BINAURAL LOCALIZATION AND SEPARATION TECHNIQUES

THÈSE N^{O} 3043 (2004)

PRÉSENTÉE À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS

Institut des systèmes de communication

SECTION DES SYSTÈMES DE COMMUNICATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Harald VISTE

Sivilingeniør degree, NTNU Trondheim, Norvège
et de nationalité norvégienne

acceptée sur proposition du jury:

Prof. M. Vetterli, Dr G. Evangelista, directeurs de thèse
Dr A. Drygajlo, rapporteur
Dr A. Härmä, rapporteur
Dr S. Launer, rapporteur

Lausanne, EPFL
2004

Binaural Localization and Separation Techniques

Harald Viste

July 16, 2004

Contents

Abstract

Based on binaural signals, i.e. the signals observed at the two ears, a listener can localize and recognize different sound sources and then focus on one of these. For decades, researchers have tried to invent a machine that can do the same under similar conditions. Despite all the efforts, the human auditory system is, by far, superior to any machine that has been devised. The topic of this thesis is computational techniques for the localization and separation of sources in binaural signals.

In order to give an overview of different areas of research that have considered the problems of source localization and separation, we start with a review of existing techniques. This provides the background for the techniques that we propose subsequently.

Binaural Localization The most important cues for localization of sound sources in binaural signals are the level and time differences between the ears. We propose a technique for the joint evaluation of these cues where noisy level difference estimates are combined with less noisy but ambiguous time difference estimates in order to provide accurate azimuth estimates. The proposed technique enables the localization of sources and the tracking of these in dynamic scenes.

Head model Based on a study of the level and time differences as function of azimuth angle for different heads, we propose a generic model that is parametrized by the distance between the ears only. This enables the use of the binaural localization technique mentioned above for a listener whose head related transfer functions have not been measured.

Binaural separation For the separation of sources we propose a method based on spatial windowing in the azimuth parameter space.

Separation of overlapping partials Finally, we propose a technique for the separation of overlapping partials in mixtures of harmonic instruments. The technique is based on the similarity of temporal envelopes between the different partials of a harmonic note.

Résumé

A partir des signaux observés avec les oreilles, une personne peut localiser et reconnaître différentes sources puis se concentrer sur l'une d'elle. Depuis des siècles les chercheurs ont essayé d'inventer une machine qui peut faire la même chose sous les mêmes conditions. Malgré tous leurs efforts, les machines sont toujours loin de la performance du système auditif humain. Le sujet de cette thèse concerne les techniques numeriques pour la localisation et la séparation de sources sonores dans les signaux observés avec nos deux oreilles.

Afin de donner une vue générale sur différents domaines de recherche qui traitent des problèmes de localisation et de séparation, nous donnons d'abord une présentation des différentes techniques existantes. Celles-ci forment la base des techniques que nous proposons dans les chapitres suivants.

Localisation Pour la localisation des sources, les différences de temps et de puissance entre les oreilles sont les informations les plus importantes. Afin d'estimer précisément les angles d'arrivé, nous proposons une technique pour leur évaluation commune. Cette technique permet de localiser des sources, et de les suivre dans les scènes dynamiques.

Modèles des différences de temps et de puissance entre les oreilles Après une étude de ces différences pour plusieurs sujets dans une base de données dont les HRTFs ont été mesurés, nous proposons un modèle générique qui permet la localisation des sources pour des personnes dont les HRTFs n'ont pas été mesurés.

Séparation Afin de séparer les sources nous présentons une technique de fenêtrage dans l'espace des paramètres spatiaux.

Harmoniques superposés Finalement, nous attaquons le problème de sources dont les énergies en temps-fréquence se superposent. Pour les instruments harmoniques nous proposons une technique pour la separation des harmoniques superposés qui se base sur la similitude des enveloppes temporelles.

Acknowledgements

This thesis work is the result of a wonderful stay in Lausanne. I am deeply grateful to my supervisors Gianpaolo Evangelista and Martin Vetterli for giving me the opportunity to come to Lausanne and the freedom to choose my own research.

I also want to thank:

Pietro Polotti, Thibaut Ajdler, Julien Cisonni, Luciano Sbaiz, and Christof Faller for all the audio related discussions and activities,

Christof Faller for all the coffee breaks, feedback and motivation,

Williams Devadason for keeping the computers running,

all the other friends and colleagues at LCAV and LCM,

FlyHigh Lowsanne for helping me keeping fit,

the Norwegian church in Geneva and the Scots Kirk in Lausanne for helping me keeping the faith,

Ingunn Marie and Isak and Noah.

Chapter 1

Introduction

1.1 Motivation

- Did you hear that?
- What?
- **Did you hear that?**
- **What?**
- **DIIID YOUUUU HEEEEAR THAT?, I said!**
- **Yes, I know! But did I hear WHAT?**
- Oh, somebody pouring potato chips into a glass bowl in the kitchen.

This conversation took place between the children that were playing in their room one Saturday. I think it nicely demonstrates how a person is able to detect and recognize different sound sources and then focus on what is considered to be important, while completely disregarding all other "unimportant" sounds. For the older child the "snack" was so important that everything else was disregarded. Surely, for his little brother it may have been unclear whether he was referring to the siren of a distant ambulance, the neighbor slamming the entrance door, or any other sound that may have been audible at the time.

The mechanism of disregarding or suppressing sources that are considered to be of little importance is an astonishing ability of the human auditory system. However, it is such a natural thing that often we are not even aware that it takes place. Since it can happen more or less unconsciously, we sometimes do not even realize that some sound sources are present. Once, the air-conditioning in the office was suddenly turned off. The immediate reaction of a colleague was:

- What happened? Where did all that silence come from?

In order to demonstrate how complex our auditory system is, Bregman (1999, pp.5-6) gives the following analogy:

"Imagine that you are on the edge of a lake and a friend challenges you to play a game. The game is this: Your friend digs two narrow channels up from the side of the lake. Each is a few feet long and a few inches wide and they are spaced a few feet apart. Halfway up each one, your friend stretches a handkerchief and fastens it to the sides of the channel. As waves reach the side of the lake they travel up the channels and cause the two handkerchiefs to go into motion. You are allowed to look only at the handkerchiefs and from their motions to answer a series of questions: How many boats are there on the lake

and where are they? Which is the most powerful one? Which one is closer? Is the wind blowing? Has any large object been dropped suddenly into the lake?"

Every day a person is exposed to a vast variety of acoustical environments. The number and types of sound sources may vary. Moreover, the original sounds emitted by sources are disturbed by reflections before they reach our ears. There may be physical obstacles and the environment can be highly dynamic, with moving sources and objects. Still, a person with normal hearing capabilities is able to communicate under strongly adverse conditions. Even though this ability is often taken for granted, the job that the auditory system performs is astonishing. Under similar conditions a hearing-impaired person may face severe difficulties and experience this as a social handicap. Similarly, man-machine interfaces, e.g. automatic speech recognition systems, face the same difficulties. It is sufficient to look at these examples in order to understand that the dream of creating a machine that can match the human performance is not new. For decades researchers have studied the problem.

The main focus of this thesis is on computational techniques for the localization and separation of sound sources. In particular, we propose techniques based on the short-time Fourier transform (STFT). We devise methods for processing of individual spectral coefficients in the STFT spectra of the observed signals, effectively enabling the tracking of sound sources in dynamic scenes and the separation of closely spaced frequency components.

1.2 Background

Auditory scene analysis

In general, the task that the human auditory system performs in order to detect, localize, recognize and emphasize different sound sources is referred to as *auditory scene analysis* (ASA). An auditory scene denotes the listener and his/her physical surroundings, including sound sources. The traditional research on ASA has been in the areas of acoustics, physiology and psychology. When referring to these areas in common, the term psychoacoustics is often used. A vast literature exists on various aspects of psychoacoustics. Bregman (1999) gives a very thorough overview of ASA related to higher psychological levels. Spatial hearing has been reviewed by Blauert (2001), giving a very thorough overview of the field. Lower level psychophysics and physiology has been reviewed by Moore (1997) and Zwicker and Fastl (1999).

In the framework of ASA the signal emitted by a physical source is called a *source event*. Each source event has a set of associated *source attributes*. These attributes are related to some characteristics of the source, such as its physical location and its temporal and spectral activity, among others.

A listener will, with each ear, observe a mixture of all the source events in an auditory scene. A source signal that reaches the ears directly is called a *direct sound*. In addition to the direct sounds, the observed signals will also contain indirect sounds, i.e. source signals that reach the ears after reflections, shadowing, etc. Each ear thus observes a superposition of filtered source signals. In these filtered mixtures, the original source events, and the attributes associated with these, are not fully preserved. However, the mixtures may contain sufficient information for the human auditory system to recover the source

events, at least partially. The core of ASA is to obtain this information. The information that the human auditory system is able to extract out of the mixtures is referred to as *auditory cues*, or simply cues. Based on these cues, the human auditory system can group components of the observed signals together into *auditory events* (Blauert, 2001). Bregman (1999, p.9) uses the term *auditory streams*: *"An auditory stream is our perceptual grouping of the parts of the neural spectrogram that go together."* Source events and their corresponding source attributes are related to the physical world, whereas auditory events and cues refer to our mental representation, or image, of the physical world.

Cues

Very simplified, there are two main groups of cues. The first group contains cues related to the characteristics of the source signals. These cues are basically related to the temporal and spectral activity of the sources. Since they only depend on the source signal, they are called *signal cues*. Examples of cues in this group are onset time, offset time, amplitude modulation, frequency modulation and harmonicity, among others. These cues, as well as the way they are employed in the human auditory system in order to group them into auditory streams, is the main subject of Bregman (1999). However, the fact that a person has two ears and the information that can be extracted based on differences between the signals observed by the ears is not as thoroughly discussed. This brings us to the other main group of cues.

The second group of cues is related to the physical conditions of the auditory scene, mainly reflecting the physical locations of the sources relative to the sensors. The basic cues in this group are level and phase differences between the ears. These cues are vital for the estimation of the original source attributes, such as elevation and azimuth angles.

Computational auditory scene analysis

The term *Computational auditory scene analysis* (CASA) denotes different computational techniques that try to mimic the behavior of the human auditory system, or at least some aspects of it. The degree to which these techniques are concerned with how the humans carry out ASA is highly varying. At one extreme, techniques that aim at modeling the processing in the human auditory system exist. At the other extreme there are purely computational techniques without any direct relation to human ASA. Rosenthal and Okuno (1998) contains a collection of articles on CASA.

Binaural signals denote signals received by our two ears. Techniques based on such signals are usually considered as CASA techniques. This is also the case of techniques working with only one signal (single channel). Purely mathematical techniques employing sensors in free-field, or a higher number of sensors, are normally not considered as CASA techniques. In the context of this thesis we do not distinguish between these types of techniques. The main focus is on localization and separation of sources in binaural signals and we draw knowledge and ideas from various types of techniques. We therefore use the specific terms source localization and source separation techniques, depending on what their primary purpose is.

Source separation techniques usually employ cues, explicitly or implicitly, in the analysis. In Chapter 2 we give an overview of different cues and review some computational techniques that are based on these cues.

From the discussion in Chapter 2 it will become clear that most existing computational methods employ only a small number of cues. This is in strong contrast to the way the human auditory system weights many different cues against each other. In fact, there are remarkably few techniques that try to exploit both groups of cues, namely signal cues and spatial cues. This seems to be the case in the context of CASA and for mathematical approaches. It is well known that the human auditory system can make use of both types of cues, and this provides the motivation for our work.

1.3 Contribution of this thesis

The focus of this thesis is on localization and separation of sound sources. We draw knowledge from existing techniques, both techniques that are motivated by human ASA and techniques that do not have a direct relation to the human auditory system. The methods we propose are purely computational and do not directly correspond to any processing that takes place in the human auditory system. Still, they are partially motivated by ASA.

In particular, a separation method based on spatial cues is used as the starting point. This is then combined with an evaluation of signal cues in order to get a more robust system and improve the quality of separation. Bregman (1999, p. 302) discusses how the different cues are used in ASA: *"It seems that spatial origin, in itself, is such a good cue to whether or not spectral components come from the same sound, that we might need no other. Yet human auditory scene analysis does not put all its eggs into one computational basket. While spatial cues are good, they are not infallible. For example, they can become unusable in certain kinds of environments. [...] For these reasons, the human auditory system does not give an overriding importance to the spatial cues for belongingness but weighs these cues against all the others."*

Motivated by this, the goal of our work is to evaluate several cues jointly. When all the cues are consistent, this confirms the decisions being made. When the different cues are inconsistent, it may in some cases be possible to determine which cues are ambiguous, and exclude these from the grouping process. For arbitrarily complex auditory scenes ambiguous situations may exist, where the ambiguities are impossible to resolve. This simply reflects the fact that source separation is, in general, an ill-posed problem. Even in the human auditory system ambiguities exist (Bregman, 1999, p. 302): *"When the cues all agree, the outcome is a clear perceptual organization, but when they do not, we can have a number of outcomes."*

Thesis layout

The organization of this thesis is as follows.

Background and motivation The discussion of cues and existing computational techniques in Chapter 2 serves as the background and motivation for our work that is presented in the subsequent chapters.

Signal cues are discussed in Section 2.1, followed by a review of source separation techniques based on these cues in Section 2.2. An overview of spatial cues is given in Section 2.3 and the related computational techniques are reviewed in Section 2.4. The discussion of these two main groups of cues is followed by a short overview of other higher level cues, that are less relevant with respect to our computational models. These cues and the related computational techniques are discussed in Section 2.5. In Section 2.6 we discuss some of the drawbacks of existing techniques.

Binaural localization In Chapter 3 we propose a method for localization of multiple sources in binaural signals (Viste and Evangelista, 2003a, 2004a,c). The computation of estimates of level and time differences between the two ears is presented in Section 3.1. Based on these estimates the source location can be found by lookup in a measured set of head related transfer functions (HRTFs). We propose a method for joint evaluation of the level and time differences, as presented in Section 3.2. In Section 3.3 we propose a parametric model for the relation between the source location and the level and time differences and compare the performance of this model with the method based on measured HRTFs. The parameters of the parametric model are compared for different subjects in a database of HRTFs leading to an average parametric model. This model only depends on the head size and can be used for any head without measuring the HRTFs for different source locations. This is presented in Section 3.4. The application of this technique to the localization of sound source in static and dynamic scenes is presented in Sections 3.5 and 3.6, respectively.

Separation by spatial weighting Chapter 4 discusses the use of the binaural localization method for the separation of sources (Viste and Evangelista, 2003a). This is introduced in Section 4.1. Experimental results are given in Section 4.2. One of the major problems is related to sources whose energies overlap in time and in frequency, causing corrupted spatial cues. Some possible approaches to this problem are discussed in Section 4.3. This provides the background and motivation for the following chapter.

Separation of overlapping partials The problem of corrupted signal cues, and overlapping energy in particular, is specially important for music and harmonic instruments. This is the topic of Chapter 5 (Viste and Evangelista, 2002, 2003b, 2004b). In Section 5.1 the problem is motivated by a discussion of harmonic instruments and musical scales. Then, we propose a method for the separation of overlapping spectral components in multi-channel mixtures in Section 5.2. Experimental results are given in Section 5.3

Conclusion Finally, in Chapter 6 we draw the conclusions and discuss some ideas for future extensions.

Chapter 2

Background and motivation

Auditory scene analysis (ASA) is the problem of detecting different sound sources and their spatial locations based on the signals that are observed by the two ears. In the human auditory system various types of processing at different levels of abstraction are involved. In the framework of source separation and computational auditory scene analysis (CASA), as is presented in this thesis, we find it instructive to distinguish between the following three types of processing: single-sensor signal analysis, inter-sensor spatial analysis, and high-level cognition. This is a highly simplified view that does not reflect the complex functioning of the human auditory system. However, for the purpose of computational models, we find it most useful. These different types of processing employ different cues. Accordingly, the cues can be grouped into *signal cues*, *spatial cues*, and *high-level cues*.

This chapter gives an overview of the different types of cues and a review of existing source separation techniques. For each type of cues, first the related cues are discussed in the context of ASA. This is followed by a review of existing CASA techniques that exploit these cues for the separation of sources.

From the discussion in this chapter, it will become clear that almost all of the existing techniques focus on only one, or a few, of the large amount of available cues. This is in strong contrast to the human auditory system, where it was argued that many cues are evaluated against each other in order to provide a more robust analysis. Therefore, this chapter does not only provide a quick overview of some existing methods. It also serves as the main motivation for the computational models that we will devise in the following chapters in which several different cue types are jointly evaluated.

Section 2.1 discusses single-sensor analysis and cues that are available in the signals of single sensors, denoted signal cues. A review and discussion of source separation techniques based on signal cues is given in Section 2.2. Section 2.3 discusses spatial cues that are available in the multiple channels of a signal observed by more than one sensor. Source separation techniques based on spatial cues are discussed in Section 2.4. High-level cues and source separation techniques based on these are briefly discussed in Section 2.5. The chapter is concluded by a discussion of corrupted cues and joint processing of different cue types in Section 2.6. This serves as the direct motivation for the techniques presented in this thesis.

2.1 Single-sensor analysis and signal cues

When a sound source is active it emits sound. For a point source the sound radiates in all directions as propagating sound pressure waves. True point sources do not exist. A general source has a physical extent where the sound waves are generated by vibrations. This means that the different parts of the surface of the source emit different acoustic waves. However, when a source is observed at a distance, its physical extent is relatively small so that it can be considered as a point source. Due to the laws of physics, the different parts of the source body are coupled. Therefore the signals emitted in different directions will be highly correlated. In a very simplified outlook, one can say that the same signal is emitted in all directions, only with different amplitude and phase. From this point of view it makes sense to talk about a single "source signal" emitted from a sound source. The distribution of the emitted energy in time and in frequency conveys information about the characteristics of the underlying source. A representation of this signal in time and in frequency is called spectrogram. From an analysis of the spectrogram of the source signal several cues can be obtained. These cues are the topic of this section.

In the human auditory system, the signal that enters the ear is decomposed into spectral components corresponding to (highly overlapping and non uniform) frequency bands. Along the basilar membrane in the cochlea, different regions react differently to different frequency components. At each location along the basilar membrane a different group of neurons encode and transmit the signal component in a particular frequency band.

Auditory filters model the frequency responses of these bands. These filters decompose a signal into *perceptual bands*. For more detailed information about how this is achieved in the human auditory system, Zwicker and Fastl (1999) gives a more thorough overview. The resulting representation of the observed signal in time and in frequency is called *neural spectrogram*, or *cochleagram* since it is carried out in the cochlea. This neural representation provides a non-uniform frequency resolution. In fact, this frequency resolution is rather limited, at least compared to the frequency resolutions that can be obtained in purely computational representations like the short time Fourier transform. From a computational point of view, any suitable time-frequency analysis may be employed. This is discussed in more details in Sections 2.2.1 and 2.6.1.

For the high-level discussion of cues in this section, the specific details of the signal representation is not of high relevance. The ideas are applicable to most time-frequency representations, i.e. perceptual models mimicking the human auditory system or purely computational models. For this discussion, it is only assumed that the signal is represented in time and frequency.

The problem of ASA consists of *grouping* together regions in a time-frequency representation that belong to the same source. Bregman (1999) notes that: *"There is a serious need for regions to be grouped appropriately. Again, it would be convenient to be able to hand the spectrogram over to a machine that did the equivalent of taking a set of crayons and coloring in, with the same color, all the regions on the spectrogram that came from the same sound source."* The mentioned "coloring problem" is the problem of detecting different sources in an observed signal. In a single signal, there are several cues that can be exploited in order to perform such grouping. These are all related to the energy distribution of the signal in time and in frequency.

A complex sound consists of several spectral components. An instrument note can be closely modeled as the sum of various sinusoids with slowly varying amplitude and frequency. The spectral components that can be closely modeled by single sinusoids are, in general, called *partials*. For harmonic instruments the frequencies of the different partials are in a harmonic relation, i.e. the frequency of each partial is an integer multiple of the fundamental frequency. In this case the partials are often called *harmonics*. Due to the physics of vibration and sound generation, these different spectral components typically have a lot in common, e.g. amplitude modulation etc. The perceptual phenomena occurring when several spectral components are perceived as one single "object" is called perceptual *fusion*. Many experiments have been conducted that suggest that coherent changes play a role in the fusion of different spectral components (Moore, 1997). This is known as the principle of *common fate* (Bregman, 1999, pp. 248). Briefly explained, this principle is the observation that in many natural sound sources, the different (spectral) components change synchronously. The chance that the coherent changes are due to components coming from the same physical source in the environment seems much more likely than the chance that this would happen for components from different sources. Under these observations, it is reasonable to group together components whose cues evolve coherently into the same auditory stream. The notion of common fate comprises cues related to amplitude and frequency modulation. Since these cues typically can be obtained from a single sensor signal, they are also known as "monaural cues" (Blauert, 2001, ch. 2.3). These signal cues are all related to properties of the observed signal and are discussed in the remainder of this section.

2.1.1 Temporal cues

When a sound source is excited, its physical mass is set into vibration. Vibrations typically occur at several different frequencies or modes. These frequencies depend on the physics of the underlying source that establishes which frequencies resonate easily and which frequencies are attenuated. Generally, as the source is excited, all the different frequency components start. The time it takes for the different components to build up depends on the physics of the source, but as a rough simplification the different components start more or less simultaneously. The same argument may also be applied when the excitation stops, depending on the type of excitation and on the attenuation of the different spectral components. For many different sound sources the temporal activity of the sound, or its duration, is common to all the different spectral components. The start and end of a sound are denoted by *onset* and *offset*, respectively. Figure 2.1 shows spectrograms of sound mixtures consisting of six spectral components. Five components are related to each other and one component is unrelated. In Fig. 2.1(a), it is clearly seen how common onset and offset make the five components appear as one group, whereas the last component stands out due to its delayed offset time. Fig. 2.1(b) shows a similar visual grouping for components with equal duration, but with different amplitude modulations. Depending on the source and on how this is excited, the temporal amplitude envelope of the different spectral components may vary in time. When the different spectral components have correlated amplitude changes, this can be exploited directly in order to group different components

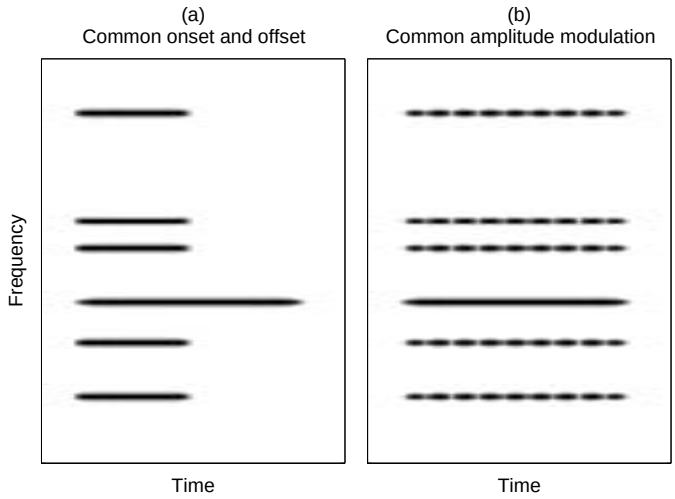


Figure 2.1: *Temporal cues, where five components are related and one stands out. (a): Common onset and offset. (b): Common amplitude modulation.*

into auditory streams.

With the rather limited spectral resolution provided by the neural spectrogram, it is clear that several spectral components may fall in the same frequency band. In that case, the individual components are not individually visible, such as those shown in Figure 2.1. However, when two spectral components fall within the same frequency band, this gives rise to strong amplitude modulations. These are called *beatings*, whose frequency equals the difference in frequency between the two spectral components. Without speculating about the functioning of the human auditory system, we simply note that these changes in amplitude can be employed in computational models in order to detect the underlying components. In Section 5 we present a method where beating is exploited in order to separate overlapping narrow-band partials.

I this thesis, all of the mentioned properties of a spectral component, namely onset, offset and amplitude changes, are described by the amplitude modulation (AM). This cue can be estimated from the spectrogram of the observed signal. Different auditory streams can then be formed by applying grouping principles based on similarity or synchrony of AM cues.

2.1.2 Spectral cues

In addition to temporal properties, a source also has properties related to its spectrum. In this context, there are two main cues. The first is related to static properties of the sound, or properties that remain constant for at least some time. The second is related to dynamic properties that change over time.

Since the different spectral components reside in different frequency regions, their relative locations in frequency provide information about the source. This is best known in relation to harmonic instruments, where the frequencies of the different partials, or harmonics, of a single note are approximately integer mul-

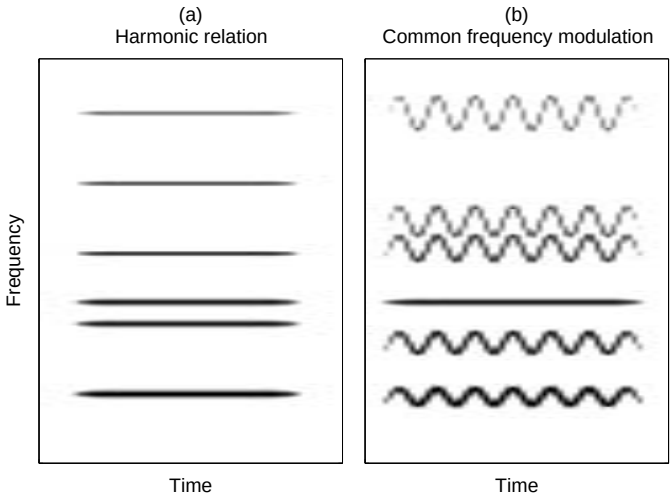


Figure 2.2: *Spectral cues, where five components are related and one stands out. (a): Harmonic relation. (b): Common frequency modulation.*

tiples of the fundamental frequency of that note. It is well known that the human auditory system tends to group spectral components that are in a harmonic relation into one auditory stream. This harmonicity principle is the first cue related to properties of the source that remain steady for some time. There exist other cues related to these static properties, such as more complex frequency relations, spectral envelope or relative intensity of the spectral components, density of spectra, etc. From the computational point of view these are more difficult to exploit and are therefore only briefly discussed in Section 2.5 on cognition and learning.

The second spectral cue is related to changes over time. Similarly to the way the amplitude fluctuates, the frequency may also vary. For a wide range of sound sources, the changes in frequency that naturally occur will typically be common to all the spectral components. One example of this is parallel changes in frequency. This naturally happens in glides, i.e. when an instrument gradually shifts frequency from a starting note to a final note. It also naturally occurs in the human voice, when the pitch changes, e.g. due to intonation. Instruments or voice with vibrato are other examples where the different partials have common frequency changes. All of these effects are covered by the term frequency modulation (FM). Auditory streams can be formed by grouping principles based on synchrony of FM cues.

Figure 2.2(a) shows five harmonically related spectral components and one that is not harmonically related. Five spectral components with similar frequency modulations are shown in Fig. 2.2(b). Visually, the modulated components are grouped together and the component without FM stands out.

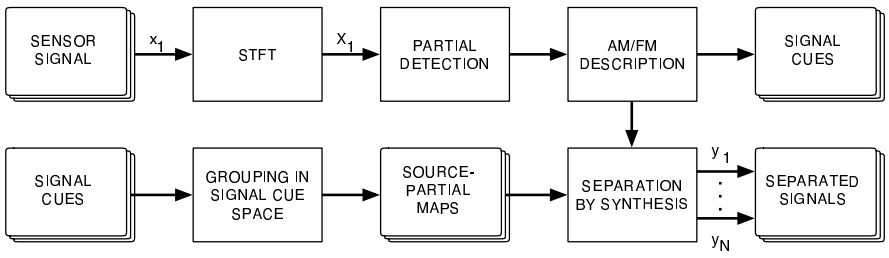


Figure 2.3: *Source separation using signal cues. Top row: Estimation of signal cues. Bottom row: Using signal cues for separation.*

2.2 Separation by signal cues

Several computational techniques exist for the separation of audio sources are based on signal cues. These techniques only require one sensor signal. As seen in Section 2.1, the basic principle is to represent the observed signal in time and in frequency and then to group spectral components in this representation together into auditory streams. This task consists of three distinct parts:

Analysis: Detection and description of different spectral components in time and frequency, yielding signal cues.

Grouping: Using signal cues in order to group spectral components into source objects, providing a mapping between sources and spectral components.

Separation: Using the descriptions and mapping found in the first two steps in order to reconstruct, or synthesize, the original sources.

This processing is illustrated in Figure 2.3. The top row shows the analysis of a single-channel signal x_1 in order to obtain the signal cues. The signal is represented in time and frequency, typically using the Short-Time Fourier Transform (STFT). In this representation the individual spectral components can be detected and extracted from the total signal. The extracted spectral components can then be described in terms of amplitude and frequency functions. The amplitude modulation (AM) and frequency modulation (FM) functions allow for the estimation of signal cues. The bottom row shows the application of these signal cues to source separation. First, the estimated signal cues are used in order to group the different spectral components into different auditory streams. This yields a map between spectral components and sources. Based on this map and on the description of the components found in the analysis stage, the different sources can be separated, or synthesized. In the following, these three processing steps are discussed.

2.2.1 Time-frequency representation

In computational models for the analysis of audio signals in time and in frequency, the Short-Time Fourier Transform (STFT) is the most commonly used signal representation. In practical implementations the windowed Discrete

Fourier Transform (DFT), the discrete equivalent of the STFT, is applied. Throughout this thesis the term STFT will be used.

For a general signal $x(l)$, with time index l , the STFT is given by

$$X(k,q) = \sum_{l} w_a(l)x(l+kH)e^{-j2\pi lq/L}, \tag{2.1}$$

where $w_a(l)$ is the analysis window of length L. The complex values of the STFT are denoted *spectral coefficients*, $X(k,q)$, and are indexed by k and q in time and frequency, respectively. The time index is related to the hop-size, H, that describes how much the window moves (in samples) between two consecutive time indexes.

Using a possibly different window for the synthesis, $w_s(l)$, the signal can be reconstructed by inverse DFT and overlap-add. For simplicity, the analysis and synthesis windows are assumed to be of equal length L. The Inverse STFT (ISTFT) is defined as

$$x(l) = \sum_{k} w_s(l-kH) \sum_{q=0}^{L-1} X(k,q)e^{j2\pi lq/L}. \tag{2.2}$$

The ISTFT results in perfect reconstruction, up to a constant scaling factor C, when the choice of analysis and synthesis windows satisfies the following condition

$$\sum_{k} w_a(l-kH)w_s(l-kH) = C, \qquad \forall l. \tag{2.3}$$

In the context of functional analysis, equations (2.1) and (2.2) can be considered as the expansion of the signal onto over-complete sets called Gabor frames (Gabor, 1946; Daubechies, 1992).

In comparison to the spectral analysis performed in the human auditory system, the STFT enables different frequency resolutions. Firstly, by choosing longer windows it is possible to get a much higher frequency resolution than that of the auditory filters. This increased frequency resolution comes at the expense of poorer time resolution. However, the increased frequency resolution can enable individual detection of closely spaced spectral components that fall within the same perceptual band. Secondly, the spectral coefficients of the STFT are complex numbers. These can be represented as magnitude and phase, giving a very intuitive physical interpretation with respect to propagating waves. Finally, the STFT is invertible. Most filter bank approaches trying to mimic the frequency resolution of the human auditory system are not invertible. In fact, it is convenient to approximate the perceptual bands by use of the STFT, grouping different spectral coefficients into perceptual bands. The resulting filters deviate somewhat from those of the human auditory system, but allow for very fast computation and perfect reconstruction.

While the human auditory system does not provide a fine enough frequency resolution for a given application, we do not want to be limited by this. From a computational point of view it may be advantageous in some cases to operate with the high frequency resolution of the STFT, rather than trying to mimic the frequency resolution of the human auditory system. Whenever perception needs to be taken into consideration, the perceptual bands can be mimicked by grouping different spectral coefficients of the STFT.

2.2.2 Detection and representation of spectral components

In ASA, the grouping problem is the problem of detecting different components in a spectrogram and then group these together. For decades, audio researchers have studied the spectrograms of many types of sound sources (Helmholtz, 1954; Plomp and Levelt, 1965). For a large group of naturally occurring sounds, such as harmonic instruments and voiced speech, the time-frequency components containing most of the total energy are spectral components whose frequencies and energies evolve relatively slowly in time. These can typically be seen as strong "horizontal" spectral lines in the spectrogram, such as those illustrated in Figures 2.1 and 2.2. Each of these spectral components constitutes an important part of the total signal and is denoted a partial.

In vision, grouping principles were put forward by the Gestalt psychologists (Bregman, 1999). Roughly explained, these principles state that we tend to group together components that are close to each other in some respect. The notion of closeness can refer to spatial proximity, similarity in shape, texture, etc. Looking at a spectrogram, the different partials of a single source are often similar in some respect. Even without knowledge of the human auditory system, it therefore seems natural that the grouping principles of the Gestalt psychologists can be applied in CASA as well. This is related to the signal cues that the human auditory system can exploit in the analysis of an auditory scene, as discussed in Section 2.1. A bit loosely formulated, the problem is to group together parts of the spectrogram that "look" similar. In order to judge the similarity, different spectral components must be defined. Many different methods for the detection and parametrization of partials have been studied in the context of audio analysis. Most of these are described in terms of amplitude and frequency functions. This is the topic of the present section.

Sinusoidal components: instantaneous amplitude, phase, and frequency

Any real-valued signal $x(t)$ can be written as the product of a real-valued function $a(t)$ and the real part of a complex exponential with time-varying phase $\varphi(t)$

$$x(t) = a(t)\,\mathcal{R}\left\{e^{j\varphi(t)}\right\}, \tag{2.4}$$

where $\mathcal{R}$ denotes the real part. It is therefore convenient to consider the analytic signal $x(t) = a(t)e^{j\varphi(t)}$. This representation is not unique. At one extreme, the signal $x(t)$ can be represented by the function $a(t)$ only, i.e. by choosing $a(t) = x(t)$ and $\varphi(t) = 0$. Similarly, at the other extreme, the signal can be represented by the phase only, with constant $a(t)$. Clearly, no unique solution exists, and infinitely many choices of $a(t)$ and $\varphi(t)$ yield the same $x(t)$. For more details, Picinbono (1997) gives an overview.

If the signal $x(t)$ is narrow-band, as is the case for many spectral components in natural sounds, and if some restrictions are made on the functions $a(t)$ and $\varphi(t)$, the representation can become more meaningful. In particular, if $a(t)$ and the time derivative of $\varphi(t)$ vary slowly in time, the representation can be related to amplitude and frequency modulation (AM-FM). In this case, the function $a(t)$ (AM) is called the instantaneous amplitude and $\varphi(t)$ (FM) is called the instantaneous phase. The instantaneous phase and instantaneous frequency

$\omega(t)$ are related by differentiation and integration with respect to time:

$$\omega(t) = \frac{d\varphi(t)}{dt}, \tag{2.5}$$

and

$$\varphi(t) = \int_0^t \omega(\tau)d\tau + \varphi_0. \tag{2.6}$$

This means that a narrow-band signal can be represented by two meaningful functions, namely a slowly varying instantaneous amplitude and a slowly varying instantaneous frequency (or phase) function. The wide band signal can be written as a sum of narrow band spectral components $x_i(t)$ where each spectral component is described by amplitude $a_i(t)$ and frequency $\omega_i(t)$ (or phase $\varphi_i(t)$)

$$x(t) = \sum_{i=1}^{I} x_i(t) = \sum_{i=1}^{I} a_i(t)e^{j\varphi_i(t)}. \tag{2.7}$$

The assumptions that $a(t)$ and $\omega(t)$ are slowly varying put restrictions on the choices of these functions and give an intuitive interpretation in the form of AM and FM. However, "slowly varying" is a rather vague term. As a rough idea, one may think of frequencies that are inaudible (e.g. below about 20 Hz) to be represented by the amplitude $a(t)$ and higher frequencies to be represented by the phase $\varphi(t)$. However, even if this is defined precisely, there is still no single unique solution in general. For instance, most natural sounds that have a strong partial structure still have energy in the partial sidebands. In other words, there is also energy in frequency ranges about the spectral peaks of the partials. Therefore, the AM-FM representation (2.7) of a narrow-band spectral component also depends on how the narrow-band signal is extracted from the total signal, or on which frequencies should be allowed to be modeled by the phase $\varphi(t)$.

2.2.3 Grouping

When source separation is based on signal cues, the different spectral components are grouped into different auditory streams by the grouping principles discussed in Section 2.1. In the AM-FM representation, it is easy to compare the different signal cues. For each component $x_l(t)$, the instantaneous amplitude $a_l(t)$ function retains the temporal cues, i.e. onset, offset, and AM, as mentioned in Section 2.1.1. Similarly, the instantaneous frequency $\omega_l(t)$ conveys useful information about the spectral cues, i.e. harmonicity and FM, as discussed in Section 2.1.2. In computational models, the harmonicity of different spectral components seems to be, by far, the most commonly used cue for the grouping (Nakatani and Okuno, 1999; Karjalainen and Tolonen, 1999; Virtanen and Klapuri, 2000, 2001, 2002; Klapuri, 2001; Drake, 2001).

For the estimation of instantaneous amplitude and frequency functions several methods exist. This is the topic of the following sections.

2.2.4 Sinusoidal models

In the mid eighties, McAulay and Quatieri (1986) and Smith and Serra (1987) independently developed similar techniques for modeling non-harmonic and pitch-varying sounds. In general, they are referred to as *sinusoidal models*.

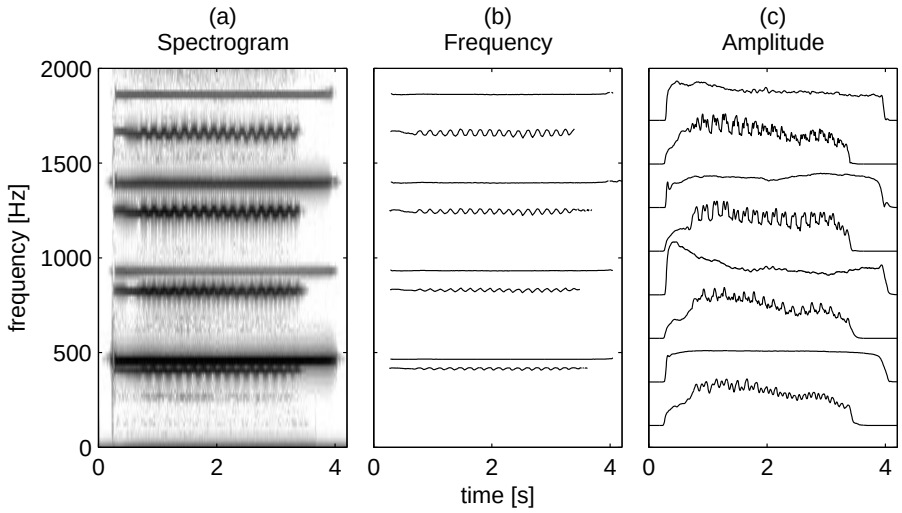


Figure 2.4: *Sinusoidal modeling, using spectral peak picking and peak matching over time. (a): Spectrogram showing first four partials of two notes. (b): Detected frequency trajectories describing the partials. (c): Amplitude of the partials along the detected frequency trajectories.*

These methods are based on spectral peak picking in the STFT spectrum. For every time index k, each spectral peak, indexed by i, corresponds to one spectral coefficient, $X(k, q_i)$. The complex value of this spectral coefficient provides estimates of the instantaneous amplitude

$$\tilde{a}(k, q_i) = |X(k, q_i)|, \tag{2.8}$$

of the instantaneous phase

$$\tilde{\varphi}(k, q_i) = \arg X(k, q_i), \tag{2.9}$$

and of the instantaneous frequency

$$\tilde{\omega}(k, q_i) = \frac{2\pi q_i}{L}, \tag{2.10}$$

where L is the DFT length (2.1). The notion of "slowly varying" is in this case based on the observation that for each spectral coefficient, the signal amplitude can be regarded as constant over the duration of the analysis window. The rate at which the amplitude is allowed to vary is thus directly related to the window length. With a chosen window length, the magnitude of a spectral coefficient is related to the "instantaneous" amplitude for the corresponding time and frequency index. This amplitude is not truly instantaneous, but averaged over the duration of the analysis window. The same holds for the "instantaneous" phase and frequency estimates. The ideas behind the sinusoidal models (McAulay and Quatieri, 1986; Smith and Serra, 1987) are based on this approximation.

