UNIVERSITE DE NICE-SOPHIA ANTIPOLIS - UFR Sciences
Ecole Doctorale des Sciences et Technologies de l'Information et de la Communication

T H E S E

pour obtenir le titre de
Docteur en Sciences
de l'UNIVERSITE de Nice-Sophia Antipolis

Discipline : Informatique

présentée et soutenue par
Emmanuel GALLO

Restitution Sonore Hiérarchique et Perceptive d'Environnements Virtuels Multi-Modaux

Thèse dirigée par George DRETTAKIS
soutenue le 19 mars 2006

Jury :

George Drettakis	Directeur de Thèse	REVES/INRIA, France
Nicolas Tsingos	Co-directeur de Thèse	REVES/INRIA, France
Jean-Christophe Lombardo	Tuteur industriel	CSTB, France
Lauri Savioja	Rapporteur	Helsinki University of Technology, Finland
Durand R. Begault	Rapporteur	NASA Ames Research Center, USA
Jean-Marc Jot	Examinateur	Creative Advanced Technology Center, USA

RESUME

Ce travail porte sur la simulation acoustique temps-réel pour des applications de réalité virtuelle ou les jeux vidéo. Ce type d'application nécessite des temps de calcul considérables, augmentant avec la complexité de la scène et impliquant des difficultés pour le rendu interactif. La simulation d'une scène sonore complexe reste encore difficile à réaliser en temps réel à cause du coût de la gestion indépendante des sources sonores. De plus, la description de la scène sonore nécessite de spécifier la nature et la position de chaque source sonore qui la compose, ce qui est une étape longue et fastidieuse. Dans ce cadre, nous avons étudié la possibilité d'effectuer la simulation acoustique en tirant parti de la puissance de calcul des cartes graphiques de dernière génération. Les résultats montrent que l'architecture hautement parallèle de ces cartes est appropriée pour ce type de calcul, augmentant grandement les performances par rapport aux processeurs actuels. Nous nous sommes intéressés par la suite à développer un algorithme exploitant l'audition humaine, permettant d'effectuer un rendu sonore de la scène en respectant un budget d'opérations donné. Pour cela, l'algorithme évalue une métrique d'importance pour chaque signal à traiter sur des intervalles de temps très fins. Puis il effectue les opérations par ordre de priorité jusqu'à atteindre le budget fixé. Une évaluation subjective a été effectuée pour comparer différentes métriques d'importance. Enfin, nous avons élaboré une méthode alternative d'acquisition de scène sonore qui évite la modélisation individuelle de chaque source. A partir d'enregistrements monophoniques simultanés d'une scène réelle, cette méthode en détache les sources qui la composent. En étudiant les différences de temps d'arrivée des enregistrements sur plusieurs bandes de fréquence, une position est extraite pour la source sonore émettrice la plus présente dans chaque bande. Les composantes de chaque source peuvent ensuite être spatialisées aux positions trouvées. En utilisant ce principe, nous pouvons également rééditer la scène acquise. Par exemple, nous pouvons déplacer ou supprimer une source, ou changer la position de l'auditeur en temps réel. Nous pouvons aussi combiner plusieurs éléments provenant de différents enregistrements tout en assurant une cohérence spatiale globale.

ABSTRACT

This thesis concentrates on real-time acoustic simulations for virtual reality applications or video games. Such applications require huge computing times, increasing with the complexity of the scene and involving difficulties for interactive rendering. In particular, the real-time simulation of a complex sound scene remains difficult due to the independent processing of each sound source. Moreover, the description of the auditory scene requires specifying the nature and the position of each sound source, which is a long and tedious process. To solve these problems, we studied the possibility of performing the acoustic simulation by leveraging the computing power of latest generation graphics processors. The results show that their massively parallel architecture is well suited to such processing, increasing significantly the performances compared to current general purpose processors. We were interested thereafter in developing an algorithm exploiting the human perception in order to render an auditory scene by respecting a target budget of operations while minimizing audible artifacts. The proposed algorithm evaluates an importance metric for each signal on very fine time-intervals. Then, it performs the required signal processing operations by descending priority order until the target budget is reached. A subjective evaluation was made to assess different importance metrics.

Finally, we developed an alternative method of sound acquisition which avoids the individual modeling of each source. From simultaneous monophonic recordings of a real scene, this method extracts the scene components. We analyze time-delay-of-arrival in the recorded signals in several frequency bands. From this information, a position is extracted for the most significant sound source in each band. The components from each source can then be re-rendered at the corresponding locations. Using this method, we can also edit the acquired scene. For instance, we can move or delete a sound source, or change the position of the listener in real-time. We can also composite several elements coming from different recordings while ensuring overall spatial coherence.

Contents

List of Figures

Restitution Sonore Hiérarchique et Perceptive d'Environnements Virtuels Multi-Modaux

Introduction

La réalité virtuelle a émergé avec la demande de simuler des phénomènes complexes afin de mieux comprendre des processus compliqués, de développer des systèmes d'entraînement ou d'analyser des comportements humains. La plupart des travaux en réalité virtuelle se sont concentrés sur les aspects visuels tandis que moins d'attention a été prêtée à d'autres informations sensorielles, telles que le son. Néanmoins, il est largement accepté que la combinaison du son et de l'image améliore le degré d'immersion et est une composante clé pour augmenter le sentiment de présence dans les environnements virtuels. En particulier, les indices auditifs nous informent sur l'environnement et nous aident à la localisation, particulièrement pour des sources invisible pour l'auditeur. Le processus pour simuler une scène virtuelle auditive s'appelle " Auralisation " [Kleiner et al., 1993]. Auraliser une source sonore virtuelle exige que l'auditeur perçoive :

1. la provenance du son,

2. les effets environnementaux dus aux occlusions, aux échos et aux réverbérations.

Ainsi, il est nécessaire de simuler les phénomènes physiques correspondants, entraînant des temps de calcul considérables. De plus, la simulation sonore dépend de la position d'écoute et, par conséquent, il est obligatoire de mettre à jour le rendu sonore lorsque l'auditeur explore l'environnement. De nos jours, le succès des applications de réalité virtuelle amène une demande croissante de simulations de haute qualité supportant des traitements et des effets plus avancés. Les environnements virtuels deviennent également de plus en plus complexes. Par exemple, les jeux vidéo actuels emploient des centaines de sources sonores simultanées, ce qui provoque un

volume de calcul très élevé qui ne peut pas être traité par les solutions logicielles ou matérielles actuelles.

D'autre part, le rendu de telles scènes complexes amène également un problème de création du contenu. Les modèles courants exigent de représenter toutes les composantes sonores par des sources ponctuelles. Chaque source sonore doit être décrite par sa position 3D, sa trajectoire et un signal monophonique associé. Quand le nombre de sources augmente, il devient de plus en plus difficile de concevoir le contenu de la scène. Ecrire des scènes complexes devient un processus pénible. D'ailleurs, créer une scène sonore virtuelle qui correspond à une scène réelle est très difficile en utilisant ce modèle. En particulier, les signaux de chaque objet sonore doivent être enregistrés individuellement, ce qui n'est pas toujours faisable. Par exemple, enregistrer séparément le moteur et la vibration engendrée sur la carrosserie du vehicule n'est pas facilement réalisable.

Le premier objectif de cette thèse est de développer de nouveaux algorithmes, permettant un rendu interactif de scènes sonores complexes comprenant un nombre important de sources sonores, typiquement des centaines voire des milliers. Pour réaliser cet objectif, nous chercherons à tirer profit des ressources logicielles et matérielles disponibles. Un de nos objectifs est que nos simulations soient réalisable sur toute plateforme, du simple ordinateur portable aux postes de travail de dernière génération. Nous proposerons donc une nouvelle méthode progressive pour rendre une scène auditive efficacement employant un compromis entre la vitesse et la qualité. Dans ce cadre, nous utiliserons notre connaissance sur l'audition humaine afin d'accélérer le processus de rendu.

La deuxième partie de cette thèse adresse le problème de création d'environnements réalistes. Comme alternative à la modélisation de scènes sonores virtuelles avec des sources ponctuelles, nous proposons une méthode innovante pour décrire une scène sonore de manière automatique, en se servant d'enregistrements pris sur le terrain, sans aucune contrainte sur la position des microphones. Cette approche s'inspire des techniques d'"image-based rendering" utilisées dans le domaine du graphique.

Résumé des Contributions

Cette thèse est divisée en trois parties. La première partie présente une vue d'ensemble des principes impliqués dans le domaine de l'acoustique et du rendu sonore 3D, ainsi que les concepts de base nécessaires pour la comprension des chapitres. L'audition

humaine est analysée dans un premier temps et, en étudiant l'anatomie de l'appareil auditif, nous expliquons divers phénomènes psycho-acoustiques. Dans la deuxieme partie, nous montrons comment nous pourrons les exploiter afin de rendre une scène sonore de manière efficace. Enfin, La troisième partie présente une nouvelle méthode pour créer automatiquement une scène sonore virtuelle à partir d'enregistrements d'une scène réelle dans sa globalité.

