karvanen and A. Cichocki, "Measuring sparseness of noisy signals," in *Proceedings of the International Symposium on Independent Component Analysis and Blind Signal Separation (ICA)*, Nara, Japan, April 2003.

[49] D. Wang, "Time frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplification*, vol. 12, no. 4, pp. 332–353, 2008.

[50] J. J. Burred, "From sparse models to timbre learning: New methods for musical source separation," Ph.D. dissertation, Technical University of Berlin, 2008.

[51] J. J. Burred and T. Sikora, "On the use of auditory representations for sparsity-based sound source separation," in *Proceedings of the 5th International Conference on Information, Communications and Signal Processing (ICICS 2005)*, Bangkok, Thailand, December 2005.

[52] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[53] S. Rickard and O. Yilmaz, "On the w-disjoint orthogonality of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, May 2002, pp. 529–532.

[54] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 2005)*, Philadelphia, PA, USA, March 2005.

[55] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Blind sparse source separation with spatially smoothed time-frequency masking," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Paris, France, September 2006.

[56] M. Küne, R. Togneri, and S. Nordholm, *Speech Recognition, Technologies and Applications.* Vienna, Austria: I-Tech, 2008, ch. Time-Frequency Masking: Linking Blind Source Separation and Robust Speech Recognition, pp. 61–80.

[57] F. Talantzis, A. G. Constantinides, and L. C. Polymenakos, "Estimation of direction of arrival using information theory," *IEEE Signal Processing*, vol. 12, no. 8, pp. 561–564, 2005.

[58] E. A. P. Habets and P. C. W. Sommen, "Optimal microphone placement for source localization using time delay estimation," in *Proc. of the 13th Annual Workshop on Circuits, Systems and Signal Processing (ProRISC 2002)*, Veldhoven, Netherlands, 2002.

[59] D. R. Campbell, "Roomsim: a MATLAB simulation shoebox room acoustics," 2007, http://media.paisley.ac.uk/ campbell/Roomsim.

[60] A. Y. Nakano, S. Nakagawa, and K. Yamamoto, "Automatic estimation of position and orientation of an acoustic source by a microphone

array network," *J. Acoust. Soc. Am.*, vol. 126, no. 6, pp. 3084–3094, 2009.

[61] M. Durković, T. Habigt, M. Rothbucher, and D. K., "Low latency localization of multiple sound sources in reverberant environments," *J. Acoustc. Soc.*, vol. 130, no. 6, pp. 392–398, 2011.

[62] M. Cobos, A. Marti, and J. J. Lopez, "A modified srp-phat functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Processing Letters*, vol. 18, no. 1, January 2011.

[63] J. Dmochowski, J. Benesty, and S. Affes, "On spatial aliasing in microphone arrays," *IEEE Transactions on Signal Processing*, vol. 57, no. 4, pp. 1383–1395, 2009.

[64] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[65] J. Benesty, S. Makin, and J. Chen, *Speech Enhancement*. Springer-Verlag, 2005.

[66] A. Levi and H. Silverman, "An alternate approach to adaptive beamforming using srp-phat," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 2726–2729.

[67] D. Barry, B. Lawlor, and E. Coyle, "Sound source separation: Azimuth discrimination and resynthesis," in *7th Conference on Digital Audio Effects (DAFX 04)*, 2004.

[68] M. Cobos and J. J. Lopez, "Stereo audio source separation based on time-frequency masking and multilevel thresholding," *Digital Signal Processing*, vol. 18, no. 6, pp. 960–976, 2008.

[69] M. Cobos, A. Marti, and J. J. Lopez, "Localization of multiple speech sources using distributed microphones," in *Audio Engineering Society*

Convention 130, 5 2011. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=15794

[70] Y. Li and D. Wang, "On the optimality of ideal binary time-frequency masks," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 3501–3504.

[71] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Processing Letters*, vol. 18, pp. 71–74, 2011.

[72] B.-H. Juang and L. Rabiner, "Speech recognition, automatic: History," in *Encyclopedia of Language & Linguistics*, K. Brown, Ed. Oxford, UK: Elsevier, 2006, pp. 806 – 819.

[73] G. Philip, "Applications of automatic speech recognition and synthesis in libraries and information services: a future scenario," *Library Hi Tech*, vol. 9, no. 35, pp. 89–92, January 1991.

[74] J. Allen, "How do humans process and recognize speech?" *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 567 –577, oct 1994.

[75] N. Roman, S. Srinivasan, and D. Wang, "Binaural segregation in multisource reverberant environments," *Journal of the Acoustical Society of America*, vol. 120, pp. 4040–4051, 2006.

[76] M. Mandel, S. Bressler, B. Shinn-Cunningham, and D. Ellis, "Evaluating source separation algorithms with reverberant speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1872 –1883, sept. 2010.

[77] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[78] K.-C. Huang, "Robust speech recognition in noisy environments," Ph.D. dissertation, Department of Electrical Engineering, National Central University, Jhongli City, Taiwan, 2003.

[79] J. C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers." *The Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510–524, Jan. 1993.

[80] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[81] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley & Sons, 2001.

[82] N. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[83] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, *Springer Handbook of Speech Processing*. Berlin Heidelberg, Germany: Springer-Verlag, ch. Convolutive Blind Source Separation Methods, pp. 1065–1084.

[84] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, 2004.

[85] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[86] M. Cobos and J. J. Lopez, "Two-microphone separation of multiple speakers based on interclass variance maximization," *Journal of the Acoustical Society of America*, vol. 127, pp. 1661–1673, 2010.

[87] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257 –286, feb 1989.

[88] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, 1980.

[89] R. Leonard, "A database for speaker-independent digit recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 9, San Diego, California, USA, mar 1984, pp. 328 – 331.

[90] "Institute for signal and information processing (ISIP) webpage," date last viewed 08/10/11. [Online]. Available: http://www.isip.piconepress.com/projects/speech/index.html

[91] Q. Lin, C. Che, D.-S. Yuk, L. Jin, B. deVries, J. Pearson, and J. Flanagan, "Robust distant-talking speech recognition," in *IEEE International Confrence on Acoustics, Speech and Signal Processing (ICASSP 1996)*, vol. 1, Atlanta, Georgia, USA, 1996, pp. 21–24.

[92] L. Couvreur and C. Couvreur, "On the use of artificial reverberation for ASR in highly reverberant environments," in *Proceedings of the 2nd IEEE Benelux Signal Processing Symposium (SPS-2000)*, Hilvarenbeek, The Netherlands, 2000.

[93] A. Marti, M. Cobos, and J. J. Lopez, "Automatic speech recognition in cocktail-party situations: A specific training for separated speech," *The Journal of the Acoustical Society of America*, vol. 131, no. 2, pp. 1529–1535, 2012. [Online]. Available: http://link.aip.org/link/?JAS/131/1529/1

[94] R. M. Stern, G. J. Brown, and W. D., *Computational Auditory Scene Analysis*. Wiley Interscience, 2006, ch. Binaural Sound Localization, pp. 147–178.

[95] K. Matsuoka, "Minimal distortion principle for blind source separation," in *SICE 2002. Proceedings of the 41st SICE Annual Conference*, vol. 4, 2002, pp. 2138–2143 vol.4.

[96] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari, "Equivalence between frequency-domain blind source separation and frequency-domain adaptive beamforming for convolutive mixtures," *EURASIP J. Appl. Signal Process.*, vol. 2003, pp. 1157–1166, Jan. 2003. [Online]. Available: http://dx.doi.org/10.1155/S1110865703305074