After the spectral peaks have been detected for a given time index, they need to be connected with the previously found peaks. This is achieved by a

peak matching between successive time indexes, effectively tracking the spectral peaks in time. For each spectral peak $X(k, q_i)$, it must be determined if this is the start of a new spectral component, or the continuation of an already existing partial, determined by the spectral peak $X(k-1, q_j)$ for some j. More information about the decisions about the birth, continuation and death of partials can be found in the articles that we referred to previously.

The tracking of spectral peaks in time yields amplitude, phase and frequency trajectories as functions of time. Figure 2.4 shows the result for a mixture of two harmonic notes, one with vibrato (AM and FM) and one without vibrato. Panel (a) shows the spectrogram of the mixture, including the first four harmonics of both notes. Panel (b) shows the frequency trajectories for the eight detected partials. The FM of the note with vibrato is clearly seen. The amplitude of the eight partial trajectories are shown in Panel (c). The order of the partials (along the ordinate axis) is the same as in Panel (b), but for visual clarity the envelopes have been normalized and placed with uniform distance along the ordinate axis. The AM of the vibrato note is clearly visible.

For low bit-rate coding of audio signals, and speech in particular, sinusoidal models have proved to be useful. The coding efficiency can be very high since each spectral component is characterized by only two parameters per time index. The data rate is directly related to the hop-size H in (2.1). By choosing a smaller H, a finer time-resolution can be obtained, at the cost of increased data rate. Indeed, for the signal represented by one spectral component, the hop-size can be viewed as the decimation rate, or the degree of down-sampling. Alternatively, the trajectories can be parametrized and interpolated, yielding even less data to code. This parametrization can also be used for signal modifications, e.g. pitch shifting, time stretching, etc.

Improved frequency resolution

In the peak picking algorithm, the frequency resolution is determined by the length of the analysis window in the STFT. The resolution can be improved by choosing a longer analysis window, but the practical usefulness of this is quite limited. As the window gets longer, the "instantaneous" amplitudes and frequencies are averaged over longer time intervals so that fine resolution variations in time are smoothed out. In order to provide better estimates of the frequency and amplitude without increasing the window length various methods have been devised.

For spectral components with little variation in frequency, the evolution of the spectral coefficients over time can improve the frequency estimates. In particular, the frequency estimates can be improved by use of the phase vocoder (Portnoff, 1976). Provided that the the frequency indexes of the spectral peaks are the same for consecutive time indexes, the phase vocoder gives the following estimate of the instantaneous frequency

$$\tilde{\omega}(k, q_i) = \frac{2\pi q_i}{L} + \frac{\arg X(k+1, q_i) - \arg X(k, q_i)}{H}. \tag{2.11}$$

This estimate is not restricted to a fixed frequency resolution, which depends on the length of the DFT. A discussion of the accuracy of this estimate for different choices of analysis windows is given in Puckette and Brown (1998).

Another possibility for improved frequency resolution is to perform processing across frequency. For each time index, the frequency and amplitude estimates can be refined by peak interpolation in the spectrum. Keiler and Marchand (2002) give an overview and comparison of different interpolation schemes of this type.

Energy separation approximation

Maragos *et al.* (1993) proposed another technique for the estimation of AM and FM for speech resonances. This method is based on the Teager energy operator (Teager and Teager, 1989),

$$\Psi[x(t)] = [\dot{x}(t)]^2 - x(t)\ddot{x}(t), \tag{2.12}$$

where the dots denote differentiation with respect to time: $\dot{x}(t) = \frac{dx(t)}{dt}$. The amplitude and frequency can then be approximated by

$$|a_i(t)| \approx \sqrt{\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]}} \tag{2.13}$$

and

$$\omega_i(t) \approx \frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}}, \tag{2.14}$$

respectively. This technique is motivated by nonlinear properties of speech production. An overview of non-linear processing techniques can be found in (Kvedalen, 2003).

Extensions for non-sinusoidal components

The sinusoidal representation that has been discussed aims at representing spectral components with slowly varying amplitude and frequency. Many signals, e.g. voiced speech and instruments, can be approximated by a sum of these components. However, a sinusoid only models the *ridge* of a partial, i.e. the spectral peak evolving in time while the sidebands are not taken into account. The quality of the reconstructed signals is limited unless other spectral components, such as transients, partial sidebands and other "non-sinusoidal" components are properly handled.

Several extensions have been proposed. The sinusoidal representation has been complemented with filtered noise (Serra and Smith, 1990; Serra, 1997), transients (Verma and Meng, 2000), and transients and noise (Levine and Smith, 1998). In a different framework employing a harmonic band wavelet representation (Polotti and Evangelista, 2001a,b) modeled signals as pseudo-sinusoidal components plus sideband noise. Like the original sinusoidal models, most of these extensions have been made with the same applications in mind, i.e. signal compression and modifications. For instance, the sidebands of a partial are modeled as filtered noise, and no connection is made to the sinusoidal trajectory that describes the spectral peaks of that partial.

Discussion

Sinusoidal models do not fully represent a signal in the sense that the original signal being modeled cannot be reconstructed perfectly. Rather, each individual spectral component is synthesized as a sinusoid with varying amplitude and phase. While this may be sufficient for intelligible speech at low bit-rates, it is not optimized for high-quality separation. In particular, sinusoidal models only consider the spectral peaks and the sidebands are not properly taken care of. By including non-sinusoidal components in the model, as discussed in the previous section, this can be improved. However most of these extensions do not make a connection between the sinusoidal trajectories and the non-sinusoidal components that belong to each other. While this can be acceptable for compression and signal modification, it is not a convenient representation for source separation.

As mentioned in the discussion of sinusoidal models, the parameters of the spectral components are only estimated every H samples in time, Equations (2.1 and 2.8–2.10). One might think of choosing a smaller hop-size H in order to capture finer details in time and to improve the quality of separation. However, since the spectra are computed from windows of length L, the amplitude and frequency estimates are not truly instantaneous, but averaged over the duration of the analysis window. The degree to which the estimated amplitude and phase are instantaneous is limited by this, regardless of the choice of H.

2.2.5 Analysis-by-synthesis/overlap-add

A slightly different approach to sinusoidal modeling and improved frequency resolution for the spectral components is based on an analysis-by-synthesis/overlap-add (ABS/OLA) principle (George and Smith, 1992, 1997). Like the sinusoidal models described above, the different spectral components are detected by spectral peak picking. However, instead of tracking these peaks over time and representing the component by a time varying amplitude and phase, the ABS/OLA techniques work on the spectrum of each individual time index independently. At time index k, for each spectral peak q_i that is detected, the sinusoid with constant amplitude a_i and constant frequency ω_i (and phase offset $\varphi_{0,i}$) that most closely resembles the signal under consideration is chosen and subtracted from the signal. The envelope of the signal is extracted by low pass filtering and taken into account in the search for constant amplitude sinusoids. This matching algorithm is recursively applied to the peaks, in descending order of magnitude.

This method is somewhat similar to the matching pursuit (Mallat and Zhang, 1993), in that it searches a dictionary of waveforms for the element that best matches the signal. However, the search is performed independently on the windowed signal for each time index and not on the entire signal as in traditional matching pursuit. Verma and Meng (1999) have developed a similar method using perceptually weighted matching.

Discussion

Like the original sinusoidal models, ABS/OLA techniques were developed primarily in the framework of low bit-rate coding and modifications of signals. For

these applications, they have an advantage in the sense that they do not need to connect the individual spectral peaks into continuous trajectories by means of peak matching over time. Reconstruction/synthesis is achieved by overlap-add and the spectral components are not represented by amplitude and phase trajectories. For source separation based on signal cues, the cues for different spectral components must be compared. This means that each spectral component must be characterized as one object. In order to obtain the description of this object, the different sinusoids must be grouped into partial objects. This is analogous to the peak matching in sinusoidal modeling. For the application to source separation, ABS/OLA is therefore similar to traditional sinusoidal modeling.

The iterative peak matching approach of ABS/OLA can model the partials more closely. In the first iterations, the spectral peaks corresponding to the ridges of the partial trajectories are modeled and subtracted. After several iterations, spectral peaks corresponding to the sidebands can also be taken into account. Even though this allows for a better description of the partials, this has some limitations. Models based on sinusoids do not lend themselves easily to non-sinusoidal components. Since the partial sidebands often consist of more noisy components, this iterative description quickly reaches its limitations, and a very high number of sinusoids is needed in order to fully characterize the partials.

In the context of source separation, ABS/OLA therefore presents little advantage over traditional sinusoidal models.

2.2.6 Instantaneous AM-FM estimation

Rather than focusing on the ridges of the spectral components, there exist methods that also take into account the sidebands. Typically, these techniques work with individual narrow-band signals, and estimate amplitudes $a_i(t)$ and phases $\varphi_i(t)$ at each time instant. Consequently, the entire narrow-band signals are preserved.

Hilbert transform

One of the oldest such methods is based on the Hilbert transform. The Hilbert transform is given by

$$\mathcal{H}[x(t)] = p.v. \int_{-\infty}^{\infty} \frac{x(t-\tau)}{\pi\tau} d\tau, \tag{2.15}$$

where $p.v.$ denotes the Cauchy principal value. From a given signal $x(t)$, a complex analytical signal is then defined by

$$z(t) = x(t) + j\mathcal{H}[x(t)] = a(t)e^{j\varphi(t)}. \tag{2.16}$$

More details can be found in (Boashash, 1992). When working with discrete time signals, special care needs to be taken with respect to the phase (Sun and Sclabassi, 1993). Shortly explained, the complex signal $z(t)$ is the same as the original signal $x(t)$, but where all the negative frequencies have been removed. As seen in (2.16), the complex signal can be expressed as an instantaneous amplitude $a(t)$ and instantaneous phase $\varphi(t)$. This particular representation gives a unique solution for the amplitude and phase in (2.4).

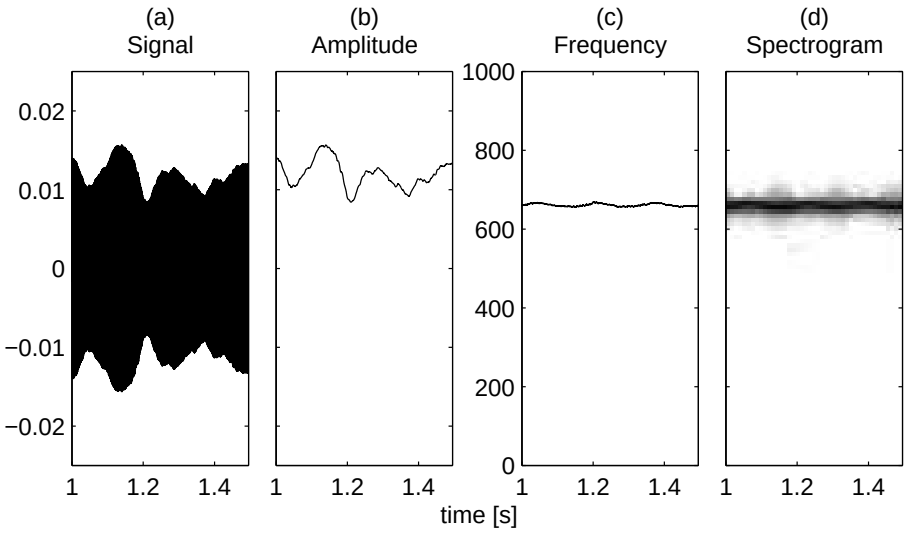


Figure 2.5: *Hilbert transform: narrow-band AM-FM estimation.*

Figure 2.5 shows this for one partial of an instrument note. In this simple case, the partial signal including sidebands was extracted from the total signal by means of a simple bandpass filter. The analytic signal was then obtained by use of DFT. Panel (a) shows the time domain signal for the extracted partial. The instantaneous amplitude and frequency are shown in Panel (b) and (c), respectively. For reference, a spectrogram of the extracted partial is shown in Panel (d). It can be seen how the instantaneous amplitude follows the envelope of the time signal. Similarly, the instantaneous frequency resembles the ridge of the partial in the spectrogram. However, both the amplitude and frequency are truly instantaneous sample by sample. On one hand, this means that the entire bandpass signal can be reconstructed perfectly. On the other hand, this means that if the signal contains more than one spectral component, the amplitude and frequency are virtually meaningless.

Time-variant filtering and frequency warping

While the described representation can yield perfect reconstruction for real signals, the amplitude and phase are only meaningful for narrow-band signals. This means that the methods must be preceded by some kind of spectral analysis that splits the entire signal into narrow-band signals, each containing one spectral component. When the frequency of the spectral component varies and when there are several closely spaced spectral components, this is not trivial.

One approach is based on instantaneous frequency warping. Wang (1994) tracks the instantaneous frequency of the individual spectral components using a frequency-locked loop (FLL). A FLL is a time-domain processing step that allows to track one single spectral component that is part of a more complex signal. Figure 2.6 shows an example of the extraction of one partial by use of FLL and instantaneous frequency warping. The technique is visualized

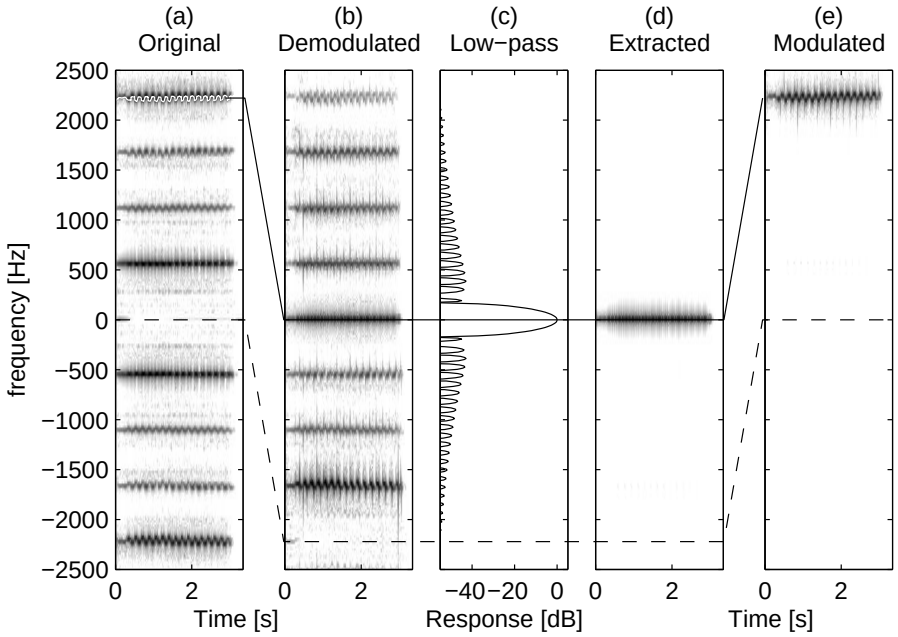


Figure 2.6: *Partial extraction by instantaneous frequency warping. (a): frequency-locked loop tracks instantaneous frequency of fourth partial (white line). (b): demodulating signal, yielding zero phase/frequency for the fourth partial. (c): time-invariant low-pass filter. (d): low-pass filtered signal yields the instantaneous amplitude. (e): modulating the fourth partial.*

in the time-frequency domain, but the actual processing is carried out in the time domain. Panel (a) shows the spectrogram of a harmonic note with vibrato (FM). A frequency-locked loop tracks the instantaneous frequency of the fourth partial (indicated with white line with FM). The estimated instantaneous frequency is used for demodulating the signal, yielding zero phase/frequency for the fourth partial. This is shown in Panel (b). All frequency variations of the fourth partial have been removed (no FM) and the partial can be extracted by a time-invariant low-pass filter, as shown in Panel (c). Panel (d) shows the extracted zero-phase partial. In the time domain, this signal is the instantaneous amplitude (for the given choice of LP-filter). Finally, Panel (e) shows the partial after modulating it again, using the estimated instantaneous frequency. This restores its FM cues.

A different approach using time-varying band-pass filters that tracks the individual harmonic components has also been proposed (Abe *et al.*, 1995). In principle, this is similar to the method based on frequency warping.

Discussion

Instantaneous AM-FM estimation based on instantaneous frequency warping and low-pass filtering can work on a single partial in more complex signals consisting of many partials. However, the frequency tracking (FLL) is quite

sensitive to noise. Whenever the instantaneous frequency is not tracked correctly, the performance degrades severely. If a signal is harmonic, it is possible to make the tracking more robust against noise by tracking several spectral components in parallel, using harmonic-locked loops (Wang, 1994).

When several spectral components are closely spaced in frequency, the tracking of the individual components is difficult. If the partial frequencies vary in time, the tracking may be erroneous or ambiguous. This is analogous to the problems seen in the peak detection and peak matching algorithms in sinusoidal models. If the frequency trajectories truly overlap or cross each other, the tracking of a single component may be impossible. The main drawback with source separation methods based on signal cues only is related to this. This is the topic of the following section that rounds off the discussion on separation techniques based on signal cues.

2.2.7 Separation based on signal cues only

Closely spaced and overlapping spectral components

Spectral components that are closely spaced in frequency can be very difficult to distinguish. In the STFT, the frequency resolution is directly proportional to the length of the analysis window. A long enough window may give a sufficiently good frequency resolution to detect the individual components. However, as the window gets longer, the frequency estimates are averaged over longer periods, and finer frequency fluctuations are lost. Moreover, the length of the window is limited by the duration of the original sounds. Several techniques for improved detection have been proposed.

Quatieri and Danisewicz (1990) proposed a method for resolving closely spaced partials. It assumes that the different components have constant frequency trajectories over the duration of the analysis window. A least squares (LS) algorithm in the spectral domain is used in order to resolve the different components. However, when the components are too close or coincide, this method becomes unstable (ill-conditioned matrix inversion) and interpolation is needed. Independently, at the same time, a very similar LS technique was proposed by Maher (1990). In addition to the LS method, for smaller frequency differences where this method is ill-conditioned, he also proposes a method that exploits the beatings introduced by the closely spaced components. Based on the assumption that the original components have constant frequencies over a time period long enough to observe the beatings, the difference between the original partials and the observed combined partial can be estimated. The frequency of the amplitude fluctuations, denoted *beating frequency*, corresponds to the frequency distance between the original components. The magnitude of the fluctuations (maximum and minimum of the envelope) relates to the relative strength of the original components. A similar technique based on a local nonlinear least squares method for the resolution of closely spaced sinusoidal components was proposed for the analysis-by-synthesis scheme (Tolonen, 1999).

Discussion

Source separation based on signal cues can work well when the different spectral components can be detected and modeled in isolation. Most existing techniques are based on this assumption. However, when several sources are concurrently active, their energy emissions in time and in frequency may overlap. In fact, the problem of overlapping partials is fundamental to source separation based on signal cues. If two or more spectral components truly overlap, no time-frequency analysis can yield a fine enough resolution in order to detect these in isolation. This is the fundamental principle of time-frequency uncertainty. A long analysis windows gives better frequency resolution, at the cost of averaging the frequency estimates over a longer time period. The methods discussed above can improve this but work only when the signals have some well known behavior. In particular, the methods assume that the partials have constant frequency over the duration of the analysis window. This limits the practical value of these techniques.

Whenever the individual spectral peaks cannot be detected in isolation, the estimates of instantaneous frequency and amplitude do not model a single partial but the combined effect of two or more partials. The partials thus modeled do not make much sense for the problem of source separation. All the methods that have been presented in this section are limited by this factor.

Rather than trying to accurately estimate the individual partial trajectories, we take a new approach where we represent the overlapping spectral components and their sidebands by a set of spectral coefficients that covers these. This is discussed in Chapter 4.1. In the further processing of these overlapping spectral components, spatial cues are employed.

As discussed in Section 2.1, the human auditory system performs a rather coarse spectral analysis of the signal observed by the ears. In many practical situations, multiple spectral components, possibly from different sources, may fall within one perceptual band. In this case the individual components cannot be distinguished by observing the output signals of individual auditory filters. This is known as the place theory of pitch perception (Moore, 1997). When different spectral components are closely spaced in frequency, this gives rise to beatings. It is possible that the human auditory system exploits the temporal variations in order to detect such cases. This is known as the temporal theory of pitch perception (Moore, 1997). The observation of beatings provides the motivation for a purely computational method that we devise in Chapter 5. In particular, the amplitude beatings are exploited in order to separate overlapping and crossing partials in multi-channel mixtures of harmonic instruments.

2.3 Inter-sensor analysis and spatial cues

So far, only cues that can be deduced from a single sensor signal have been considered. It is clear that the human auditory system can use a single ear for ASA, but the results are not as good as if both ears are available. By taking into account the interaural differences, i.e. differences between the two ear entrance signals, the performance can be significantly improved. In a general computational setup there may be a much higher number of sensors and, accordingly, the term inter-sensor analysis is used in the general case. The cues

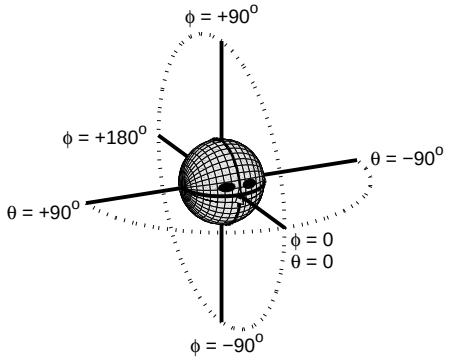


Figure 2.7: *Interaural polar coordinates. Grid of azimuth (θ) and elevation (ϕ) angles at 10 degrees resolution superimposed on a head. A given azimuth and elevation describes the line that passes through the center of the head and the corresponding point on the sphere.*

related to the differences between the sensor signals are very important for the localization of sources. As mentioned in the introduction, these cues are called spatial cues. In the following the two most important spatial cues are studied, namely time and level differences. The discussion is made with emphasis on binaural signals but some of the principles also apply to more general sensor arrays.

When a wave propagates from a source to either ear, it is subject to filtering due to reflections and shadowing in the physical environment surrounding the listener. In addition, the head and torso of the listener also causes reflections and shadowing. The effect of the head and torso on a signal coming from a particular direction is described by the head related transfer function (HRTF). For each ear, the HRTF is a filter whose transfer function is a function of the source location. In polar coordinates, the source location is described by the coordinate (θ, ϕ, ρ), where θ is the azimuth angle, ϕ is the elevation angle, and ρ is the distance from the source to the center of the head. In this thesis the interaural polar coordinate system, as shown in Figure 2.7, is used.

When considering the location of a source, there are two common simplifying assumptions that can be made. When the source distance is large relative to the distance between the ears, the sound waves reaching the head can be closely approximated as plane waves. This means that beyond a few meters, the source distance is of little importance for the interaural level and time differences. The other simplification is related to the region in space that is of most interest. In everyday life, many sources are in, or close to, the horizontal plane, i.e. zero elevation. For sources in this horizon of interest, the elevation is relatively small and can be neglected. This is the reason why source localization methods often only consider the estimation of azimuth angle, or direction of arrival (DOA).

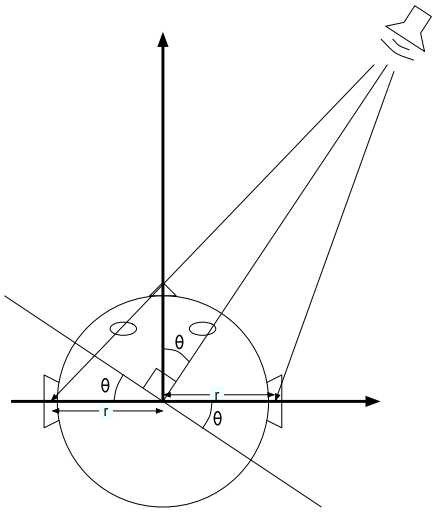


Figure 2.8: *Free-field wave propagation, or transparent head, with sound waves propagating through the head.*

2.3.1 Time differences

One of the simplest model of the head is a sphere where the ears are located on each side with a distance $d = 2r$ between the ears. The length of the wave propagation path between a source and an ear depends on the source location and is, in general, different for the two ears. If the head were transparent, the physical setup would correspond to two sensors in free-field. Under the assumption that the source distance is relatively large compared to the distance between the sensors, i.e. $\rho \gg d$, the distances between the source and sensors are approximately $\rho_1 = \rho + \Delta\rho$ and $\rho_2 = \rho - \Delta\rho$, for the left and right sensors, respectively. By simple geometric considerations, $\Delta\rho \approx r \sin\theta$, as seen in Figure 2.8. In this free-field case, there is no dependency on elevation. The difference in arrival times for this simple scenario is given by

$$\Delta T = \frac{\rho_1 - \rho_2}{c} = \frac{2r \sin\theta}{c}, \tag{2.17}$$

where c is the wave propagation speed. This is similar to the "sine law" proposed by von Hornbostel and Wertheimer (1920)

$$\Delta T = \frac{\kappa \sin\theta}{c}, \tag{2.18}$$

where $\kappa = 21cm$. The value κ does not represent the actual distance $d = 2r$ between the two ears. However, d and κ are related by the empirical formula

$$\Delta T = \frac{\kappa' d \sin\theta}{c}, \tag{2.19}$$

that was later introduced, with $\kappa' = 1.2 - 1.3$. In this simple empirical formula ΔT is directly proportional to the free-field value (2.17). The proportionality

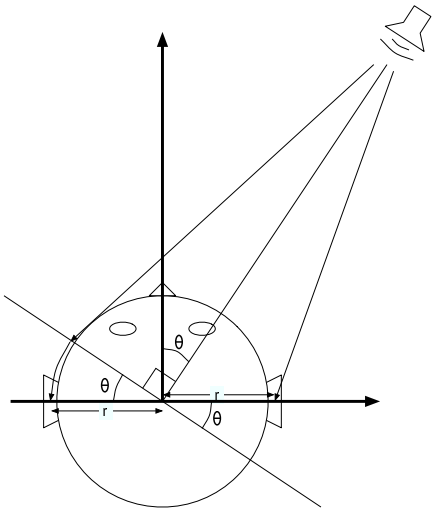


Figure 2.9: *Head shadowing, with sound waves propagating around the head.*

factor κ' accounts somewhat for the facts that the head is not perfectly spherical and that the ears are located slightly towards the back.

However, the effect of shadowing by the head is ignored in these formulas. For a source located towards one of the sides, the propagation path is a straight line to the closest ear. For the sensor on the opposite side, the wave needs to propagate around the head. The length of this propagation path is longer than in the free-field case. Woodworth and Schlosberg (1954) proposed a model that takes this into account:

$$\Delta T = \frac{r(\sin\theta + \theta)}{c}. \tag{2.20}$$

The geometry of this model is shown in Figure. 2.9. Also in this case, the head is considered to be spherical and symmetrical with one ear on each side. Furthermore, the formula is therefore of elevation.

In reality, the head is not purely spherical and, as mentioned, the ears are located slightly towards the back. More sophisticated models take this into account. However, each individual head is different. The advantage of the simple models presented in this section is the fact that they capture the principal effect of wave propagation around a generic head. More complex models exist, that can better approximate a particular head. However, when applied to a different head, these models are not necessarily more accurate than the simple models presented here.

2.3.2 Level differences

The effect of head shadowing not only affects the length of the propagation paths, but also effects the level of the signal that reaches the ear. This shadowing is highly frequency dependent. For instance, at low frequencies, the wavelength is of the order of the size of the head. Consequently, the head

shadowing has little effect. For higher frequencies, the head shadowing becomes highly significant and can cause level differences of several tens of dB between the two ears. The dependencies on source position and frequency is rather complex and there are, to our knowledge, no simple models that try to describe this in general. However, the very fundamental behavior can be described intuitively. In general, a source that is located to one of the sides of the head yields a stronger signal at the ear facing the source than at the ear on the opposite side. Also, the larger the level difference is between the two ears, the farther to the side the source is located. Beyond these simple considerations, the use of level differences in computational models becomes much more difficult. In ASA, it is likely that the human auditory system is trained to the particular dependency on frequency and location for the individual head.

2.3.3 The duplex theory

The two different types of spatial cues that have been discussed, namely time and level differences, provide important information that can be used in grouping auditory streams. Clearly, the level and time differences convey similar information, namely that of the direction of a source. Since the two cues are related to the same information, it is not surprising that the human auditory system combines them in some way. Almost a hundred years ago, Lord Rayleigh (Strutt, 1907) studied the relative importance of interaural level and time differences as a function of frequency, for the localization of narrow-band sources. He found that, at low frequencies, e.g. below about 500 Hz, the level difference between the ears is negligible and the localization is dominated by the time differences. At higher frequencies the level difference is the primary cue. These results are well known as the *duplex theory*. Later experiments have shown similar results for the localization of pure tones. Psychophysical experiments have shown that the role of the time difference is limited to frequencies below about 1.5 kHz. It is then natural to consider the importance, or weighting, of the level and time differences for different frequencies. Traditionally, various trading ratios have been proposed in order to describe the weighting of the cues, e.g. relating level (in dB) and time (in μsec) differences to each other, or relating them to the perceived source locations. Dimensionless weights have been proposed (Macpherson and Middlebrooks, 2002). For wide-band sound sources the picture is more complex and the duplex theory does not apply in general. In particular, the time differences play a role at higher frequencies too. However, for mixtures containing several sources with intermixed spectral component, processing across frequency bands in computational models is not trivial.

In computational methods for the localization and the separation of individual narrow-band signals, the duplex theory is still interesting. In Chapter 3 we propose a simple empirical model for joint evaluation of level and time differences in narrow-band binaural signals. The model is in correspondence with the duplex theory at low and high frequencies, and also copes with the transition from low to high frequencies.

2.3.4 Dynamic spatial cues

As was the case for temporal and spectral cues in Section 2.1, one can distinguish between static and dynamic cues. In a static scene, where neither the sources nor the sensors move, level and time differences are static. When sources start to move, in general, source localization becomes more complicated.

For a perfectly spherical head, with the sensors diametrically opposite, the shadowing of the head does not depend on the elevation. In this discussion only the head shadowing is considered, while the effects of the torso and ears are ignored. This means that, for a given azimuth and distance, all elevations give rise to the same level and time differences between the two ears. Under the far-field assumption, the dependency on distance is small and consequently there is an entire hyperplane, shaped like a cone, on which all positions yield the same interaural cues. Clearly, as the source distance gets smaller, the assumption about plane waves no longer holds and the hyperplane deviates from the cone. Similarly, if one takes into account the real shape of the head and ears, as well as of the torso, the positions yielding equal interaural cues differs from the ideal cone. However, due to the symmetrical properties of the head and ears, the general shape is still "cone-like". This hyperplane is called the *cone of confusion*. In other words, there exists a large number of possible positions yielding the exact same level and time differences between the ears. For example, a source directly in front, a source directly above, and a source directly behind a listener all have zero level and time difference between the ears (for a symmetric head). In a static scene, a listener may not be able to distinguish between these source positions. However, as a matter of fact, the human auditory system can exploit head movements in order to improve the localization. The listener performs conscious and subconscious movements of the head (sensors), and even small movements can help resolve ambiguities. By rotating the head slightly, the front source will cause a level and time difference (reaching one ear before the other), the source behind will yield a similar change, but with opposite sign (reaching the other ear first), and the source above retains zero level and time difference. Blauert (2001) explores the types and magnitudes of the head movements in more details. Needless to say, in ASA the dynamics in the scene can be of help rather than representing an additional difficulty.

2.4 Separation by spatial cues

In a situation where there is more than one sensor, the inputs are described by the sensor signals $x_m(l)$, where $(1 \leq m \leq M)$ is the sensor index and l is the time index. In a general setup with M sensors and N sources, each sensor records a superposition of filtered source signals. This can be modeled as

$$x_m(l) = \sum_{n=1}^{N} h_{mn}(l) * s_n(l), \tag{2.21}$$

where x_m denotes the signal at sensor m, s_n are the source signals and h_{mn} are the filters modeling the sound propagation from source n to sensor m, including

the direct path and room reflections. In matrix notation, this can be written as $\mathbf{x} = \mathbf{h} * \mathbf{s}$:

$$\begin{bmatrix} x_1(l) \\ \vdots \\ x_M(l) \end{bmatrix} = \begin{bmatrix} h_{11}(l) & \cdots & h_{1N}(l) \\ \vdots & \ddots & \vdots \\ h_{M1}(l) & \cdots & h_{MN}(l) \end{bmatrix} * \begin{bmatrix} s_1(l) \\ \vdots \\ s_N(l) \end{bmatrix}. \tag{2.22}$$

When the signals are represented in the STFT domain, the convolution operator is replaced by multiplication. The mixing model can then be written in matrix notation as $\mathbf{X} = \mathbf{HS}$:

$$\begin{bmatrix} X_1(k,q) \\ \vdots \\ X_M(k,q) \end{bmatrix} = \begin{bmatrix} H_{11}(k,q) & \cdots & H_{1N}(k,q) \\ \vdots & \ddots & \vdots \\ H_{M1}(k,q) & \cdots & H_{MN}(k,q) \end{bmatrix} \begin{bmatrix} S_1(k,q) \\ \vdots \\ S_N(k,q) \end{bmatrix}, \tag{2.23}$$

where $X_m(k,q)$ and $S_n(k,q)$ are the STFT spectra of the source and sensor signals, respectively, and $H_{mn}(k,q)$ are the STFT spectra of the mixing filters h_{mn}. These are usually assumed to be time-invariant. In this general framework, source separation is often considered equivalent to the problem of estimating the mixing matrix $\mathbf{H}$ or its inverse.

2.4.1 Beamforming

Beamforming techniques are some of the oldest techniques employed for the separation by spatial cues in multi-channel mixtures. They do not try to estimate the mixing matrix (2.23) but rather make use of the spatial information explicitly in designing directional filters. While these techniques are not truly separation techniques, they are able to enhance signals coming from particular directions. Originally, these techniques were developed in analog radar and sonar systems in order to focus a beam in a particular direction. In audio, similar techniques can be employed in order to form "listening" beams in space. Digital beamformers built on such principles are known as delay-and-add beamformers (Frost, 1972). Several sensors are placed equidistantly along a line. Depending on the target direction of arrival, i.e. in what direction the beam should point, the signal observed at each sensor is delayed in order to provide in-phase signals at all sensors. After this delay correction, the sensor signals are summed together, yielding a spatially filtered signal.

Figure 2.10 shows the spatial response for a signal at a given frequency as a function of the arrival angle θ. The three panels show the responses for different numbers of sensors and for different target angles. The number of sensors are 8, 16, and 24, and the target angles are 0, -45, and -30 degrees, respectively. Beamformers are often also called spatial filters.

The spatial response of beamformers can also be represented as a function of the inter-sensor delay, $d \sin\theta$, cf. (2.17) and Figure 2.8, as opposed to the angle of arrival θ. In that case, the narrow band (single frequency) delay-and-add beamformer with equidistant sensors, as shown in Fig. 2.10, is analogous to an FIR filter in the time domain. The FIR filter response is a function of frequency (temporal frequency), and the spatial response of the beamformer is a function of the inter-sensor delay. This analogy means that existing FIR filter design techniques can be easily applied in beamforming. For instance, all

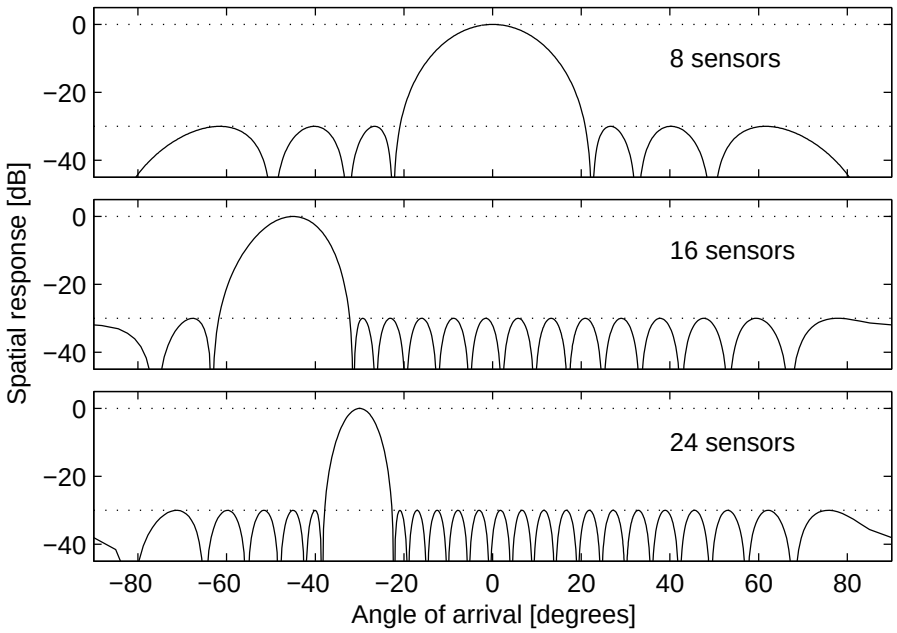


Figure 2.10: *Spatial responses of beamformers for different numbers of sensors and different target angles.*

the beamformer responses shown in Figure 2.10 were designed using Chebychev filters with constant side lobe attenuation of 30 dB. It can also be seen how the width of the main lobe (in space) is related to the filter order (number of sensors). Since the graphs in Figure 2.10 are plotted as functions of arrival angle rather than delay, the axis is warped.

More sophisticated beamformers have been proposed. The generalized side lobe canceler (Griffiths and Jim, 1982) employs different beamformers in parallel, where the additional beamformers aim at canceling the side lobes of the main beamformer. Other beamforming techniques employ filter banks. In particular, the STFT enables easy delay compensation and individual filtering of the different frequency bands. A more thorough overview of beamforming techniques is given in (van Veen and Buckley, 1988).

However, the spatial response of these beamformers highly depends on the signal frequency and the sensor spacing. In fact, the relation between the maximum delay between neighboring sensors (i.e. the sensor spacing d when the full range of directions from -90^o to 90^o is considered) and the wavelength of the signal is of high importance. This relation can be interpreted as sampling in space. The traditional sampling theorem (temporal sampling) states that a signal can only be exactly represented if the highest (temporal) frequency in the signal is less than half the (temporal) sampling rate f_s, i.e. $f_{max} < f_s/2$. This means that the shortest wavelength is twice the distance a wave propagates in one sample interval, $\lambda_{T,min} \geq 2\frac{1}{f_s c}$. If the signal contains frequencies above the Nyquist frequency $f_s/2$, spectral aliasing occurs. Analogously, the spatial "Nyquist frequency" is the frequency whose wavelength is twice the sensor dis-

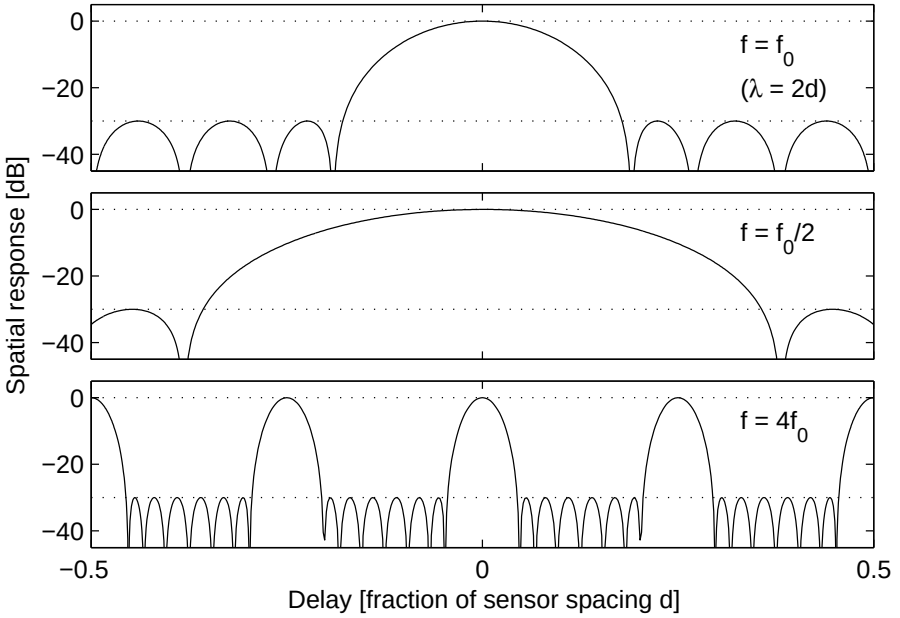


Figure 2.11: *Beamforming for target angle of 0 degrees, as function of signal frequency. Top: Spatial filter designed for a given frequency, f_0. Middle: At lower frequencies the spatial resolution is much lower. Bottom: At higher frequencies spatial aliasing occurs, seen as grating lobes at angles different from the target angle.*

tance d, i.e. $\lambda_{S,min} \geq 2d$. If the signal frequency is higher than this frequency, spatial aliasing occurs. For a fixed sensor spacing, when the signal frequency changes, this is analogous to changing the sampling rate in temporal sampling. This means that if a spatial filter has been designed for a given frequency, lower frequencies give a poorer spatial resolution and higher frequencies may cause spatial aliasing.