Technique de Rendu Sonore Efficace

Notre première approche est d'exploiter le traitement parallèle fourni par le processeur des cartes graphiques afin de pouvoir rendre un grand nombre de source efficacement. Une simulation sonore 3D en temps réel est présentée comme exemple d'application comprenant des effets standard comme le ré-echantillonage du au délai, l'atténuation par la distance et une égalisation sur plusieurs bandes de fréquence pour prendre en compte les indices spectraux de localisation (HRTF). Les résultats ont prouvé que l'architecture de ces processeur est bien adaptée pour effectuer ce type de traitement audio. Cependant, bien que nous obtenions une amélioration significative des performances, l'algorithme dépend toujours directement du nombre de sources.

De ce fait, nous examinons la possibilité de réduire la quantité de données à traiter en utilisant les propriétés de la perception humaine, telles que le masquage auditif et l'illusion de continuité. Nous avons étudié plusieurs métriques d'émergence généralement employées pour classifier ou trier diverses composantes auditives selon leur importance perceptive. Nous les utilisons pour choisir des sous parties de plusieurs signaux progressivement afin d'optimiser le traitement audio temps réel. Nous effectuons une étude subjective pour évaluer quelle métrique réalise le mieux la reconstruction en utilisant seulement une quantité limitée de données. Nos études montrent que le niveau RMS, utilisé en tant que métrique d'importance, offre un bon compromis pour tous les types de signaux. Nos résultats montrent également que des sous parties des signaux d'origine peuvent être omises dans la plupart des cas, sans dégradation apparente dans les mélanges générés, ce qui valide cette approche pour des applications temps réel. Enfin, nous proposons une approche de type niveau de détail semblable à celles utilisées dans le domaine du graphique pour rendre une scène sonore selon les ressources disponibles.

Description Automatique de Scene Sonore

La troisième partie présente une nouvelle méthode pour créer automatiquement une scène auditive virtuelle à partir d'enregistrements de la scène dans sa globalité, semblables aux techniques utilisées pour l' "image based rendering" dans le domaine graphique. Utilisant un ensemble de microphones standard, distribués dans un environnement réel, nous avons enregistré le champ sonore simultanément à plusieurs endroits sans aucune contrainte sur leur emplacement. Cette méthode peut être employée sur des scènes extérieures, ainsi que des scènes d'intérieur ayant peu de réverbération. Après une calibration spatial, nous segmentons dans ces enregistrements un certain nombre de composantes sonores, ainsi que leur provenance. Dans ce contexte, nous évaluons différentes techniques pour estimer la différence de temps d'arrivée entre les paires de microphones et nous présentons un algorithme hiérarchique pour localiser les sources sonores. Notre approche extrait automatiquement une description haut niveau de la scène sonore à partir des enregistrements, basée sur l'endroit d'émission des signaux et leurs contenus fréquentiels, qui peut être ensuite spatialisé avec un modèle de source ponctuelle. En utilisant la représentation obtenue, nous pouvons éditer et re-rendre la scène sonore acquise avec different types de système de restitution. En particulier, nous pouvons nous déplacer ou changer les différentes sources sonores et arbitrairement choisir une position d'écoute. Nous pouvons également combiner plusieurs éléments de différentes scènes en préservant une cohérence spatiale. Nous montrons également un éventail d'applications possibles concernant les jeux, la réalité virtuelle/augmentée et la post-production cinématographique.

Cependant, l'algorithme suppose que le champ sonore est émis par des sources ponctuelles, ce qui n'est pas toujours vrai. Nous présentons donc une nouvelle approche pour résoudre ce problème. Nous séparons les éléments sonores de premier plan, constitué d'événements bien localisés, de ceux d'arrière plan, constitué de sons plus diffus, en utilisant des hypothèses sur la stationnarité de l'arrière plan. Nous adaptons, par conséquent, le rendu sonore en utilisant une stratégie différente suivant le plan sonore. Nous reconstruisons la scène sonore de n'importe quel point de vue en re-spatialisant le premier plan avec l'approche précédente alors que l'arriere plan est encodé avec des harmoniques sphériques d'ordre réduit afin de fournir un rendu spatial diffus. Nous montrons également qu'une déformation des enregistrements originaux permet de simuler des changements fluides du point d'écoute et des positions des sources. Anfin de valider notre méthode, nous présentons une étude subjective, comparant notre approche avec des enregistrements binauraux servant de référence, ainsi que des enregistrements B-format montrant que notre approche réalise un bon rendu spatial, tout en gardant une flexibilité sur le rendu et offrant de nombreuses possibilités d'éditions.

Conclusion

Dans cette thèse, nous nous sommes intéressé aux problèmes liés au rendu sonores de scènes complexes, contenant un nombre considérable de sources. Nous avons identifié deux difficultés qui empêchent le rendu interactif de telles scènes. Les temps de calcul engendrés par le traitement audio requis par ce type de simulation sont au delà des capacités des processeurs actuelles et le processus de création de telles scènes est long et pénible. Dans cette thèse, nous avons proposé des solutions à ces deux problèmes.

Afin d'exécuter le nombre massif d'opérations requis par le rendu des simulations sonores d'environments virtuels, nous avons proposé de tirer profit de l'architecture parallèle fournie par les cartes graphiques (GPU). Bien que le GPU soit conçu pour des applications graphiques, sa flexibilité et son architecture apportent une solution alternative qui surpasse clairement les processeurs courants (CPU). D'ailleurs, les performances des GPUs ont augmenté considérablement ces trois dernières années par rapport aux CPUs et ils tendent à devenir de véritables processeurs universels. Nos études ont montré que cette architecture est bien appropriée pour le traitement audio. Prochainement, de telles architectures sont susceptibles de devenir des standards de facto et, de ce fait, nous sommes convaincu que les cartes sonores peuvent tirer bénéfice d'inclure le même type d'architecture et de programmabilité.

Afin de simplifier une scène sonore et fournir une approche de type progressive au rendu audio, nous avons proposé un nouvel algorithme qui exploite les propriétés de l'audition humaine, comme le masquage auditif et l'illusion de continuité. La méthode proposée fournit un niveau de détail, traitant progressivement les composantes sonores importantes d'une scène. Cet algorithme apporte un compromis entre la vitesse et la qualité, ce qui est bien adaptée aux applications temps réel, procurant une solution pouvant être utilisée sur toute plateforme. La solution proposée dans cette thèse permet de rendre des milliers de sources sonores sur tout type de plateformes, allant des ordinateurs dernière génération aux ordinateurs portables.

Dans la deuxième partie de cette thèse, nous avons présenté une méthode pour créer automatiquement des scènes sonores virtuelles basées sur des enregistrements de scènes réelles. Cette technique complète les méthodes d'enregistrement sonore spatial existantes. Cette approche contourne le problème de capturer chaque source individuellement tout en offrant un niveau semblable d'interaction avec la scène. En outre, la représentation de la scène obtenue fournit un codage compact du champ sonore qui est indépendant du système de restitution. Ce travail suggère également que des scènes sonores réelles peut être efficacement codées en utilisant peu d'information spatiale.

Une évaluation subjective de la qualité de la reconstruction spatiale, comparant notre méthode avec d'autres techniques spatiales d'enregistrement telles que binaural et le B-format ont été effectué. Il montre que notre approche surpasse l'enregistrement B-format et peut obtenir une localisation proche de l'enregistrement binaural. Notre approche permet une auralisation interactive des scènes sonores réelles tout en maintenant un rendu flexible. Nous sommes convaincus que notre approche peut offrir de nouvelles perspectives pour la post-production.

Directions de Recherche

Cette thèse ouvre beaucoup de directions prometteuses pour de futures recherches. Nous avons vu que le GPU est bien adapté au traitement de l'audio 3D. Malheureusement, nous n'avons pas pu évaluer nos algorithmes sur les derniers processeurs G80 qui résolvent les quelques problèmes soulignés par notre étude et qui amélioreraient certainement les performances. Il serait intéressant d'examiner d'autres algorithmes acoustiques tels que le filtrage par réponse impulsionnelle infinie (RII) ou par réponse impulsionnelle finie (RIR). En effet, ces algorithmes sont des outils de base pour inclure des effets de réverbération dans les applications de réalité virtuelle. Dans ce cas, les paramètres de réverbération pourraient sans doute être calculés directement en analysant la géométrie de la scène 3D avec le GPU. Une évaluation sur la performance de notre algorithme avec d'autres processeurs de type DSP pourrait être également intéressante. D'autre part, l'approche de type progressive pourrait être amélioré en ajoutant une sélection des grains plus fine ou en employant une autre métrique d'importance comme une sonie pour les signaux variant dans le temps. D'ailleurs, la théorie sur l'illusion de continuité n'a pas été explicitement employée dans la métrique d'importance, et pourrai améliorer les résultats. Une extension intéressante pourrait être d'utiliser cet algorithme avec d'autres représentations de signal comme celles obtenues avec un algorithme de type "sparse coding" [Lewicki, 2002]. Dans ce cas, nos grains élémentaires seront les atomes de la décomposition. L'algorithme présenté pourrait être également employé dans d'autres applications, pour réduire le trafic réseau ou compresser seulement les parties importantes du signal.