Figure 2.11 shows the spatial response of the beamformer with eight sensors, as seen in the top panel of Figure 2.10, but as function of inter-sensor delay. The top panel shows the response for a signal of frequency f_0, for which the beamformer was designed, i.e. $\lambda = 2d$. The middle panel show the response for a signal of half the frequency $f_0/2$. In the FIR analogy, this corresponds to doubling the sampling frequency, or up-sampling by a factor of two. Clearly, the main lobe is twice as wide. The bottom panel shows the spatial response for a higher frequency $4f_0$. The spatial aliasing is seen as multiple replica of the main lobe at different angles, called "grating lobes". Intuitively explained, the aliasing is caused by "phase wrapping". This means that the phase difference between two sensors is bigger than π (in magnitude). Different angles that give the same phase difference between the sensors (up to an integer multiple of 2π) cannot be distinguished and yield the same response.

For the application to separation of wide band audio sources, all beamformers with equidistant sensors suffer from being narrow-band by nature. In fact, for a fixed uniform sensor spacing, it is not possible to have main lobes with

fixed width over several octaves. The design of wide band beamformers has been considered by several researchers. Some methods and further references can be found in (Goodwin and Elko, 1993) and (Gerven *et al.*, 1995). In general, wide band beamformers with more homogeneous beams lead to non-uniform sensor spacing, such as logarithmic spacing (Brandstein and Ward, 2001).

Discussion

Beamforming is a powerful tool for the separation of sound sources. However, a rather large number of sensors is needed for good spatial resolution. Drake (2001) has derived a mathematical measure for comparison of the performance of beamformers with that of the human auditory system. The conclusion is that a high number of sensors is needed in order to match the performance of the human auditory system which is based only on two sensors (ears).

The human auditory system performs surprisingly well with its only two sensors. One reason for this may be that the human auditory system also may make use of signal cues as discussed in Section 2.2. The use of signal cues in the context of beamforming has been studied (Drake, 2001), but with only little interaction between signal cues and spatial cues. Another explanation why beamformers need a much higher number of sensors is that the sound waves are assumed to be planar, i.e. the distance between the sensors is much smaller than the distance to the sources. Beamformers have typically been designed for sensors in free-field. This means that there is virtually no intensity differences between the sensors and the filters of the mixing matrix (2.22) are pure delays.

2.4.2 Statistical methods

A different group of techniques for separation based on spatial cues works by exploiting the statistics of the source signals. These techniques are called statistical methods. Most of these aim at estimating the mixing matrix (2.22) by optimizing some statistical measure. For separation, the sensor signal vectors are left multiplied by the "inverse" of this matrix in order to obtain the original source signals. Since the elements of the mixing matrix are FIR filter an not scalars, the "matrix inversion" involves FIR filter matrix algebra (Lambert, 1996). In general, a mixing matrix can only be inverted if the rank of the matrix is larger than or equal to the number of sources. This means that there must be at least as many sensors as there are sources and the different sources must have different mixing filters (independent columns in the mixing matrix). First, some techniques for separation of instantaneous mixtures (i.e. the elements of the mixing matrix $\mathbf{H}$ are constants) are reviewed. Then the more general problem with convoluted mixtures is discussed. It is shown how the mixing problem in the time domain can be split up into a set of instantaneous mixtures in the frequency domain.

Principal component analysis

Principal component analysis (PCA) is a statistical technique for the analysis of multi-dimensional data. It is known in some fields as the Karhunen-Loeve transform or the Hotelling transform. Starting with an M-dimensional data set such as the sensor signals $\mathbf{x}$ as function of time index l (2.22), the vector in

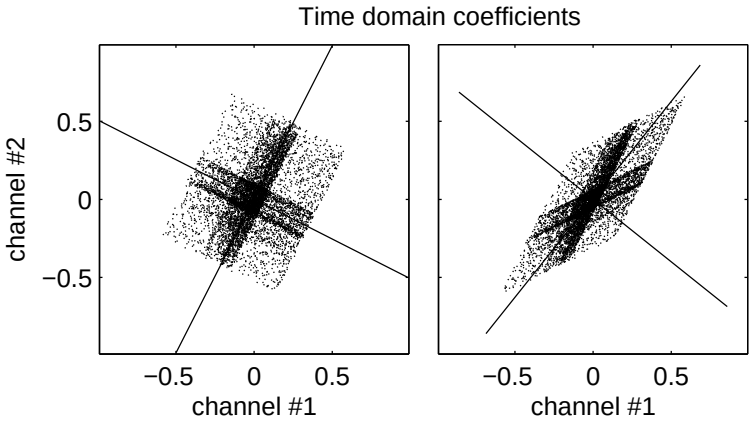


Figure 2.12: *Principal component analysis for two-channel mixtures of two sources: sampled amplitudes of the two sensor signals plotted against each other for two different cases.*

this space that accounts for the maximum amount of variance is found. The projection of the data set onto this vector is then subtracted from the original data set, reducing the dimension by 1 and the same technique is repeated recursively, yielding orthogonal vectors representing decreasing order of variance. These vectors are the eigenvectors of the covariance matrix of $\mathbf{x}$ and the principal components are the projections of the data set onto these vectors.

PCA is not commonly associated with audio source separation. Rather it is a powerful technique for dimension reduction in high dimensional data sets. It has mainly been included in this discussion for illustrational purposes, in order to provide an intuitive understanding of some of the techniques that are presented later in this chapter.

Figure 2.12 shows the scatter plot of a two-channel mixture of two different instrument notes. The amplitudes of the samples of the first channel $x_1(l)$ are plotted against those of the second channel $x_2(l)$. For the left panel the columns of the mixing matrix were $[1, 2]^T$ and $[2, -1]^T$, respectively. The directions of these column vectors are clearly seen as the dark cross in the graph within which most data points are located. Applying PCA on this data set, the two directions indicated by the solid lines are found. These correspond to the column vectors of the mixing matrix. The line from the lower left to the upper right corner corresponds to the eigenvector with highest eigenvalue, i.e. describing the most variance. Each of these lines corresponds to one source, and each source can be separated out of the mixture by projection onto the corresponding vector.

While the signal can be properly separated in this case, this is not true in general. The fact that all the eigenvectors are orthogonal makes this technique useless for most mixtures, except artificial constructions where the columns of the mixing matrix are orthogonal, as was the case in this example. The sign inversion of one coefficient in the second column is not very probable to occur in a real situation. A more realistic setup is shown in the right panel, with mixing matrix columns $[1, 2]^T$ and $[2, 1]^T$. Again, these vectors can be seen in the graph, but the PCA analysis clearly fails to detect them, and is

consequently useless for separation purposes.

Even though this method is of little use for the separation of audio signals, this discussion gives a geometrical interpretation of the separation problem that can be useful in the following discussion of other techniques.

Independent component analysis

Similarly to PCA, independent component analysis (ICA) deals with instantaneous mixtures and tries to estimate the mixing matrix by means of optimization of some statistical measure. Rather than looking for the principal components accounting for most of the variance, ICA aims at finding the components that are most independent, taking into account higher order statistics. Over the last decade, several methods for ICA have been developed in different areas of research. Although slightly different approaches exist, most methods are based on the assumption that the sources are statistically independent. The separation problem consists of finding the mixing matrix $\mathbf{H}$ or its inverse. Whether it is the mixing matrix or its inverse that is to be estimated, an iterative scheme is usually used. We call the matrix to be estimated $\mathbf{W}$. In general, at each iteration i, the previous estimate is updated by taking into account new mixture samples (in time). Without going into details, this can be written as $\mathbf{W}_i = \mathbf{W}_{i-1} + \xi f(\mathbf{W}_{i-1}, \mathbf{x})$, where ξ is the adaptation rate, $\mathbf{W}_i$ is the estimated matrix at step i, $\mathbf{x}$ is the sensor signal vector, and f is some function. A large variety of functions f have been proposed, in different approaches. ICA was pioneered by Jutten and Herault (1991). Inspired by neural networks, odd non-linear functions of $\mathbf{x}$ were used for updating of the non-diagonal elements of $\mathbf{W}$. The updating rule was developed for a least squared error measure and the nonlinearities served to give statistical independence (higher order statistics, not only the variance as in PCA). Later, similar algorithms were developed in other fields. From an information theoretical approach Bell and Sejnowski (1995) developed an updating rule based on nonlinear functions related to the cumulative density functions of the signals. In statistics nonlinear functions directly related to higher order statistics of the signals are employed (Comon, 1994). More thorough overviews of the vast number of existing techniques can be found in (Torkkola, 1999) and (Hyvärinen, 1999).

Figure 2.13 shows the result that can be obtained by ICA when applied to the same example as was shown for PCA in Figure 2.12. The correct independent components are detected in both cases, and perfect separation is achieved by left-multiplying the sensor signals with the inverse of the mixing matrix. In a general mixture, the ICA algorithms can find the independent components also when they are not as "visible" as in this figure, which is the case for most mixtures.

Multi-channel blind deconvolution

In many real mixtures of audio sources, the mixing filters $h_{mn}(l)$ in (2.21) are neither pure delays, as in the beamformers, nor pure scaling factors, as in PCA and ICA. ICA has been extended to deal with mixing matrices having FIR filters as elements. These techniques are called multi-channel blind deconvolution (MCBD) techniques. Similarly to ICA, usually iterative algorithms are used for updating the elements of the mixing matrix, i.e. the coefficients of the

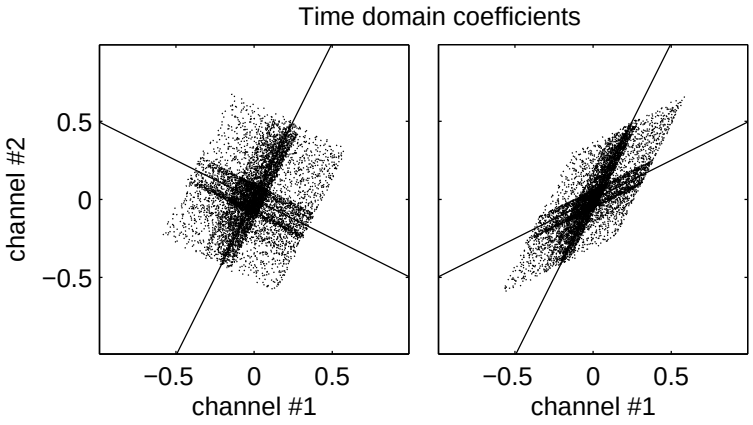


Figure 2.13: *Independent component analysis for two-channel mixtures of two sources: sampled amplitudes of the two sensor signals plotted against each other for two different cases.*

FIR filters $h_{mn}(l)$ (Lambert, 1996; Bell and Sejnowski, 1995; Torkkola, 1996; Lee *et al.*, 1997). The computational complexity of these methods is very high. In fact, for realistic scenarios, the mixing filters may be quite long and are excessively computationally complex.

For very long filters, even small changes in the scene can severely affect the performance. In a static scene where the correct mixing matrix has been estimated, even small source movements of a few centimeters can degrade the performance. The effect of uncertainty of the source positions has been evaluated for parametric source demixing (room modeling) in (Balan *et al.*, 2001).

Frequency domain ICA

By use of the STFT, the temporal convolution is replaced by multiplication (2.23). This was exploited by Smaragdis (1997) in order to turn the problem with convolved sources into a set of problems of complex valued instantaneous mixtures, one for each spectral coefficient of the STFT.

In practice, the analogy between time-domain convolution and frequency domain multiplication only holds when the window length in the STFT is (much) longer than the length of the mixing filters. For shorter window lengths, the frequency domain multiplication yields a circular convolution. This makes frequency domain ICA less useful for reverberant environments where the mixing filters can be in the order of hundreds of milliseconds and longer (Araki *et al.*, 2003; Balan *et al.*, 2001).

In addition, the separated signals for each spectral coefficient are ordered differently, introducing a permutation ambiguity. This makes it non-trivial to put together the separated narrow-band signals across the frequency bands. Some methods for resolving these ambiguities exist, based on temporal envelopes (Ikeda and Murata, 1999) and the statistics of these (Mitianoudis and Davies, 2003). The use of beamforming techniques for resolution of these permutation ambiguities has also been proposed (Parra and Alvino, 2002). Other

authors have pointed out the relation between frequency domain ICA and beamforming (Araki *et al.*, 2002).

ICA is based on the assumption that the sources are statistically independent. When narrow-band partials of different harmonic instruments overlap, the signals in the corresponding bands of the STFT may not satisfy this assumption. This is discussed further in Chapter 5, where we present a different method for the estimation of the mixing matrix, based on similarity of temporal envelopes.

Discussion

As mentioned in the introduction, many techniques for the separation of sources are, in a practical sense, better described as source enhancement techniques. Statistical methods are among the few methods that can truly separate sources in the mathematical sense, under some assumptions. If all the assumptions are satisfied, the unknown mixing matrix can be found and inverted, yielding perfect separation. Statistical methods have proved to be useful in other domains than audio. However, for the separation of audio signals, there are several restrictions related to the assumptions on which these methods are based.

Statistical methods assume that the mixing matrix is time-invariant. In other words, these methods are not able to deal with dynamic scenes where sources and sensors move. This is in strong contrast to the human auditory system which can actually exploit such dynamic changes, as discussed in Section 2.3.4. This constraint is related to the convergence speed and behavior of the iterative algorithms that these methods are based on. Therefore the statistical methods are of little value for audio signals in many real-world situations.

Another restriction of statistical methods is that, in general, there cannot be more sources than sensors. This is comparable to the performance of beamforming, where few sensors yield poor spatial resolution. This is a rather severe constraint in comparison to the performance of the human auditory system that can distinguish many sources with only two ears.

2.4.3 Time-frequency weighting

The initial motivation of this thesis was to devise source separation techniques inspired by the human ability to detect and focus on one single sound source in a complex auditory scene. However, it is not clear whether the auditory system actually separates the different sources or whether it enhances the source(s) of interest while suppressing the other sources. Methods exist that aim at modeling the peripheral processing of the human auditory system. Other methods take into account higher level perception in order to focus on sources of interest. While these methods are not strictly "separation" methods, their goals are similar to the goals of source separation methods.

Denoising and source enhancement motivated by auditory scene analysis (ASA)

Lyon (1983) presented a computational model for binaural localization and separation of sources. Based on similar ideas, Bodden (1993) later proposed his

cocktail-party processor, a computational method for enhancement (or separation) of sources in binaural signals. This method is based on models of how the human auditory system localizes sources. The sensor signals are decomposed into perceptual bands. Each band is then processed by a model of the processing of the inner ear. In the resulting approximation of the neural excitation, interaural cross-correlation and extended processing algorithms, modeling the human auditory system (Lindemann, 1986; Gaik, 1993), provide spatial cues. The peaks in the overall neural excitation as a function of azimuth represent candidate source positions. Windowing in the spatial cue space around the location of one source yields a time-varying weighting factor for each band. The extracted source is built as the sum of weighted perceptual band signals. This method can be seen as non-linear beamforming. In fact, with only two sensors, it can perform narrow beamforming over the whole audible frequency range.

Similar methods based on the STFT and grouping of STFT bands into perceptual bands were later proposed (Peissig, 1992; Wittkop *et al.*, 1997). These methods are based on similar principles, but use standard mathematical tools rather than trying to mimic the processing of the human auditory system. The spatial cues are estimated by short-time auto- and cross-correlation in the STFT bands. The weighting factors are computed as a function of the coherence and phase and level differences between the two channels. Different variations of this technique exist. An overview is presented in (Wittkop *et al.*, 1997).

Time frequency masking

A very similar approach was later proposed in a more theoretical framework, called the Degenerate Unmixing and Estimation Technique (DUET) (Jourjine *et al.*, 2000; Yılmaz and Rickard, 2002). Rather than weighting the signals in perceptual bands, this method is based on assumed disjointness of the sources in the STFT representation. In the discussion of signal cues in Section 2.1 it was seen how the STFT provides a much sparser representation of signals than the time-domain representation. This means that many types of sources can be closely approximated by a small number of spectral coefficients. If the different sources do not overlap, i.e. if each spectral coefficient only has significant energy contribution from one source, the sources are said to be disjoint. Obviously, the notions of overlap and disjointness are highly dependent on the choice of time-frequency representation for the signals. Different choices of analysis window and window length in the STFT yield different degrees of overlap. However, due to the time frequency uncertainty, the sources will always overlap to some extent, except for trivial cases or artificially constructed signals. In practice, however, disjointness of the sources may hold to a high degree for some types of sources, such as speech sources (Rickard and Yılmaz, 2002).

If each spectral coefficient contains significant energy from only one source, separation is easily obtained by simply assigning the individual spectral coefficient to its corresponding source. This is the idea behind the DUET method. For each spectral coefficient, spatial cues are computed between the two sensor channels. The phase difference between the two channels is

$$\Delta P(k,q) = \arg \frac{X_1(k,q)}{X_2(k,q)}, \tag{2.24}$$

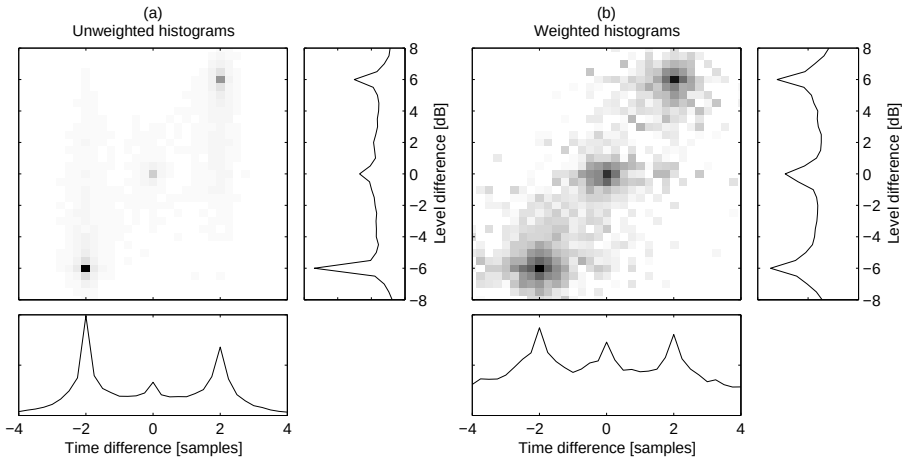


Figure 2.14: *Two-dimensional histograms of spectral coefficients in spatial cue parameter space. (a): Unweighted, linear scale. (b): Weighted by the magnitude of the spectral coefficients, logarithmic scale.*

and the time difference between the channels is given by the phase delay difference:

$$\Delta T(k,q) = -\frac{\Delta P(k,q)}{q}\frac{L}{2\pi}. \tag{2.25}$$

Similarly, the level difference is given by

$$\Delta L(k,q) = \left|\frac{X_1(k,q)}{X_2(k,q)}\right|. \tag{2.26}$$

Based on the statistics of these cues, the source locations are estimated by peak picking in spatial cue parameter space. Each spectral coefficient is then assigned to the source that lies closest in this space. This is equivalent to the original time-frequency weighting, but with the weights restricted to a binary decision (0 or 1). For each source, this yields a binary STFT mask that indicates which spectral coefficients belong to a specific source and which do not. Separation of each source is achieved by multiplying the sensor signal STFT spectra with the corresponding mask. The time domain signals are obtained by applying the inverse STFT with overlap-add.

Figure 2.14 shows statistics of the spatial cues for a two-channel mixture of three sources. Level and time differences between the two sensor channels were computed for each spectral coefficient by use of equations (2.25) and (2.26). The large figures show two-dimensional histograms of these. The marginal one-dimensional histograms are shown just below (time difference) and to the right (level difference). Panel (a) shows the unweighted histograms. The number of spectral coefficients whose spatial cues fall within a given range of time and level differences are counted and shown on a linear scale. Panel (b) shows the weighted histogram, where each spatial cue has been weighted with the magnitude of the corresponding spectral coefficient. The weighted histograms are shown in logarithmic scale. The three prominent peaks in these histograms

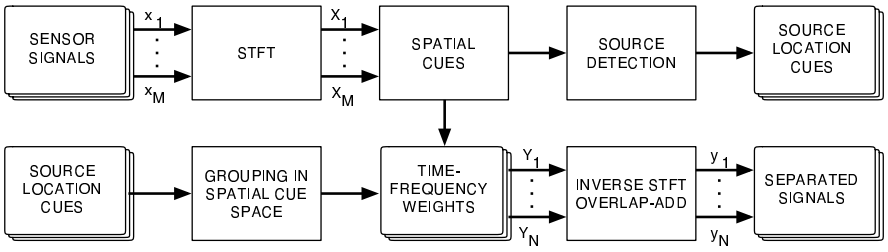


Figure 2.15: *Source separation using spatial cues. Top row: Estimation of spatial cues. Bottom row: Separation by grouping and time-frequency weighting.*

correspond to the different sources. In this spatial cue parameter space the locations of these peaks have a very intuitive interpretation with respect to the physical location of the sources in relation to the sensors. For the artificial mixture shown in this figure, the weighted and unweighted histograms convey similar information. For real-world mixtures the use of these histograms is discussed in more detail in Chapters 3 and 4.

The DUET method is based on two assumptions. The first assumption is that the sources are "W-disjoint orthogonal". This means that for a chosen windowing function (W), the overlaps between the source STFT spectra are negligible. The second assumption is that any delay between the sensors can be (uniquely) described as a phase shift. This is called the "narrow band assumption" in array processing. In practice this means that the distance between the sensors must be shorter than half the shortest wavelength.

Discussion

The time-frequency weighting methods have in common that the signals are represented in time and in frequency and the different coefficients in this representation are weighted based on spatial cues. The general processing steps are visualized in Figure 2.15. The top row shows the estimation of source locations based on the statistics of spatial cues. The bottom row shows the combination of these source location cues with the spatial cues of the individual spectral coefficients in order to determine the weights. Separation is obtained by weighting of the sensor signal spectra, followed by inverse STFT with overlap-add.

Similarly to beamforming techniques, these methods are not truly "separating" the sources. In fact, they can be seen as spatial filters, or generalized beamformers. Nevertheless, these techniques are extremely versatile, and exhibit several advantages:

- There can be more sources than sensors.
- The methods can work with binaural signals yielding narrow beams with only two sensors.
- The spatial cues and their histograms have an intuitive interpretation.
- Computations are fast since the methods can be based on the STFT.

However, there are some difficulties that must be considered in practical applications, in connection with the two assumptions on which the DUET method is based. These are strongly related to the estimation of the level and time differences.

Narrow band assumption: In practice, this assumption is a severe constraint. If the full range of audible frequencies is considered, i.e. up to about 20 kHz, this means that the sensors must be less than 1 cm apart. While this may be fine (using small microphones), it has some implications on the grouping in spatial cue parameter space. In fact, with such a small sensor spacing, the level difference between the sensors is virtually zero. In this case, all the histogram peaks lie on the horizontal line corresponding to zero level difference. By choosing a very large sensor spacing, the opposite happens. The level differences provide useful information, whereas the time differences are useless due to the phase ambiguity. The histogram would in this case center around the vertical line of zero time difference. These two cases are just the extreme cases for sensors placed in free-field. In binaural signals both the level and time differences are important, as discussed in Section 2.3.

With the "sensor spacing" of the ears, the narrow band assumption only holds below about 1500 Hz. The drawback of this method for binaural signals is that level and time differences are treated independently. In practice, the level and time differences convey similar information. When a source signal reaches one sensor before the other sensor, it usually is stronger also at that sensor. In other words, the level and time differences are not completely independent. This means that in most natural situations, the main energy in the histograms will be restricted to a subregion of the level/time difference parameter space, e.g. along the diagonal as seen in Figure 2.14. This subregion corresponds to a curve in this parameter space. If the curve is parametrized, the estimation of source locations and time-frequency weights can be performed in this one-dimensional parameter space. In Chapter 3 we devise an empirical model for the human head. This model relates the level and time differences to the azimuth angle, or direction of arrival (DOA) in the horizontal plane, as functions of frequency. This yields a parametrization of the curve in the spatial parameter space, where the positions along the curve corresponds to the DOAs. In particular, by joint evaluation of level and time differences, the phase ambiguities can be resolved. This effectively enables narrow band processing (individual spectral coefficients) in binaural signals.

W-disjointness: The other practical challenge with time-frequency weighting is related to the disjointness of the sources, i.e. to what degree their STFT representations overlap. In the example in Figure 2.14, the strong peaks corresponding to the source locations are easily observable since almost all spectral coefficients yield consistent spatial cues. This is due to the fact that the chosen sources are disjoint in the STFT representation, at least to a high degree of approximation. For a spectral coefficient where one of the sources has a strong partial, the other sources have little energy. Effectively, the level and time differences are not corrupted by other overlapping sources. In more complex mixtures, where the sources

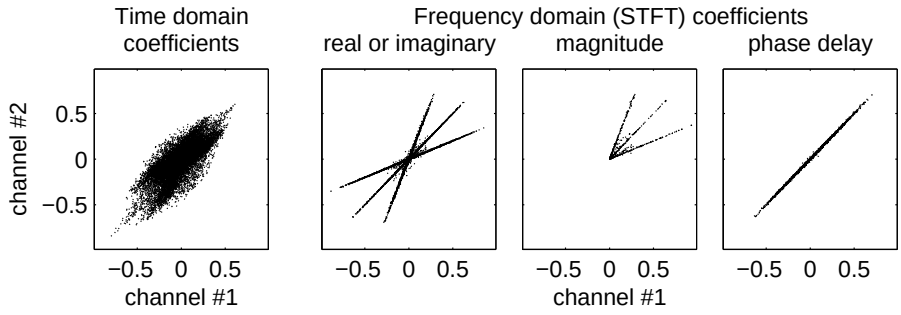


Figure 2.16: *Instantaneous, underdetermined mixture: sparsity in time and STFT domains.*

overlap to some extent, this yields spatial cues also between the peaks (i.e. cues not corresponding to any of the source locations). This makes both the detection and the separation of sources more difficult. Our work towards a solution to this problem is presented in Chapters 4 and 5.

2.4.4 Underdetermined mixtures and sparse decompositions

As seen in the last section, the time-frequency weighting techniques are not truly separating the sources, but are better characterized as source enhancement techniques. In fact, when there are more sources than sensors, $M < N$, the separation problem is underdetermined, i.e. the mixing matrix in (2.23) is non-square and does not have a left inverse. In this case true separation of the sources is, in general, impossible. This is a major limitation of the traditional statistical methods. However, in a representation where at most M sources overlap at any point, it is possible to achieve separation also in the underdetermined case. For the general case, such a representation does not exist. However, even though more than M sources may overlap in some spectral regions, this method is already a relaxation over the DUET method and its strong assumption about disjointness.

Source separation in underdetermined mixtures, based on the sparsity of the source signals in the STFT domain was proposed in (Bofill and Zibulevsky, 2000) for instantaneous mixtures. A bit loosely formulated, the method can be thought of as an ICA subproblem of order M applied to each spectral coefficient individually. For each spectral coefficient an $M \times M$ mixing submatrix must be estimated. The separation of the sources is then obtained by left multiplying this spectral coefficient in all sensor signals with the inverse of this matrix. The spectral coefficient is thus distributed amongst at most M sources and the remaining $N - M$ sources are excluded. Clearly, a single spectral coefficient does not provide enough data for a statistical analysis, such as the statistical independence that ICA is based upon. The principles for the estimation of the mixing submatrix is the topic of this section. The method involves two steps.

First, the original $M \times N$ mixing matrix is estimated. The N columns of this matrix are estimated by statistical processing. In analogy to ICA, this corresponds to finding the directions of the independent components. In the illustrations of the PCA and the ICA in Figures 2.12 and 2.13, the time do-

main coefficients of the two sensor signals were plotted against each other. Figure 2.16 shows a two-channel instantaneous mixture of three sources. The columns of the mixing matrix were $[1, 2]^T$, $[1, 1]^T$, and $[2, 1]^T$ respectively. The left panel shows the time domain samples of the two sensor signals plotted against each other. This is similar to Figure 2.13, but there is no clear structure in the cloud of data points. The three panels on the right show all the spectral coefficients of the two sensor signals plotted against each other, for real or imaginary parts, magnitude, and phase delay, respectively. Each spectral coefficient yields one data point, i.e. one dot in the scatter plots and the different sensor channels corresponds to the different dimensions of the data point. In this example the two sensor channels correspond to the x and y-coordinates, respectively. For higher dimensions, the principles are the same, but more difficult to visualize. The three different sources are easily observable as strong clusters, or "lines", in the scatter plots in the second and third panels of Figure 2.16. Each cluster corresponds to one source, indexed by n. The direction of the n^{th} cluster in the M-dimensional space corresponds to the coefficients of the n^{th} column the mixing matrix. In the two-sensor case, the slopes of these clusters correspond to the level differences of the peaks in Figure 2.14, at -6, 0, and 6 dB respectively.

Once the columns of the $M \times N$ mixing matrix have been estimated, separation is achieved by finding a $M \times M$ submatrix for each data point. Each data point is the value of a particular spectral coefficient evaluated in all M sensor signals, yielding an M-dimensional coordinate. This coordinate can be described as a linear combination of M (or less) of these columns. Each column is related to one source (2.23). The weight of this column in the linear combination is the "contribution" of this source in the considered spectral coefficient. Since the number of sensors (i.e. the dimension of the data point) is smaller than the number of sources (i.e. the number of estimated columns), $M < N$, several linear combinations are possible. In (Bofill and Zibulevsky, 2000), the $M \times M$ submatrices are estimated based on mathematical minimization. In particular, for each data point, the M column vectors that describe the shortest path to the data point are chosen for the submatrix that is used for separation. This corresponds to a minimization of the l_1 norm of the separated sources. For the two-dimensional example shown in Figure 2.16 this corresponds to selecting the two vectors (independent components) that lie closest to a given data point. This choice of columns for the submatrix is based on a purely mathematical argument, and the practical significance is not clear.

A method that also takes into account delays in the mixture model has been proposed by Bofill (2003). Figure 2.17 shows the same mixture as in Figure 2.16, but where delays between the sensors have been introduced for the first and third sources. This is the same example as the one shown in Figure 2.14. The scatter plot of the STFT magnitudes clearly retains the original information, as seen in the third panel. However, the real (or imaginary) parts of the STFT coefficients provide less information than they did in the instantaneous mixture, as seen in the second panel. Notably, the clusters corresponding to the delayed sources have been "blurred", or spread out more. The proposed method is to estimate the delays by maximizing the clustering over phase. Briefly explained, for each cluster found in the magnitude scatter plot, the method searches for the delays that best explain the "blur" in

the real/imaginary scatter plot. As in the original method, this is a purely mathematical approach and the practical significance is not considered.

The phase delays between the sensor channels give a much more intuitive interpretation of the phase, as seen in the DUET method. For any spectral coefficient, the phase delay is defined as the phase divided by the frequency

$$\delta_i(k,q) = -\frac{\arg X_i(k,q)}{q}\frac{L}{2\pi}. \tag{2.27}$$

This is shown in the scatter plots in the rightmost panels of Figures 2.16 and 2.17, with the two sensor channels plotted against each other. In the instantaneous mixture, since the sources are amplitude scaled only, the phase of the spectral coefficient is the same for both sensor channels. This yields equal phase delay for both channels, $\delta_1 = \delta_2$, visible as a single straight line with unit slope in the right panel of Figure 2.16. When the mixing model includes delays, this introduces a constant difference in phase delay between the sensor channels. In the example shown in Figure 2.17, delays (of opposite signs) were imposed on the first and third sources. In the rightmost panel, these delayed sources are seen as two new lines that are slightly offset (in opposite directions) from the original diagonal line. These offsets, i.e. the distances that the lines have been shifted, represent the delays. This corresponds to the peaks in Figure 2.14, as function of phase delay difference. In comparison to the phase estimation method outlined above, this observation provides a simpler and much more intuitive way for the estimation of the delays.

Discussion

Like the time-frequency weighting techniques, the techniques reviewed in this section have several advantages over the statistical methods. Most notably, they can deal with more sources than sensors, at least to some extent. In addition, since the estimation of the demixing matrices is not based on statistical independence, but on the basis of individual spectral coefficients, these methods are better suited to track changes over time, e.g. in dynamic scenes. Finally, the fact that they are based on the STFT makes them easy to combine with techniques based on signal cues.

However, there are some impracticalities. These are related to the mathematical approaches to the estimation of the submatrices and the delays, respectively. There is a mismatch between these mathematical methods and the intuitive, practical interpretation of spatial cues that was seen in the DUET method.

When the STFT spectra of sources overlap, this yields data points in between the data clusters. For one such data point the energy is then distributed amongst the sources whose mixing column vectors describe the shortest path to the data point. This choice of sources is simply motivated by mathematical minimization of the norms of the separated signals. In reality, this may not be the optimal or correct choice of sources. In order to illustrate this, consider a two-channel mixture of four sources, with mixing columns $[1,4]^T$, $[1,2]^T$,$[2,1]^T$, and $[4,1]^T$, respectively. If the first and fourth sources have overlapping energy of equal strength in one frequency band, all the data points in that band is proportional to $[1,1]^T$. The mathematical method that has been discussed would in this case distribute the resulting energy equally among the second and

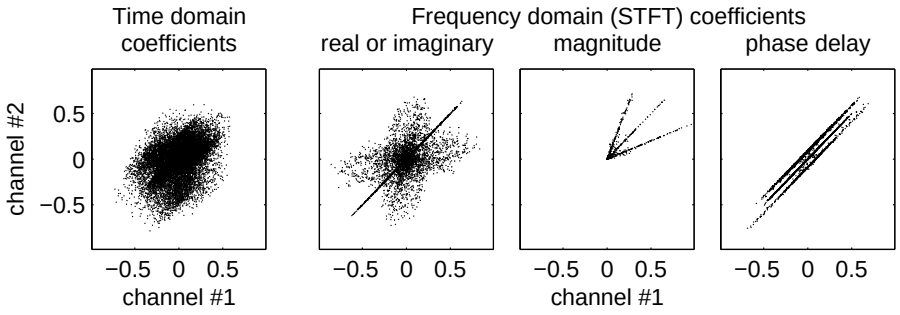


Figure 2.17: *Underdetermined mixtures: sparsity of convolved mixtures.*

third source, since the corresponding column vectors in the mixing matrix are the ones that describes the shortest path to the data points. In this case the real sources, namely the first and the fourth, would be attributed no energy. Clearly, if signal cues had been taken into account, it would be possible to make a more well-founded choice of which sources overlap, i.e. which M sources are to be considered in the separation submatrix. This is one of the main motivations for jointly exploiting spatial cues and signal cues, as proposed in this thesis. In separation methods based on spatial cues, the signal cues can be used in order to improve and assure the quality of the separation. Similarly, for separation based on signal cues, the spatial cues can yield useful additional information. Based on this observation, in Chapter 4, we propose a method for determining regions in the STFT spectra that corresponds to different spectral components. Each region may contain overlapping components from different sources. By employing signal cues, the underlying sources can be determined, and effective separation can be achieved.

2.5 High-level cues and processing

Three groups of cues for the analysis of an auditory scene were mentioned in the introduction of this chapter. The first two groups, signal cues and spatial cues, have already been discussed. For the purpose of making a computational technique for localization and separation of sources, both these groups contain cues that are relatively easy to estimate and to exploit in the grouping of streams. The third group, high-level cues, contains cues that can provide additional information that the human auditory system can exploit, but are somewhat more difficult to implement in a simple computational model. Many of these cues are not easily exploitable for the grouping process (in mathematical terms). Rather, some previously acquired knowledge is required. This is highly related to pattern recognition with respect to a known set of data. In the context of ASA, this is in the field of cognitive psychology. In order to illustrate this, consider the following: a listener can recognize different instruments and knows what they sound like. In addition, depending on the music style, the further development of a music piece may be anticipated. It is even possible (to some extent) to tell if a musician made a mistake. Similarly, it is possible

to recognize a person merely by a brief excerpt of his voice, even under adverse conditions. The meaning of a spoken sentence may be understood even if it is impossible to hear some of the sound. Clearly, the human auditory system can be considered a huge database of acquired knowledge where cognition plays an important role.

When a listener hears a sound for the first time, it may not be clear what it is. However, if the sound is repeated, the listener may start to recognize the sound. There are several cues related to the recognition and classification of sound sources. Basically these depend on the time and frequency characteristics of the sources.

For the identification of musical instruments, Martin and Kim (1998) describe a large set of spectro-temporal features that can be exploited. The identification is based on a pattern-recognition approach in the feature space. A similar approach based on cepstral coefficients and temporal features is given in (Eronen and Klapuri, 2000). Higher level models can also take knowledge about the temporal structure of music into account. A method based on a database of chord-note relations and dictionaries of chord transitions and note transitions is described in (Kashino and Hagita, 1996). This model exploits statistical information on the progression of chords and notes in a particular genre. A somewhat simpler model based on probabilistic models of note transitions can be found in (Kashino and Murase, 1998).

Cues related to cognition are also important for source localization. Once a source has been recognized as described in the above, this may put some restrictions on the location of the source. For instance, if the source is recognized as an airplane, this clearly gives an indication about the location of the source, e.g. the airplane is likely to pass above the listener. In addition, it is well known that the human auditory system exploits spectral cues (among other ones) in order to determine the elevation of a source (Duda, 1997; Blauert, 2001). However, the relation between elevation and spectral cues seems to be quite complex and subjective. This is difficult to exploit in a purely computational model. The spectral cues are also useful for determining the distance of a source.

Many of the cues in this group are quite complex and take into account information about language, linguistics, music theory, anticipation, etc. These constitute separate fields of research. Some cues related to cognition were presented just for reference, in order to show how the human auditory system can take advantage of additional information that is not taken into account in the computational models presented in this thesis.

2.6 Combined processing of cues

2.6.1 Corrupted cues

Thus far, the original source signals have been considered. Not much attention was given to the fact that the sensors observe a mixture of modified source signals. The original source signals convey important information about the underlying source. However, this information may be (partly) corrupted in the signal observed by a sensor. There are mainly two reasons for this:

1. As the sound wave propagates in all directions, it is subject to reflections,

refractions, occlusions, and other effects. Due to this, in general, there are many paths leading from a single source to a sensor. Along each of these paths the original source signal is differently affected. The signals along the different paths superimpose and the resulting signal at the sensor can be considered as a filtered version of the original source signal. Some of the information inherent in the source signal may be corrupted in the sensor signals. The degree of corruption depends on the physical conditions of the scene. However, the degree of corruption is, in general, higher for spatial cues than for signal cues. When the spatial cues are ambiguous, signal cues can provide useful additional information.

2. When there are more sources, each of the observed signals is a superposition of several (filtered) source signals. When multiple sound sources are simultaneously active, their spectrograms may overlap. In the regions where this overlap is significant, it is no longer possible to estimate all the cues accurately. This applies to both signal and spatial cues. Nevertheless, by jointly evaluating thw spatial and signal cues this can be improved.

2.6.2 Separation by both spatial cues and signal cues

As it has been argued in this chapter, the vast majority of existing techniques and literature only consider one type of cues, i.e. either spatial cues or signal cues. Indeed, by considering the knowledge about the human auditory system and how different cues are jointly evaluated or weighted against each other, it is natural to consider joint evaluation of cues in CASA as well. Lyon (1983) mentioned this combined processing more than 20 years ago: *"Collectively, these and other techniques will allow versatile signal separation based on frequency content, time of occurrence, direction, pitch, and higher-level cues."* Still, there are surprisingly few techniques that have followed up on this.

Woods *et al.* (1996) proposed a method for the joint evaluation of several cue types. The method takes a purely mathematical approach, blindly weighting the cues against each other. While this may be useful for some cues, it is not applicable in general. For some cues, the interactions between the cues are more complex and are not easily expressed by a simple weighting function. One example is the relative importance of level and time differences (as function of frequency) for localization of sources. We present a computational technique for the joint evaluation of level and time differences in binaural signals. This is discussed in Chapter 3. In Chapter 4 we present the application of this to source separation.

One of the few methods that exploit both spatial cues and signal cues was developed by Nakatani and Okuno (1999). This method is similar to source separation methods based on signal cues only (exploiting harmonicity, in particular). In addition, the spatial cues of the individual spectral components are taken into account in the grouping into streams. In Chapter 4 we present a related technique based on "spectral regions". This is somewhat similar to sinusoidal modeling, but where the AM-FM description has been replaced by a definition of regions in the spectrogram. Both signal and spatial cues are employed for the definition of these regions and for the grouping of these into streams. Techniques for the separation of overlapping partials in such spectral

regions are presented in Chapter 5. These methods take into account both signal and spatial cues.

Chapter 3

Binaural localization

Binaural source localization is the problem of estimating the location of a sound source based on the signals observed at the entrances of the two ears. In order to determine source distance and elevation, the cues that can be estimated from each of these signals, called signal cues, are primarily used. The head related transfer functions (HRTFs) describe the relation between these cues and the position of a source in terms of azimuth, elevation and distance. For the estimation of the azimuth angle, the differences between the two ear signals are most important. These are described by the relation between the HRTFs for the two ears. Here, we only consider the problem of estimating source azimuth angles.

Background

Over the past decades, several computational models for the estimation of source azimuths in binaural signals have been proposed. Many of these (Joris *et al.*, 1998; Blauert, 2001) are based on the coincidence model proposed by Jeffress in 1948. This is a model of the neural system where nerve impulses from each of the two ears propagate along delay lines in opposite directions. At the position along the delay lines where the impulses coincide a nerve cell is excited, effectively transforming time information into spatial information. This model corresponds to evaluating ITDs by means of a running short-time cross-correlation function between the ear input signals. Based on Jeffress model, several extensions have been proposed that take into account ILDs, such as the models by Lindemann (1986) and Gaik (1993). An overview of these and other models of binaural perception is given in (Stern and Trahiotis, 1997). Most of these models work by decomposing the ear input signals into perceptual bands and estimate the interaural cues in these bands. When several sources at different locations have significant energy within a given perceptual band, the resulting azimuth estimates for that band will not, in general, correspond to any of the actual azimuths of the sources. In some cases, therefore, it can be advantageous not to be limited by the frequency resolution of the human auditory system, but rather to estimate the azimuths in individual narrow frequency bands. In particular, in relation to the source separation methods based on time-frequency weighting, as discussed in Section 2.4.3, spatial cues estimated from the more frequency selective STFT spectra of the two sensor

signals are of high interest.

Motivation

When HRTFs have been measured at several different azimuth angles for each of the two ears, the differences between these two sets of HRTFs describe the ILDs and ITDs as functions of azimuth (and frequency). This means that, in an observed signal, an ILD estimate can be compared with the HRTF data sets in order to obtain an estimate of the source azimuth. This is referred to as *HRTF data lookup*, yielding azimuth estimates based on ILD only. Similarly, the ITDs can be used for HRTF data lookup of azimuths based on ITD only.

In this chapter, we propose a method for the estimation of source azimuths through the joint evaluation of ILDs and ITDs. For each spectral coefficient in the STFT spectra the ILD and ITD is estimated. On one hand, the azimuth estimates based on ILDs have a relatively large standard deviation. On the other hand, the azimuth estimates based on ITD have smaller standard deviation, but are ambiguous. By jointly evaluating these quantities the ILDs are used in order to resolve the ITD ambiguities, effectively improving the azimuth estimates. Since the method is based on the STFT it is computationally fast and has a simple reconstruction scheme that is highly useful in source separation applications.

We also propose a parametric model for the relation between azimuth angle and interaural cues (ILD and ITD). In this *individual parametric model* the parameters are optimized with respect to an individual head for which the HRTFs have been measured. Similarly to the azimuth estimation by HRTF data lookup, it is possible to lookup the azimuths by use of this model. This is referred to as *individual model lookup*. This method has the advantage that the azimuth lookup is faster. However, the azimuth estimates are not as accurate as those obtained by HRTF data lookup. In addition, it still requires the HRTFs to be measured for the individual head in order to determine the parameters.

Based on the study of the parameters of the individual model for each of the 45 subjects in the CIPIC database of HRTFs (Algazi *et al.*, 2001) we propose a generic model for the relation between azimuth angle and ILDs and ITDs that only depends on one parameter, namely the distance between the ears. This "average parameter model" can be used with any head and does not require the measurement of HRTFs. The performance of this model is comparable to that of the individual parametric model.

In Section 3.1 the estimation of ILDs and ITDs for individual spectral coefficients in the STFT spectra is discussed. In Section 3.2 we propose a method for joint evaluation of these cues. The parametric model for ILD and ITD is presented in Section 3.3. In Section 3.4 the parameters of the individual model are studied and the average model is proposed. The application to source localization is studied in Section 3.5. Some experimental results in dynamic scenes (head movements) are presented in Section 3.6.

3.1 Cue estimation

In a binaural signal, the STFT spectra of the two ear input signals $X_{\text{left}}(k, q)$ and $X_{\text{right}}(k, q)$ are obtained by Equation 2.1. The spatial cues can be estimated

for each individual spectral coefficient in these spectra, indexed by k and q in time and frequency, respectively. The ILDs in dB are given by

$$\Delta L(k,q) = 20 \log_{10} \left| \frac{X_{\text{right}}(k,q)}{X_{\text{left}}(k,q)} \right|. \tag{3.1}$$

This is simply the ratio, measured in dB, of the STFT magnitudes of the right and left ear signals. Similarly, the ITDs are estimated as

$$\Delta T_p(k,q) = \frac{-\Delta P_p(k,q)}{q} \frac{L}{2\pi}, \tag{3.2}$$

where L is the window length in the STFT, given in (2.1). The interaural phase differences $P_p(k,q)$ are given by

$$\Delta P_p(k,q) = \arg \frac{X_{\text{right}}(k,q)}{X_{\text{left}}(k,q)} + 2\pi p. \tag{3.3}$$

Each spectral coefficient represents a periodic and narrow band signal. The phase of a periodic signal can only be estimated up to an integer multiple of 2π. This is reflected by the integer parameter p in the estimates of ITD and interaural phase difference. The practical significance of this is that, for a given frequency, several different source locations yield the same phase difference between the two ear input signals. This is equivalent to the spatial aliasing seen in beamforming techniques. The parameter p indexes these positions, with $p = 0$ corresponding to the source position closest to zero azimuth, $\theta = 0$. A negative value of p corresponds to a position on the left side (negative θ). Positive p corresponds to positions on the right side. In this case, possible values of p depend on the physical layout of the sensors and sources. The frequency whose period equals twice the largest possible delay between the two ears corresponds to the highest frequency for which the phase can be estimated without ambiguity. Below this frequency only $p = 0$ is physically realizable. For an average head size the phase ambiguities occur for frequencies above approximately 1500 Hz. The spatial cue estimates given by equations (3.1)–(3.3) are similar to those used in time-frequency weighting techniques, (2.24)–(2.26), but taking into account the phase ambiguity.

3.2 Estimation of azimuth angles

In order to relate the ILDs to the ITDs, a common reference frame is needed such as the azimuth angle. Using measured HRTFs, the azimuth can be estimated from ILDs and ITDs by HRTF data lookup.

3.2.1 HRTF data lookup

Based on the HRTFs measured at different azimuth angles, the ILD and ITD can be described as function of azimuth and frequency. Since the HRTFs are assumed to be time-invariant, there is no dependency on the time index k. Instead, the HRTFs depend on the azimuth angle θ. By changing the role of the time index k with that of the azimuth angle θ and by using the left and right HRTFs as functions of azimuth and frequency, $H_{\text{right}}(\theta,q)$ and $H_{\text{left}}(\theta,q)$, as the

signals in Equations (3.1) and (3.2), we obtain the HRTF data lookup models for level difference, $\Delta L(\theta, q)$, and time difference, $\Delta T(\theta, q)$, as functions of azimuth angle θ and frequency index q. In the computation of the ITD lookup model special care must be taken to "unwrap" the phase, i.e. to determine the correct choice of p for all frequencies and azimuths.

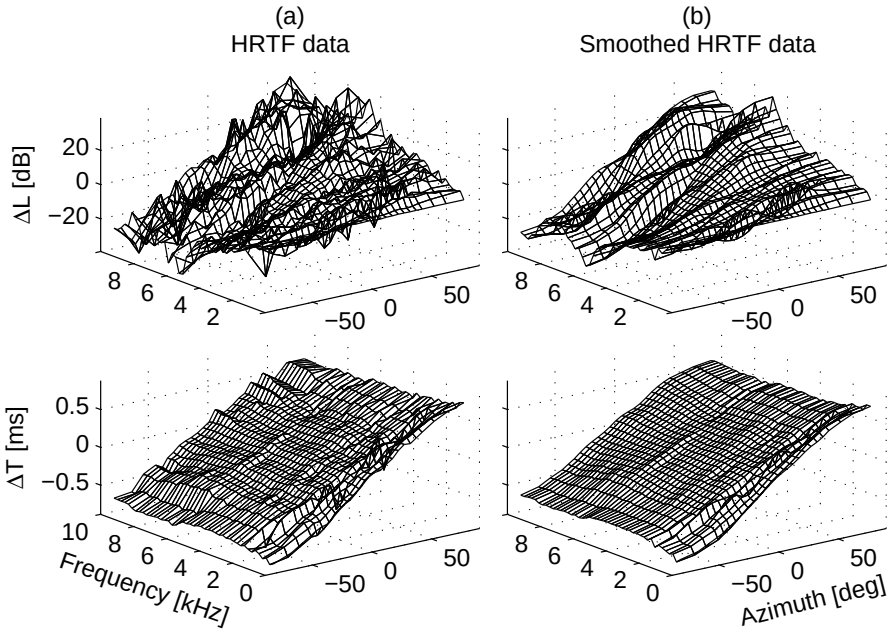


Figure 3.1: *Interaural level and time differences as functions of azimuth angle and frequency. (a): HRTF data lookup. (b): HRTF data lookup smoothed across azimuth.*

The ILD and ITD as functions of azimuth and frequency for one particular head are shown in Figure 3.1. The panels in the left column show the ILD and ITD as computed from the measured HRTFs. In the right column, the same functions are shown after smoothing in the azimuth direction. No processing across frequency was performed.

When a sound source is located on one side of the head, the source signal will arrive first at the ear on the same side. Also, the signal level will normally be stronger at the ear facing the source than at the ear on the opposite side. Intuitively, the farther to the side a source is located, the larger the level and time differences between the ears should be. The smoothed HRTF data confirm this intuitive facts. Both functions are relatively monotonic in azimuth.

Based on the ILD estimate (3.1) for a given left/right pair of spectral coefficients, $X_{\text{left}}(k, q)$ and $X_{\text{right}}(k, q)$, the azimuth angle can be looked up in the smoothed HRTF data $\Delta L(\theta, q)$. This yields azimuth estimates based on ILD only, denoted $\theta_L(k, q)$. Similarly, each ITD estimate (3.2) can be looked up in $\Delta T(\theta, q)$. Due to the phase ambiguity, this results in multiple possible azimuth estimates, denoted $\theta_{T_p}(k, q)$, indexed by p.

The ITD is usually a relatively smooth function of azimuth, as seen in Fig-

ure 3.1. This means that the standard deviation of the azimuth estimates based on ITD is relatively small. However, there may be several possible azimuth candidates due to the phase ambiguity. The ILD is a more complex function of azimuth and must be smoothed across azimuth in order to become useful for azimuth lookup. Consequently, the azimuth estimates based on ILDs have a much larger standard deviation than those based on ITD. In addition, the ILDs as function of azimuth are not, in general, monotonic for all frequencies. When this is the case, the azimuth lookup is non-unique, yielding multiple possible azimuth estimates. For the examples in this chapter the azimuth estimates closest to 0 degrees were chosen whenever this is the case.

3.2.2 Joint evaluation of ILD and ITD

Since both, the ILD and the ITD, are related to the azimuth, they can also be related to each other. We propose a method for the joint evaluation of these quantities in order to improve the azimuth estimates. Briefly explained, the noisy $\theta_L(k,q)$ provides a rough estimate of the azimuth for each left/right spectral coefficient pair. Then, this estimate is refined by choosing the $\theta_{T_p}(k,q)$ that lies closest. The combined azimuth estimate is then given by

$$\theta(k,q) = \theta_{T_p}(k,q)\Big|_{p=\arg\min\left(\left|\theta_L(k,q)-\theta_{T_p}(k,q)\right|\right)} . \tag{3.4}$$

Effectively, the ILD estimate is used in order to choose the "correct" parameter p in the ITD estimate. The azimuth estimate based on ITD is chosen since this estimate is "more precise", i.e. the standard deviation of these estimates are smaller. The general processing is illustrated in Figure 3.2.

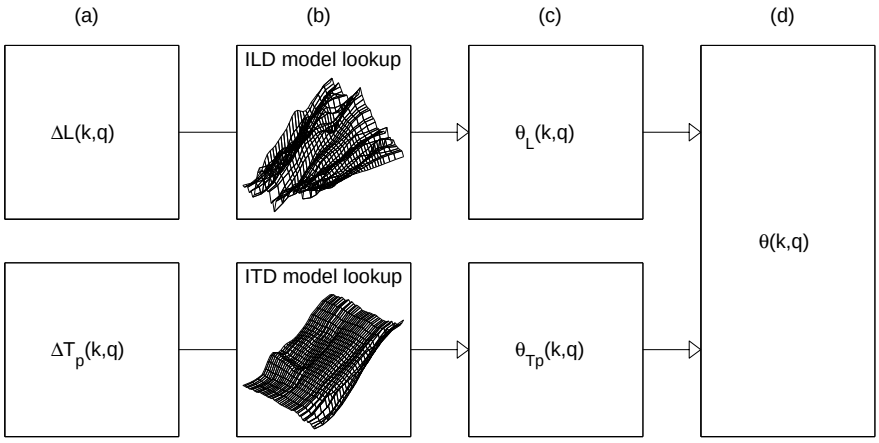


Figure 3.2: *Combined evaluation of ILDs and ITDs for the estimation of azimuths. (a): The interaural cues are estimated by use of the STFT. (b): The relation between these cues and azimuth angle is described by the ILD and ITD in the smoothed HRTF data. (c): Azimuth estimates are found by lookup of the interaural cues in the HRTF data models. (d): Final azimuth estimates are obtained by combined evaluation of the azimuths estimated from the ILDs and ITDs.*

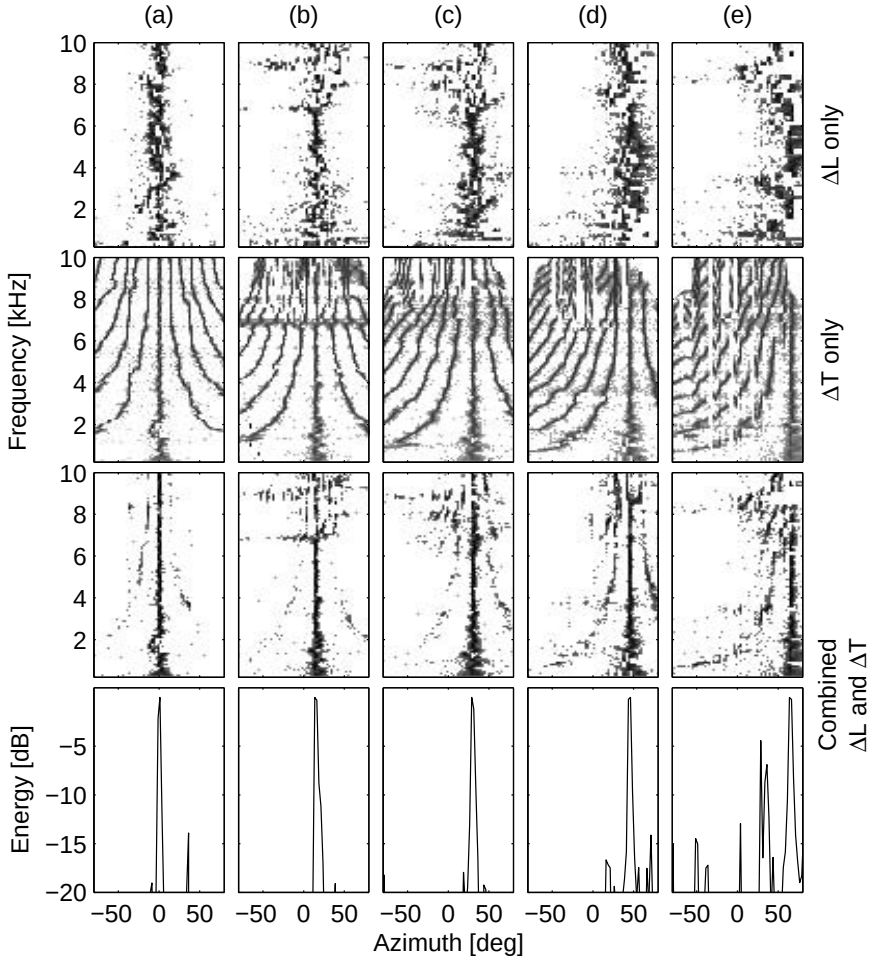


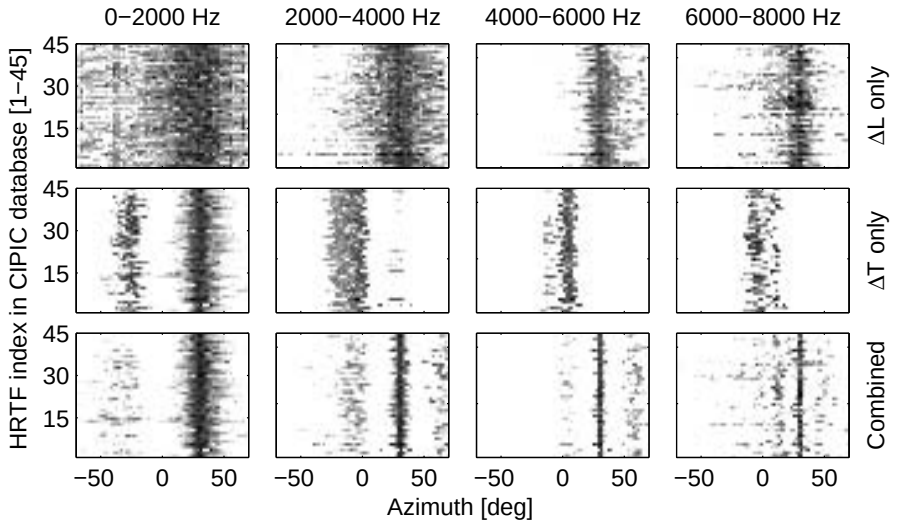
Figure 3.3: *Histograms of azimuth estimates for 5 different heads and azimuth angles, at 0, 15, 30, 45 and 65 degrees, (a)–(e), respectively. First row: based on ILD only. Second row: based on ITD only. Third row: based on our method for combined evaluation of ILD and ITD. Bottom row: marginal histograms for our method.*

Figure 3.3 shows some experimental results to further illustrate the processing. In order to assess the performance of the proposed method for different frequencies, a white noise signal was chosen as source signal. This signal was then filtered with HRTFs for five different heads, each using a different azimuth angle for the source position. A window length of about 10 ms was chosen in the computation of the STFT spectra, and the interaural cues were estimated by means of (3.1) and (3.2). The five different situations are shown in the different columns in the figure, with azimuth angles of 0, 15, 30, 45 and 65 degrees (on the right side of the head), respectively. The panels in the first row show two-dimensional histograms as function of azimuth and frequency, based on azimuths estimated from ILD only. These histograms imply, as mentioned in Section 3.2.1, that the azimuth estimates based on ILD have a larger standard deviation than those based on ITD. The precision decreases with increasing azimuth. In addition, at low frequencies, the ILDs are small and are virtually useless for the estimation of the azimuth. The panels in the second row of Figure 3.3 show similar histograms based on azimuths estimated from ITDs. Above approximately 1-2 kHz, these azimuth estimates are ambiguous. For a given frequency this is seen as several equally strong peaks at different azimuths. These correspond to different choices of p in (3.2). As the frequency increases, more values of p are possible. Additionally, the distance between the peaks decreases with increasing frequency, making the ITD estimates less useful at higher frequencies. In the third row the results that were obtained by applying the proposed method are shown. Visually explained, the estimates based on ILDs (first row), are used in order to select the right p in the estimates based on ITDs (second row). Note that this is done independently for each spectral coefficient in the STFT spectra: no processing is performed across frequency.

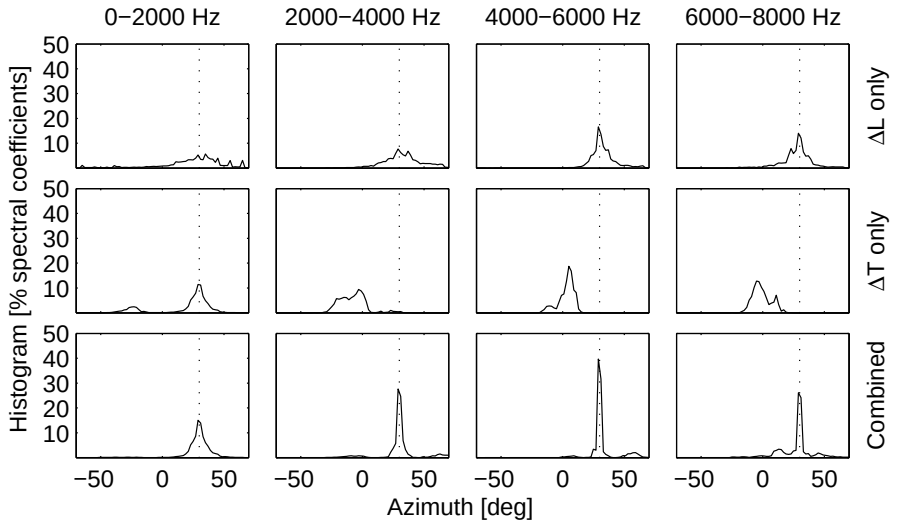
The last row of panels shows the one-dimensional marginal histograms as function of azimuth (i.e. summing over all frequencies), based on the azimuths computed by the proposed method. The strongest peaks are found at the true azimuth angles of 0, 15, 30, 45 and 65 degrees, respectively.

The example with a source located at 30 degrees azimuth, as shown in Figure 3.3(c), was repeated for all the 45 heads in the CIPIC database. For each head, the one-dimensional marginal histograms, as function of azimuth, were computed in four different frequency ranges, namely 0–2 kHz, 2–4 kHz, 4–6 kHz and 6–8 kHz. The results are shown in Figure 3.4(a). The four columns correspond to the different frequency ranges. The three rows correspond to the different azimuth estimates, i.e. based on ILD only (top), ITD only (middle), and the proposed method for combined ILD and ITD (bottom). Each panel shows the marginal histogram, as seen in the bottom row in Figure 3.4, for all the 45 different heads along the ordinate axis. For all the frequency ranges and for most of the heads the azimuth estimates have been improved by application of the proposed method.

Figure 3.4(b) shows the unweighted histograms, i.e. the percentage of spectral coefficients yielding different azimuth estimates, averaged over all 45 subjects. The layout of the panels is the same as in Figure 3.4(a). In the azimuth estimates based on the proposed method most of the ambiguous azimuths based on ITD only have been resolved. In addition, the standard deviation is smaller than for the azimuth estimates based on ILD only.



(a) Energy weighted histograms for all 45 subjects.



(b) Unweighted histograms (percentage of spectral coefficients) averaged over all 45 subjects.

Figure 3.4: *Histograms of azimuth estimates in different frequency bands, using measured HRTF data for azimuth lookup.*

3.3 Parametric ILD and ITD models

In the previous section, we described how to apply HRTF data in order to lookup azimuths. In this section we propose a parametric model for the relation between azimuth angles and ILDs and ITDs. This model allows a simpler lookup of azimuths, but at the cost of decreased accuracy. More importantly, it serves as the basis and motivation for the generic model that will be introduced in the next section.

When the distance of a source is large compared to the distance between the ears the source is said to be in the "far-field". This means that the sound waves reaching the sensors are close to planar waves. For a source in the far-field the source distance is of little significance for the interaural differences. The sound pressure level at different points on a rigid sphere as function of source position has been studied in (Cooper and Bauck, 1980). Under the far-field assumption the level differences between the two ears are mainly caused by head shadowing only. The effect of the source distance can be largely neglected, as indicated in (Duda and Martens, 1998, 1997).

3.3.1 Interaural time differences

Based on simple geometric considerations, as shown in Figure 3.5, the following formula for the ITD was proposed in (Woodworth and Schlosberg, 1954):

$$\Delta T(\theta) = \frac{r(\sin\theta + \theta)}{c}, \tag{3.5}$$

where r is the "head radius", and c is the wave propagation speed. In reality, the ITD is slightly larger than this due to the fact that the head is not perfectly spherical. In addition, the time difference is also somewhat larger at low frequencies, as has been observed in (Wightman and Kistler, 1997). In order to take this into account, we use a frequency dependent scaling factor α_q,

$$\Delta T(\theta, q) = \alpha_q \frac{r(\sin\theta + \theta)}{c}. \tag{3.6}$$

3.3.2 Interaural level differences

As implied by the data shown in Figure 3.1, the ILD is a much more complex function of azimuth and frequency. Based on a study of the HRTFs in the CIPIC database, we propose the following model:

$$\Delta L(\theta, q) = \beta_q \frac{\sin\theta}{c}, \tag{3.7}$$

with frequency dependent scaling factor β_q. Based on the observation that the ILD is periodic in θ, a Fourier series expansion of the ILD was proposed in (Duda, 1997). Our model is similar to this (single-term expansion), with the exception that we only consider the range $-90° \leq \theta \leq 90°$.

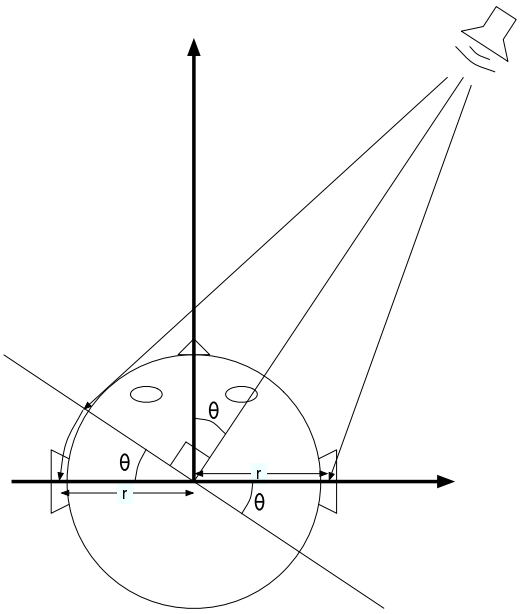


Figure 3.5: *A simple geometrical model of interaural time difference.*

3.3.3 Frequency dependent scaling factors

The ILD and ITD models in (3.7) and (3.6) can be optimized for a given head by finding the scaling factors α_q and β_q that give the closest match to the smoothed HRTF data. For the head whose HRTF data is shown in Figure 3.1, the best matching parametric models were found. The independent parametric models and the model errors are shown in Figure 3.6.

For the estimation of azimuth based on ILDs and ITDs, (Figure 3.2(b)), the HRTF data can be replaced by the parametric model. The experiment shown in Figure 3.4(a) was repeated, but with the use of parametric models. For each of the 45 heads, the best matching parametric model was found and used for azimuth lookup. The results are shown in Figure 3.7(a). In this case, the azimuth estimates are not as accurate as when the HRTF data were used for azimuth lookup. In particular, the estimation errors are significant at higher frequencies. The estimation error is mainly due to the model error in the ILD model, as shown in Figure 3.6. For most heads, however, the model is useful up to about 6 kHz. In any case, the use of the proposed method for the joint evaluation of ILD and ITD yields sharper peaks in the azimuth histograms.

3.4 Average ILD and ITD models

In order to obtain ILD and ITD models for a given head HRTFs must be measured for a range of different azimuth angles. This can be a tedious task. Also the parametric models that were proposed rely on the HRTF data for estimating the frequency dependent scaling factors. In this respect, the para-

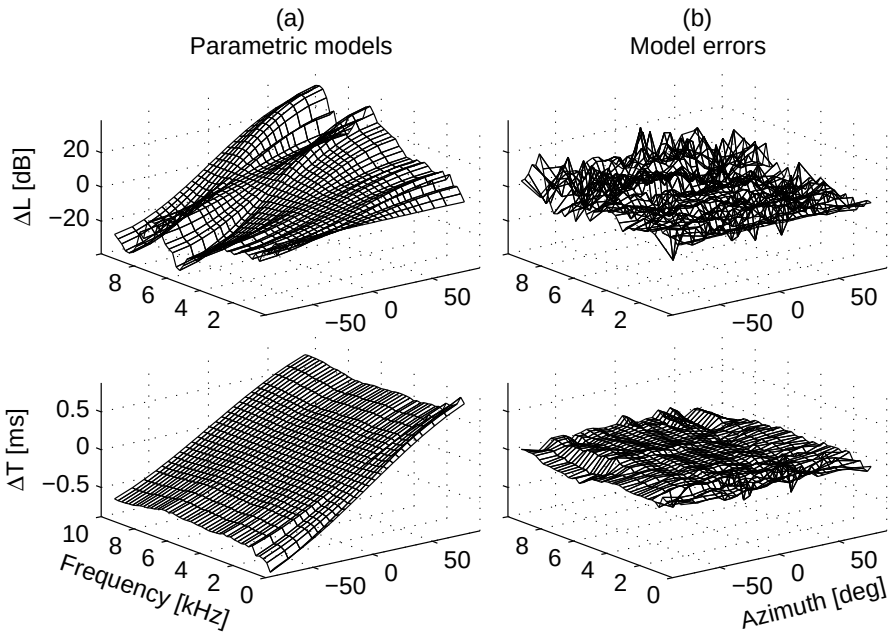


Figure 3.6: *Interaural level and time differences as functions of azimuth angle and frequency. (a): Parametric models. (b): Model errors.*

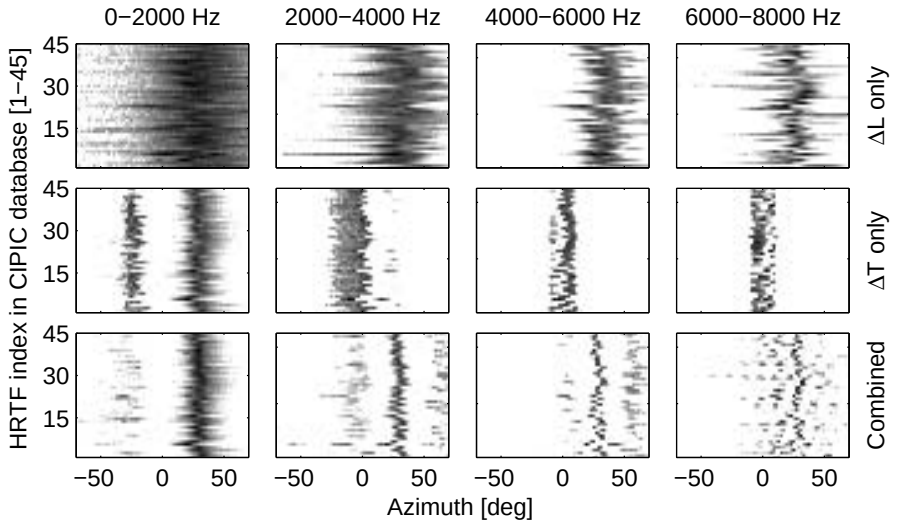
metric models do not present an advantage over the HRTF data. In fact, the parametric models are less accurate.

The scaling factors for the ILD model, β_q, and ITD model, α_q, were computed as functions of frequency for the 45 subjects in the CIPIC database. These are shown in gray in Figure 3.8. Qualitatively, all these quantities follow the same trend, at least up to about 7 kHz. Above this frequency, the ILD scaling factors vary highly among the different heads.

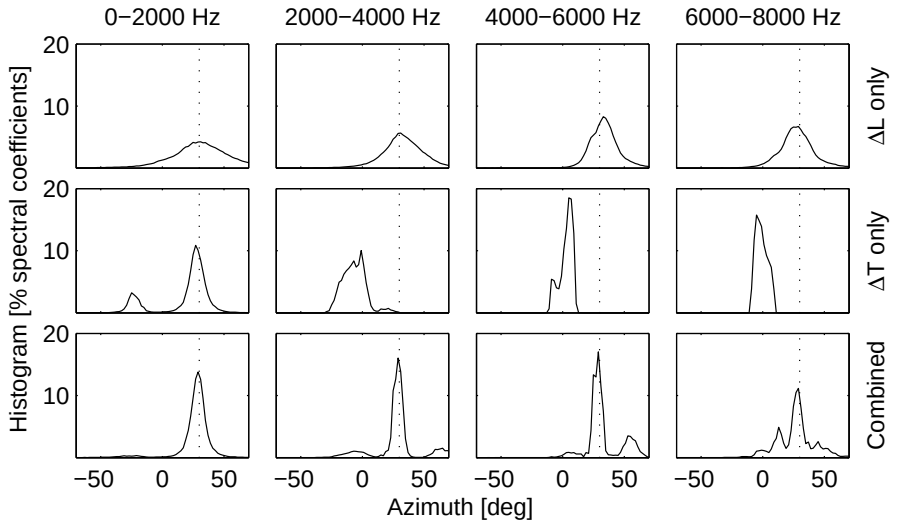
The black lines in Figure 3.8 show the scaling factors averaged over all heads. If the average scaling factors are used in the parametric ILD and ITD models, these models only depend on one parameter, namely the head radius r. This can easily be measured, and consequently these *average parameter models* can be employed for any head without the need to measuring the specific HRTFs. The resulting average models and the model errors are shown in Figure 3.9.

In order to compare the accuracy of the individual parametric models and the average models, the model errors were computed for each head (relative to the smoothed HRTF data). The absolute value of these errors were then averaged over all azimuths and heads. Figure 3.10 shows these errors as functions of frequency. The ITD models are shown in the top row, and the ILD models in the bottom row. Columns (a) and (b) correspond to the individual parametric models and the average parameter models, respectively. The average models (b) are almost as good as the individual models (a), and only a slight increase in error can be observed. However, above about 6-8 kHz the accuracy of all the models is significantly worse than that attained by the smoothed HRTF data.

The performance of the average models for estimation of azimuths is shown



(a) Energy weighted histograms for all 45 subjects.



(b) Unweighted histograms (percentage of spectral coefficients) averaged over all 45 subjects.

Figure 3.7: *Histograms of azimuth estimates in different frequency bands, using individual parametric models for azimuth lookup.*

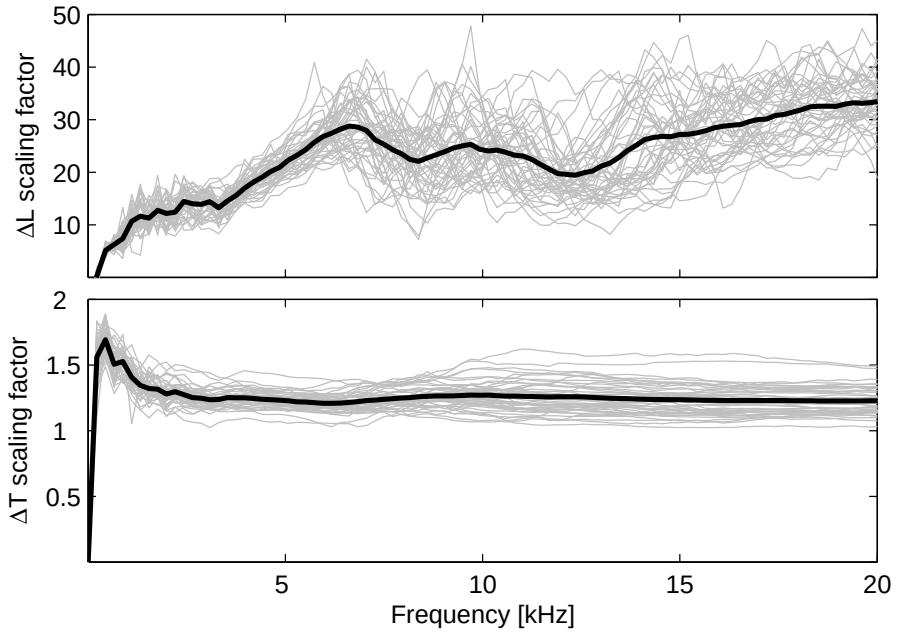


Figure 3.8: *Frequency-dependent scaling factors. Top: ILD model scaling factor, β_q. Bottom: ITD model scaling factor, α_q. The factors optimized for each of the 45 heads are shown in gray, and the average over these is shown in black.*

in Figure 3.11. The accuracy of azimuth estimates based on the average models are comparable to those based on the individual parametric models, as shown in Figure 3.7.

3.5 Localization in static scenes

So far, only a single source was considered and white noise was used as the source signal in order to study the accuracy of azimuth estimates at different frequencies. Since the proposed method for joint evaluation of ILDs and ITDs is based on interaural cues in narrow bands independently it is also applicable in situations where several sources are located at different locations.