La deuxième partie de la thèse pourrait être également améliorée de plusieurs manières : Nous avons utilisé une hypothèse de "W-Disjoint orthogonalité" dans le domaine de fréquence pour la séparation de source mais le domaine de Fourier n'est en général pas assez parcimonieux pour des scénarios réels complexes. Travailler

dans un autre domaine plus “creux” pourrait améliorer la séparation de source. Dans chaque trame temporelle, nous cherchons une position de source indépendamment de la qualité de l'évaluation. Nous pourrions garder uniquement les bonnes estimations pendant un lapse de temps et les interpoler quand aucune évaluation satisfaisante n'est trouvée. Dans ce cas, la difficulté serait de trouver une bonne mesure de qualité pour nos évaluations. Dans son état actuel, l'algorithme emploie un nombre fixe de bandes. Une stratégie alternative serait d'optimiser ce découpage pour chaque trame. La localisation des microphones est basée sur des photographies. Il serait préférable de trouver les positions des microphones en utilisant une technique utilisant la différence de temps d'arrivée d'un signal aux microphones permettant de calibrer le système directement sur place. Ceci pourrait être utile pour des applications de télédiffusion. Enfin, dans la composante segmentée de l'arrière plan, il reste quelques parties du premier plan dans le signal. Une amélioration possible serait de choisir les parties du signal de l'arrière plan pour lesquelles l'intensité du signal de premier plan correspondant est importante et les remplacer entièrement par une approche de type synthèse de texture sonore.

En conclusion, nous pensons que cette thèse a pu réaliser les buts présentés dans l'introduction: accélérer la vitesse du traitement audio pour la simulation de scène sonore complexes et améliorer le processus de création de scène. Nous pensons que les résultats ainsi que les directions décrites pour les futurs travaux illustrent le fort potentiel de cette thèse.

Chapter 1

Introduction

In the last 30 years, virtual reality appeared with the demand to simulate complex and interactive phenomena in order to help us understand complicated processes, develop training applications or analyze human behavior. Most work on virtual reality has focused on visual aspects whereas less attention has been paid to other sensory information, such as sound. However, it is now widely accepted that the combination of sound and images improves realism and is a key component to increasing the sense of presence in virtual environments. In particular, auditory cues inform us about the surrounding environment and help the localization, especially for sources not directly visible to the listener.

The process of simulating a virtual auditory scene is called "Auralization" [Kleiner et al., 1993]. Auralizing a virtual sound source requires that the listener perceives:

1. the correct location of the sound

2. the correct environmental effects due to occlusions, echoes and reverberations

As a result, it is necessary to simulate related physical phenomena, resulting in a high computational cost. Moreover, the sound simulation depends on the listening position and, consequently, it is mandatory to update the rendering parameters while the listener explores the environment. Nowadays, the success of virtual reality applications has lead to growing demand for high quality simulations and support for more advanced processing and features. Virtual environments are also becoming increasingly complex. For instance, current video games use hundreds of simultaneous sound sources. This results in a very high computational load which cannot be supported by current brute-force hardware or software solutions.

On the other hand, the rendering of such complex scenes also leads to an authoring problem. Current models require representing all sounding components using point sound sources. Each sound source has to be described by its 3D location or trajectory

path and an associated monophonic signal. As the number of sources grows, the contents of the scene becomes increasingly difficult to design. Authoring complex scenes is thus a tedious process. Moreover, creating a virtual auditory scene to match a real scene is complicated using this model. In particular, signals have to be recorded individually for each sounding object, which is not always feasible. An example would be recording the engine and the vibration caused on the body of a car separately.

The first goal of this thesis is to develop new algorithms, enabling the interactive rendering of complex auditory scenes including a massive number of sound sources, typically hundreds to thousands. To achieve this goal, we will seek to benefit from all software and hardware resources available. Our simulations have to run on every kind of personal computer, ranging from a simple laptop to last-generation workstations, and thereby, we will propose a new scalable method to render an auditory scene efficiently using a speed versus quality trade-off. In this context, we will leverage our knowledge of human perception to accelerate the rendering process.

The second goal of this thesis addresses the problem of authoring realistic environments. As an alternative to describing virtual auditory scenes with point sound sources, we propose an innovative method for authoring an auditory scene automatically from field recordings inspired from image-based rendering techniques in graphics using an unconstrained setup of microphones.

1.1 Thesis Outline

This thesis is divided in three parts. The first part (Chapter 2 and Chapter 3) will present an overview of the principles involved in the field of 3D audio rendering and will introduce the basic concepts necessary to understand the following chapters. Human hearing will be studied first and, by investigating the anatomy of the auditory apparatus, we will explain various perceptual phenomena. We will see thereafter how we can exploit them to render a scene efficiently. Finally, we will review possible representations of an audio signal as well as standard simulation techniques used for auralization.

The second part contains two chapters which present new methods to render a complex audio scene in real-time. Recently, powerful graphics processor architectures have appeared. These processors are massively parallel and programmable for general-purpose applications. However, algorithms have to be designed to fit this architecture. Chapter 4 will present this architecture and will describe how to use this novel hardware to effectively perform audio processing operations as required for 3D sound simulations.

As human listeners are not able to perceive every sound source with the same degree of accuracy, Chapter 5 will explain how to exploit the properties of human

perception, such as auditory masking and the continuity illusion in order to simplify an auditory scene. Several metrics, generally used to categorize and sort various auditory components according to their perceptual importance, will be assessed in order to optimize the rendering. This chapter will also propose a level of detail approach similar to those used in graphics to render an auditory scene according to the available resources.

The third part will present a new method to automatically author a virtual auditory scene from field recordings, similar to *image-based rendering* in computer graphics. In chapter 6, we will present a practical pipeline to convert real-world multi-track recordings into a high level representation suitable for interactive auralization. By analyzing the recordings in several frequency bands, we are able to obtain the location from which the signals were emitted. We also extract the different components for each source in order to re-render them from their extracted positions. This method offers the possibility of re-editing the scene and move or delete sounds in a specified area. The proposed algorithm also allows us to freely change the listening point and to combine different elements coming from various recordings while ensuring global spatial coherence. The proposed method could also be used in motion-picture audio production. In this context, the visuals are generally acquired from many points of view and the audio engineer has to make recordings consistent with the visuals. The proposed method can re-render the acquired soundtrack from any point of view and direction, thus simplifying this task. Chapter 7 will discuss improvements to this approach. Where the previous method fails to segment correctly diffuse sound components, the improved algorithm separates the audio in foreground and background components under an assumption of stationarity of the background components. Both foreground and background components will be spatialized using dedicated strategies.

1.2 Publications

The body of this thesis is part of the following publications :

Efficient 3D Audio Processing on the GPU. Emmanuel Gallo and Nicolas Tsingos. GP2, ACM Workshop on General Purpose Computing on Graphics Processors, August 2004

Prioritizing Signals for Selective Real-time Audio Processing. Emmanuel Gallo, Guillaume Lemaitre, Nicolas Tsingos. International Conference on Auditory Display (ICAD'05) , July 2005

3D-Audio Matting, Post-editing and Re-rendering from Field Recordings. Emmanuel

Gallo, Nicolas Tsingos, and Guillaume Lemaitre. EURASIP Journal on Applied Signal Processing, 2007 (Special Issue on Spatial Sound and Virtual Acoustics)

Extracting and Re-rendering Structured Auditory Scenes from Field Recordings. Emmanuel Gallo and Nicolas Tsingos. AES 30TH International Conference, 2007

Part I

Background

Chapter 2

Overview of Human Perception

Sound is an auditory sensation caused by the presence of an acoustic traveling wave generated by a vibration and which propagates through a medium. The human auditory apparatus captures the vibration and transmits the stimuli to the brain which informs us about the nature of the sound and its localization. In this context, audio rendering aims at creating the illusion of a real sound scene to the ears of the listener. Studies in human hearing provide the knowledge for recreating all the cues needed by our brain in order to fool the senses of the listener.

In this chapter,we will present some general concepts on human hearing. The anatomy and the physiological mechanism of the auditory system will be described. We will see thereafter how these concepts can be useful to reduce the rendering calculations by processing only the necessary data.

2.1 Human Hearing

2.1.1 Auditory System

The auditory apparatus, shown in Figure 2.1, has three main elements which have different but complementary functions: the outer ear, the middle ear and the inner ear.

Outer ear

The outer ear is the first part of the auditory apparatus and it is the only part visible. It has two components : the pinna and the auditory canal. The pinna receives the sound vibration and transmits it to the middle ear through the auditory canal to vibrate the eardrum (or tympanic membrane). It also protects the middle ear in order to prevent damage to the eardrum. The shape of the pinna is primarily responsible for

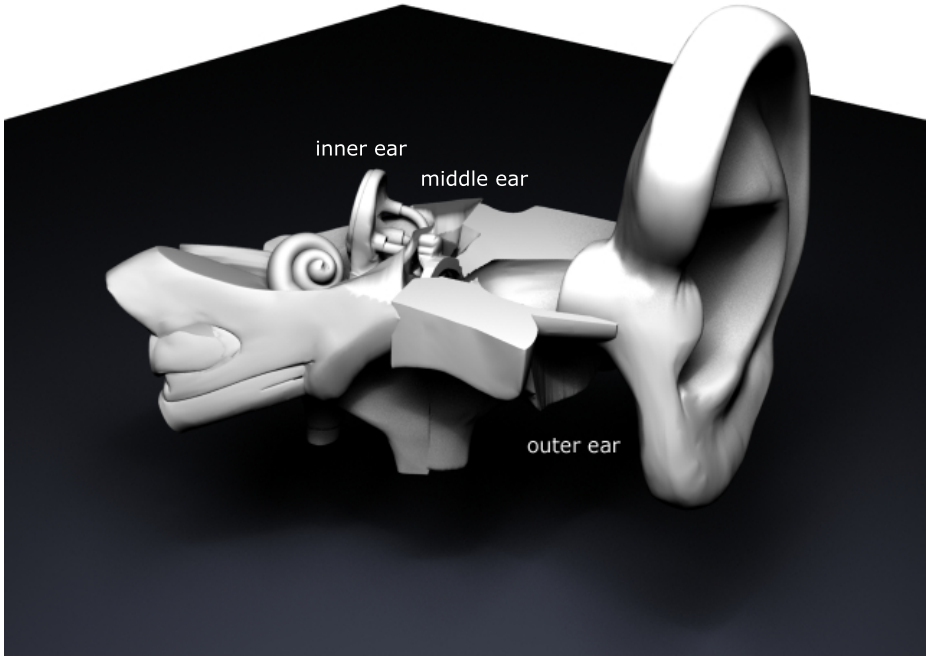

Figure 2.1 3D model of the auditory apparatus divided into three parts : Outer, middle and inner ear.

the sound localization by reflecting and diffracting the sound waves, hence modifying their spectral components according to the incoming direction. The auditory canal acts as a closed tube resonator, enhancing sounds in the range of 2000Hz-5000Hz, which corresponds to human speech (see Figure 2.2). It is 2.5 cm long on average allowing the middle and the inner ear to be located near the brain.

Middle Ear

The second part of the auditory apparatus carries the vibrations toward the inner ear. As the inner ear contains fluids, an adaptation of the impedance is made to avoid large losses of energy. The transmission is achieved by three ossicles (malleus, incus and stirrup) which amplify the energy by means of a lever system. This transfer is most efficient at middle frequencies (500-4000Hz).

Inner Ear

The function of the inner ear is to transform the mechanical energy into bio-electric energy. This part is composed of the cochlea and the vestibular apparatus which is

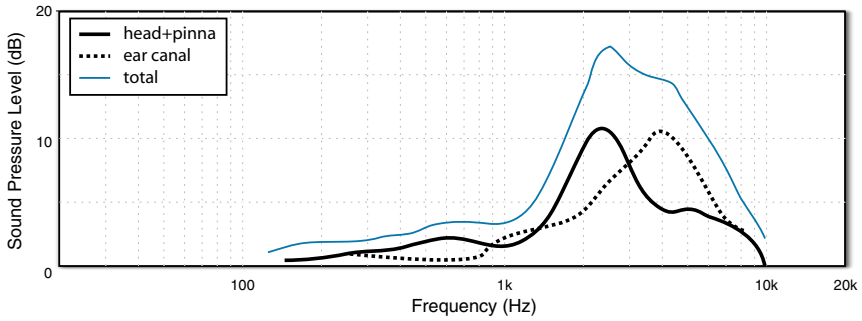

Figure 2.2 Transfer function of the outer ear from a sound at 45 degrees in the horizontal plane, related to the head and pinna and to the ear canal. This figure also show the combined result which emphasizes the frequency range of speech.

used as a balance system but does not intervene for hearing. The cochlea has the shape of a shell, measures 2/3 cm unrolled and is filled with an incompressible fluid called perilymph. It is separated along its length by an osseous blade on which two membranes are attached: the basilar membrane and the tectorial membrane. The incoming variations of pressure cause progressive traveling waves which deform the membranes from the base toward the apex (the extremity of the cochlea). The vibration of the basilar membrane involves the deformation of the organ of Corti, grouping of auditory sensory cells or "hair cells". In response to the pressure variations, the hair cells send electrical impulses to the brain via the auditory nerve.

2.1.2 Auditory Filters

There is a tonotopic mapping of frequencies on the basilar membrane which resonates at different regions along its length according to the frequency contents of the sound. Low frequency sounds produce vibrations across the entire basilar membrane with a maximum amplitude at the apex. High frequency sounds produce vibrations close to the base of the basilar membrane.

Fletcher suggests that the auditory system can be modeled as a bank of band pass filters, each region of the basilar membrane responding to a range of frequencies with maximal amplitude [Fletcher and Munson, 1933]. He named these regions "critical bands" and their bandwidth "critical bandwidth". Each critical band is about 1.3 mm long on the basilar membrane and embraces about 1300 neurons. Later, based on Zwicker's work, the bark scale (in memory of Barkhausen) was introduced where each bark spans one critical band [Zwicker et al., 1957]. Zwicker measured 25 critical

bands and defined a table with corresponding central frequency, lower and upper limit covering the whole audible frequency range. An analytical expression was later introduced [Zwicker and E.Terhardt, 1980] :

$$z = 13 \arctan(\frac{0.76f}{1000}) + 3.5 \arctan(\frac{f}{7500})^2 \tag{2.1}$$

where f is the frequency and z the corresponding Bark index.

More recently, Traunmüller introduced a more accurate expression [Traunmüller, 1990] :

$$z = \frac{26.81f}{1960+f} - 0.53 \left| \begin{array}{ll} z' = z + 0.15(2-z) & z < 2 \\ z' = z + 0.22(z-20.1) & z > 20 \end{array} \right. \tag{2.2}$$

where f is the frequency and z' the corresponding Bark.

This critical band scale is largely adopted as a psychoacoustic model. However, an alternative was introduced by Moore based on different experiments called ERB for Equivalent Rectangular Bandwidth [Moore et al., 1997]:

$$e = 21.4 \log_{10}(\frac{4.37f}{1000} + 1) \tag{2.3}$$

where f is the frequency and e the corresponding ERB.

The measured widths of the corresponding filters were smaller than those obtained by Zwicker (see Figure 2.3). Finally, Patterson introduced the gammatone auditory filter bank [Patterson et al., 1992]. He modeled the human auditory filter using gammatone filters keeping an ERB spacing between each filter.

2.1.3 Auditory Masking

Masking Overview

The mechanism of the basilar membrane produces an interesting psycho-acoustic phenomenon called auditory masking which appears when a sound becomes inaudible due to the presence of another sound. Two types of auditory masking exist : simultaneous (or frequency) masking and temporal masking. Simultaneous masking occurs when two concurrent sounds emit at different frequencies but in the same critical band. One of them (the maskee) can be totally inaudible due to the presence of the other (the masker). If the vibration in regions of the basilar membrane produced by a masker sound is large, the vibration of the maskee sound in the same regions will not be perceived unless the excitation produced by the maskee exceeds that of the

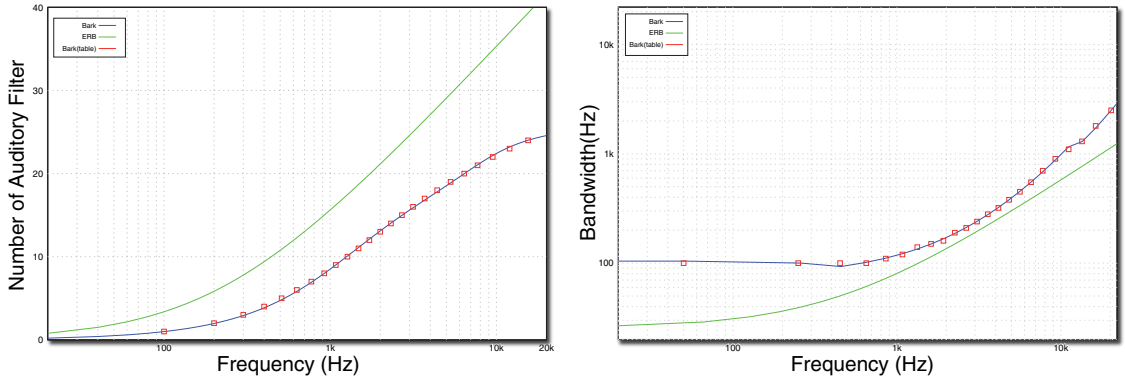

Figure 2.3 Comparison between BARK and ERB scale. The right plot shows the bandwidth as a function of the central frequency. The left plot shows the number of auditory filters upto each frequency value.

masker by a given minimum threshold. This masking threshold depends on the sound pressure level, the frequency and some characteristics of the masker (see Figure 2.4).