In our experiment three different harmonic tones were chosen as source signals, i.e. three consecutive half-notes played by an alto trombone. The binaural signal was obtained by filtering these sources with the HRTFs at different azimuth angles. Figure 3.12 shows histograms of the estimated azimuth angles in this mixture. The histograms have been energy weighted, i.e. each azimuth estimate is weighted by the energy of the corresponding spectral coefficient. The three columns show the results obtained by use of HRTF data, parametric models, and average models, respectively. The panels in the first row show the results for azimuth estimates based on ILD only. Since the histograms are weighted, the different harmonics can be observed as strong peaks (dark horizontal lines). However, these are quite wide and do not provide very accurate estimates of the azimuth. In the second row, the estimates based on ITD are

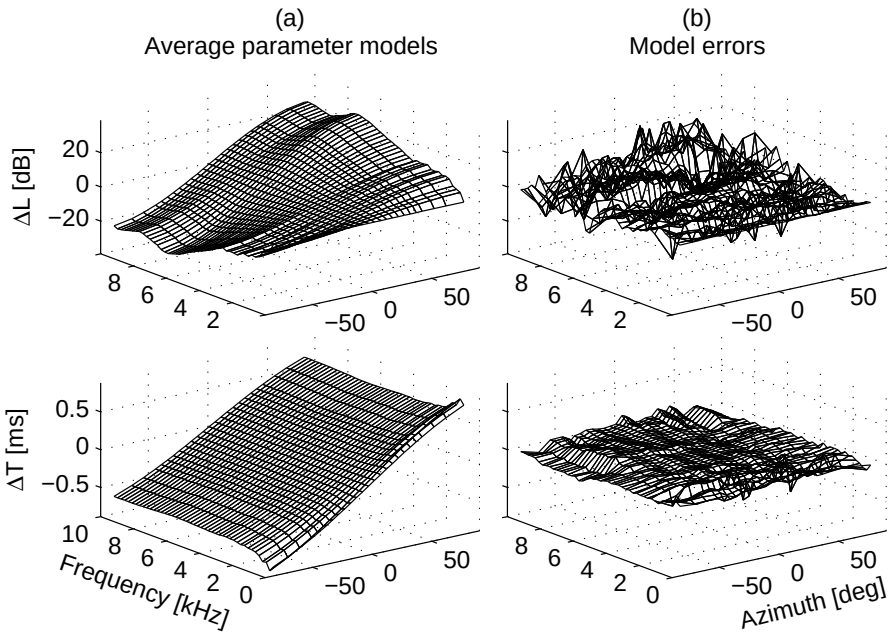


Figure 3.9: *Interaural level and time differences as functions of azimuth angle and frequency. (a): Average parameter models. (b): Model errors.*

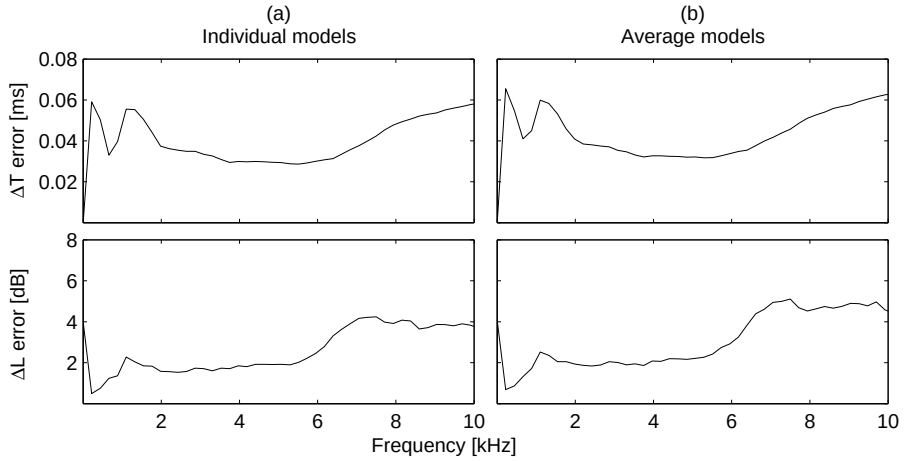
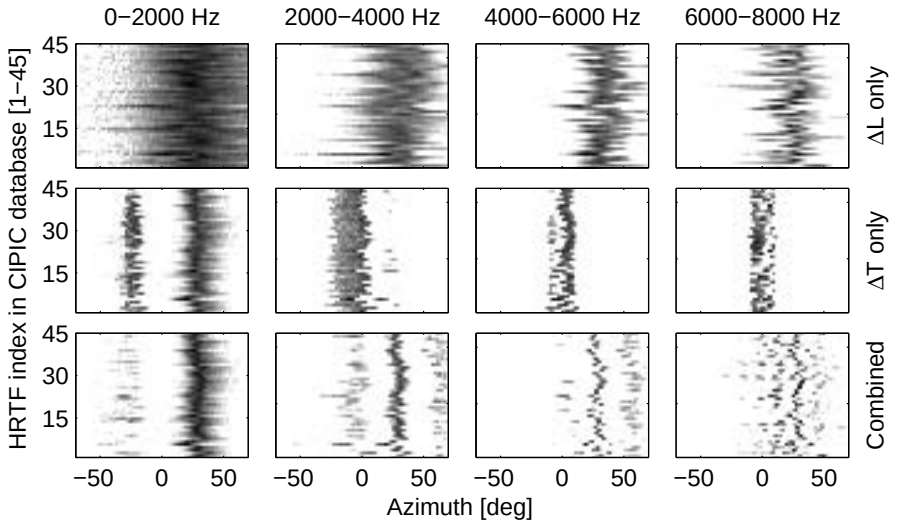
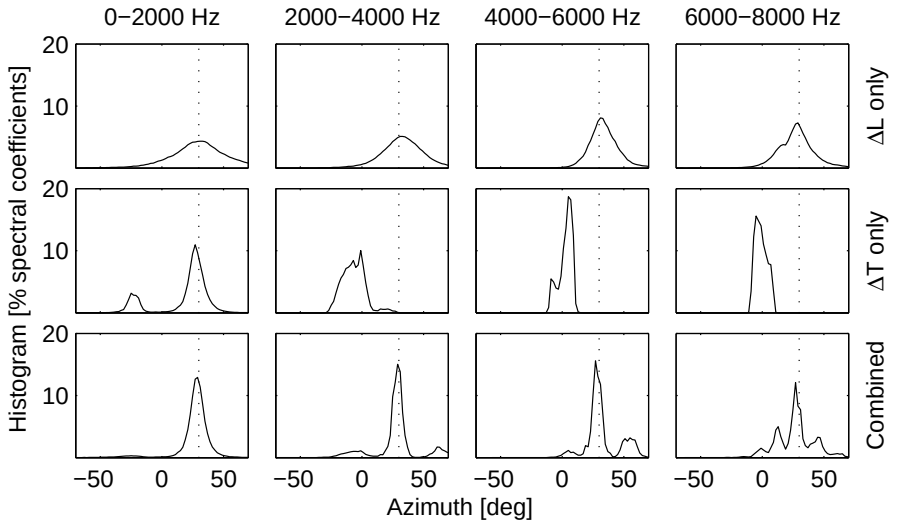


Figure 3.10: *Absolute model errors averaged over all azimuths and heads, as function of frequency. (a): Individual models. (b): Average parameter models.*

shown. The peaks corresponding to the harmonics are much narrower, but ambiguous above about 1500 Hz. In the third column the results obtained by the proposed method for joint evaluation of ILD and ITD are shown. For all three models, the different harmonics are well aligned about the true source



(a) Energy weighted histograms for all 45 subjects.



(b) Unweighted histograms (percentage of spectral coefficients) averaged over all 45 subjects.

Figure 3.11: *Histograms of azimuth estimates in different frequency bands, using average parameter models for azimuth lookup.*

azimuths of −30, 15 and 45 degrees, respectively. The marginal histograms are shown in the bottom row.

3.5.1 Relative importance of ITDs and ILDs

In the human auditory system, the relative importance of the ITDs and ILDs for localization of narrow-band sources depends on frequency. At low frequencies, below about 1 kHz, the ITDs are dominant. At higher frequencies, the ILDs are dominant. This is known as the "Duplex Theory", as discussed in Section 2.3.3. A more detailed discussion can be found in (Moore, 1997). More generally, the relation between level and time differences have been described by "trading ratios", e.g. relating level differences (in dB) to time differences (in μsec) (Blauert, 2001), relating them to the perceived source locations, or relating them through dimensionless weights (Macpherson and Middlebrooks, 2002). A discussion of the relative salience of sound localization cues can be found in Wightman and Kistler (1997).

The proposed method for joint evaluation of ILD and ITD is compatible with the duplex theory. At low frequencies the azimuth estimates are determined uniquely by ITDs since there are no ambiguities. At higher frequencies, the ambiguous azimuth estimates based on ITDs are closely spaced. In this case the ITDs are virtually useless and the ILDs are dominant for azimuth estimation. In addition, the proposed method handles gracefully the transition from low to high frequencies by exploiting both ILDs and ITDs. The relative importance of these changes smoothly from one extreme (ITD only) to the other extreme (ILD only). This can be considered as a relative weighting of the ILDs and ITDs. However, the azimuth estimates are obtained by a joint evaluation of the estimates based on ITD and ILD, and not computed blindly as a weighted sum of these.

The ILDs and ITDs convey similar information about the source azimuth. The parametric models that were proposed in Equations (3.6) and (3.7) in this chapter describe the relation between these cues. Figure 3.13 shows this relation in 24 different perceptual frequency bands. Each panel corresponds to one perceptual band and the ITDs are plotted against the ILDs for different azimuths. The dotted gray lines show the relation between the individual parametric models for each of the 45 subjects at different frequencies within the band. The solid black lines show the relation for the average parameter model at the center frequency of the band. This relation can be interpreted as a frequency dependent trading ratio (also depending on azimuth). It also provides a parametrization of the "curve" of naturally occurring combinations of level and time differences, as discussed in the context of time-frequency weighting techniques in Section 2.4.3 and shown in Figure 2.14. Figure 3.13 is consistent with Fig. 4 in (Gaik, 1993), where the "naturalness" of a given combination of ITD and ILD was studied based on HRTFs measured at various azimuths and elevations.

The ILDs and ITDs depend on elevation. However, their dependency on azimuth remains qualitatively the same. The experiment with localization of three sources, as shown in Figure 3.12, was repeated with the same sources located at an elevation of 30 degrees. The azimuths were then estimated with HRTF data and individual and average models for the horizontal plane. The results are shown in Figure 3.14. In this case the azimuth estimates are less

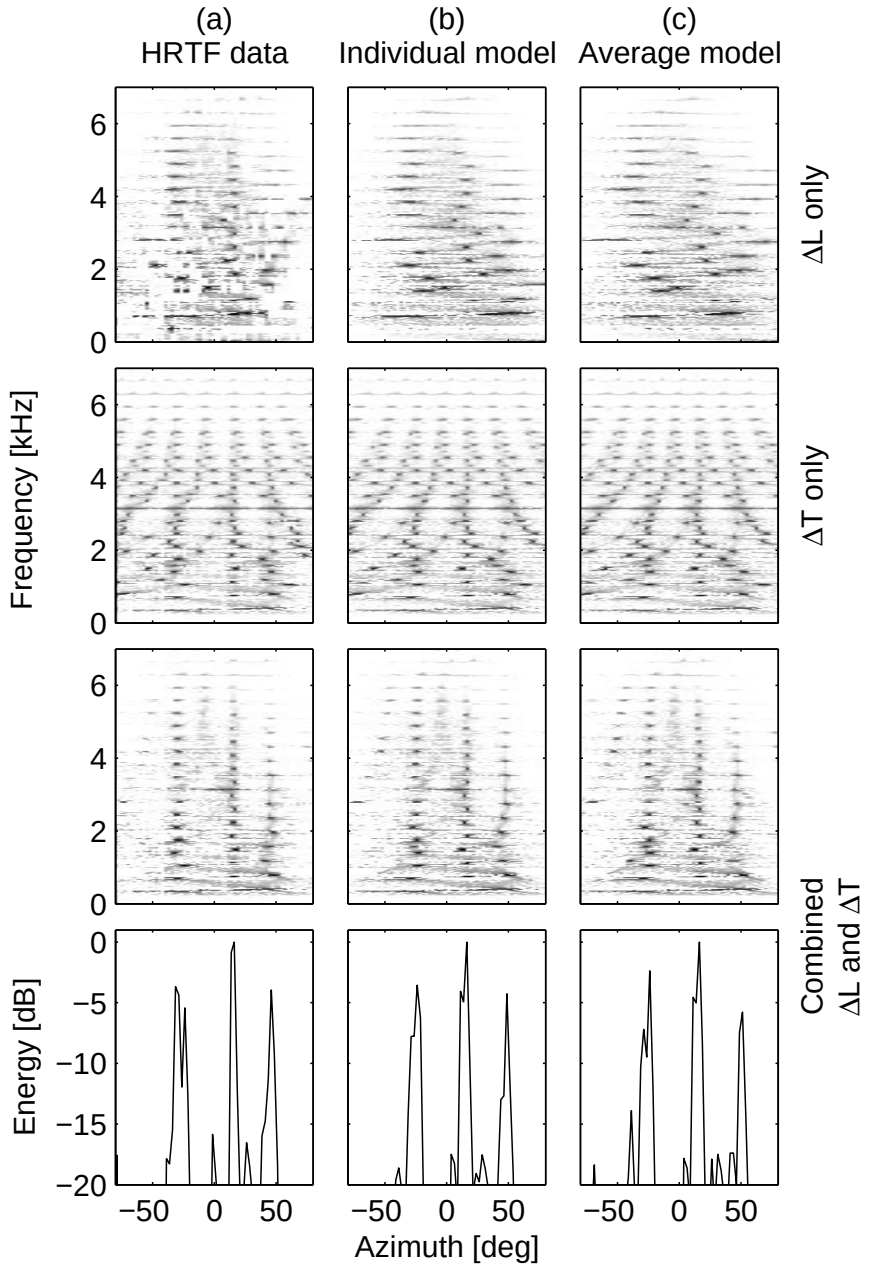


Figure 3.12: *Histograms of azimuth estimates in different frequency bands, comparing the three different models for azimuth lookup. (a): HRTF data. (b): Individual parametric model. (c): Average parameter model.*

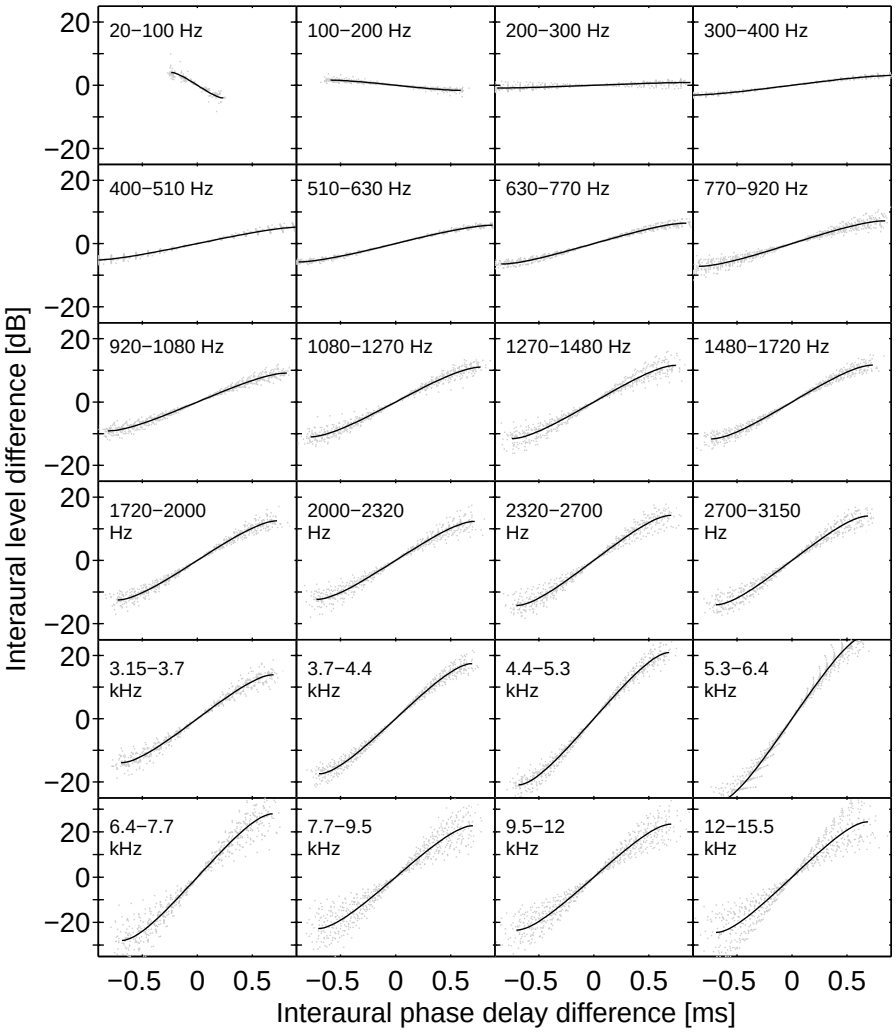


Figure 3.13: *Relation between ILD and ITD as function of azimuth in different frequency bands. The gray dotted lines show the individual parametric models for all the 45 subjects. The solid black line shows the average parameter model.*

accurate. At low frequencies the azimuths are overestimated, e.g. as in the partials at about 1 kHz. This yields additional peaks in the marginal histograms. However, qualitatively, the three different source locations are still visible as three distinct groups of peaks.

3.6 Localization in dynamic scenes

The examples that have been presented were all based on static, artificial mixtures, i.e. the binaural signals were obtained by filtering the source signals with left/right HRTF pairs for different azimuth angles. In these examples the histograms were computed for each individual frequency index, e.g. Figure 3.12, or in different frequency bands, e.g. Figure 3.11. These examples were primarily made in order to visualize the performance of the proposed method for different frequencies. These histograms were computed for the entire duration of the sounds, i.e. averaged over all time indexes. Due to this fact, they are are not applicable to real-time processing. Moreover, if the sources or sensors move, the time-varying azimuths will be averaged over time, yielding virtually useless histograms without any clear peaks. Finally, if one source is active only for a short time compared to the other sources, its corresponding peak in the azimuth histogram will be very weak or not observable at all. In practice, therefore, the histograms must be made over all frequencies in short periods of time.

Figure 3.15 shows the localization of two sources in a static scene. One source (left) consisted of two sinusoids and was located directly in front. The other source was a much weaker white noise signal and was located at 20 degrees to the right. Panels (a) and (e) show these source signals as function of time. The difference in strength is clearly visible. Panels (b)–(d) show histograms of the azimuth estimates over all frequencies as function of time (for each time index k in the STFT). Panel (b) shows weighted histograms in logarithmic scale. Each azimuth estimate was weighted by the total energy in the corresponding left/right pair of spectral coefficients. Due to the relative strength of the signals, only the left source is detectable. Panel (c) shows the unweighted histogram. Due to the wide band nature of the right source, most of the azimuth estimates correspond to its position at about 20 degrees. The position of the left source is barely detectable. Both types of histograms carry useful information about the source positions. The relative importance of these histograms depends on the number and on the nature of the sources. In order to take into account both weighted and unweighted histograms we propose to use their product. Panel (d) shows the product of weighted and unweighted histograms. Both source positions are clearly visible.

3.6.1 Head movements

In order to test the performance in a more realistic scenario, experiments were carried out in our audio lab. Two small microphones were fitted at the ear entrances of the author and different sound sources were played back at different loudspeakers. Since the HRTFs of the author had not been measured, the average parameter model was employed. This means that only one parameter was needed, namely the distance between the ears.

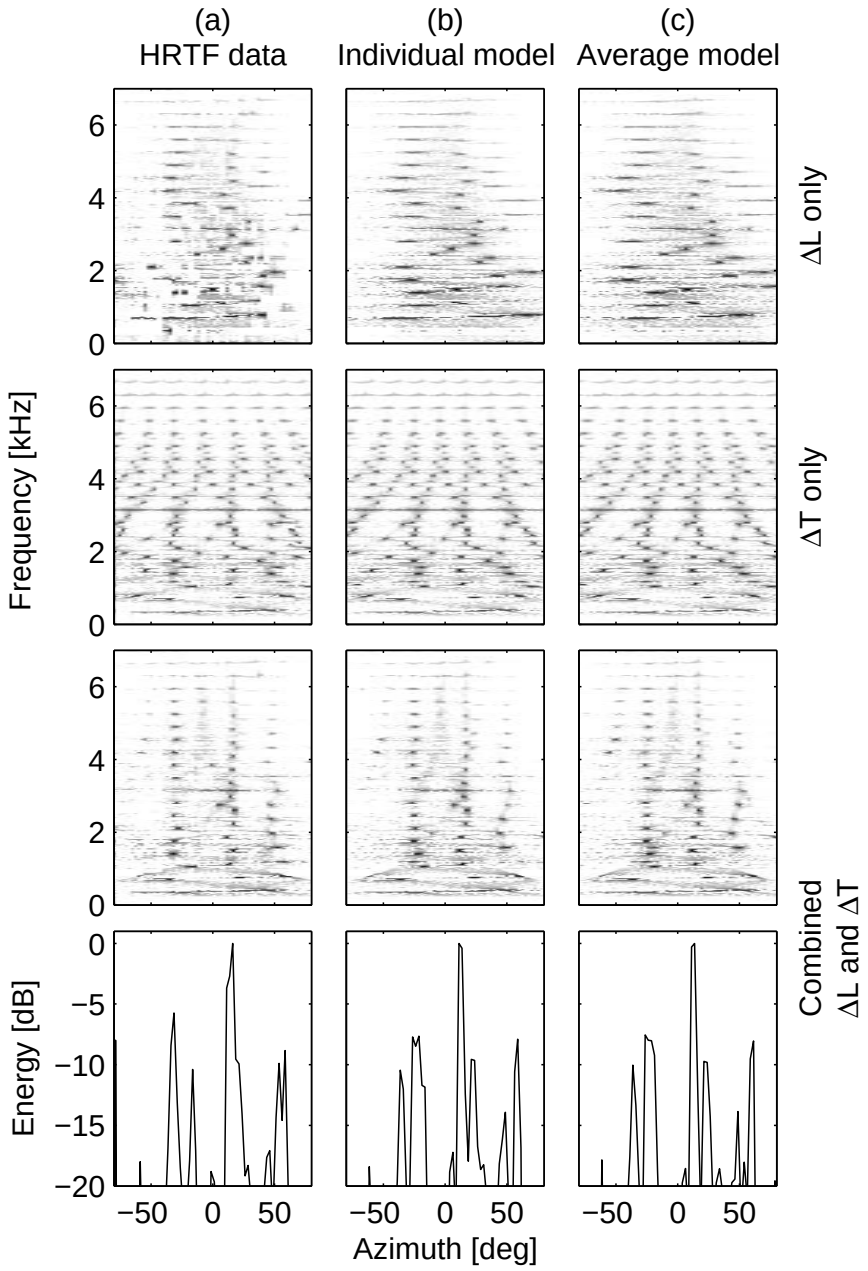


Figure 3.14: *Histograms of azimuth estimates in different frequency bands for sources at 30 degree elevation, comparing the three different models for azimuth lookup (measured/optimized for 0 degree elevation). (a): HRTF data. (b): Individual parametric model. (c): Average parameter model.*

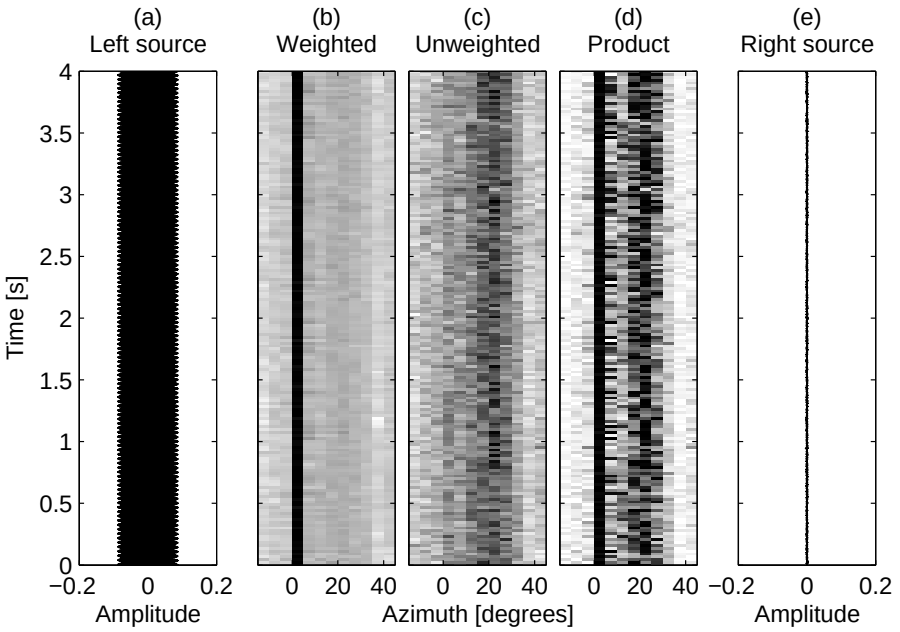


Figure 3.15: *Localization of two sources in binaural signal, showing the source signals and histograms of the azimuth estimates as function of time. (a): Left source signal at zero degree azimuth, consisting of a superposition of two relatively strong sinusoids. (b): Energy weighted azimuth histograms. (c): Unweighted histograms. (d): Product of weighted and unweighted histograms. (e): Right source signal at about 30 degrees azimuth, consisting of weak white noise.*

Figure 3.16 shows the result for localization of one sound source located about 45 degrees to the right. Panels (a) and (e) show the signals observed at the left and right ear, respectively, as function of time. The figure layout is the same as in Figure 3.15 except that the signals observed at the ears are shown in Panels (a) and (e), as opposed to the original source signals. The first two seconds, the head was kept still. Effectively, the peaks in all three histograms remain constant at about 50 degrees. Then the head was turned slowly to the right (about 60 degrees total rotation). Relative to the head, the source moves to the left until it ends up on the left side. After 6-7 seconds, the head was turned back to the original position.

In this example the source position relative to the head is visible in all three histograms as a strong histogram peak that evolves in time. This forms a trajectory that describes the source azimuth as function of time. In addition, there is a weaker parallel trajectory further to the right. We found that this is due to model mismatch at higher frequencies, i.e. above 5 kHz the ILD based azimuths are overestimated. The experiment was repeated with larger head movements. Initially the listener faced the source, then he turned the head to the left and then to the right until the source ended up at about 50 degrees to the left. Figure 3.17 shows the resulting histograms. Similarly to the first example a parallel trajectory is visible due to overestimated azimuths

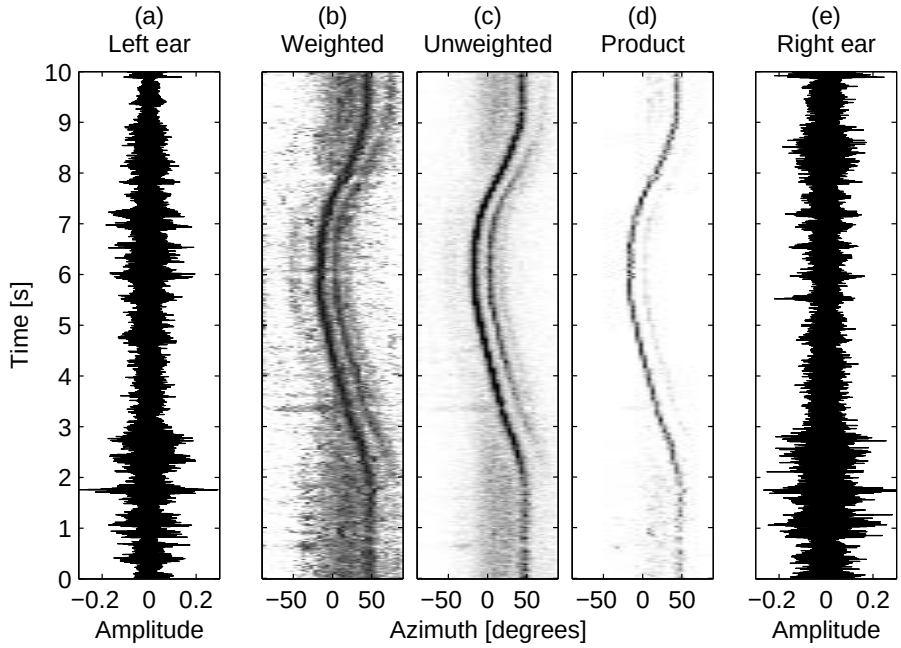


Figure 3.16: *Localization of one source in binaural signal, including head movement. The panels show the observed mixture signals and histograms of the azimuth estimates as function of time. (a): Signal observed at left ear. (b): Energy weighted azimuth histograms. (c): Unweighted histograms. (d): Product of weighted and unweighted histograms. (e): Signal observed at right ear.*

above 5 kHz. We note that for azimuths of magnitude larger than about 50-60 degrees no sharp peak is detected and the localization of the source degrades significantly. The reason for this is unknown. One possible explanation is that at extreme azimuths the direct sound to the ear farthest away is strongly attenuated and the indirect sounds (reflections) may be most important for the estimation of spatial cues. In this case, the combinations of level and time differences between the ears do not correspond to the free-field situation for which the parametric model was developed. However, a further study would be necessary in order to determine this phenomenon.

The previous experiment shows how well the average parameter model can perform for a single source in a moderately reverberant environment. This establishes the reference for the best performance that can be expected with the use of the average parameter model. In situations where there is more than one source, in general, the performance degrades. This depends on the sources and on the degree to which the cues are corrupted, as discussed in Section 2.6.1. Figure 3.18 shows an example with two sources. The listener was located in the middle of the room with one loudspeaker playing music directly in front. After 2-3 seconds a short excerpt of male voice was presented at about 20 degrees to the right. After this the listener turned the head to face this source whose signal was repeated at about 7 seconds. When the speaker is active its position can be seen in the histograms. However, due to the fact

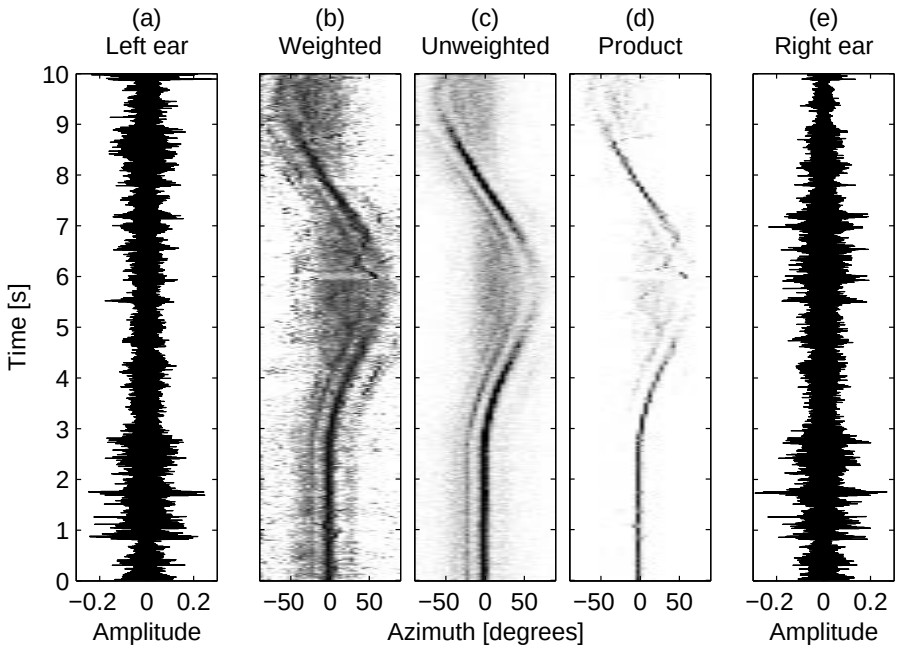


Figure 3.17: *Localization of one source in binaural signal, including head movement. The panels show the observed mixture signals and histograms of the azimuth estimates as function of time. (a): Signal observed at left ear. (b): Energy weighted azimuth histograms. (c): Unweighted histograms. (d): Product of weighted and unweighted histograms. (e): Signal observed at right ear.*

that this signal was stronger than the music, the peak corresponding to the position of the music source is not visible when the speaker is active. This could be improved by tracking source positions over time. We emphasize that all the work in this thesis is based on histograms of azimuth estimates that were computed individually for each left/right pair of spectral coefficients. No processing across frequency or time has been performed.

3.6.2 Front-back discrimination

The technique that has been proposed in this chapter is based on azimuth estimation for individual left/right pairs of spectral coefficients. This is a powerful property with respect to the localization of sources in dynamic scenes, as demonstrated in the examples including head movements. As discussed in Section 2.3.4, head movements are of vital importance for the localization of source on "cones of confusion". The application of head movements in order to resolve front-back ambiguities is illustrated in Figure 3.19. In this example two music pieces were used as source signals. One was located directly in front of the listener and the other one directly behind. Both these positions yield zero level and time differences between the ears. The first 1–2 seconds, the head was kept still. In this case, the two sources cannot be distinguished. Subsequently, the head was turned slightly to the left and then returned to the original po-

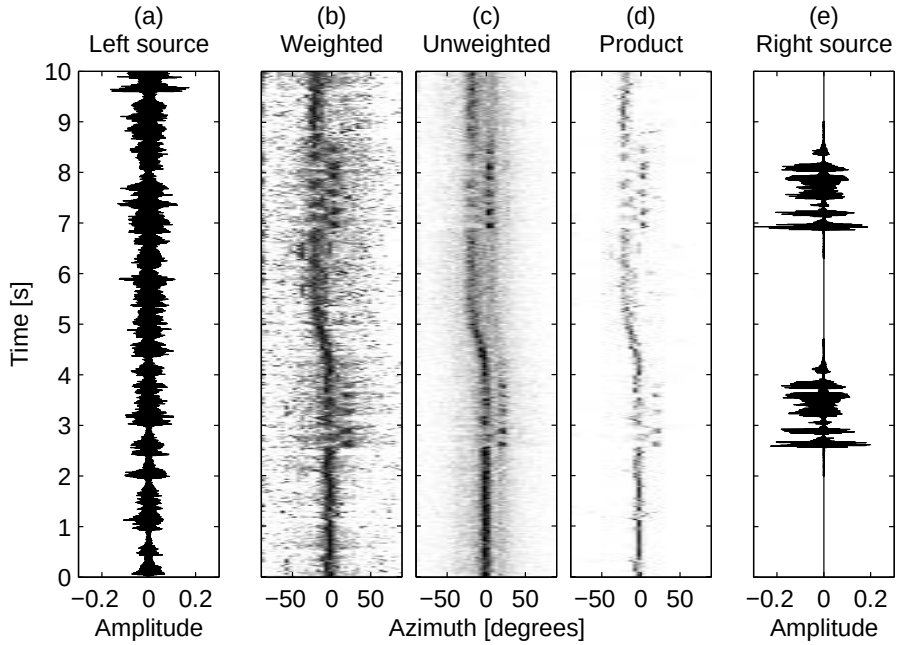


Figure 3.18: *Localization of two sources in binaural signal, including head movement. The panels show the observed mixture signals and histograms of the azimuth estimates as function of time. (a): Left source signal, jazz piece with three instruments. (b): Energy weighted azimuth histograms. (c): Unweighted histograms. (d): Product of weighted and unweighted histograms. (e): Right source signal, monophonic piano with two strong drum beats.*

sition. This was repeated twice. Once the head started to turn, the source behind the listener got closer to the left ear, effectively moving towards the left. At the same time, the source in front got closer to the right ear, moving in the opposite direction. The peaks in the histograms at zero azimuth divide into two distinct trajectories, left and right, respectively.

3.7 Discussion

In this chapter, a method has been proposed that enables the detection of sources in dynamic environments based on the signals observed by the two ears. As seen in the examples, the method can work well in many cases. Nevertheless, there are several aspects of the proposed technique that can be improved. This section gives a short discussion of some of these drawbacks.

ITD estimates based on phase delays: The examples showed that the azimuth estimation is only reliable below about 6 kHz. Above this frequency the relation between ILDs and azimuth angle becomes more complex. In addition, the variation in level difference between different heads increases significantly, as seen in Figure 3.8. At high frequencies the ITDs based on narrow-band phase delays are of little use. In order to extend the method

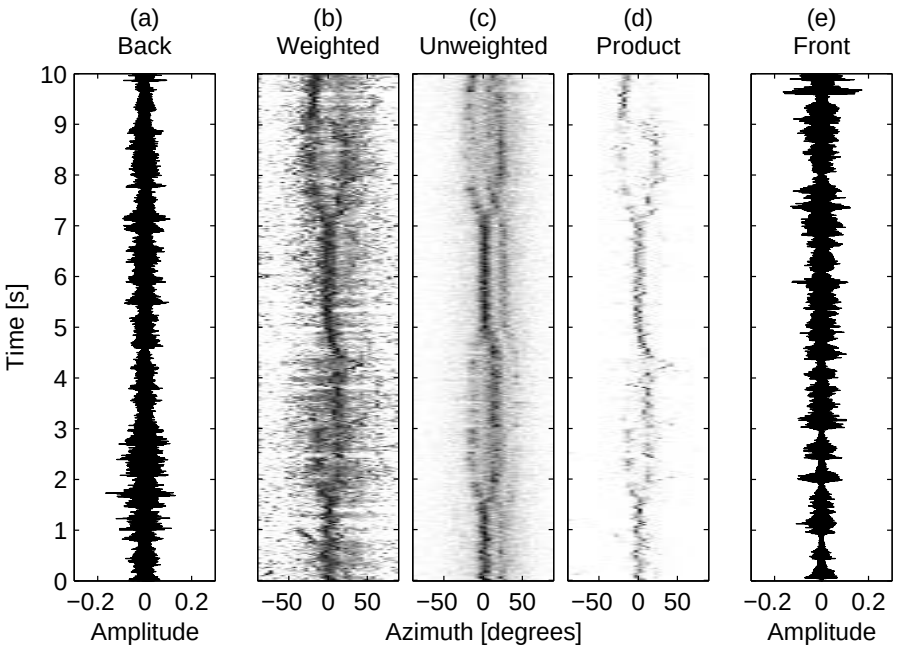


Figure 3.19: *Localization of two sources in binaural signal, resolving front-back confusion by head movements. The panels show the two source signals (back/front) and histograms of the azimuth estimates as function of time. (a): Source signal for source behind listener (an excerpt of a jazz piece). (b): Energy weighted azimuth histograms. (c): Unweighted histograms. (d): Product of weighted and unweighted histograms. (e): Source signal of source in front (an excerpt of a different jazz piece).*

to work better with high frequencies one possible approach is to employ ITD estimates based on group delays, rather than on phase delays. An example of this can be found in (Sontacchi *et al.*, 2002). Envelope delay estimates must be computed in wider bands and therefore do not provide as fine spectral resolution as that of the phase delay estimates presented in this chapter. However, the use of envelope delays can be justified from a perceptual point of view (Blauert, 2001). Since envelope delays are widely used it is not discussed further in this thesis but only mentioned as a possible extension to the presented work.

Equalization: In Section 3.6 different types of histograms were discussed. In particular it was shown how the product histograms combined the information conveyed by both weighted and unweighted histograms. However, for many naturally occurring sources, such as harmonic instruments, the energy decays with frequency. This means that the lower frequencies are predominant in the histograms. The technique for localizing sources is based on the assumption that the majority of azimuth estimates are accurate and such that the peaks are detectable in the histograms. The parametric models and the average parameter models that have been

presented in this chapter are by no means perfect. Since many azimuth estimates will be erroneous it is not favorable to emphasize the high energy low frequencies. In order to weight on all frequencies more equally the sensor signals would need to be equalized.

Corrupted estimates: Some azimuth estimates are erroneous due to model mismatches. However, an equally important reason for wrong azimuth estimates is related to sources whose energies overlap in time and in frequency, as discussed in Section 2.6.1. This is discussed further in the following chapters.

Chapter 4

Binaural separation

In the previous chapter, a technique for localization of sources in binaural signals was proposed. It was based on the short-time Fourier transform (STFT) spectra of the signals observed at the two ears, where the azimuth was estimated for each left/right pair of spectral coefficients. Each azimuth estimate is thus related to a narrow-band signal component in a short time interval. The trade-off between frequency and time resolution is controlled by the choices of window function and window length in the STFT, given in Equation (2.1). The narrow-band and short-time properties of the azimuth estimates make them applicable to source separation and enhancement. The goal is to get "valid" binaural signals where the spatial information is preserved, but where only one source is present. This is the topic of the present chapter.

4.1 Separation by spatial windowing

The source localization technique that was proposed in Chapter 3 provides a convenient means for the separation of sources. For each time index k and frequency index q, the azimuth $\theta(k,q)$ is estimated. In the histograms of these estimates for a given time index k, the peaks provide estimates of the source locations. The tracking of these peaks over time yields estimates of the source locations relative to the head as a function of time. When the sources have different physical locations and are detected individually in the histograms, each source corresponds to one distinct trajectory, as was illustrated in Figures 3.16–3.18. We denote these time varying source azimuths by $\vartheta_i(k)$, where i is the source index.

Given the azimuth estimates $\theta(k,q)$ and the source azimuths $\vartheta_i(k)$, a weighting factor is computed for each spectral coefficient. For the i^{th} source we denote this weighting factor $G_i(k,q)$. For each source, the binaural separated signal (left and right ear) is obtained by multiplying this weighting factor by the STFT spectra of the left and right sensor signals $X_{\text{left}}(k,q)$ and $X_{\text{right}}(k,q)$, respectively. The i^{th} reconstructed source signal is given by:

$$Y_{i,\text{left}}(k,q) = G_i(k,q)X_{\text{left}}(k,q) \tag{4.1}$$

and

$$Y_{i,\text{right}}(k,q) = G_i(k,q)X_{\text{right}}(k,q), \tag{4.2}$$

for the left and right ears, respectively.

4.1.1 Relation to existing separation techniques

The weighting factors $G_i(k,q)$ determine the relative weighting of the spectral coefficients for each separated source signal. This method is closely related to the existing separation techniques based on time-frequency weighting that were discussed in Section 2.4.3. Different approaches can be considered for the estimation of these weights.

Binary weights

The DUET method (Jourjine *et al.*, 2000), discussed in Section 2.4.3, is based on the W-disjointness assumption, i.e. on the assumption that the STFT spectra of the different sources do not overlap significantly. Under this assumption, only one source contributes significant energy to a given spectral coefficient and each spectral coefficient is assigned to one source exclusively. In order to separate the sources, the use of binary weights was proposed. Shortly explained, each spectral coefficient is assigned to the source that is nearest in the two-dimensional parameter space of level and time differences shown in Figure 2.14.