Temporal masking occurs when two sounds are close in time. There is a loss of sensitivity around the sound's frequency during a few milliseconds (see Figure 2.5 a). Thus, the ear does not perceive the sounds preceding or immediately following a sound of strong intensity. The premasking is short, about 5ms, contrary to the postmasking which persists longer, depending of the duration of the sound (see Figure 2.5 b).

Masking Threshold Model

The auditory masking concept is widely used in audio compression to calculate the optimal quantization level without introducing audible artefacts [Painter and Spanias, 1997]. Masking thresholds can be computed as a function of the critical band energy and tonality index for each frequency band of the auditory model [Johnston, 1988]. In more details, the critical band spectrum is computed as the sum of the energy in the band convolved by a spreading function to simulate the spreading of vibrations into neighboring bands.

An analytical expression of the spreading function is given by [Painter and Spanias, 1997] :

$$SF_i = 15.81 + 7.5(i + 0.474) - 17.5\sqrt{1 + (i + 0.474)^2} \tag{2.4}$$

where SF_i is the spreading function in dB in function of the critical band index i.

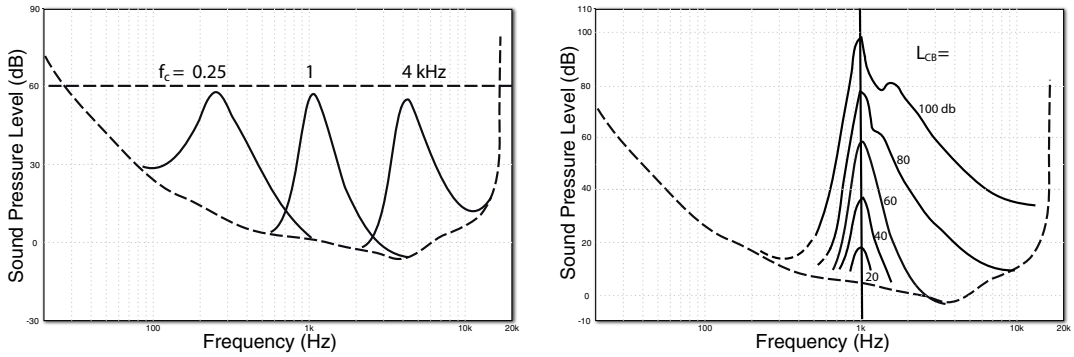

Figure 2.4 Simultaneous masking : (a) Level of test tone just masked by critical-band noise with level of 60 dB, and center frequencies of 0.25, 1 and 4 kHz. The broken curve is the threshold in quiet. (b) Level of test tone just masked by critical-band white noise with center frequency of 1kHz and different levels as a function of the frequency of the test tone [Zwicker and Fastl, 1999].

From the Fourier representation of the signal, the critical band spectrum can be computed by summing the energy in each critical band :

$$B_i = \sum_{\omega=bl_i}^{bh_i} P(\omega) \tag{2.5}$$

where B_i the energy in the critical band i, bl_i is the lower boundary of the critical band i, bh_i is the upper boundary of the critical band i and $P(\omega) = Re(\omega)^2 + Im(\omega)^2$ is the instantaneous power at the frequency index ω.

The spread spectrum can be computed as :

$$C_i = SF_i * B_i \tag{2.6}$$

where C_i is the spread spectrum in dB in critical band i, SF_i is the spreading function in critical band i and B_i the energy in the critical band i.

The tonality index in $[0, 1]$ is an indication of the signal noisiness, low values indicating a noisier component. This index can be computed using the "Spectral Flatness Measure"(SFM), which is the ratio between the geometric mean (Gm) and the arithmetic mean (Am) of the spectral components. In dB, the SFM can be estimated as:

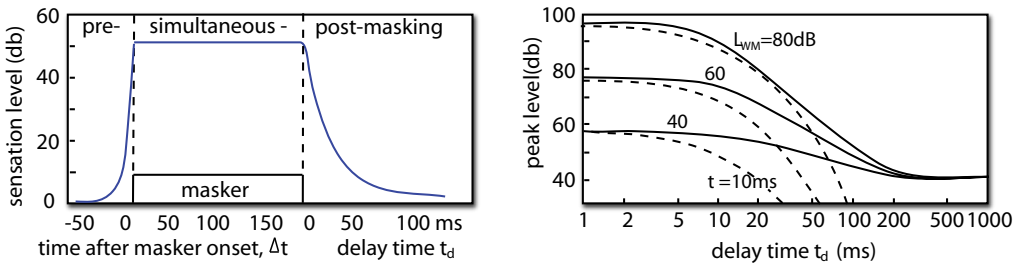

Figure 2.5 Temporal masking : (a) Regions where premasking, simultaneous masking and postmasking occur. (b) Peak level of a 20-μs Gaussian pressure impulse as a function of the delay time after the offset of white noise maskers of a given level [Zwicker and Fastl, 1999].

$$SFM_{dB} = 10log_{10}(\frac{Gm}{Am}) = 10\left[(\frac{1}{N}\sum_{n=1}^{N} log_{10}(P(n))) - (log_{10}(\frac{1}{N}\sum_{n=1}^{N}(P(n))))\right] \quad (2.7)$$

where P is the power spectrum, N is the FFT size and n is the frequency bin.

The tonality index can then be estimated as :

$$\alpha = min(\frac{SFM_{dB}}{SFM_{dBmax}}, 1) \quad (2.8)$$

where SFM_{dB} is spectral flatness measure and $SFM_{dBmax} = -60dB$.

The threshold offset O_i of the masking energy for the band i is given by linear combination of a tonal and noise threshold depending on the tonality :

$$O_i = \alpha(14.5 + i) + (1 - \alpha)5.5 \quad (2.9)$$

where i is the index of the critical band.

Finally, the masking threshold T_i is estimated by subtracting the threshold offset from the energy of the critical band :

$$T_i = 10^{log_{10}(C_i) - \frac{O_i}{10}} \quad (2.10)$$

All frequency components below masking threshold are inaudible.

2.1.4 Illusory Continuity

Illusory continuity belongs to the auditory illusion field which fools our hearing sense. When a tone signal is replaced by a louder noise for a short time, the tone signal is still perceived, although not present in the stimulus, and appears continuous and unbroken. It is generally observed when there is no evidence of discontinuity and the level of the second signal is large enough [Warren et al., 1988]. An interesting case is the illusory continuity of interrupted speech. When some parts of speech sentences are removed and replaced by a louder sound, they appear non-deleted from the signal [Bashford et al., 1988].

2.1.5 Perception of Sound Intensity and Loudness Models

Harvey Fletcher, a pioneer on auditory perception, defined loudness in 1933 as : *"a psychological term used to describe the magnitude of an auditory sensation"* [Fletcher and Munson, 1933]. In this section, after some basics of auditory perception, some models of loudness will be presented. We will see that loudness varies with intensity, but also with the duration and the spectral composition of the signals.

Sound Propagation and Pressure Level

Sound corresponds to a vibration that propagates through a medium in time and space. It travels as waves of alternating pressure with celerity c which is dependent on the propagation medium (air, liquid, ..). The unit of pressure commonly used in acoustics is the Pascal (Pa). To define the sound pressure level L_{db}, we generally express it in decibels of sound pressure level (or dB_{SPL}) :

$$L_{db} = 20log10(\frac{p}{p_0}) \tag{2.11}$$

where p_0 is the reference sound pressure (generally $20\mu Pa$ in air) and p is the effective sound pressure.

Hearing Area

Figure 2.6 shows the hearing area. The audible frequency range spans from 20 Hz to 20 kHz and the dynamic range is about 130 dB between a sound which is just audible (threshold of hearing) and the discomfort level (threshold of pain). This auditory curve is the result of statistical means based on user studies [Zwicker and Fastl, 1999]. Music roughly extends from 50Hz to 10kHz, while speech covers areas from 200 Hz to 5 kHz with a limited dynamic range.

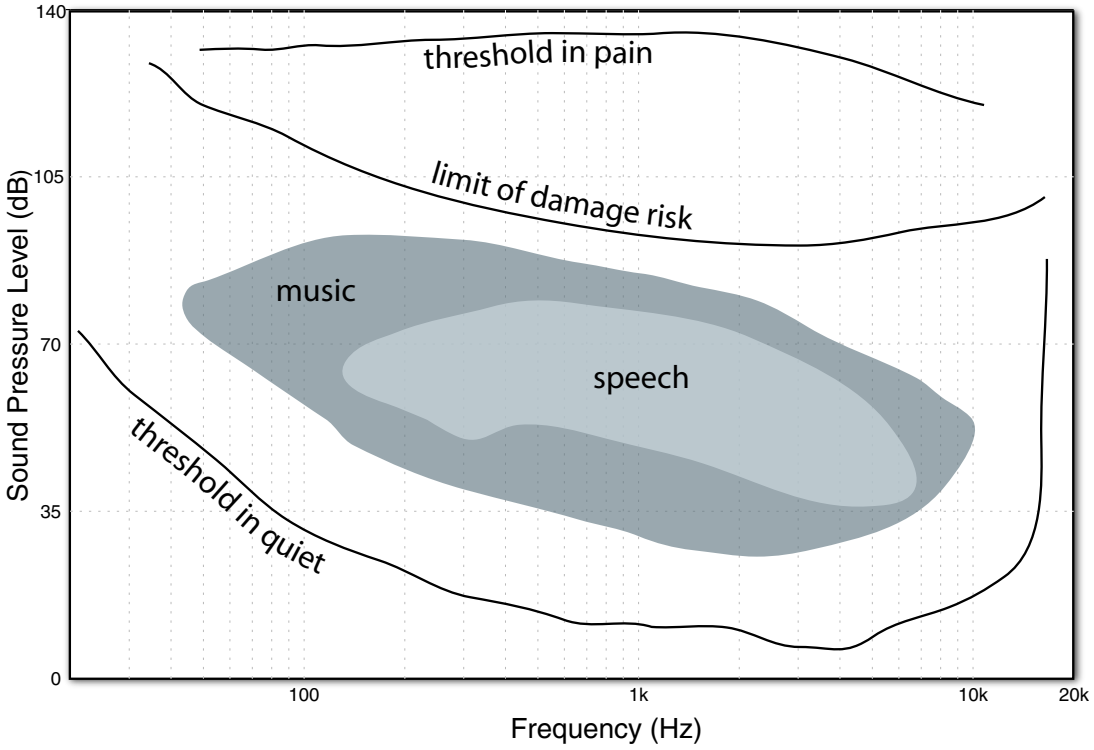

Figure 2.6 Hearing area between threshold in quiet and threshold of pain [Zwicker and Fastl, 1999]. This figure denote regions of speech and music and limit of damage risk.

Threshold of Hearing

The threshold of hearing is the minimum sound amplitude that the average ear with normal hearing can detect in a quiet environment. This threshold varies with the frequency. Two thresholds are generally measured: "Minimum Audible Field" (MAF) and "Minimum Audible Pressure" (MAP).

In the case of the MAP, the pressure is measured at the ear canal whereas in the case of the MAF, the pressure is measured in free field with an artificial head. At low frequencies, the MAF is 5/10dB lower than the MAP measure due to the masking from vascular origin. Additional differences are due to head and pinna filtering (see Figure 2.7).

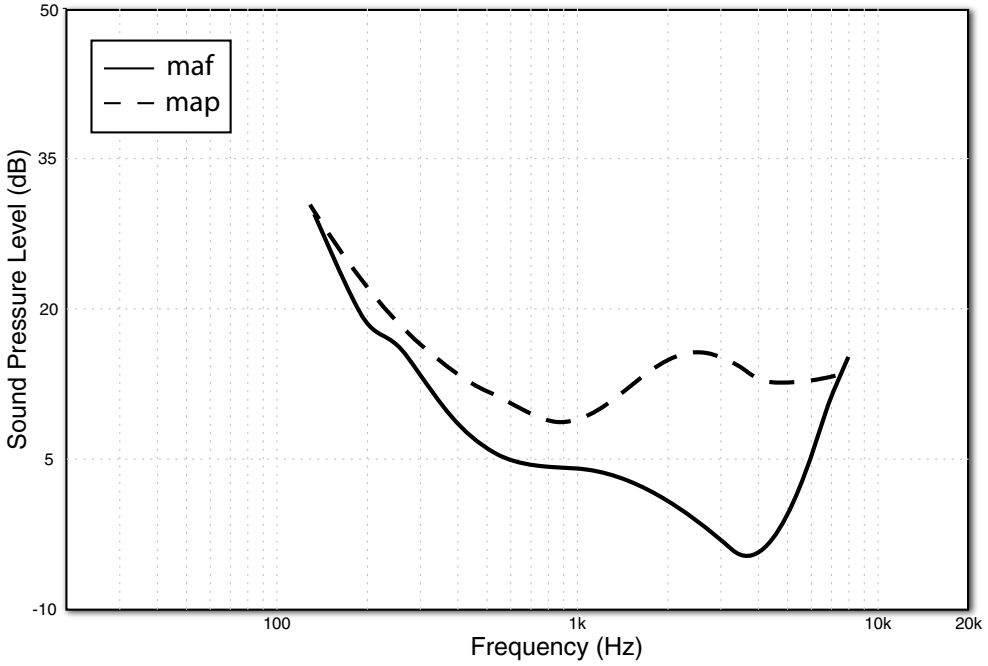

Figure 2.7 Minimum sound level as a function of the frequency [Moore, 1997]. The solid curve show the minimum audible pressure at the eardrum (MAP). The dashed curve show the minimum audible pressure in a free sound field (MAF).

Equal Loudness Contours

The perception of intensity varies with the frequency. For this reason, a subjective scale of loudness level was proposed by Barkhausen: the phon. By definition, at 1 kHz, the phon scale is the same as the dB scale. The equal-loudness-level contours are a measure of sound pressure over the frequency spectrum, for which a constant loudness is perceived. Figure 2.8 show the equal loudness contour as a function of the frequency [Fletcher and Munson, 1933]. This curve reflects the limits of the hearing area. The lowest contour corresponds to the absolute threshold of hearing (see Figure 2.8) whereas the highest is the threshold of pain. The contours are lowest in the medium frequency range indicating that the ear is most sensitive to frequencies in this range, due to the resonance of the auditory canal. The shape of the contour shows that the ear is less sensitive to frequencies which are higher or lower. In 1987,

equal-loudness-level contours were revised and presented in the ISO 226 standard.

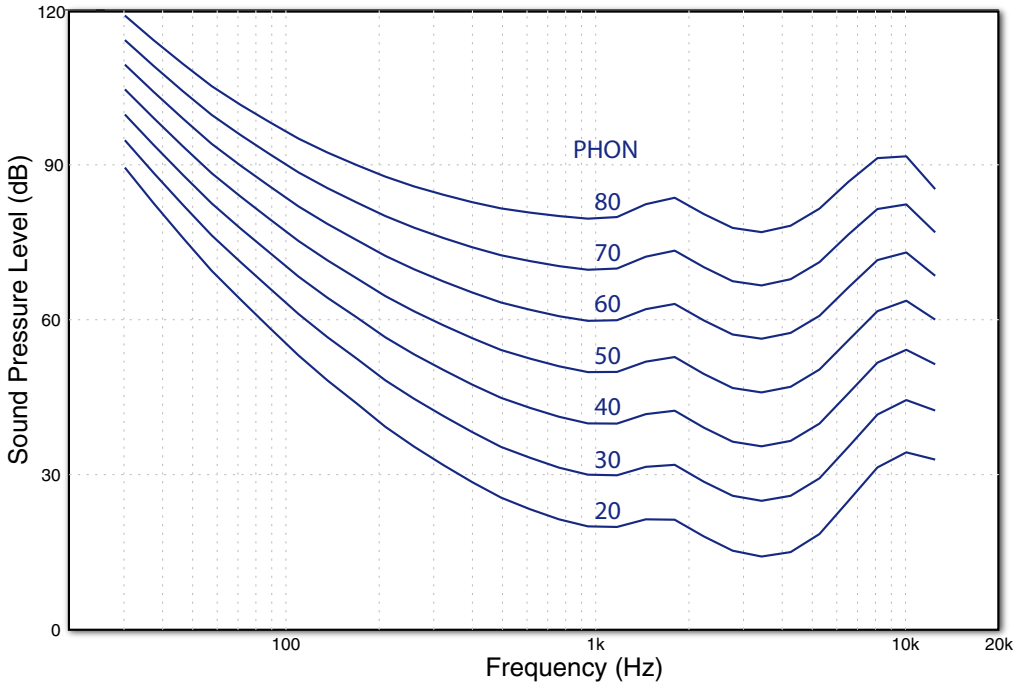

Figure 2.8 Equal loudness contour as define by the ISO 226 for various loudness level (in Phon).

A-Weighting

Based upon his work on loudness measurement, Fletcher [Fletcher and Munson, 1933] proposed a filter to compensate the signal frequency in order to account for human hearing sensitivity. The filter weighted the signal frequencies according to the equal-level-contour and was adopted in a standard for measuring loudness and usually incorporated in sound level meter. However, recent study from Aarts [Aarts, 1992] have showed that it could produce inadequate or even misleading results because it was originally designed for quiet sounds (40 phon).

Sone and the Power Law

In 1936, the sone was introduced by Stevens [Stevens, 1936,Stevens, 1975]. Unlike the phon scale, loudness is expressed on a linear scale. Stevens defined a sone to be equal as 40 phons and to be directly proportional to the perceived loudness (see Figure 2.8). He proposed that loudness perception could be described as a mathematical function, "the power law", defined with the formula :

$$L = kI^a \tag{2.12}$$

where L is the loudness in sones, I the intensity, k a constant depending on the subject and a was defined as 0.3. This corresponds to 10 dB per doubling in loudness.