This approach is also applicable to the separation based on azimuth estimates. The azimuth estimates obtained by the technique proposed in Chapter 3 provide a one-dimensional parameter space. The binary masks can be computed by assigning each spectral coefficient to the source whose azimuth estimate is closest. This is given by

$$G_i(k,q) = \begin{cases} 1, & \arg\min\limits_j \; |\vartheta_j(k) - \theta(k,q)| = i, \\ 0, & \text{otherwise.} \end{cases} \tag{4.3}$$

For instance, if two sources are detected at 10 and 30 degrees azimuth, respectively, the binary masks for these sources are obtained by assigning the spectral coefficients whose azimuth estimates are smaller than 20 degrees to the first source, and the spectral coefficients whose azimuth estimates are larger than 20 degrees to the second source. The one-dimensional azimuth parameter space has an intuitive interpretation with respect to the significance of the weights. In particular, the binary weights can be considered as spatial windowing of the spectral coefficients using a rectangular spatial windows. This procedure is similar to the spatial filtering in beamforming techniques. However, a higher spatial resolution can be achieved since the level differences between the sensors are also taken into account in the azimuth estimates.

Perceptually motivated weights

The source enhancement techniques that were discussed in Section 2.4.3 also employ weighting factors. In these techniques, the weights can take on continuous values in the range from zero to one and are not restricted to a binary zero or one decision. However, these weights are not computed for the individual spectral coefficients. Rather, the sensor signals are decomposed into perceptual bands by use of STFT and grouping of spectral coefficients into perceptual bands (Peissig, 1992) or by use of an auditory filter bank (Bodden, 1993). For

each perceptual band a time-varying weight is computed. These weights can be considered as non-rectangular spatial windows.

Spatial windowing

The continuous-valued weights that were used in the source enhancement techniques are more general than the binary weights that were used in the DUET method. For source enhancement techniques based on perceptual bands these weights (and the resulting spatial windowing) are motivated by perception. This means that not only they provide smooth spatial windows but they also ensure smooth spectral and temporal envelopes in the processed signals. On one hand, the perceptual processing aims at avoiding audible artifacts such that the quality of the separated signals is high. On the other hand, the aim of these techniques is to enhance one source in a binaural signal and not to truly separate it out of the mixture. Indeed, when one source is separated (enhanced) the other sources may still be audible.

For separation methods that are based on individual spectral coefficients, non-rectangular spatial windows do not provide any advantage over binary weights. Due to the fact that each spectral coefficient is treated independently, the smoothness of the spatial window has little influence on the smoothness of the spectral and temporal envelopes. In this thesis we have used rectangular spatial windows. For each source azimuth trajectory $\vartheta_i(k)$ we use the following weights/spatial windows:

$$G_i(k,q) = \begin{cases} 1, & |\vartheta_i(k) - \theta(k,q)| < \frac{W}{2} \\ 0, & \text{otherwise,} \end{cases} \tag{4.4}$$

where W is the width (in degrees) of the spatial window. In order to deal with the artifacts that are introduced by the independent processing of the spectral coefficients, spectral and temporal smoothing can be applied as a post-processing step. This is discussed in Section 4.3.1.

4.2 Experimental results

Experiments were conducted in order to study the performance of binaural separation employing rectangular spatial windowing of azimuth estimates. Binaural signals were generated by superposition of filtered source signals using measured HRTFs from the CIPIC database (Algazi *et al.*, 2001). Each source signal was filtered with the HRTFs for the left and right ears at the corresponding source azimuth.

In the first example, two sources located at -30 and 30 degrees azimuth, respectively, were chosen. Two short excerpts of different jazz pieces were used as source signals: one rapid piece played by a trio of piano, bass, and percussions (left multi-instrument source) and one slow monophonic piano tune with two strong drum beats (right source). Figure 4.1 shows the resulting signals at the left and right ears in Panels (a) and (e), respectively. From the STFT spectra of these two signals the azimuths were estimated. This was achieved by employing the average parameter model proposed in Section 3.4 that only depends on one parameter, namely the distance between the ears. Panels (b)–(d) shows the histograms of the azimuth estimates, i.e. weighted histograms,

unweighted histograms, and the product of these, respectively. The different histogram types were discussed in Section 3.6. Most of the time, the left source is dominant, visible as a strong source azimuth trajectory at about -30 degrees. The right signal is much weaker, except for the two drum beats at about 3 and 6 seconds and the piano signal after about 6–8 seconds. For these time ranges, its position of about 30 degrees can be observed.

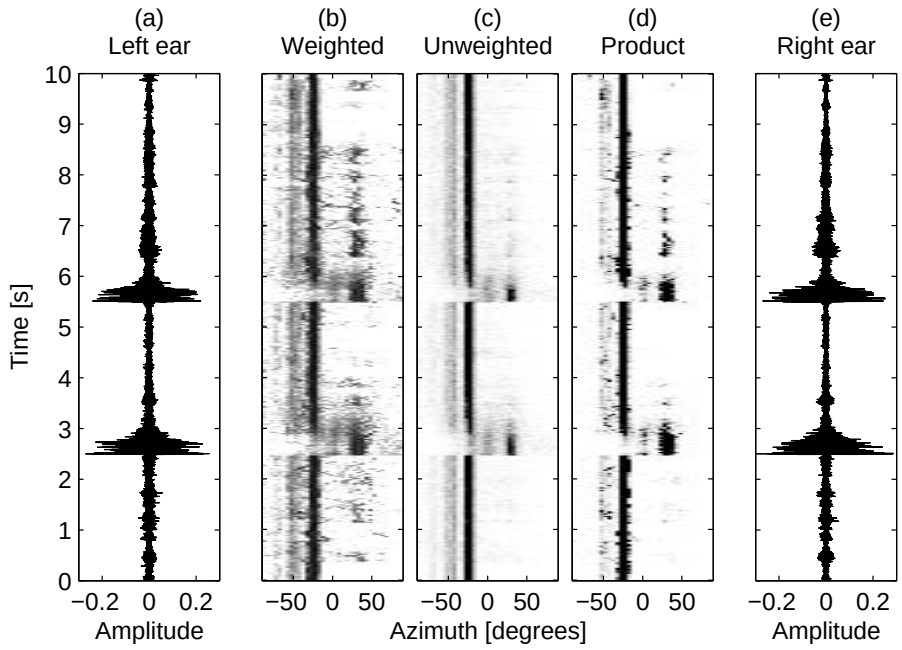


Figure 4.1: *Mixture of two sources, showing the observed mixture signals and histograms of the azimuth estimates as function of time. (a): Signal observed at left ear. (b): Energy weighted azimuth histograms. (c): Unweighted histograms. (d): Product of weighted and unweighted histograms. (e): Signal observed at right ear.*

Based on the azimuth estimates obtained by the average parameter model, the two sources were separated by spatial windowing. A rectangular window of width $W = 30$ degrees was used, i.e. all spectral coefficients whose azimuth estimates were between -45 and -15 degrees were assigned to the left source. The right source was obtained similarly (15–45 degrees). The spectrograms of the separated sources are shown in Figure 4.2. Panel (a) shows the mixture of the two sources as observed by the left ear. The left and right separated sources are shown in Panels (b) and (c), respectively. These are the separated signals for the left ear, obtained by Equation (4.1). Similar spectrograms can be obtained for the right ear by Equation (4.2). The individual harmonics of the monophonic piano notes and the two strong drum beats are clearly visible in the right source. The majority of the remaining spectral coefficients have been assigned to the left source.

The time domain signals shown in Figure 4.3 are obtained by inverse STFT with overlap add. Panel (a) shows the observed signal at the left ear. The left

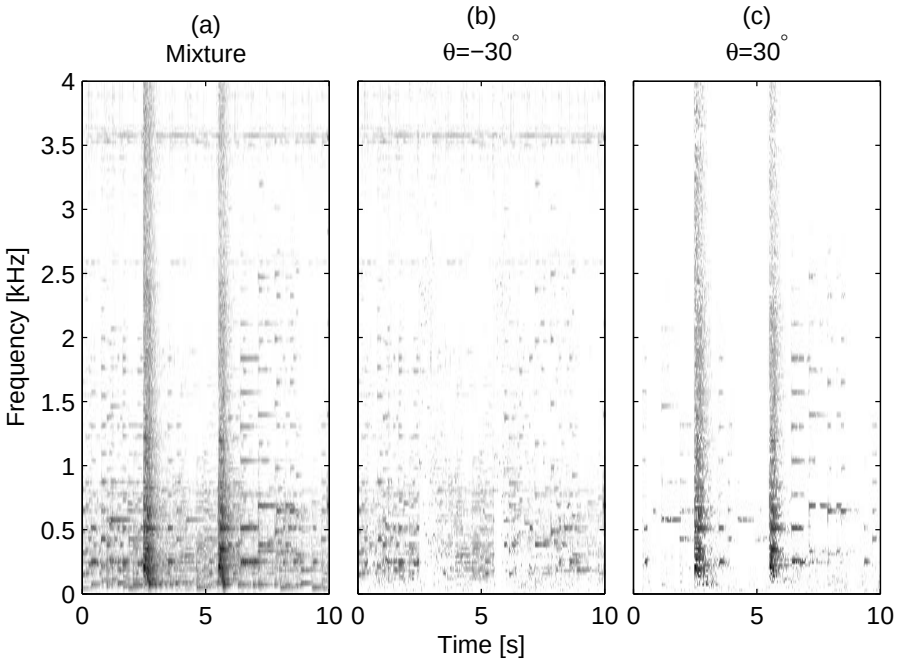


Figure 4.2: *Spectrograms showing the separation of two sources from a binaural signal. (a): Signal observed at left ear. (b): Separated signal for source at* -30 *degrees azimuth. (c): Separated signal for source at* 30 *degrees azimuth.*

and right separated signals (for the left ear) are shown in Panels (b) and (c), respectively. The strong drum beats have been correctly removed in the left signal.

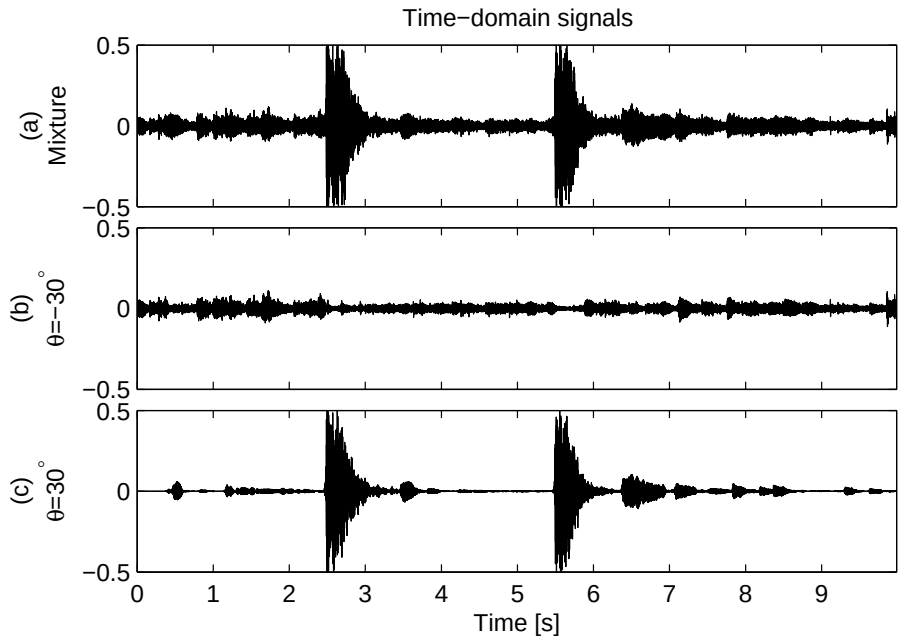


Figure 4.3: *Separation of two sources from a binaural signal, showing time domain signals. (a): Signal observed at left ear. (b): Separated signal for source at* -30 *degrees azimuth. (c): Separated signal for source at* 30 *degrees azimuth.*

A similar experiment was made with 5 sources located at azimuths of -60, -30, 0, 30 and 60 degrees, respectively. Figure 4.4 shows the observed mixtures and the histograms. Depending on the temporal characteristics, the positions of all five sources can be observed in the product histograms for some time indexes. The rightmost source is the same as in the previous example. Its weak monophonic piano notes are not detectable in the histograms but the source position of about 60 degrees is visible at the times of the strong drum beats.

The spectrograms of the left sensor signal and the separated sources are shown in Figure 4.5, and the time domain signals are shown in Figure 4.6. In both these figures it can be seen how the strong drum beats are mainly assigned to the rightmost source. In the STFT spectra, the drum is visible as the strong vertical lines spanning all frequencies. However, due to the number of sources and their characteristics, the STFT spectra of the original sources overlap significantly. This yields corrupted cues, as discussed in Section 2.6.1. As a result, some of the energy of the drum beats have corrupted spatial cues and are assigned to other sources, e.g. the middle source shown in Figure 4.5(d).

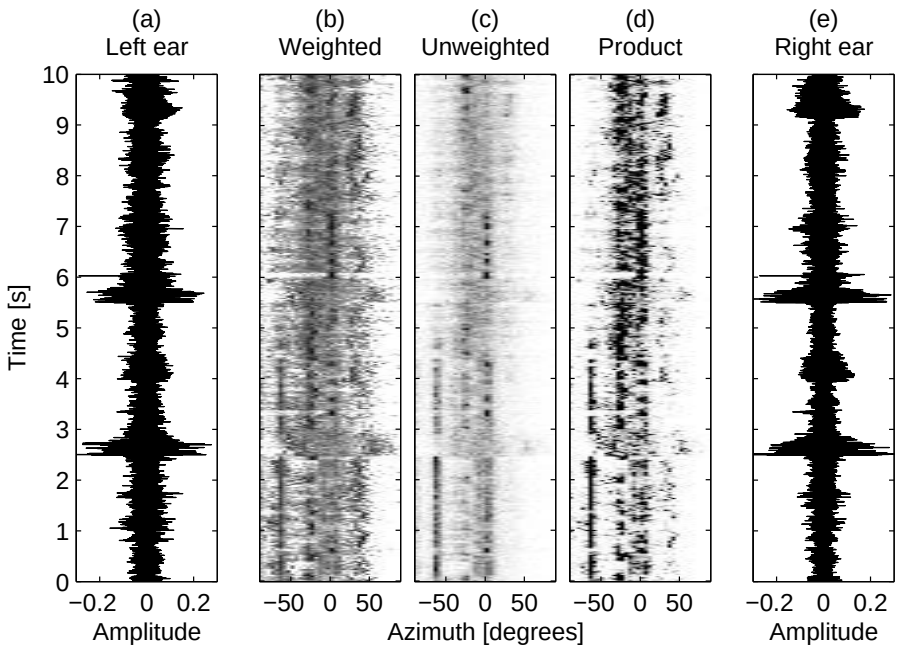


Figure 4.4: *Mixture of five sources, showing the observed mixture signals and histograms of the azimuth estimates as function of time. (a): Signal observed at left ear. (b): Energy weighted azimuth histograms. (c): Unweighted histograms. (d): Product of weighted and unweighted histograms. (e): Signal observed at right ear.*

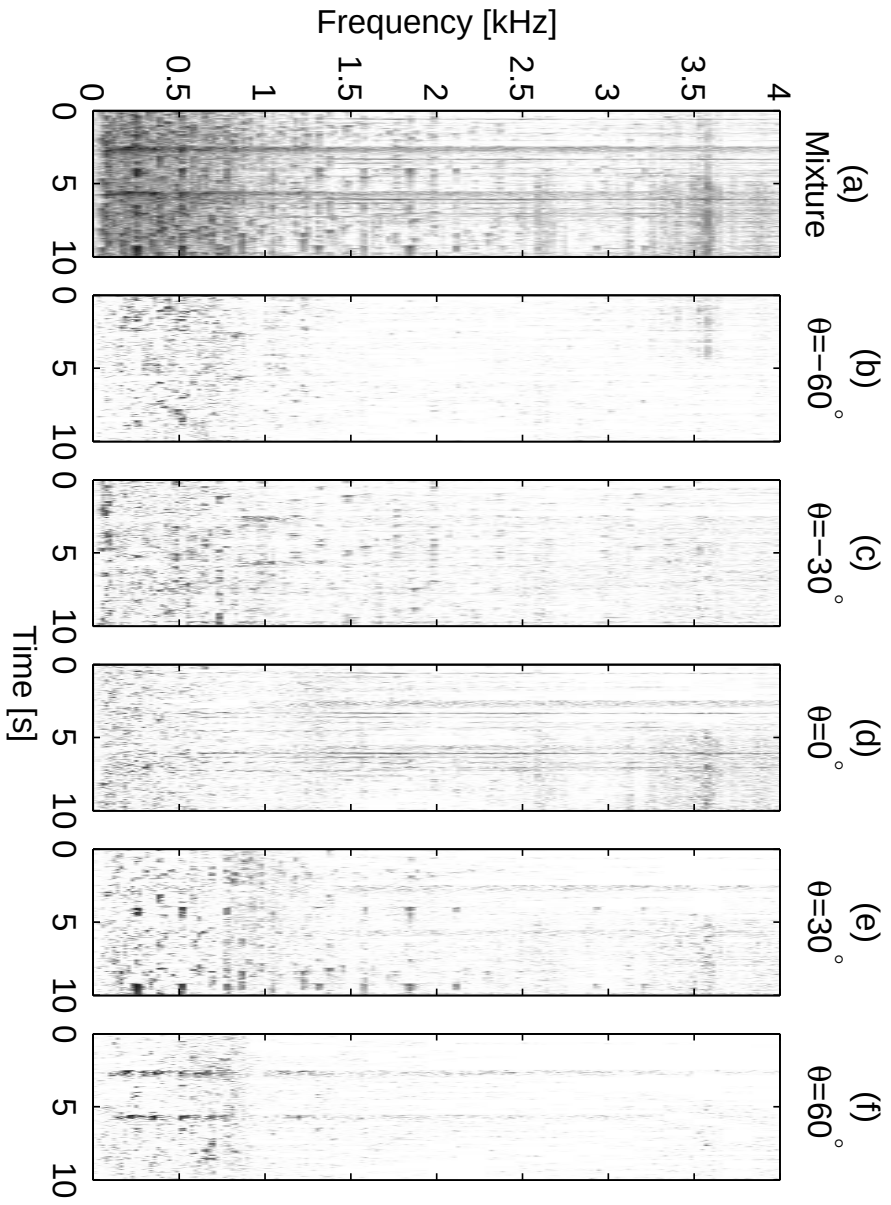


Figure 4.5: *Spectrograms showing the separation of five sources from a binaural signal. (a): Signal observed at left ear. (b)–(f): Separated signals for sources at* -60, -30, 0, 30, *and* 60 *degrees azimuth, respectively.*

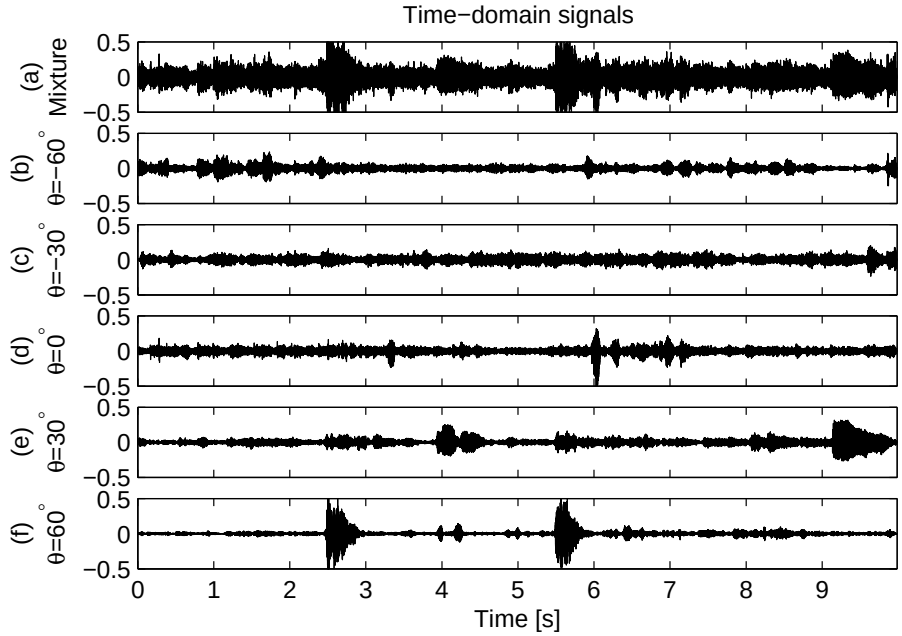


Figure 4.6: *Separation of five sources from a binaural signal, showing time domain signals. (a): Signal observed at left ear. (b)–(f): Separated signals for sources at* -60, -30, 0, 30, *and* 60 *degrees azimuth, respectively.*

4.3 Discussion

The separation technique that has been presented works with individual spectral coefficients. The major motivation was to obtain a sufficiently fine frequency resolution to achieve the detection and separation of closely spaced spectral components. In this respect, the method is different from the enhancement techniques working with perceptual bands. The experiments that have been conducted have shown that, in general, the sources have been separated in the sense that each source contains little energy from the other sources. Naturally, this depends on the number of sources, on their locations and on their temporal and spectral activity. However, even though there is little leakage between the separated sources, there may be significant artifacts. This is the topic of the present section.

4.3.1 Processing of individual spectral coefficients

The fact that each spectral bin is processed independently gives rise to some difficulties that must be taken into consideration. In particular, due to the time-frequency uncertainty principle, the spectral coefficients are not independent.

Choice of STFT windows

In the STFT the trade-off between time resolution and frequency resolution is determined by the choice of analysis window $w_a(l)$ and its length. For many applications it is preferred to use a smooth analysis window in order to have good localization in frequency, i.e. a strong main lobe with strongly attenuated side lobes. Moreover, it is desirable to have analysis and synthesis windows that enable a perfect reconstruction, as given by Equation (2.3). Figure 4.7 shows different windows of equal length in column (a) and their spectra in column (b). The rectangular window gives perfect reconstruction when used both for analysis and synthesis and for any hop-size H smaller than or equal to the window length L. This window and its spectrum are shown in the top row. The rectangular window has the narrowest main lobe but quite significant side lobes due to discontinuities. In the second row, the von Hann window is shown, also known as the raised cosine window. This is given by

$$w_{\text{Hann}}(l) = \frac{1}{2}\left(1 - \cos\frac{2\pi l}{L}\right), \qquad 0 \leq l < L, \tag{4.5}$$

where L is the window length. This window yields perfect reconstruction when used together with the rectangular window (one for analysis, the other one for synthesis) and with a hop-size of an integer sub-multiple of the window length, $H = L/G$ for integer G. This window is smooth and consequently the attenuation of the side lobes is much stronger than in the rectangular window. However, the width of the main lobe is wider due to the decreased "effective" length of the window. In the last row the half-wave sine window is shown. This is the square root of the von Hann window. The window is continuous but its derivative is discontinuous at the edges. Since the von Hann window gives perfect reconstruction when used together with the rectangular window the sine window gives perfect reconstruction when used for both analysis and

synthesis, by Equation (2.3):

$$\sum_k w_{\text{Hann}}(l-kH)w_{\text{Rect}}(l-kH) = \sum_k \sqrt{w_{\text{Hann}}(l-kH)}\sqrt{w_{\text{Hann}}(l-kH)} = C, \tag{4.6}$$

with hop-size $H = L/G$ for integer G dividing L.

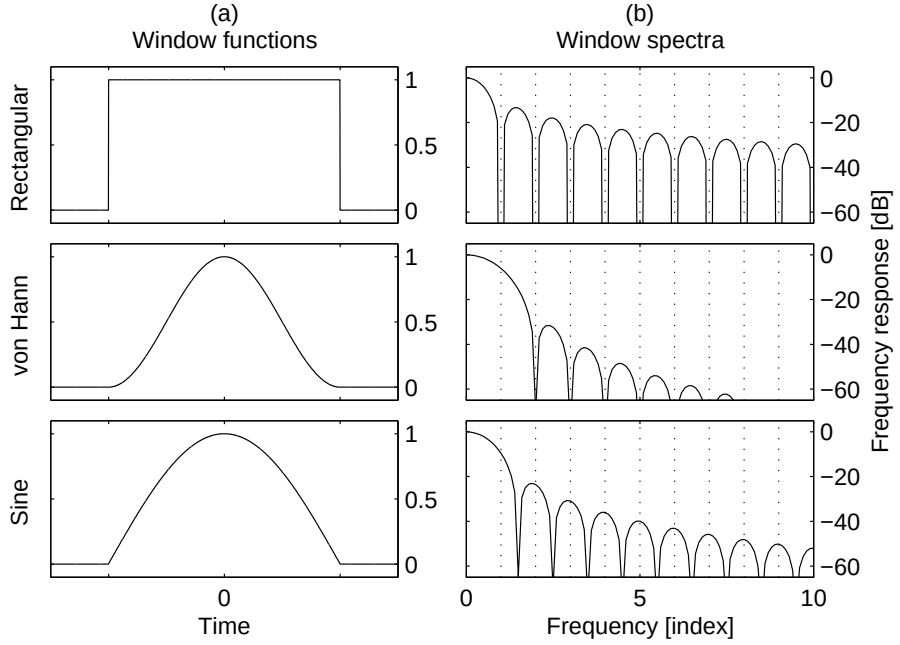


Figure 4.7: *Window functions and their spectra.*

The fact that each spectral coefficient is processed independently in the proposed separation method makes the choice of windows in the STFT very important. Figure 4.8 illustrates this fact. The top row shows the analysis of a signal for a given time index k. The signal is shown in column (a). This signal is windowed with the smooth von Hann window shown in column (b). The resulting windowed signal and its magnitude spectrum are shown in columns (c) and (d), respectively. The bottom row shows the synthesis for a single spectral coefficient in this spectrum, from right to left. In the magnitude spectrum in column (d) only one spectral coefficient is kept. When this spectral coefficient is treated in isolation, the spectral shape of the analysis window is not preserved. This also affects the shape of the analysis window in the time domain. A single spectral coefficient corresponds to a rectangular analysis window. The time signal corresponding to the single spectral coefficient is shown in column (c), where the envelope has become rectangular. This artifact can be explained as time domain aliasing due to down-sampling in the frequency domain. Finally, the rectangular synthesis window shown in column (b) is applied yielding the final signal component shown in column (a), which corresponds to one spectral coefficient.

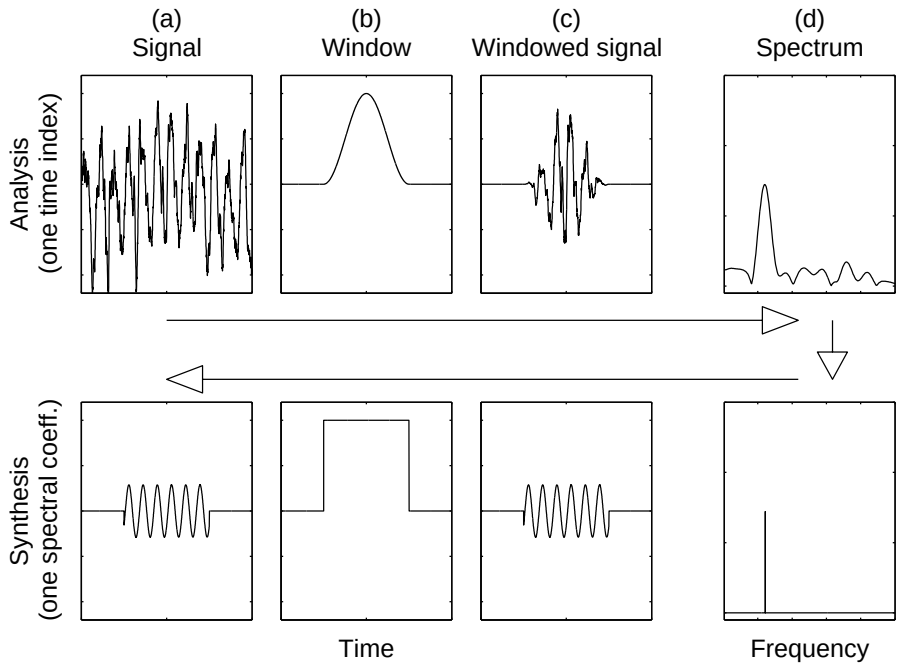


Figure 4.8: *Processing of individual spectral coefficients may introduce discontinuities. Top row: analysis of a signal for a given time index. (a): Original signal. (b): Window function at time index k. (c): Windowed signal. (d): Spectrum of windowed signal. Bottom row: synthesis of the signal component corresponding to a single coefficient. (d): Spectrum with only one non-zero spectral coefficient. (c): Corresponding signal component in the time domain. (b): Synthesis window. (a): Signal component resulting from a single spectral coefficient, discontinuous at the edges.*

The given choice of analysis and synthesis windows enables perfect reconstruction. In other words, if the processing depicted in Figure 4.8 is repeated for all time indexes k and for all spectral coefficients q then the superposition of the resulting signal components is equal to the original signal. When a rectangular synthesis window is employed, each spectral coefficient corresponds to an "unwindowed" time-domain signal. This signal may have strong discontinuities at the edges, as seen in column (a) in the bottom row. When all spectral coefficients are taken into account, these discontinuities are canceled by other signal components and perfect reconstruction can be obtained. However, when the spectral coefficients are processed individually, as in the separation technique, these discontinuities will in general lead to strong audible artifacts in the resulting separated signals. The result can be somewhat improved by choosing a very short hop-size of only a few samples. However, a better solution is to avoid the rectangular window in both the analysis and the synthesis. The half-wave sine window shown in the bottom row in Figure 4.7 has several advantages. Most importantly it is continuous. When used for the analysis it leads to a better side lobe attenuation than that of the rectangular window. In the synthesis of a single spectral coefficient it ensures that there are no discontinuities. Figure 4.9 illustrates this property. In addition, as already mentioned, it enables perfect reconstruction when used for both analysis and synthesis.

When using the sine window for both analysis and synthesis, the separated signals do not have discontinuities. However, the separated signals still have audible artifacts.

Spectral smoothing

As mentioned in the introduction to this section, the spectral coefficients are not independent. A sinusoid with frequency equal to the center frequency of a given spectral coefficient yields only one non-zero spectral coefficient when no windowing is applied. When the sinusoid is windowed, the resulting spectrum is the convolution of the single coefficient spectrum of the signal with the spectrum of the window, as shown in Figure 4.10. The top row shows the sinusoid signal in column (a) and its spectrum in column (b). In the bottom row the same graphs are shown for the signal windowed with the sine window. In the spectrum in column (b) (linear scale), the discrete spectral coefficients indexed by q are indicated with black points. Due to windowing the spectral coefficients in the neighborhood of the frequency of the sinusoid are also significant. The same discrete frequency indexes are indicated with dotted vertical lines for the different windows in Figure 4.7(b).

In order to avoid the artifacts introduced by time domain aliasing, spectral smoothing can be useful. We conducted informal experiments where the binary weighting functions were smoothed over frequency. For the smoothing we used a filter whose coefficients were chosen as the magnitude spectrum of the windowing function in a small neighborhood around the main lobe. The weights were upper limited by 1. Effectively, this avoids the abrupt changes in the spectral envelope. In the experiments, the aliasing artifacts were smaller. However, this was at the cost of introducing more leakage between the sources, which is a bothersome artifact.

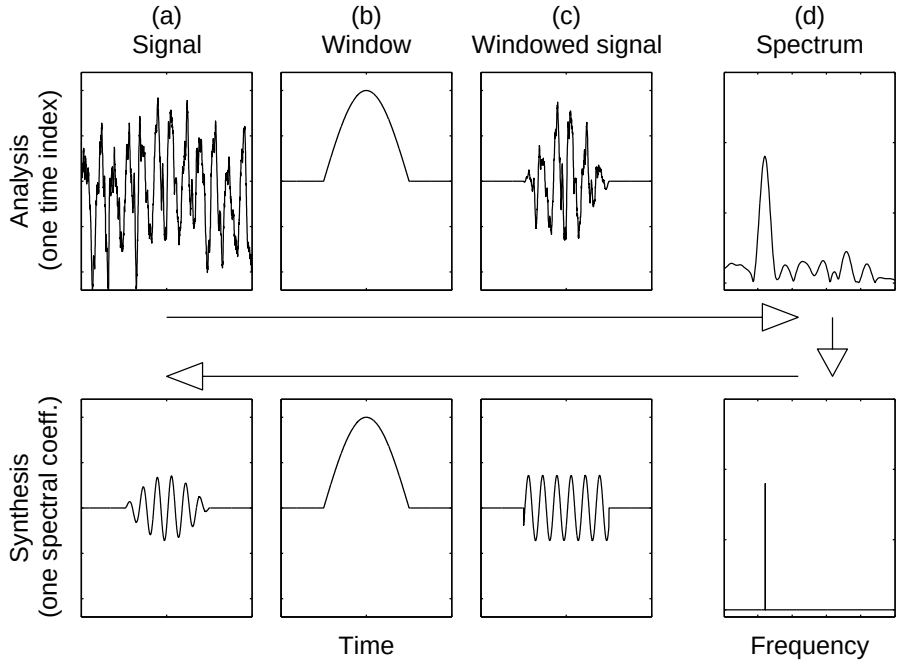


Figure 4.9: *Avoiding discontinuities when processing individual spectral coefficients by choosing continuous analysis and synthesis windows. Top row: analysis of a signal for a given time index. (a): Original signal. (b): Window function at time index k. (c): Windowed signal. (d): Spectrum of windowed signal. Bottom row: synthesis of the signal component corresponding to a single spectral coefficient. (d): Spectrum with only one non-zero coefficient. (c): Corresponding signal component in the time domain. (b): Synthesis window. (a): Signal component resulting from a single spectral coefficient without discontinuities due to the continuous synthesis window.*

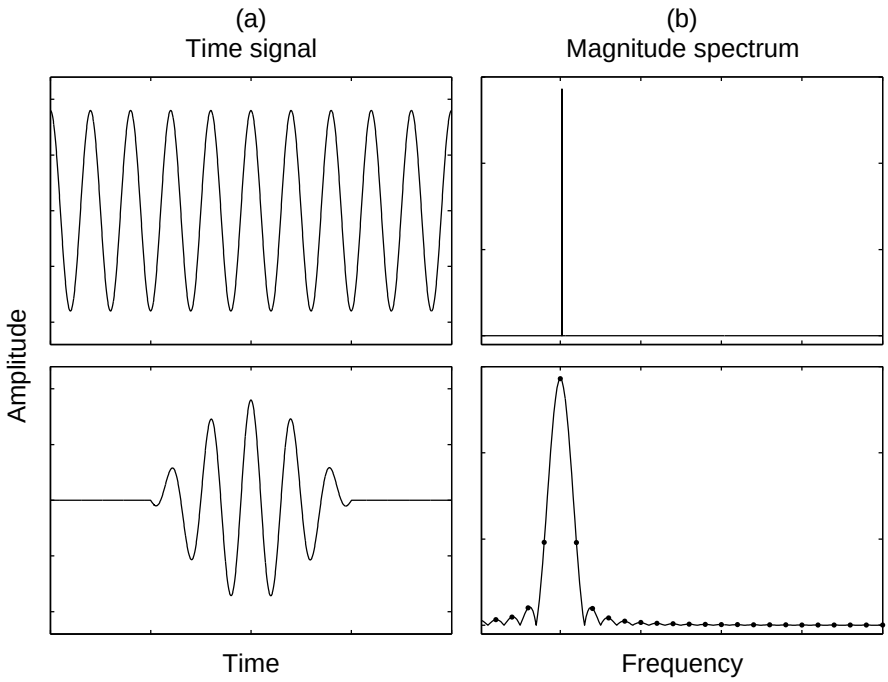


Figure 4.10: *Effect of windowing: signal spectrum is convolved with window spectrum.*

Suppression versus separation

If some leakage between the sources is acceptable, it is possible to avoid some of the artifacts and improve the quality of the separated signals. This can be done by taking a suppression approach instead of aiming at true separation. In other words, rather than applying a rectangular spatial window about the position of the source of interest ("spatial bandpass filter"), inverse spatial filters can be applied to the other sources that should be suppressed ("spatial band-stop filters"). These "suppression windows" can be narrower than those used for separation. All the spectral coefficients that have azimuth estimates close to the azimuth estimate of one of the unwanted sources are suppressed. The non-suppressed signal will contain energy also from the other sources. However, the sound will be more natural and rich since this is a more liberal approach that allows more spectral coefficients to be part of the separated signals.

4.3.2 Corrupted cues

In the previous discussion and examples, the problem of corrupted cues introduced in Section 2.6.1, has not been considered. A brief discussion is given in the present section. This serves mainly as an introduction and motivation to the following chapter that deals with separation of overlapping harmonics.

Overlapping energy

When the STFT spectra of different sources overlap significantly, we obtain corrupted cues. This phenomenon is not unique to our computational model for azimuth estimates, but it also occurs in the human auditory system. For instance, if the same signal is played back with the same strength at two loudspeakers that are located symmetrically at 30 degrees to the left and to the right of the head, in general, only one source will be perceived whose position will appear in the middle between the loudspeakers. In this example, the resulting signals at the two ears will be identical (for a perfectly symmetrical head), and the azimuth estimation proposed in Chapter 3 will yield most estimates at about zero azimuth. If one loudspeaker is stronger, the perceived location will move towards that loudspeaker. This "amplitude panning" is the basic principle that is exploited in stereo recordings and playback.

In a mixture of two sources, many azimuth estimates will correspond to the positions of the two sources. This was clearly visible as the two distinct source trajectories in Figures 4.1 and 3.19. In both these cases the source signals consisted of different instruments and were concurrently active. However, the signals were chosen from different jazz pieces, playing in different keys. Figure 4.11 shows an example of two sources located 30 degrees to the left and right, respectively. The figure layout is the same as that of Figure 4.1. However, in this example, two different excerpts of the same jazz piece (piano and percussions) were chosen as source signals. In the left signal the piano and percussions are active. The right signal has a strong piano note at about 4 seconds and the percussions start after about 4.5 seconds. In the beginning only two source trajectories can be observed. However, once the percussions starts in the right signals, these overlap significantly with those of the left signal. This leads to a high number of azimuth estimates that reflects none of the source

positions but fall somewhere in between, visible as a third "phantom" source trajectory at about zero degrees azimuth.

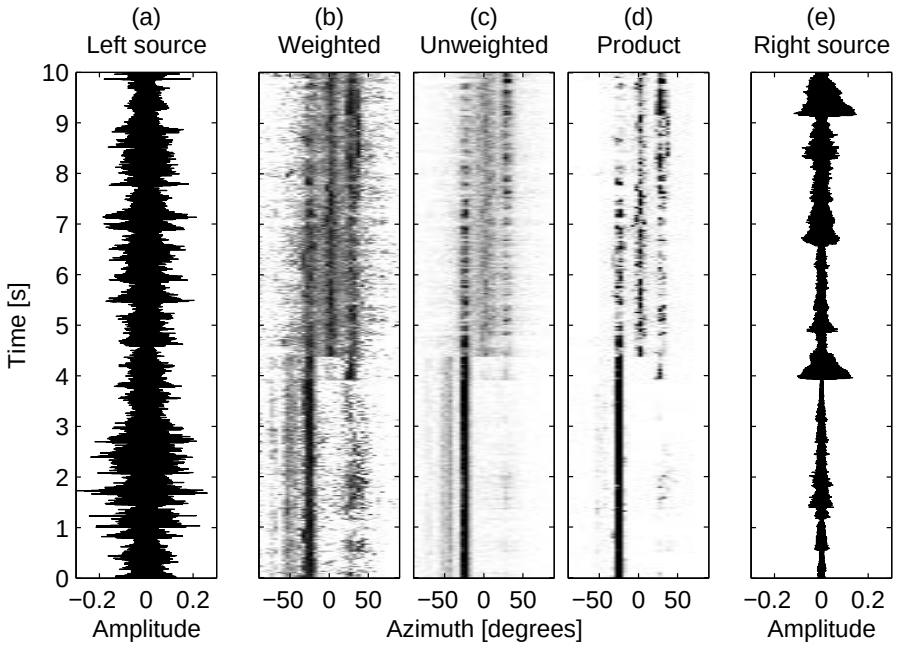


Figure 4.11: *Localization of two sources in binaural signal, showing the source signals and histograms of the azimuth estimates as function of time. (a): Left source signal at about -30 degrees azimuth, consisting of a superposition of two relatively strong sinusoids. (b): Energy weighted azimuth histograms. (c): Unweighted histograms. (d): Product of weighted and unweighted histograms. (e): Right source signal at about 30 degrees azimuth.*

Combined evaluation of signal cues and spatial cues

If it is known that only two sources are active it is possible to estimate the source azimuths even when there is significant overlap. When the source locations are known it is possible to estimate the frequency dependent mixing coefficients in the mixing model given in Equation (2.23) by use of HRTF data or the ILD/ITD models proposed in Chapter 3. For each frequency index and source azimuth the ILD and ITD can be looked up in these models, effectively yielding the complex mixing coefficients $H_{mn}(k, q)$. Ideally, applying the inverse of the estimated mixing matrix should result in true separation of the sources, as discussed in Section 2.4.4. However, due to the relatively large level difference that may occur between the ears, even small errors in the estimated mixing coefficients may result in amplification of the signal that should be canceled. Indeed, informal experiments with both HRTF data and the average parameter model have shown that this approach is not very useful.