Stevens Loudness Model (International Standard ISO 532A)

The power law leads to a computational model of loudness calculated from a linear equation. The model first determines a loudness index for each frequency bands according to a predefined table which describes the perceived magnitude as a function of pressure level. The total loudness is then computed using the formula:

$$S_l = S_m + F(\sum S - S_m) \tag{2.13}$$

where S_m is the greatest of the loudness indices, $\sum S$ is the sum of loudnesses indices in all bands and F depends on the bandwidth used in the analysis of the signal.

The total loudness is converted into loudness level by:

$$P = 40 + 10\log_2(S_l) \tag{2.14}$$

Nevertheless, this loudness model is only valid for diffuse field.

Zwicker Loudness Model (International Standard ISO 532B)

In 1960, Zwicker presented a method to compute loudness for stationary sounds which is certainly one of the most popular methods [Zwicker, 1960,Zwicker and Fastl, 1999]. From the signal, a filter is applied to correct the middle and outer ear response. The signal is analyzed through an auditory filter bank (ISO uses a 1/3 octave band pass filter [Zwicker et al., 1991], but other auditory filter banks can be applied). The loudness of each band is computed from the excitation level of the band using the Stevens power law. The total loudness is integrated across each band.

Moore Loudness Model

In 1996, Moore and Glasberg introduced a new method to compute loudness as a revised version of Zwicker's model [Moore and Glasberg, 1996], following the same structure as his method :

1. Fixed Filter for transfer of outer / middle ear
2. Transform spectrum to excitation pattern
3. Transform excitation to specific loudness
4. Calculate area under specific loudness pattern

The Zwicker model was improved in several ways: the Moore loudness model works both for monaural and binaural loudness. If the same sound is presented to both ears, the loudness is doubled compared to single ear presentation. It also takes into account the correct threshold of hearing for monaural or binaural estimations. It used the equal-loudness-contours defined in the ISO 226 and an ERB scale for the auditory filter bank. The excitation level of the band is transformed to loudness using a modified power law, taking into account the background noise.

Time-Varying Loudness Model

Although the previous loudness models provide reasonable estimates for a variety of signals, there are supposed to be applied on stationary sounds, limiting their domain of application. Zwicker proposed an improved version of his loudness model to take into account time-varying signals using temporal masking when computing the excitation [Zwicker, 1977]. Glasberg and Moore improved their model as well [Glasberg and Moore, 2002], introducing a difference between instantaneous loudness, short term loudness and long term loudness. The excitation was predicted using multiple Fast Fourier Transforms with different sizes in parallel to get a finer temporal resolution at high frequency and good spectral resolution at low frequency. Instantaneous loudness was predicted using their earlier model. Short term loudness is computed using a form of temporal integration of the instantaneous loudness as well as the long term loudness using a form of temporal integration of short term loudness. Both loudness models seem to be suitable for time varying sounds.

2.2 Spatial Hearing

2.2.1 Coordinate System

To describe the localization of a sound source, it is easier to reference the location of the sensation with angles and distance instead of using an Euclidean coordinate system. The origin lies at the center of the head of the listener. The azimuth angle is used to localize a sound in the horizontal plane located at ear level. The median-sagittal plane is the plane splitting the body into right and left halves and the frontal plane splits body into front and back halves. The elevation is the angle between the horizontal plane and the source (see Figure 2.9).

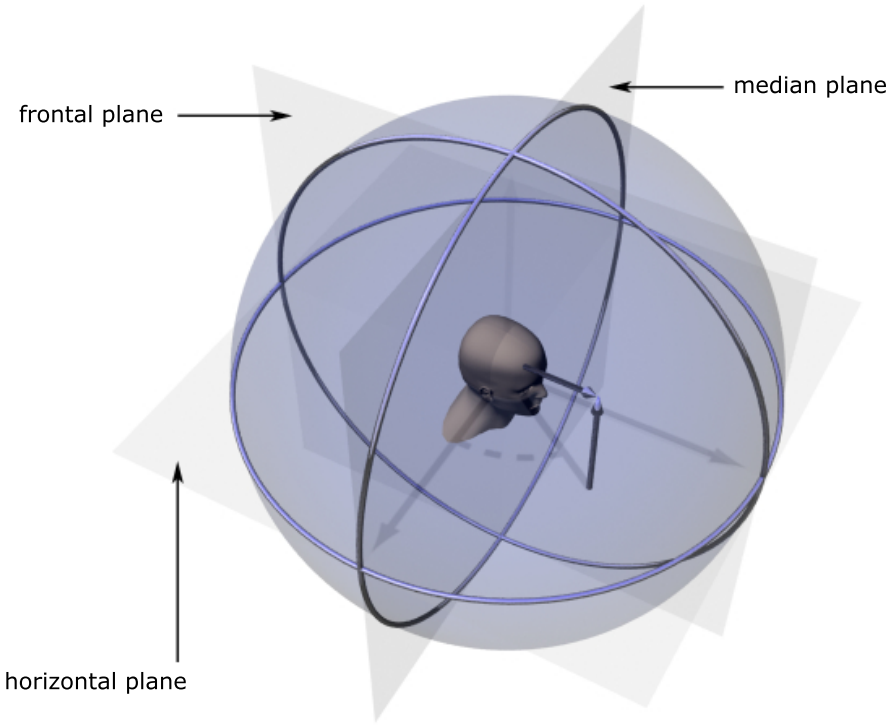

Figure 2.9 Coordinate system generally involved in auditory experiments.

2.2.2 Localisation Cues

Interaural Cues

In the beginning of the century, Lord Rayleigh introduced the "Duplex theory" [Raleigh and Strutt, 1907]. This theory demonstrates that the interaural time difference (ITD) and the interaural intensity difference (IID) are the principal cues for source localization in the horizontal plane. However, these cues are not equivalent according to the frequency components of the incoming sound. Interaural time differences prevail at low frequency, where the IID is small, whereas the interaural intensity difference dominates at high frequencies due to the difficulties of the brain to correlate high frequency sounds. Wightman suggested that the ITD cue is primarily used to localize, while IID and spectral cues are used to resolve possible confusions of location [Wightman and Kistler, 1992].

In order to evaluate sound localization accuracy, the discrimination of difference in azimuth has been studied [Mills, 1958]. Depending on the frequency and the azimuth, the minimum audible angle (MAA) varies. In Figure 2.10, the MAA is plotted as a function of frequency for four reference directions (0°, 30°, 60° and 75°). The MAA is smallest for a sound coming in front. Around 1.5 kHz-1.8 kHz, the accuracy is very low and consistent with the "Duplex theory". The localization with the ITD works up to 1.5 kHz, whereas the IID are small up to 1.8kHz. The MAA at an azimuth of 90 deg was always more than 40 deg. This is due to the cone of confusion.

This cone is centered on the interaural axis (see Figure 2.11). In space, each source location belonging to the cone of confusion has the same ITD and approximately the same IID. Therefore, localization in this area is ambiguous, but generally resolved using head movements [Young, 1931, Wallach, 1940].

Spectral Cues

The "Duplex theory" model is validated for the localization of sources on the horizontal plane, but fails for the localization in elevation, especially on the median sagittal plane, where the ITD and IID are identical. Head movements help to resolve ambiguous localization. Yet, short bursts can be localized in elevation, even if their duration is too short to allow movements of the head. Many experiments demonstrate that the form of the head, pinna and torso is required for the localization [Batteau, 1967,Roffler and Butler, 1968, Gardner and Gardner, 1973]. Indeed, a complex filtering is realized by our physionomy and modifies the spectral components of the sound according to its direction of incidence [Wiener and Ross, 1946, Shaw, 1966]. The corresponding pair of filters for both ears is called "Head related transfer functions" (HRTF). The HRTF is dependent on each person and using non-individualized HRTF can result in poor localization accuracy with front-back and up-down confusions [Wightman and

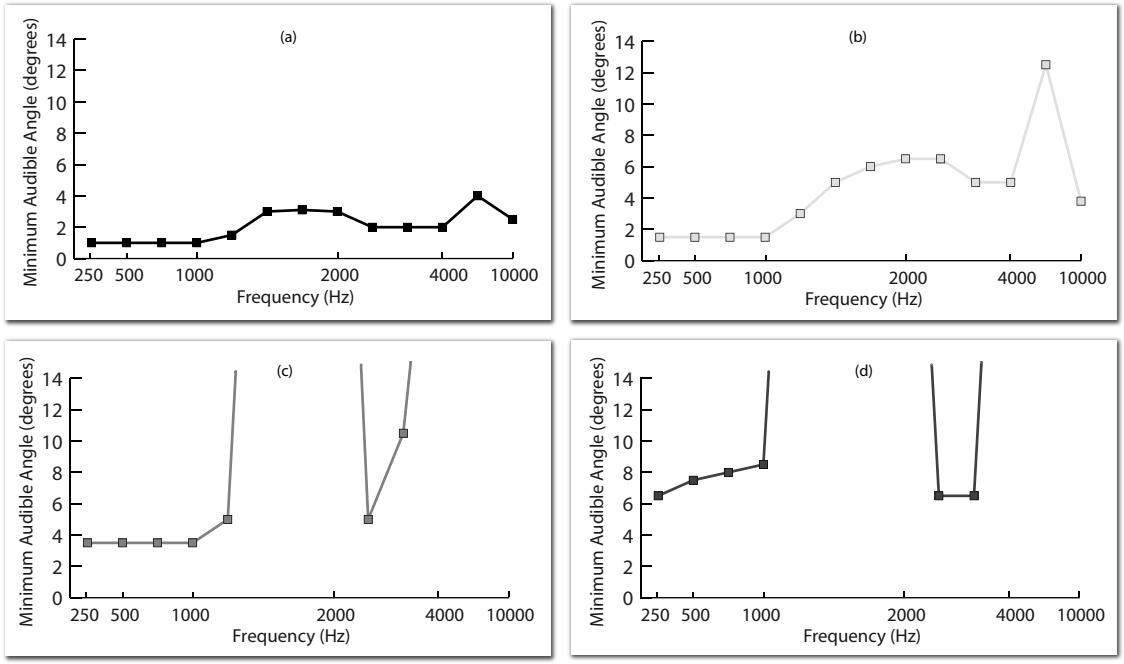

Figure 2.10 The minimum audible angle (MAA) as a function of the frequency for four reference directions (a) 0°, (b) 30°, (c) 60° and (d) 75° [Mills, 1958].