The ideas of truly separating overlapping sources based on the mixing model in Equation (2.23) are still applicable. In the next chapter we propose a method

for separation of overlapping harmonics based on similarity of temporal envelopes. This method combines spatial cues and spatial separation with signal cues, i.e. temporal envelopes. It is similar to the statistical methods discussed in Section 2.4.2 in that it only works in static scenes where the mixing filters are time invariant. However, since it is based on similarity of temporal envelopes rather then on statistical independence, it can successfully work with overlapping partials in narrow bands.

Chapter 5

Separation of overlapping harmonics

In a situation where multiple sound sources are concurrently active, the signals of the individual sources often overlap in time and in frequency. This is particularly likely for voiced instruments where the frequencies of some of the partials of one single note coincide with the frequencies of some of the partials of another instrument playing a harmonically related note. A source separation algorithm suitable for musical applications must address the problem of overlapping partials.

We propose a method for separation of overlapping partials in multi-channel mixtures. The method consists of an analysis stage and a separation stage. In the analysis stage, distinct regions in the time-frequency plane are identified, denoted "partial regions", such that each of these regions covers one or more partials. Furthermore, a mapping between these regions and the various sources is established. This is achieved by employing grouping principles on both spatial cues (multi-channel) and signal cues (single channel). The information provided by the analysis is then exploited in the separation stage. Each partial region contains either overlapping partials or single (non-overlapping) partials. The partial regions that contain single partials provide information about the characteristics of the underlying sources. The separation uses this information and applies an iterative search for a frequency domain demixing matrix. When applied on regions that contain overlapping partials, this matrix yields separated partials whose envelopes resemble the envelopes of the given non-overlapping partials.

5.1 Harmonic instruments and musical scales

Many instruments exhibit a strong sinusoidal nature. For a single excitation, or single note, the corresponding signal can be closely modeled as a sum of sinusoids whose amplitudes and phases vary slowly in time. When instruments play together in an ensemble there is a high chance that their energy distributions overlap in time and in frequency. Depending on the types of instruments being played, and on the tuning of the various instruments, some of the partials of one instrument may overlap with the partials of other instruments.

5.1.1 Overlapping energy in music signals

It is important to note that the notion of "overlap" depends on the type of time-frequency analysis being applied. For instance for long stationary sounds, where some partials are closely spaced, it may be possible to sacrifice time resolution in favor of frequency resolution in order to detect the individual partials. Obviously, this is not practical for partials that are only a few Hz apart, since the typical duration of a single note is shorter than the analysis time interval length needed for sufficient frequency resolution. Even worse, when there are frequency fluctuations, such as vibrato, the partial trajectories of different notes and instruments may cross each other. Due to the uncertainty principle, no time-frequency analysis will be fine enough to detect those individual partials accurately.

Harmonic instruments are instruments where the frequencies of the different partials of a single note are in a harmonic or quasi-harmonic relation to each other. In other words, the frequency of each partial is approximately an integer multiple of the fundamental frequency of the corresponding note. In this case the partials are typically called *harmonics*. Other instruments such as drums, bells, but also the piano in the low register, feature partials that are not harmonically related.

Two tones having a frequency relation that is a ratio of small integers are known to produce a pleasing harmony. A study of psychoacoustical explanations for this "pleasingness" can be found in (Plomp and Levelt, 1965). The discovery is traditionally credited to Pythagoras, and such relations are called exact intervals, or consonances. Most of the music scales that have been used throughout the ages are built on these consonances. A scale in which the notes are related through exact intervals is called *just intonation*. These scales cannot be transposed to a different key. Another type of scale is *equal temperament*, where all the intervals are powers of a constant (irrational) frequency ratio. An equal-tempered scale is a compromise that allows transposition to any key, at the cost of introducing deviations from just intonation. This means that the intervals in the equal-tempered scale are only approximations of the exact consonance intervals. This can be seen in Table 5.1 in which the frequency ratios for the tones in two 12 tone scales are shown. The first scale is the harmonic duodene (Helmholtz, 1954), which is a 12 tone scale in just intonation. The other scale is the 12 tone equal-tempered scale that is used in almost all modern western music.

When harmonic instruments play in these scales, it is very likely that some of their partials will overlap. Each harmonic note consists of partials whose frequencies are (close to) integer multiples of the fundamental frequency and the frequency ratio between the fundamental frequencies of the different notes are (close to) the ratio of two integers. In many music genres the most commonly used intervals are those with high consonance. The frequency relations for these intervals are approximately ratios of small integers, such as a fifth (3:2), a third (5:4), a fourth (4:3), and an octave (2:1). As a consequence, a high number of partials overlap.

Table 5.1: Frequency ratios in two 12 tone scales: just intonation and equal temperament, respectively.

Note # in octave	Just intonation ratio of integers	Equal temperament fixed ratio	Deviation (%)
0	$\frac{1}{1} = 1.000$	$2^{\frac{0}{12}} = 1.000$	0.00
1	$\frac{16}{15} = 1.067$	$2^{\frac{1}{12}} = 1.059$	0.68
2	$\frac{9}{8} = 1.125$	$2^{\frac{2}{12}} = 1.122$	0.23
3	$\frac{6}{5} = 1.200$	$2^{\frac{3}{12}} = 1.189$	0.91
4	$\frac{5}{4} = 1.250$	$2^{\frac{4}{12}} = 1.260$	-0.79
5	$\frac{4}{3} = 1.333$	$2^{\frac{5}{12}} = 1.335$	-0.11
6	$\frac{45}{32} = 1.406$	$2^{\frac{6}{12}} = 1.414$	-0.56
7	$\frac{3}{2} = 1.500$	$2^{\frac{7}{12}} = 1.498$	0.11
8	$\frac{8}{5} = 1.600$	$2^{\frac{8}{12}} = 1.587$	0.79
9	$\frac{5}{3} = 1.667$	$2^{\frac{9}{12}} = 1.682$	-0.90
10	$\frac{9}{5} = 1.800$	$2^{\frac{10}{12}} = 1.782$	1.02
11	$\frac{15}{8} = 1.875$	$2^{\frac{11}{12}} = 1.888$	-0.68
12	$\frac{2}{1} = 2.000$	$2^{\frac{12}{12}} = 2.000$	0.00

5.2 Partial separation

5.2.1 Motivation

In the human auditory system, aspects of both groups of methods discussed in Sections 2.2 and 2.4 are exploited in order to analyze an auditory scene. Since a person has two ears he/she is able to focus on sounds coming from particular directions. In addition, as a complement, the person is also able to exploit the structure of the individual sources in time and in frequency. It therefore seems plausible to combine aspects of these methods, i.e. spatial analysis and time-frequency analysis.

Partial fusion denotes the phenomena occurring when several partials are perceived as one single auditory event, or as a single note. Some of the cues that are important in order to achieve partial fusion are harmonicity and synchronicity of onset, offset, frequency and amplitude modulation (Bregman, 1999, ch.3). In other words, the different partials of one single note typically have these cues in common. Otherwise, they would not be perceived as one sound. If a few partials for a given note are known, they give a rough idea of what the other partials should resemble. Thus, if some non-overlapping partials can be detected, these can be used as models for the unknown partials that need be separated out of a superposition of overlapping partials. This is the motivation

behind the method proposed in this chapter. In particular, it can be observed that, for a single note, the temporal envelopes of the partials have quite similar onset/offset times and amplitude modulation. In synthesis techniques this is well known in the form of a model for attack, decay, sustain and release (ADSR model). Column (a) in Fig. 5.1 shows the envelopes of the first six partials of a single note being played by an alto trombone. The partial envelopes change somewhat with frequency. For instance, the higher partials have delayed onset times as well as faster decay. However, there is a noticeable similarity between all the partials, and in general the similarity is higher between partials whose frequencies are close to each other.

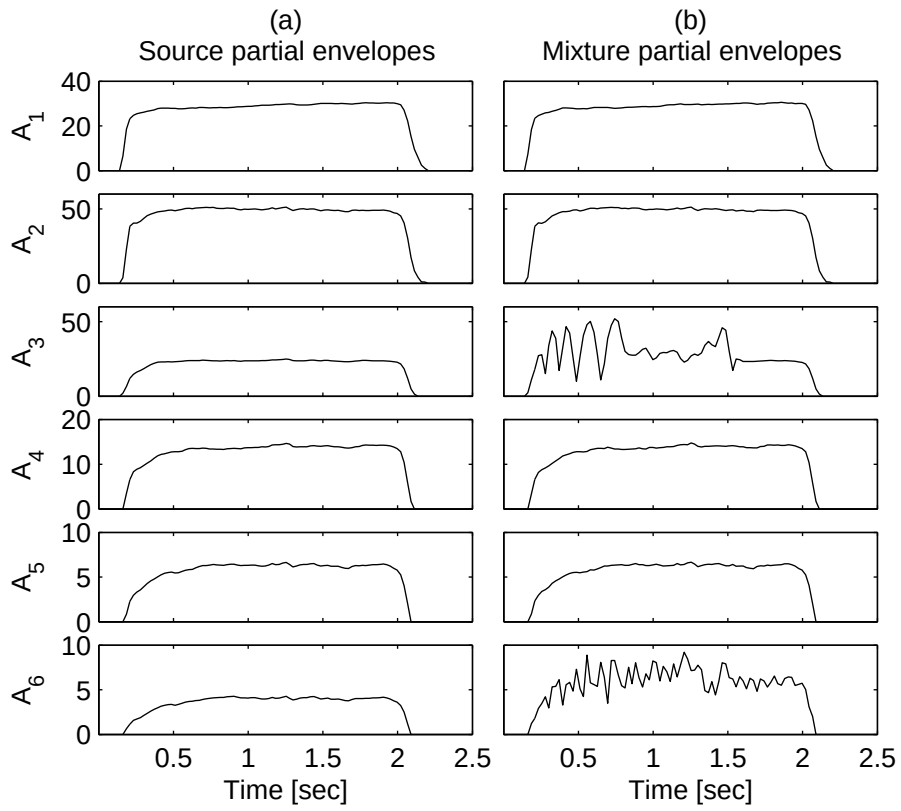


Figure 5.1: *Amplitude envelopes of the first 6 partials of an A note played by an alto trombone. (a): Original source partials. (b): Partials in a mixture, where the third and sixth partials overlap with partials of other instruments.*

The proposed method consists first of a time-frequency analysis where distinct regions, denoted partial regions, are detected in the time-frequency plane such that each region covers the main energy of one or more (overlapping) partials. Then grouping principles based on harmonicity, spatial cues, and temporal envelope shapes of the partials are employed in order to find a mapping between the sources and these regions. This provides information about which sources contribute with significant energy in each of the different regions,

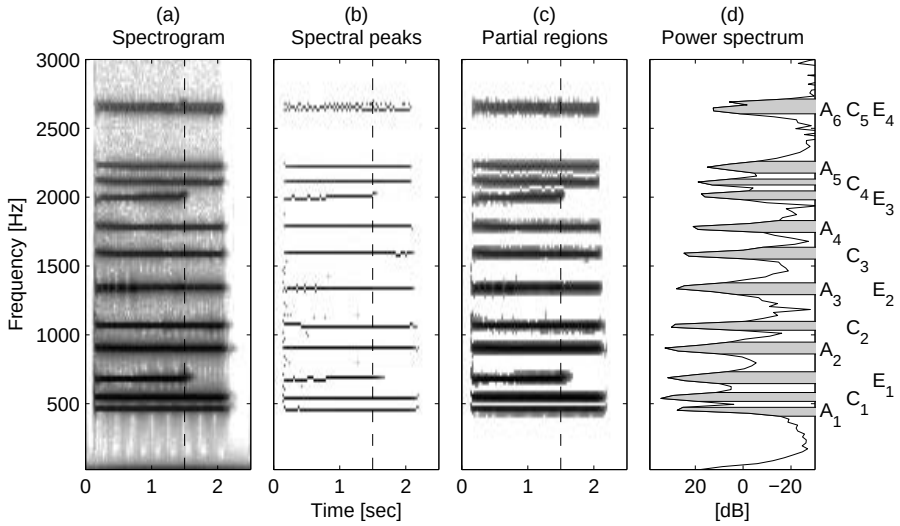


Figure 5.2: *Three notes (A minor chord) played by alto trombones, from left to right: (a): Spectrogram, (b): Spectral peaks, (c): Partial regions around the spectral peaks, (d): Power spectrum and detected partial regions for one particular time index (indicated by dashed lines in the other panels).*

as well as which regions contain only one partial (non-overlapping). The latter are used as models for the unknown overlapping partials. Finally, for each region containing overlapping partials the demixing matrix is estimated. This is accomplished by an iterative search for the matrix that gives separated partials whose temporal envelope shapes most closely resemble those of the model partials.

5.2.2 Definition of partial regions

The STFT is a well suited tool for the detection of partials and for the estimation of the parameters in sinusoidal models. As in most sinusoidal modeling techniques, the peaks in the STFT are used in order to describe the partial trajectories. However, unlike sinusoidal modeling, we are not trying to characterize each of these partials by their amplitude and phase trajectories. Trying to do this to a high degree of accuracy would be useless since each spectral peak may actually be the net result of several overlapping partials in its close neighborhood. In addition, the sidebands of the partials also contain information that may be important for the naturalness of the sounds. Therefore, the goal is not to estimate the amplitude and phase trajectories to a high degree of accuracy. Rather, the aim is to define regions around these peaks (partial regions), such that any underlying partial is covered by these regions. At each time k, for each detected spectral peak, a frequency range is formed by including neighboring spectral coefficients at lower and higher frequencies, as long as the amplitude is strictly decreasing (up to some maximum region width).

Figure 5.2 shows an example of a signal consisting of three notes, namely the notes A, C, and E, which together constitute an A minor chord. Panel

(a) shows the spectrogram of the signal (one of the sensor channels). The sinusoidal nature of the partials, as well as the harmonic structure, can be clearly seen. Panel (b) shows the spectral peaks that were detected by applying a standard peak picking method. Panel (c) shows the partial regions that have been identified around these spectral peaks. Panel (d) shows the power spectrum of the signal for a particular time index. Its temporal position is indicated by vertical dashed lines in the three other panels. Around the most prominent peaks in the power spectrum partial regions have been identified. These are shown as horizontal, filled bars. The frequency widths of the partial regions at the given time instant equal the extent of these bars on the frequency axis (ordinate). Finally, the partials are annotated with the names of the underlying notes and their partial indexes, i.e. A_2 denotes the second partial of the A note. The relative fundamental frequencies of the three notes are approximately $\frac{1}{1}$, $\frac{6}{5}$, and $\frac{3}{2}$, respectively. The partials A_3 and E_2 overlap, introducing amplitude fluctuations, denoted beatings. This can be seen as intensity variations at the beginning of the corresponding partial region in Panel (a) in Fig. 5.2. Similarly, the partials A_6, C_5 and E_4 overlap. The peak picking algorithm is naturally not able to detect each of them individually. This can be seen in Panel (b), where the multiple peaks do not form a single clear partial trajectory. The effect of the overlapping partials can also be observed in column (b) in Fig. 5.1 where the envelopes of the partial regions corresponding to the first six partials of the A note are shown. The overlapping partials with strong beatings and erroneous envelopes are clearly seen in the third and sixth panel from the top, respectively. Column (a) shows the envelopes of the uncorrupted partials of the original source signal in the same partial regions.

Due to the fact that traditional sinusoidal models aim at describing partials by their amplitude and phase trajectories, the individual spectral peaks need to be connected in time as trajectories. This is typically achieved by means of frame-to-frame peak matching techniques. However, in general, there can be ambiguous situations in which the peak matching algorithm needs to make a hard decision. When there are several possible candidate peaks for the continuation of a partial, only one of these peaks can be chosen. The other peaks must be either discarded, or modeled as separate partials. The example with the three overlapping partials in Fig. 5.2 illustrates this problem. For some time index k, two or more peaks are detected (Panel (b)) in the narrow band region where these partials overlap (about 2600 Hz). Neither is it possible to connect these peaks into one single partial trajectory (without ambiguities), nor do these peaks provide estimates of the phase and amplitudes of the underlying partials.

When working with regions covering the partial trajectories, rather than trying to accurately characterize them, multiple candidate peaks no longer constitute a problem. Ambiguities are gracefully handled by employing a frame-to-frame partial region matching technique, as opposed to the conventional peak matching techniques. The partial region matching connects together individual partial regions (one particular time index k) into resulting partial regions evolving over time. Whenever there are two (or more) possible candidate regions for the continuation of a given partial region, any number of these candidate regions, as well as any non-covered spectral coefficients between them, can be included in the given partial region. In Panel (d) in Fig. 5.2 it can be seen

how two candidate peaks (about 2600 Hz) have been included in the same partial region. In Panel (c) it can be seen how the corresponding partial region effectively covers the multiple candidate peaks.

Partial regions provide a means for dividing the entire time-frequency plane into non-overlapping regions Ω_i, indexed by i, where each region captures the main energy from one or more partials. Each partial region can be described by the corresponding indicator function $I_i(k,q)$, which equals 1 for $(k,q) \in \Omega_i$ and 0 elsewhere. For a given signal, each region may contain a single partial or a superposition of overlapping partials. For notational convenience, we use the term partial to denote the energy of any signal that is contained in a single partial region, even when the region actually contains a superposition of two or more overlapping partials. The term *sensor partial* is used to denote the part of a sensor signal that is contained in a given partial region. A sensor partial may consist of a single partial, or a combination of several overlapping partials. Similarly, for the original (unknown) source signals, the term *source partial* is used.

Even if the selected partial regions do not cover the entire time-frequency plane, they capture most of the energy in the signal. It can therefore be assumed that the entire plane is covered in such a way that any residue, or spectral coefficients that are not part of any partial region, is neglected. The residue is better handled by other separation techniques, such as those based on time-frequency weighting, as discussed in Section 2.4. Under the assumption that the entire time-frequency plane is covered, each of the sensor signals may be written as a sum of sensor partials:

$$X_m(k,q) = \sum_i P_{im}(k,q), \tag{5.1}$$

where each sensor partial is simply the product of the corresponding sensor signal and indicator function:

$$P_{im}(k,q) = X_m(k,q) I_i(k,q). \tag{5.2}$$

5.2.3 Partial temporal envelope similarity

In order to determine how similar two partial envelopes are, a measure of partial similarity is needed. For the signal contained in a partial region A, the temporal envelope $E_A(k)$ is defined as follows:

$$E_A(k) = \sqrt{\sum_q |A(k,q)|^2}. \tag{5.3}$$

The temporal envelope of the signal in a partial region (5.3) contains information about onset, offset, and amplitude modulation, which are the fusion cues mentioned in Section 5.2.1. It does not contain frequency modulation information. It is important to note that since the definition of the temporal envelope involves summing over a range of spectral coefficients, it is only truly meaningful when the time-frequency representation is a tight frame. For the STFT, this means that for any signal $s(l)$ with STFT $S(k,q)$, the following must hold:

$$\sum_k \sum_q S(k,q)^2 = C \sum_l s(l)^2, \tag{5.4}$$

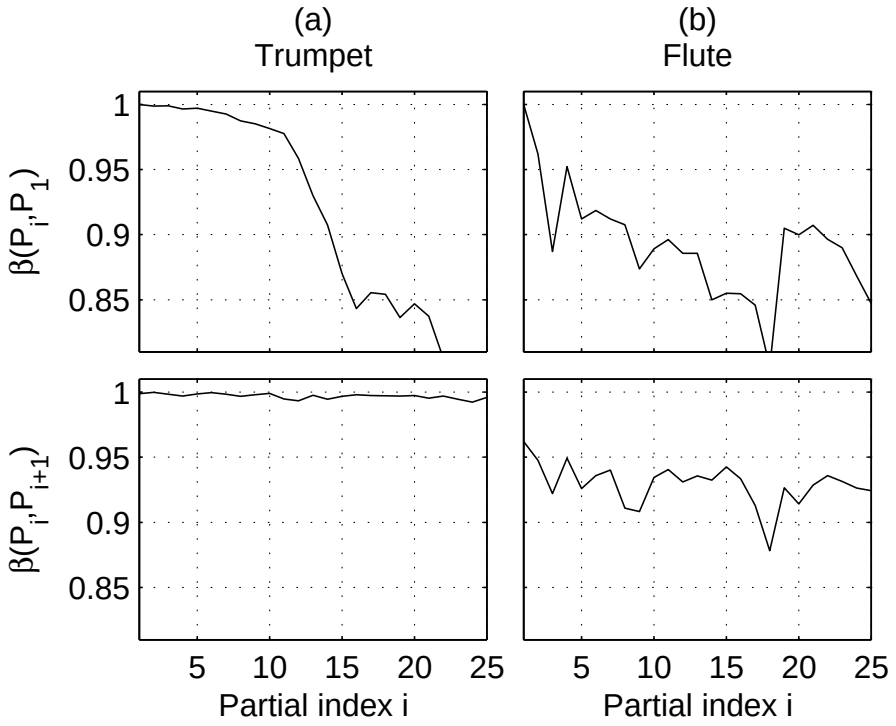


Figure 5.3: *Partial envelope similarity for two notes, (a): Trumpet, (b): Flute. Top row: similarity relative to the first partial, Bottom row: similarity relative to the neighbor partial.*

for some constant C. This can also be written as a constraint on the choice of analysis window $w(l)$ in the STFT:

$$\sum_k w^2(l - kL) = C, \tag{5.5}$$

where L is the hop size. For the examples in this chapter, we have used the half-wave sine window, with 50% overlap. A window length of about 50 ms was chosen.

The temporal envelope shape similarity β between the signals in two partial regions, A and B, can be defined as the inner product between the normalized temporal envelopes:

$$\beta(A, B) = \frac{\sum_k E_A(k) E_B(k)}{\sqrt{\sum_k |E_A(k)|^2} \sqrt{\sum_k |E_B(k)|^2}}. \tag{5.6}$$

This is a measure in a range between 0 and 1, where 1 means that the two normalized envelopes are identical. Figure 5.3 shows this similarity measure for two different notes. In column (a) the similarity of the first 25 partials $(P_1, P_2, \ldots, P_{25})$ of a trumpet note are shown. The top graph shows the similarity of each partial relative to the first partial, $\beta(P_i, P_1)$. The bottom graph

shows the similarity of each partial relative to its neighbor partial, $\beta(P_i, P_{i+1})$. Column (b) shows similar graphs for a flute note with vibrato. The amplitude modulations make the partials less similar, but the general trend is the same. Even though the similarity relative to the first partial decreases as the partial index increases, the local similarity remains high. Of course this depends on the type of instrument and on the note being played, but for many instruments, and harmonic instruments in particular, there is a significant correlation between the envelopes of the different partials of a single note.

5.2.4 Partial grouping

Each of the defined sensor partials contains one or more partials. In order to process them, the number of sources N and a mapping between the sources and the partial regions must be known. In other words, for each source it must be known in which partial regions its partials lie and for each partial region it must be known which are the sources that contribute to a partial. Existing techniques can provide this information. In particular, we obtain the number of sources N by peak picking in histograms of spatial cues, as discussed in Chapter 3 and Section 2.4.3. The mapping between sources and partial regions is obtained by harmonicity considerations (Virtanen and Klapuri, 2000), as discussed in Section 2.2.3.

Once this information has been established it is straightforward to find all the partial regions that contain only a single partial, as well as which source each partial is part of. For each source that has at least one non-overlapping partial, this provides a rough estimate of the temporal envelope shapes of all the other partials of that source. These rough estimates are called *model partials*, $Q_n(k, q)$, as they serve as models for the partials that need to be separated out of a partial region where they overlap.

5.2.5 Partial demixing

For each of the sensor partials containing overlapping partials from several sources, the mixing matrix $\mathbf{H}$ (or its inverse) in (2.23) needs to be estimated in order to recover the original source partials. This is achieved by an iterative search in the space of mixing matrices. For each candidate matrix, its inverse is applied to the sensor partials and the envelopes of the resulting "separated" partials are computed. The matrix that gives separated partials with envelopes whose shapes most closely resemble those of the model partials is chosen as the estimate of the mixing matrix for that corresponding partial region.

For most harmonic instruments the partials are narrow band. This means that the partial regions are also narrow band. Given the mixing model (2.23) it is possible to treat the different partial regions independently. The individual filters of the mixing matrix, $H_{mn}(q)$, depend on frequency. In other words, they vary over the frequency range of a given partial region. However, when the partial regions are narrow band, the filters change little over the actual frequency range. Therefore they can be closely approximated by complex constants (filters with constant amplitude and constant phase). The separation problem is then equivalent to the problem of estimating the complex elements of a constant mixing matrix $\mathbf{H}_i$ for each partial region I_i. For larger frequency ranges, it is possible to use filters with linear phase rather than constant phase. This

does not increase the number of unknowns or the complexity of the problem. For notational simplicity only the former case is discussed in this chapter.

Combining (2.23) and (5.2) gives

$$\begin{bmatrix} P_{i1}(k,q) \\ \vdots \\ P_{iM}(k,q) \end{bmatrix} = \begin{bmatrix} H_{11} & \cdots & H_{1N} \\ \vdots & \ddots & \vdots \\ H_{M1} & \cdots & H_{MN} \end{bmatrix} \begin{bmatrix} S_{i1}(k,q) \\ \vdots \\ S_{iN}(k,q) \end{bmatrix}, \tag{5.7}$$

where $S_{in}(k,q) = I_i(k,q)S_n(k,q)$ are the source partials and the mixing filters H_{mn} are complex constants. These filters are different for each partial region I_i, but do not depend on frequency within the partial regions. When the rank of the mixing matrix for a given partial region is equal to or greater than the number of partials in that partial region, separation of these partials is possible. This, in general, means that there must be as many sensors as there are overlapping partials in that region. However, the sources to which the overlapping partials belong to must be in different locations (giving M independent rows in the mixing matrix). For any estimate $\hat{\mathbf{H}}_i$ of the mixing matrix $\mathbf{H}_i$, separation is achieved by applying its (pseudo-)inverse $\hat{\mathbf{H}}_i^+$ on the known sensor partials. This is similar to frequency based blind source separation techniques. The difference lies in the way the mixing matrix is estimated.

Left multiplying (5.7) with the estimated pseudo-inverse gives:

$$\begin{bmatrix} R_{i1}(k,q) \\ \vdots \\ R_{iN}(k,q) \end{bmatrix} = \hat{\mathbf{H}}_i^+ \begin{bmatrix} P_{i1}(k,q) \\ \vdots \\ P_{iM}(k,q) \end{bmatrix} = \hat{\mathbf{H}}_i^+ \mathbf{H}_i \begin{bmatrix} S_{i1}(k,q) \\ \vdots \\ S_{iN}(k,q) \end{bmatrix}, \tag{5.8}$$

where S_{in} are the source partials, and R_{in} are the *separated partials*. Each R_{in} represents the contribution of a single source S_n in the partial region I_i. It is obvious that under the given assumptions and with a correct estimate $\hat{\mathbf{H}}_i$ of the mixing matrix, the separated partials in (5.8) are identical to the source partials, i.e. perfect separation has been achieved.

For each partial region I_i the mixing matrix $\mathbf{H}_i$ is estimated by a search for the estimate $\hat{\mathbf{H}}_i$ that gives a best match between the separated partials R_{in} and the model partials Q_n. We achieve this by applying a standard multi-dimensional optimization technique that attempts at maximizing the following similarity vector in L_1 norm:

$$\beta_i = (\beta(R_{i1}, Q_1), \cdots, \beta(R_{iN}, Q_N)). \tag{5.9}$$

5.2.6 Practical considerations

Reducing complexity by disregarding reverberation

In general, the mixing matrix for a system with N sources and M sensors has $M \times N$ (complex) unknown elements. For a 3×3 system this gives an 18-dimensional space of candidate mixing matrices (9 complex elements).

A technique that can reduce the order of unknowns is to force all the elements of one row in the mixing matrix $\mathbf{H}$ to 1 (Jutten and Herault, 1991; Jourjine *et al.*, 2000). This can be done without loss of generality, as long as

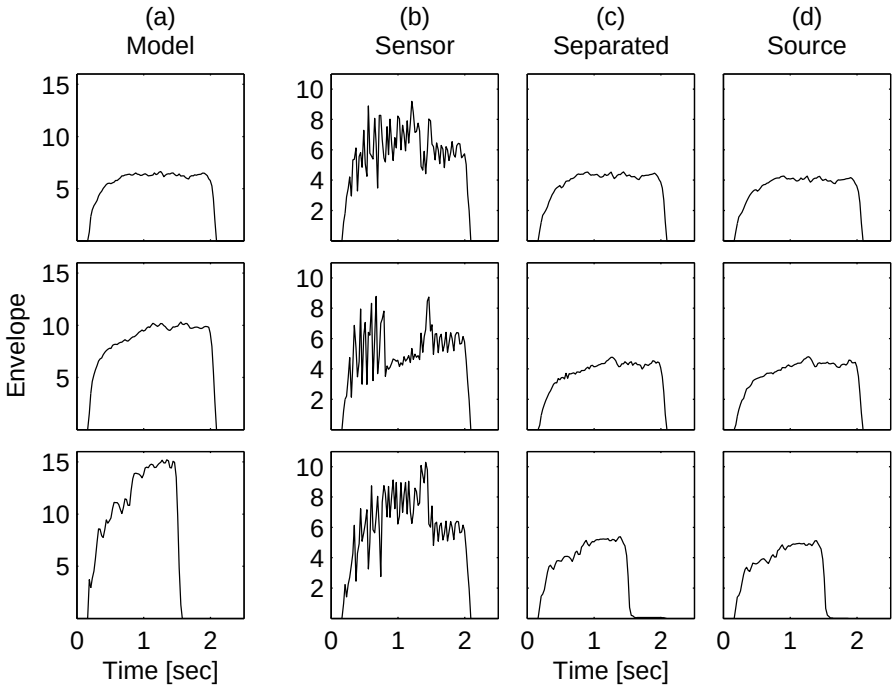


Figure 5.4: *Partial envelopes for mixture of alto trombones, from left to right: (a): Model partials, A_5, C_4, and E_3 respectively (closest neighbors). (b): Sensor partials each containing overlapping A_6, C_5, and E_4. (c): Separated partials, A_6, C_5, and E_4 respectively. (d): Original source partials, shown for comparison.*

the real mixing filters H_{1n} contains no zeros. Then, (2.23) can be written as follows:

$$\begin{bmatrix} X_1(k,q) \\ \vdots \\ X_M(k,q) \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ \frac{H_{21}}{H_{11}} & \cdots & \frac{H_{2N}}{H_{1N}} \\ \vdots & \ddots & \vdots \\ \frac{H_{M1}}{H_{11}} & \cdots & \frac{H_{MN}}{H_{1N}} \end{bmatrix} \begin{bmatrix} H_{11}S_1(k,q) \\ \vdots \\ H_{1N}S_N(k,q) \end{bmatrix}. \tag{5.10}$$

In the mixing matrix, the number of (complex) unknowns has been reduced by N. The formulation of the problem is the same as before (see e.g. (2.23)). The difference is that the real source signals $S_n(k,q)$, as emitted by the sources, have now been replaced by the same source signals as they would be received by the first sensor, $H_{1n}S_n(k,q)$. This means that no attempt is made to remove echoes or reverberation. However, in the context of source separation this can in general be accepted. The problems of echo cancellation, echo suppression, and dereverberation will not be discussed in this thesis. It can be noted that other source separation techniques use STFT based techniques (Smaragdis, 1998; Ikeda and Murata, 1999; Jourjine *et al.*, 2000). Our method can therefore easily be used as an extension to these existing methods.

Convergence

Even after applying the dimension reduction technique, the dimension of the space of mixing matrices remains high. When there are 3 sources and 3 sensors, the optimization algorithm trying to maximize (5.9) performs a search in a 12-dimensional space. The norm of the similarity vector (5.9) is a function in this space. Needless to say, it can take an arbitrarily complex form, and is not, in general, a concave function. The optimization algorithm may get trapped in some local maxima and not converge. In general, this depends on the starting estimate of the mixing matrix that is used in the iterative optimization algorithm. We have conducted several experiments on different sources and setups. In these experiments, although not being extensive, the choice of the L_1 norm in the optimization of Equation (5.9) gave the best convergence behavior. We also experienced that when several sources have very similar temporal envelope shapes, this yields more local maxima. In this case the starting estimate is more critical for the convergence.

If the sensors and sources are in free field, each of the mixing filters will consist of a simple scaling factor and a pure delay. In this case it is possible to estimate these parameters from the model partials, and directly compute the complex elements of the matrix $\hat{\mathbf{H}}_i$ for any other partial regions. Its pseudo-inverse $\hat{\mathbf{H}}_i^+$ can be computed directly, and the partials can be separated as in (5.8). Even though this is not feasible in real situations, the estimated free-field matrix $\hat{\mathbf{H}}_i$ can be used as the starting estimate in the iterative optimization algorithm.

In specific physical setups, it may be possible to further reduce the dimension of the problem. For instance, if the sensors are very closely spaced, the scale factor in the mixing filters can be approximated by the same constant for all the filters, leaving only the phases as unknowns. This effectively reduces the dimension by a factor of 2.

Special case

When there are only 2 sensors (and maximum 2 simultaneously overlapping partials), each mixing matrix contains only two (complex) unknowns, since

$$\mathbf{H}_i = \begin{bmatrix} 1 & 1 \\ H_{21} & H_{22} \end{bmatrix}. \tag{5.11}$$

In this case, the inverse matrix is very simple:

$$\mathbf{H}_i^+ = \frac{1}{H_{22} - H_{21}} \begin{bmatrix} H_{22} & -1 \\ -H_{21} & 1 \end{bmatrix}. \tag{5.12}$$

Each of the rows of the demixing matrix has only one complex unknown, up to a scaling factor of $H_{22} - H_{21}$. Since our method is based on normalized envelopes, this scaling factor can be disregarded. This gives two independent equations, each with one unknown parameter. Thus, the original problem of dimension 4 has been split into 2 individual problems of dimension 2. This provides a fast computational method not involving any matrix inversions. The separation formula (5.8) consists of only 2 complex multiply-add operations for each time index k in the sensor partial envelopes. This allows for a more extensive, iteratively refined, search for the global maximum, in each of the two separate problems.

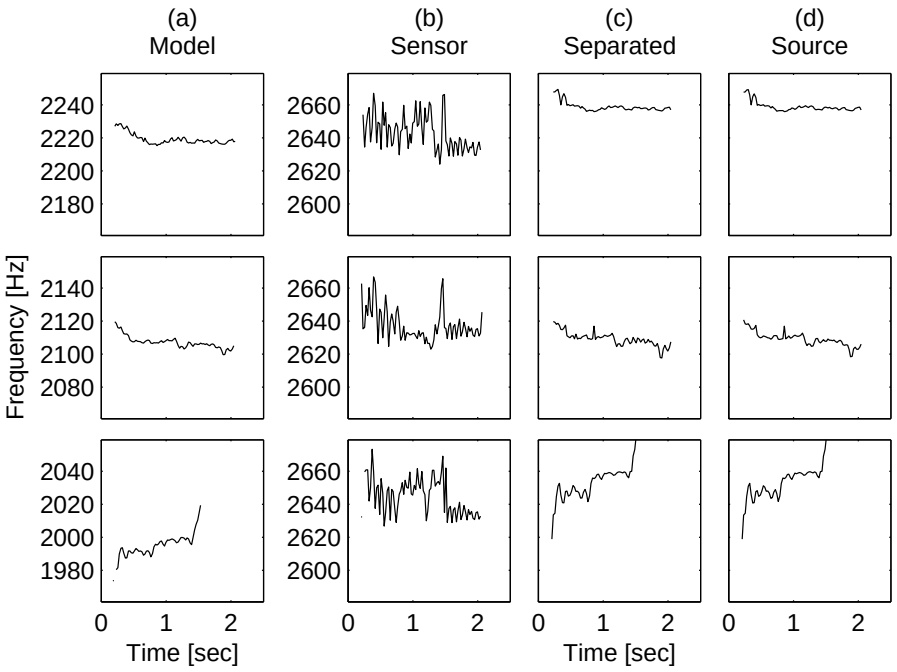


Figure 5.5: *Partial frequency trajectories for mixture of alto trombones, from left to right: (a): Model partials, A_5, C_4, and E_3 respectively (closest neighbors). (b): Sensor partials each containing overlapping A_6, C_5, and E_4. (c): Separated partials, A_6, C_5, and E_4 respectively. (d): Original source partials, shown for comparison.*

5.3 Results

In order to demonstrate the performance of the presented method, a number of numerical simulations were carried out, which will be described in the following paragraphs.

5.3.1 Three overlapping partials

The first example is a simulated situation of three sources and three sensors in free-field. The three notes A, C and E, all played by an alto trombone, were chosen as source signals. These are the same notes as shown in Fig. 5.2. All the source signals were present at all the sensors. However, for each sensor, the source signals were scaled and delayed in order to simulate the wave propagation path. This setup provides us with three sensor signals. These sensor signals are quite similar since they all record the same scene. When the time-frequency analysis and grouping principles of section 5.2 are applied to either of these, the information shown in Fig. 5.2 is obtained. This provides us with a set of partial regions, as well as a mapping between these regions and the different sources (the partial labels that have been annotated in the figure).

Figure 5.4 shows the separation of the three overlapping partials in this

signal. In column (a) the envelopes of the model partials are shown. Each of these model partials, $Q_n(k,q)$, corresponds to a different source, $S_n(k,q)$. In this example the closest non-overlapping partials were chosen as model partials, namely A_5, C_4, and E_3, respectively. Two of the model partials have quite similar envelopes, whereas the duration of the last model partial is shorter. These panels clearly show the smooth temporal envelopes of the original alto trombone notes, without beatings and strong amplitude modulations. In column (b), the envelopes of the overlapping partials are shown for each of the three sensor signals. None of these are very regular or smooth since the three original partials interact and create heavy beatings, or amplitude modulations. Only toward the end, when one of the partials is silent (after about 1.5 sec), the envelopes are somewhat more regular. In column (c), the envelopes of the separated partials are shown. These are A_6, C_5, and E_4, respectively. The three overlapping partials have been well separated, and the resemblance to the model partials is striking. In particular, we note that the beatings have disappeared and that the shorter partial (E_4) has been correctly recovered: after about 1.5 seconds the energy of the other two partials (longer duration) has vanished. Finally, in column (d), the original (unknown) source partials are shown for comparison. The separated partials have accurately retained the amplitude of the original signals. Both the scale and the shape of the separated partials are more similar to the original source partials than they are to the model partials that were employed in the demixing algorithm.

Figure 5.5 shows the frequency trajectories for the same partials. The figure layout is the same as in Fig. 5.4. In column (a) the frequency trajectories of the model partials are shown. These are quite smooth, and relatively constant since the alto trombone exhibits no pronounced frequency modulation. If one disregards the initial segment, where the attack transients affect the frequency estimates and the final segment, where the signal energy is very low, the frequency estimates are indeed almost constant. The ordinates also show how these model partials were chosen in different frequency regions. In column (b), the frequency trajectories for the three sensor signals are shown. For each time frame the strongest spectral peak (interpolated) in the partial region was selected in order to form these trajectories. As previously mentioned, they are erroneous due to the fact that only one candidate peak can be chosen in each time frame and anyway all the peaks are the result of a superposition of overlapping partials. Consequently, the estimated frequency trajectories are noisy, as seen in the figure. Column (c) shows the frequencies of the separated partials. Qualitatively, these show the same behavior as the model partials. Both the constant frequency during the steady-state and the trend at onsets and offsets have been recovered. Finally, in column (d), the frequency trajectories of the original source partials are shown. The trajectories of the separated partials accurately recover those of the original source partials.

Effectively, three different partials whose frequency trajectories are less than 40 Hz apart and occasionally cross each other have been accurately separated from a 3 channel mixture.

We repeated the same separation example with another choice of model partials. When choosing model partials whose frequency bands lie farther away from the partial region containing overlapping partials, the envelopes of the model partials and the original source partials are in general less similar (see

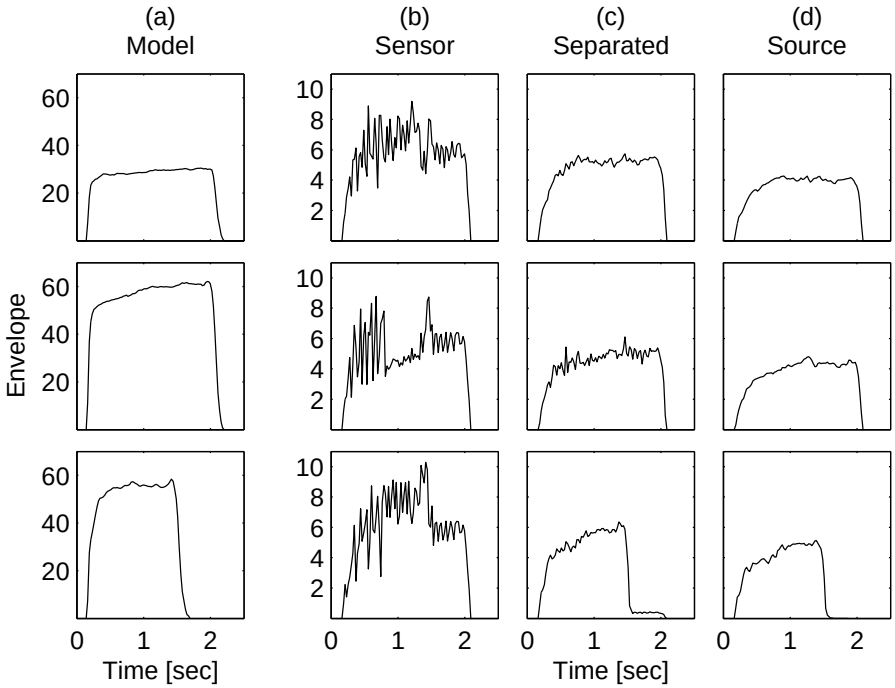


Figure 5.6: *Partial envelopes for mixture of alto trombones, from left to right: (a): Model partials, A_1, C_1, and E_1 respectively (fundamental frequencies). (b): Sensor partials each containing overlapping A_6, C_5, and E_4. (c): Separated partials, A_6, C_5, and E_4 respectively. (d): Original source partials, shown for comparison.*

Fig. 5.3). In the new example the partial regions corresponding to the fundamental frequencies of the three sources, i.e. A_1, C_1, and E_1, were chosen as model partials. The separation result can be seen in Fig. 5.6. As in the first example, most of the beating has been removed. However, for the sources under analysis, the duration (offset) of the partials decreases with frequency, as can be seen in Fig. 5.1. This means that the chosen model partials slightly overestimate the duration of the overlapping partials. In this case, the estimate of the mixing matrix becomes slightly erroneous, resulting in some leakage between the separated partials. For example, some of the energy of the long duration A_6 and C_5 is still present in the separated E_4. This can be seen as a "tail" at the end of the envelope for the separated E_4. A possible explanation is that the tail increases the duration of this separated partial and yields higher similarity in (5.6) to the model partial whose duration was longer.

5.3.2 Frequency and amplitude modulation

A similar experiment, where the notes were played by a violin, is shown in Fig. 5.7-5.9. The notes A and E were played on open strings, without any pronounced frequency modulation. The C tone was played with vibrato, and the frequency modulations on its partials can be seen in Fig. 5.7. The vibrato

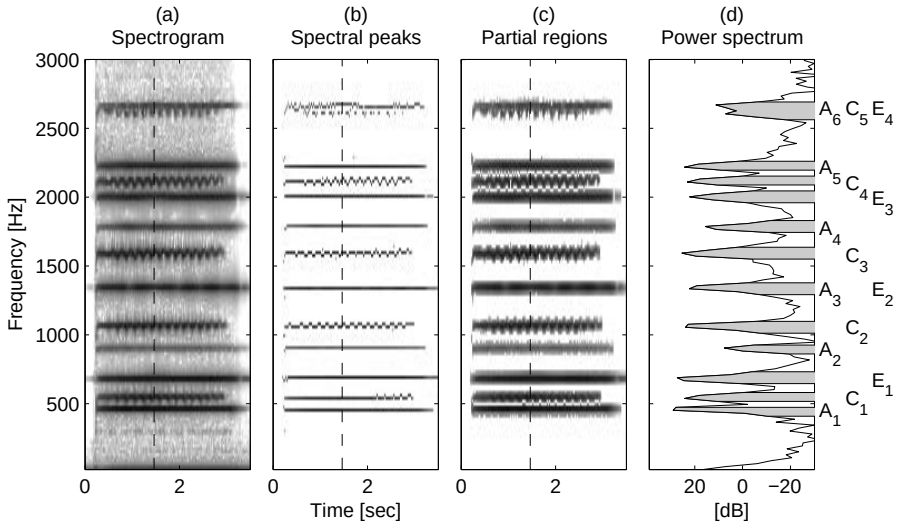


Figure 5.7: *Three notes (A minor chord) played by violins, from left to right: (a): Spectrogram, (b): Spectral peaks, (c): Partial regions around the spectral peaks, (d): Power spectrum and detected partial regions for one particular time index (indicated by dashed lines in the other panels).*

also introduces amplitude modulations, as can be seen in the second row in Fig. 5.8. The amplitude modulation, although slightly different from that of the original source, is retained in the separated C_5. In the other two separated partials, which were without vibrato, the amplitude modulations have been correctly removed. Figure 5.9 shows the frequency trajectories for the partials. In the separated partials the frequency modulation of the vibrato has been retained in the C tone (with some error), whereas is has been removed in the other two partials.

When a note has strong amplitude modulations, the performance of the separation technique may degrade. This depends on the physics of the underlying instrument. For many instruments, frequency modulations are produced by varying the excitation, e.g. shaking the finger on the violin string board. In this case, the frequency modulations of the different partials are synchronized (due to the change in effective length of the string imposed by finger movements). However, the amplitude modulations of the various partials are strongly related to the body modes of the instrument (Mellody and Wakefield, 2000), that are, in general, not synchronized. This means that, even though different partial envelopes may look similar, their amplitude modulations may be out of phase, or even have different modulation frequencies. In such cases, the similarity measure in (5.6) is less meaningful. In the example above, the two notes without vibrato have led to correct estimates of the corresponding rows in the demixing matrix. For the note with vibrato, the estimate of the corresponding row in the demixing matrix is slightly erroneous because of the asynchrony in envelope modulations between the model and source partials. The difference in temporal envelope shape can be seen in Fig. 5.8. The er-

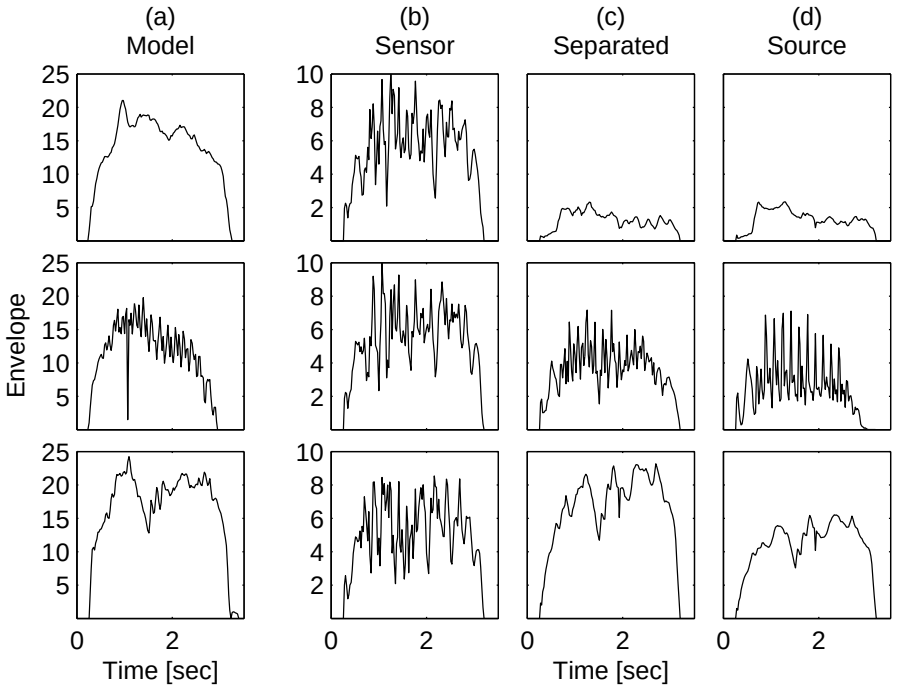


Figure 5.8: *Partial envelopes for mixture of violins, from left to right: (a): Model partials, A_5, C_4, and E_3 respectively (closest neighbors). (b): Sensor partials each containing overlapping A_6, C_5, and E_4. (c): Separated partials, A_6, C_5, and E_4 respectively. (d): Original source partials, shown for comparison.*

roneous row in the demixing matrix leads to scaling errors in the two other separated partials (correct shapes, but incorrect scale/phase) and errors in both the amplitude and frequency modulation of the separated vibrato partial.

It is possible to smooth the partial envelopes before computing the similarity in (5.6) in order to remove any strong amplitude modulations. In this case, however, most of the envelope information is lost and only the overall duration of the partials is significant in the separation. A better solution is to use similarity of frequency modulations as the similarity measure in the optimization method, as opposed to similarity of envelopes. For instruments where the frequency modulations of the different partials are in synchrony, like the violin, this can improve the separation quality, as we achieved in some manual experiments. However, it is not clear how different similarity measures, namely similarity of envelopes and similarity of frequency modulations, should be combined in the similarity vector (5.9). In addition, there may be convergence issues, since the space in which this vector is to be maximized is rather complex.

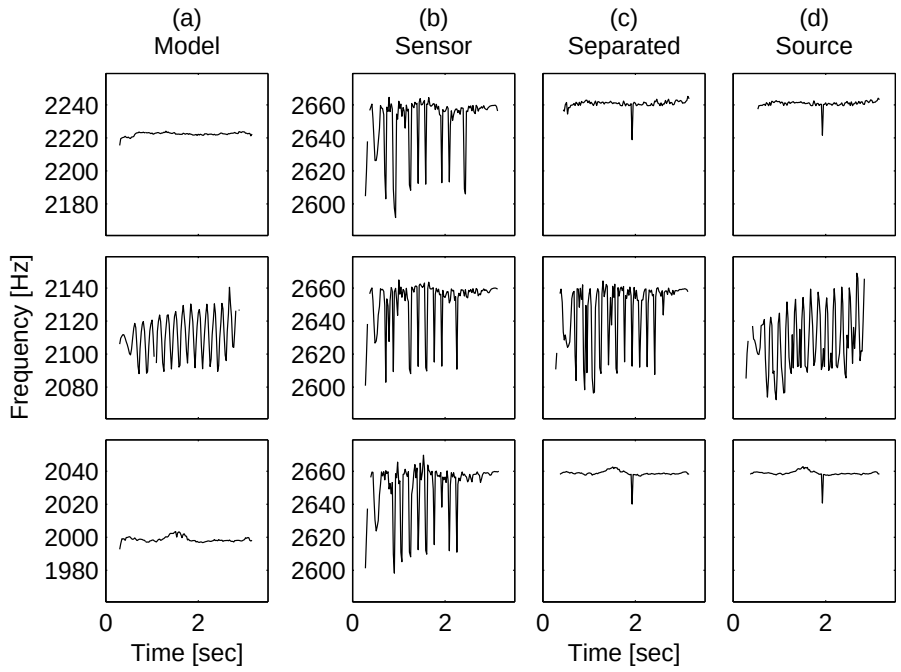


Figure 5.9: *Partial frequency trajectories for mixture of violins, from left to right: (a): Model partials, A_5, C_4, and E_3 respectively (closest neighbors). (b): sensor partials each containing overlapping A_6, C_5, and E_4. (c): Separated partials, A_6, C_5, and E_4 respectively. (d): Original source partials, shown for comparison.*

5.3.3 Two overlapping partials

For a more realistic example, an artificial two-channel mixtures was generated by using head related impulse responses (HRIR) from the CIPIC database (Algazi *et al.*, 2001) as the mixing filters. These impulse responses are about 4.5 ms long filters (200 samples at 44.1kHz sampling rate), measured for a range of azimuth and elevations around various heads. Figures 5.10 and 5.11 show the envelope and frequency trajectories in the separation of two overlapping partials in a two-channel (binaural) mixture of two notes. One of the notes is played on a violin with vibrato, and the other note on a trumpet without vibrato. The figure shows the separation of the first overlap, i.e. the partial region (with overlapping partials) at lowest frequency. The closest neighbor partial regions were chosen as models. The figure layouts are the same as in previous examples, but with two panel rows instead of three. The amplitude and frequency modulations of the violin are seen in the top panel rows in Fig. 5.10 and Fig. 5.11, respectively. In this case, for the violin partials containing vibrato, the amplitude modulations of the model partial and the original source partial are relatively synchronous and the envelope similarity measure gives nice separation. The separated partials have recovered the shapes of the original source partials to a high degree of accuracy, in both amplitude and in frequency.

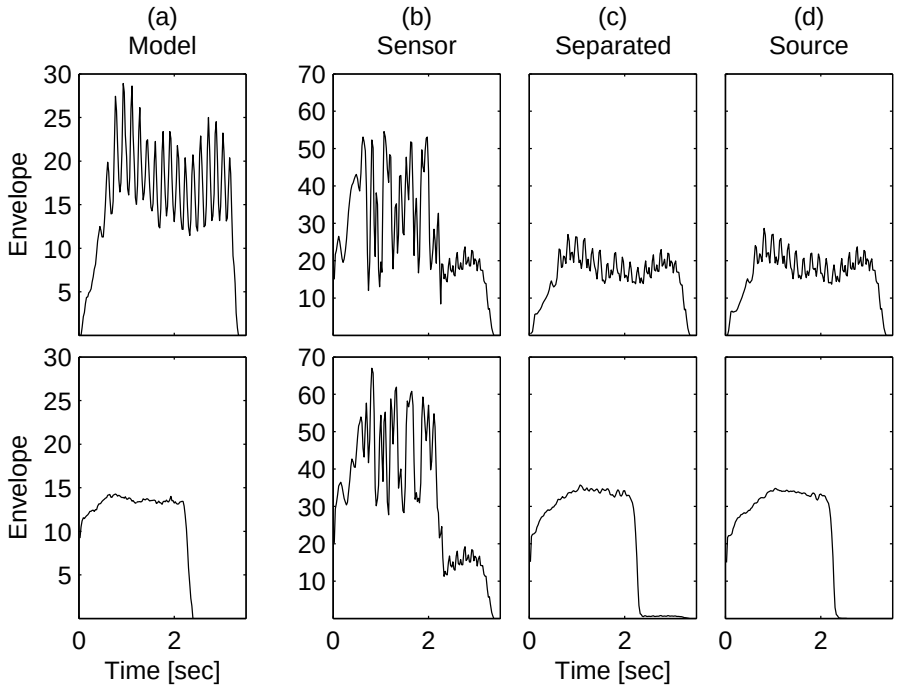


Figure 5.10: *Partial envelopes, from left to right: (a): Model partials (closest neighbors). (b): Sensor partials. (c): Separated partials. (d): Original source partials.*

We emphasize that the envelopes of the separated partials (column (c)) are better than what it could be expected. Their shapes are closer to those of the perfect source partials (column (d)) than to those of the model partials (column (a)). Furthermore, even though the method works with normalized envelopes, the separated partials have retained not only the shapes of the original source partials, but also the correct scaling.

5.4 Conclusion

We have presented a method for separation of overlapping partials in multi-channel audio mixtures. The method combines aspects of time-frequency analysis and spatial demixing techniques. It is based on the maximization of the similarity of normalized envelopes of the partials and is able to accurately recover the amplitude and frequency modulation of the original source partials from the sensor signals where they overlap. The method works on partials individually, and can therefore also work (to some extent) in scenarios where there are more sources than sensors. Finally, it can easily be used in conjunction with several of the existing source separation methods.

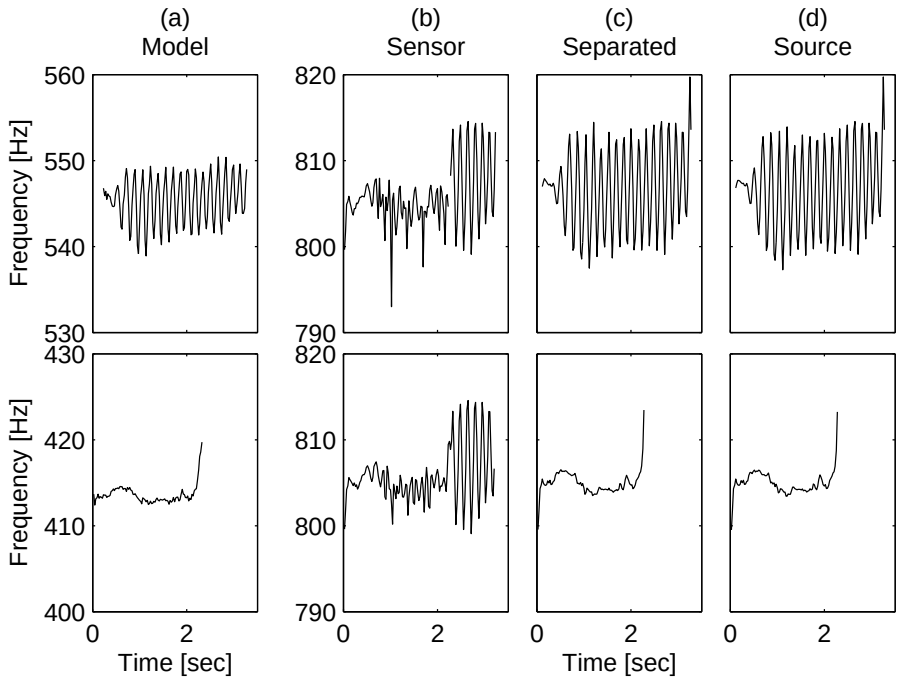


Figure 5.11: *Partial frequency trajectories, from left to right: (a): Model partials (closest neighbors). (b): Sensor partials. (c): Separated partials. (d): Original source partials.*

Chapter 6

Conclusion

6.1 Summary

The goal of this thesis was to devise computational techniques for the localization and separation of sound sources in binaural signals. This was motivated by the performance of the human auditory system, as discussed in the introduction chapter. An overview of existing techniques was given in Chapter 2. Drawing knowledge from these, we have proposed new techniques for localization and separation of sound sources.

Binaural localization and separation

The most important cues for binaural localization of sound sources are the interaural level and time differences. In Chapter 3, we proposed a technique for the joint evaluation of these cues for each individual left/right pair of spectral coefficients in the binaural STFT spectra. In this technique, the noisy azimuth estimates based on level differences were combined with the less noisy but ambiguous azimuth estimates based on time differences, effectively improving the accuracy of the azimuth estimates. We saw how the histograms of the azimuth estimates enabled the localization of sources and the tracking of these in dynamic scenes.

We also studied the relations between source locations and level and time differences for several different subjects in a database of measured head related transfer functions (HRTFs). Based on this study we proposed generic models for the interaural level and time differences as function of azimuth angle. These models were parametrized by a single parameter, namely the distance between the ears. Effectively, they enable the localization of sources in binaural signals observed by a listener whose HRTFs have not been measured.

Based on the source location estimates, we saw how separation could be obtained by spatial filtering in the azimuth parameter space. This was discussed in Chapter 4.

Separation of overlapping partials

In Chapter 5 we proposed a technique for the separation of overlapping partials of harmonic instruments in static scenes. The technique was based on the

similarity of the temporal envelopes between the different partials of a harmonic note.

6.2 Future research

The techniques that have been proposed in this thesis do not provide a complete solution to the problems of source localization and separation. However, they are powerful and can serve as motivation and basis for further work towards a more complete system. In particular, the proposed techniques are versatile since they are based on the STFT and can be combined with several of the existing techniques. We briefly mention two ideas for extensions that may possibly improve the proposed techniques, leading one further step towards a robust system for source localization and separation.

Source localization above 6 kHz

The accuracy of the proposed technique for binaural source localization degraded significantly for frequencies above about 6 kHz, as discussed in Chapter 3. This limits the frequencies for which the separation technique based on azimuth estimates is useful. At high frequencies, however, the human auditory system is less sensitive to phase. Moreover, the width of the perceptual bands are larger. For time delay estimation above 6 kHz it can be advantageous to estimate the group delays in perceptual bands, as opposed to the phase delays for individual spectral coefficients that were employed in Chapter 3. The group delay estimates can be combined with the phase delay estimates in a hybrid system.

Combined analysis of spatial and signal cues

Another possible extension is related to the artifacts due to independent processing of the spectral coefficient pairs, as was discussed in Section 4.3.1. The source separation by spatial windowing of the azimuth estimates yields binary time-frequency weights. By employing signal cues, overlapping spectral components can be detected similarly to the analysis described in Section 5.2. This analysis can be used for the estimation of non-binary weights for overlapping spectral components, effectively performing a "smart" spectral smoothing. In addition, it can work in dynamic scenes.

Bibliography

Abe, T., Kobayashi, T., and Imai, S. (1995). Harmonics tracking and pitch extraction based on instantaneous frequency. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, pages 756–759.

Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. (2001). The CIPIC HRTF database. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 99–102, Mohonk, New York, USA.

Araki, S., Makino, S., Mukai, R., Hinamoto, Y., Nishikawa, T., and Saruwatari, H. (2002). Equivalence between frequency domain blind source separation and frequency domain adaptive beamforming. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, pages 1785–1788, Orlando, Florida, USA.

Araki, S., Mukai, R., Makino, S., Nishikawa, T., and Saruwatari, H. (2003). The fundamental limmitation of frequency domain blind source separation for convolutive mixtures of speech. *IEEE Transactions on Speech and Audio Processing*, **11**(2), 109–116.

Balan, R., Rosca, J., and Rickard, S. (2001). Robustness of parametric source demixing in echoic environments. In *Proceedings of International workshop on Independent Component Analysis and Blind Signal Separation*, pages 144–149, San Diego, USA.

Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, **7**(6), 1129–1159.

Blauert, J. (2001). *Spatial Hearing*. MIT press.

Boashash, B. (1992). Estimating and interpreting the instantaneous frequency of a signal. *Proceedings of the IEEE*, **80**(4), 520–568.

Bodden, M. (1993). Modeling human sound-source localization and the cocktail-party-effect. *acta acustica*, **1**, 43–55.

Bofill, P. (2003). Underdetermined blind separation of delayed sound sources in the frequency domain. *Neurocomputing*, **55**(3-4), 627–641.

Bofill, P. and Zibulevsky, M. (2000). Blind separation of more sources than mixtures using sparsity of their short-time Fourier transform. In *Proceedings of International workshop on Independent Component Analysis and Blind Signal Separation*, pages 87–92, Helsinki, Finland.

Brandstein, M. and Ward, D., editors (2001). *Microphone Arrays: Signal Processing Techniques and Applications*. Springer Verlag, Berlin.

Bregman, A. S. (1999). *Auditory Scene Analysis*. MIT Press.

Comon, P. (1994). Independent component analysis - a new concept? *Signal Processing (Elsevier)*, **36**(3), 287–314.

Cooper, D. H. and Bauck, J. L. (1980). On acoustical specification of natural stereo imaging. In *66th AES Convention*.

Daubechies, I. (1992). *Ten lectures on wavelets*. SIAM, Philadelphia.

Drake, L. A. (2001). *Sound Source Separation via CASA-enhanced Beamforming*. Ph.D. thesis, Northwestern University, Evanston, Illinois.

Duda, R. O. (1997). *Elevation Dependence of the Interaural Transfer Function*, chapter 3, pages 49–75. In (Gilkey and Anderson, 1997).

Duda, R. O. and Martens, W. L. (1997). Range-dependence of the HRTF for a spherical head. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk, New York, USA.

Duda, R. O. and Martens, W. L. (1998). Range dependence of the response of a spherical head model. *Journal of the Acoustical Society of America*, **104**(5), 3048–3058.

Eronen, A. and Klapuri, A. (2000). Musical instrument recognition using cepstral coefficients and temporal features. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, pages 753–756, Istanbul, Turkey.

Frost, O. L. (1972). An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE*, **60**(8), 926–935.

Gabor, D. (1946). Theory of communication. *IEE Proceedings*, **93**(3), 429–457.

Gaik, W. (1993). Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling. *Journal of the Acoustical Society of America*, **94**(1), 98–110.

George, E. B. and Smith, M. J. T. (1992). Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones. *Journal of the Audio Engineering Society*, **40**(6), 497–516.

George, E. B. and Smith, M. J. T. (1997). Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model. *IEEE Transactions on Speech and Audio Processing*, **5**(5), 389–406.

Gerven, S. V., Compernolle, D. V., Wauters, P., Verstraeten, W., Eneman, K., and Delaet, K. (1995). Multiple beam broadband beamforming : Filter design and real time implementation. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk, New York, USA.

Gilkey, R. H. and Anderson, T. R., editors (1997). *Binaural and Spatial Hearing in Real and Virtual Environments*. Lawrence Erlbaum Associates.

Goodwin, M. M. and Elko, G. W. (1993). Constant beamwidth beamforming. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, pages 169–172.

Griffiths, L. J. and Jim, C. W. (1982). An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, **30**, 27–34.

Helmholtz, H. (1954). *On the Sensations of Tone*. Dover Publications.

Hyvärinen, A. (1999). Survey on independent component analysis. *Neural Computing Surveys*, **2**, 94–128.

Ikeda, S. and Murata, N. (1999). A method of ICA in time-frequency domain. In *Proceedings of International workshop on Independent Component Analysis and Blind Signal Separation*, Aussois, France.

Joris, P. X., Smith, P. H., and Yin, T. C. T. (1998). Coincidence detection in the auditory system: 50 years after jeffress. *Neuron*, **21**, 1235–1238.

Jourjine, A., Rickard, S., and Yılmaz, Ö. (2000). Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, pages 2985–2988, Istanbul, Turkey.

Jutten, C. and Herault, J. (1991). Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing (Elsevier)*, **24**, 1–10.

Karjalainen, M. and Tolonen, T. (1999). Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, pages 929–932, Phoenix, Arizona, USA.

Kashino, K. and Hagita, N. (1996). A music scene analysis system with the mrf-based information integration scheme. In *IEEE International Conference on Pattern Recognition*, pages 725–729.

Kashino, K. and Murase, H. (1998). Music recognition using note transition context. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, pages 3593–3596.

Keiler, F. and Marchand, S. (2002). Survey on extraction of sinusoids in stationary sounds. In *Proceedings of 5th International Conference on Digital Audio Effects*, pages 51–58, Hamburg, Germany.

Klapuri, A. (2001). Multipitch estimation and sound separation by the spectral smoothness principle. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, Salt Lake City, Utah, USA.

Kollmeier, B., editor (1996). *Psychoacoustics, Speech and Hearing Aids*. World Scientific Publishing.

Kvedalen, E. (2003). *Signal processing using the Teager Energy Operator and other nonlinear operators*. Master's thesis, University of Oslo, Norway.

Lambert, R. H. (1996). *Multichannel blind deconvolution: FIR matrix algebra and separation of multipath mixtures*. Ph.D. thesis, University of Southern California.

Lee, T.-W., Bell, A. J., and Lambert, R. H. (1997). Blind separation of delayed and convolved sources. *Advances in Neural Information Processing Systems, MIT Press*.

Levine, S. N. and Smith, J. O. I. (1998). A sines+transients+noise audio representation for data compression and time/pitch scale modifications. In *AES 105th Convention*, San Francisco, USA.

Lindemann, W. (1986). Extension of a binaural cross-correlation model by contralateral inhibition. I. simulation of lateralization for stationary signals. *Journal of the Acoustical Society of America*, **80**(6), 1608–1622.

Lyon, R. F. (1983). A computational model of binaural localization and separation. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, pages 1148–1151, Boston, USA.

Macpherson, E. A. and Middlebrooks, J. C. (2002). Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *Journal of the Acoustical Society of America*, **111**(5), 2219–2236.

Maher, R. C. (1990). Evaluation of a method for separating digitized duet signals. *Journal of the Audio Engineering Society*, **38**(12), 956–979.

Mallat, S. G. and Zhang, Z. (1993). Matching pursuit with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, **41**(12), 3397–3415.

Maragos, P., Kaiser, J. F., and Quatieri, T. F. (1993). Energy separation in signal modulations with application to speech analysis. *IEEE Transactions on Signal Processing*, **41**(10), 3024–3051.

Martin, K. D. and Kim, Y. E. (1998). Musical instrument identification: A pattern-recognition approach. In *136th meeting of the Acoustical Society of America*.

McAulay, R. J. and Quatieri, T. F. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **34**(4), 744–754.

Mellody, M. and Wakefield, G. H. (2000). The time-frequency characteristics of violin vibrato: Modal distribution analysis and synthesis. *Journal of the Acoustical Society of America*, **107**(1), 598–611.

Mitianoudis, N. and Davies, M. E. (2003). Audio source separation of convolutive mixtures. *IEEE Transactions on Speech and Audio Processing*, **11**(5), 489–497.

Moore, B. C. (1997). *An introduction to the psychology of hearing*. Academic Press.

Nakatani, T. and Okuno, H. G. (1999). Harmonic sound stream segregation using localization and its application to speech stream segregation. *Speech Communications (Elsevier)*, **27**, 209–222.

Parra, L. C. and Alvino, C. V. (2002). Geometric source separation: merging convolutive source separation with geometric beamforming. *IEEE Transactions on Speech and Audio Processing*, **10**(6), 352–362.

Peissig, J. (1992). *Binaurale Hörgerätestrategien in komplexen Störschallsituationen*. Ph.D. thesis, Universität Göttingen, Germany.

Picinbono, B. (1997). On instantaneous amplitude and phase of signals. *IEEE Transactions on Signal Processing*, **45**(3), 552–560.

Plomp, R. and Levelt, W. (1965). Tonal consonance and critical bandwidth. *Journal of the Acoustical Society of America*, **38**, 548–560.

Polotti, P. and Evangelista, G. (2001a). Analysis and synthesis of pseudo-periodic1/f-like noise by means of wavelets with applications to digital audio. *EURASIP Journal on Applied Signal Processing*, **2001**(1), 1–14.

Polotti, P. and Evangelista, G. (2001b). Fractal additive synthesis by means of harmonic-band wavelets. *Computer Music Journal*, **25**(3), 22–37.

Portnoff, M. R. (1976). Implementation of the digital phase vocoder using the fast fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **24**(3), 243–248.

Puckette, M. S. and Brown, J. C. (1998). Accuracy of frequency estimates using the phase vocoder. *IEEE Transactions on Signal Processing*, **6**(2), 166–176.

Quatieri, T. F. and Danisewicz, R. G. (1990). An approach to co-channel talker interference suppression using a sinusoidal model for speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **38**(1), 56–69.

Rickard, S. and Yılmaz, Ö. (2002). On the approximate w-disjoint orthogonality of speech. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, pages 529–532, Orlando, Florida, USA.

Roads, C., Pope, S., Picalli, A., and de Poli, G., editors (1997). *Musical Signal Processing*. Swets and Zeitlinger Publishers.

Rosenthal, D. F. and Okuno, H. G., editors (1998). *Computational auditory scene analysis*. Lawrence Erlbaum Associates.

Serra, X. (1997). *Musical Sound Modeling with Sinusoids plus Noise*. In (Roads *et al.*, 1997).

Serra, X. and Smith, J. (1990). Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, **14**(4), 12–24.

Smaragdis, P. (1998). Blind separation of convolved mixtures in the frequency domain. In *Int. Workshop on Independence and Artificial Neural Networks*.

Smaragdis, P. J. (1997). *Information theoretic approaches to source separation*. Master's thesis, MIT.

Smith, J. O. and Serra, X. (1987). PARSHL: An analysis/synthesis program from non-harmonic sounds based on a sinusoidal representation. In *Proceedings of International Computer Music Conference*, San Francisco, USA.

Sontacchi, A., Noisternig, M., Majdak, P., and Höldrich, R. (2002). An objective model of localisation in binaural sound reproduction systems. In *AES 21st International Conference*, St.Petersburg, Russia.

Stern, R. M. and Trahiotis, C. (1997). *Models of Binaural Perception*, chapter 24, pages 499–531. In (Gilkey and Anderson, 1997).

Strutt, J. W. (1907). On our perception of sound direction. *Philos. Mag.*, **13**, 214–232.

Sun, M. and Sclabassi, R. J. (1993). Discrete-time instantaneous frequency and its computation. *IEEE Transactions on Signal Processing*, **41**(5), 1867–1880.

Teager, H. M. and Teager, S. M. (1989). Evidence for nonlinear speech production mechanisms in the vocal tract. In *NATO Advanced Study Institute on Speech Production and Speech Modeling*, Bonas, France.

Tolonen, T. (1999). Methods for separation of harmonic sound sources using sinusoidal modeling. In *AES 106th Convention*, Munich, Germany.

Torkkola, K. (1996). Blind separation of convolved sources based on information maximization. In *Proc. IEEE Workshop on Neural Networks for Signal Processing*, pages 423–432, Kyoto, Japan.

Torkkola, K. (1999). Blind separation for audio signals – are we there yet. In *Proceedings of International workshop on Independent Component Analysis and Blind Signal Separation*, pages 239–244, Aussois, France.

van Veen, B. D. and Buckley, K. M. (1988). Beamforming - a versatile approach to spatial filtering. *IEEE ASSP Magazine*, pages 4–24.

Verma, T. S. and Meng, T. H. Y. (1999). Sinusoidal modeling using frame-based perceptually weighted matching pursuits. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, pages 981–984, Phoenix, Arizona, USA.

Verma, T. S. and Meng, T. H. Y. (2000). Extending spectral modeling synthesis with transient modeling synthesis. *Computer Music Journal*, **24**(2), 47–59.

Virtanen, T. and Klapuri, A. (2000). Separation of harmonic sound sources using sinusoidal modeling. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, pages 765–768, Istanbul, Turkey.

Virtanen, T. and Klapuri, A. (2001). Separation of harmonic sounds using multipitch analysis and iterative parameter estimation. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 83–86, Mohonk, New York, USA.

Virtanen, T. and Klapuri, A. (2002). Separation of harmonic sounds using linear models for the overtone series. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, Orlando, Florida, USA.

Viste, H. and Evangelista, G. (2002). An extension for source separation techniques avoiding beats. In *Proceedings of 5th International Conference on Digital Audio Effects*, pages 71–75, Hamburg, Germany.

Viste, H. and Evangelista, G. (2003a). On the use of binaural cues to improve binaural source separation. In *Proceedings of 6th International Conference on Digital Audio Effects*, pages 209–213, London, UK.

Viste, H. and Evangelista, G. (2003b). Separation of harmonic instruments with overlapping partials in multi-channel mixtures. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk, New York, USA.

Viste, H. and Evangelista, G. (2004a). Binaural source localization (invited paper). In *Proceedings of 7th International Conference on Digital Audio Effects*, Naples, Italy.

Viste, H. and Evangelista, G. (2004b). A method for separation of overlapping harmonics based on similarity of temporal envelopes in multi-channel mixtures. *accepted to IEEE Transactions on Speech and Audio Processing*.

Viste, H. and Evangelista, G. (2004c). Narrow band DOA estimation in binaural signals. *in preparation*.

von Hornbostel, E. M. and Wertheimer, M. (1920). Über die wahrnehmung der schallrichtung [on the perception of the direction of sound]. *Sitzungsber. Akad. Wiss. Berlin*, pages 388–396.

Wang, A. L.-C. (1994). *Instantaneous and Frequency-Warped Signal Processing Techniques for Auditory Source Separation*. Ph.D. thesis, Stanford University.

Wightman, F. L. and Kistler, D. J. (1997). *Factors Affecting the Relative Salience of Sound Localization Cues*, chapter 1, pages 1–23. In (Gilkey and Anderson, 1997).

Wittkop, T., Albani, S., Hohmann, V., Peissig, J., Woods, W. S., and Kollmeier, B. (1997). Speech processing for hearing aids: Noise reduction motivated by models of binaural interaction. *ACUSTICA united with acta acustica*, **83**, 684–699.

Woods, W. S., Hansen, M., Wittkop, T., and Kollmeier, B. (1996). A simple architecture for using multiple cues in sound separation. In *IEEE Fourth International Conference on Spoken Language Processing*, pages 909–912.

Woodworth, R. S. and Schlosberg, H. (1954). *Experimental psychology*. Holt, New York.

Yılmaz, Ö. and Rickard, S. (2002). Blind separation of speech mixtures via time-frequency masking. *Submitted to IEEE Transactions on Signal Processing*.

Zwicker, E. and Fastl, H. (1999). *Psychoacoustics: Facts and Models, 2nd edition*. Springer, New York.

Curriculum Vitae

Name:	Harald Viste
Date of birth:	June 21, 1972.
Nationality:	Norwegian
WWW:	http://lcavwww.epfl.ch/∼viste

Assignment history

2000-currently	**PhD studies/Research assistant** in the Audiovisual Communications Laboratory at the Swiss Federal Institute of Technology, Lausanne (EPFL), Switzerland.
1997-2000	**Project Engineer** in Schlumberger, Reservoir Evaluation Seismic, R&D department in Oslo, Norway.
1992-1997	**Sivilingeniør in industrial mathematics** at the Norwegian University of Science and Technology (NTNU), Trondheim, Norway. Major: Numerical mathematics. Minor: Digital signal processing. NTNU considers the sivilingeniør degree equivalent to the Master of Science degree in engineering.
1992-1996	First four years at NTNU at the Faculty of Physics and Mathematics, Department of Mathematical Sciences.
1996-1997	Diploma Thesis at Centre de Physique Théorique (CPT) at Centre Nationale de Recherche Scientifique (CNRS) in Marseille, France. Title: "Wavelet transforms: reconstructions based on extrema on dyadic scales and on ridges."
1991-1992	**Military service**, Norway.
1990-1991	**Surveyor assistant**, Kristiansand, Norway.

Publications

Journals and manuscripts

Harald Viste and Gianpaolo Evangelista,
"Narrow band DOA estimation in binaural signals,"
in preparation.

Harald Viste and Gianpaolo Evangelista,
"A generic HRTF model for binaural source localization and separation,"
in preparation.

Harald Viste and Gianpaolo Evangelista,
"A method for separation of overlapping partials based on similarity of temporal envelopes in multi-channel mixtures,"
accepted to IEEE Transactions on Speech and Audio Processing, 2004.

Conferences

Harald Viste and Gianpaolo Evangelista,
"Binaural source localization," (invited paper)
to appear in Proceedings of 7th International Conference on Digital Audio Effects (DAFx-04), October 2004, Naples, Italy.

Harald Viste and Gianpaolo Evangelista,
"On the use of spatial cues to improve binaural source separation,"
in Proceedings of 6th International Conference on Digital Audio Effects (DAFx-03), September 2003, pp. 209-213, London, UK.

Harald Viste and Gianpaolo Evangelista,
"Separation of harmonic instruments with overlapping partials in multi-channel mixtures,"
in Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-03), October 2003, New Paltz, New York, USA.

Harald Viste and Gianpaolo Evangelista,
"An extension for source separation techniques avoiding beats,"
in Proceedings of 5th International Conference on Digital Audio Effects (DAFx-02), September 2002, pp. 71-75, Hamburg, Germany.

Harald Viste and Gianpaolo Evangelista,
"Sound Source Separation: Preprocessing for hearing aids and structured audio coding,"
in Proceedings of 4th International Conference on Digital Audio Effects (DAFx-01), December 2001, pp. 67-70, Limerick, Ireland.