Kistler, 1989, Wenzel et al., 1993]. Whereas the azimuth localization is robust using interaural cues, the spectral cues help to determine the elevation and disambiguate the front-back localization and work best for complex stimuli [Roffler and Butler, 1968]. The monaural localization paradigm has been investigated to emphasize the contribution of the spectral cues to the duplex theory. One problem is that complete monauralisation of a listener is difficult to realize and achieving such studies is quite complex because very low sound levels provide access to interaural cues [Wightman and Kistler, 1997].

Cues for distance

Although the localization of a sound source is done using both interaural cues and spectral cues, others cues are employed for distance perception. The principal cue for distance perception is based on the intensity [Gamble, 1909, Coleman, 1963] which gives a notion of absolute distance to familiar sounds. In the case of unfamiliar sounds, the intensity cues only carry information about relative distance. In the presence of multiple sound sources, the relative distance is improved and a difference of distance

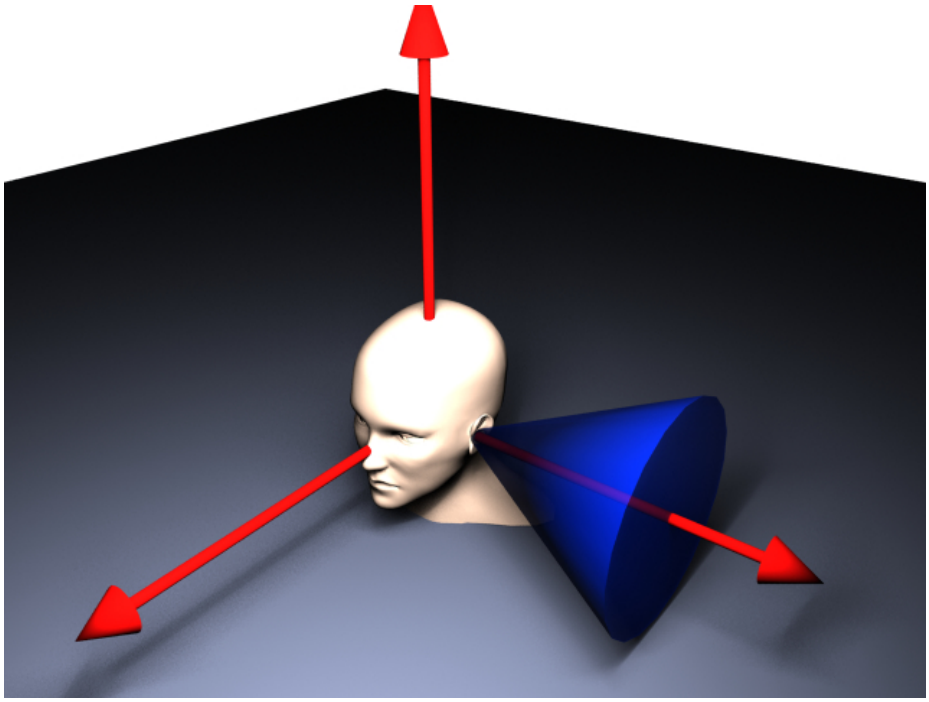

Figure 2.11 The cone of confusion : Different spatial locations produced null ITD and ILD in this area.

of 6% is sufficient to be discriminated [Ashmead et al., 1990]. But the intensity cues are not the only important information for auditory distance perception. The reverberation also influences perceived distance. The sound arrives at the listener from a direct path, but the environment alters the sound propagation by scattering, absorbtion or reflection on the various obstacles. Sound sources in reverberant environments are judged more distant. However, the ratio of direct sound to reverberant sound provides another major cue for distance perception [Mershon and King, 1975].

2.2.3 Precedence Effect

In complex environments, sound is reflected and arrives from multiple directions. But, even in this case, the localization of sound can be performed correctly. Wallach investigated this effect and showed that when two sounds arrive with a short delay, they are perceived fused into a single auditory image [Wallach et al., 1949]. The location of this image correspond to the first wavefront that reaches the listener. This "precedence effect" occurs when the delay between the two sounds is in the range of 1 to 5ms. Below 1ms, the localization is perceived as the means of the source locations and above 5ms, both sources become audible. However, the precedence

effect disappears if the second sound is louder than the first one (10-15dB louder) or if the sounds are not similar.

Chapter 3

Overview of Rendering Techniques

In this chapter, we will present some possible representations of an audio signal. From a continuous signal, we will sample it in discrete time to be able to process the data on a computer. Then, we will show that, according to the chosen representation, information can be coded in a more sparse manner. This representation is important because it directly correlates with the amount of data to process. The second section of this chapter covers techniques used to simulate a virtual auditory scene. We will review common acquisition and resynthesis models, their principles, their performance and their limits.

3.1 Representation of Sound

This section will describe some possible representations of an audio signal, from time-domain waveform to sparse signal representation. Some signal representations are more suitable for auditory perception. The representation also influences the manner of analyzing, transmitting, storing, querying or displaying the data. In the last part of this thesis, we will aim to obtain a higher level representation of an auditory scene. This representation must lead to an efficient way to handling the data, to separating various sources and to providing a good reconstruction of the signal without any artifact.

3.1.1 From Analog to Digital

The typical representation of sound is generally the amplitude of the pressure level as a function of time. In order to analyze and manipulate a continuous signal on a computer, it has to be sampled at discrete instants in time spaced by a duration τ called sampling period, whose inverse is the sampling frequency or sampling rate.

From a mathematical point of view, the discretization of the signal $x(t)$ can be done by applying a Dirac comb with unit amplitude and sampling period τ.

$$x(n) = \int_{-\infty}^{\infty} x(t)\delta(t - n\tau)dt \tag{3.1}$$

Fourier Transform

The Fourier transform was introduced by a French mathematician, Fourier, in 1822 as a tool to study heat conduction. He demonstrated that an arbitrary periodic function can be decomposed into an infinite weighted sum of sinusoidal functions. The Fourier transform is a powerful tool used to explore the frequency content of the signal. The Fourier transform is given by the formula:

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt \tag{3.2}$$

And the inverse transform :

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega)e^{j\omega t}d\omega \tag{3.3}$$

Since its development, its capability to analyze and manipulate the frequency domain of a signal, and also its fast implementation have lead to its widespread use in all scientific domains.

Cross Correlation

A useful application of the Fourier transform is to measure the similarity between two signals. The cross correlation is given by a sliding dot product over the two signals x_1 and x_2.

$$R_{x_1x_2}(\tau) = \int_{-\infty}^{\infty} x_1(t)x_2(t - \tau)dt \tag{3.4}$$

where τ is the lag and $R_{x_1x_2}$ is maximum when the signal are identical.

A special case is the cross-correlation of a signal with itself, called auto-correlation, and can be useful, for instance, for finding repeating patterns in a signal. In order to compute the cross correlation efficiently, the Fourier transform can be used, giving the cross spectral density (or cross spectrum).

$$R_{x_1x_2}(\tau) = \int_{-\infty}^{\infty} X_1(\omega)X_2^*(\omega)e^{j\omega\tau}d\omega \tag{3.5}$$

3.1.2 Data Sparseness

Data sparseness is the percentage of empty cells in signal data. A sparse dataset has most components equal to zero. Therefore, the low number of non-zero coefficients provide important signal characteristics. As a consequence, in the case of linear combination of signals, the probability that a coefficient belongs to a single signal is high. For instance, when mixing two sparse signals, the probability that one component of the two signals is zero is high. Since counting the number of zeros of the signal is not practical, the degree of sparseness is measured using the sum of a cost function. This function should be non-convex such as $f(x) = \beta \log(1 + (x/\sigma)^2)$ or $f(x) = \beta \left|x/\sigma\right|^q, q < 1$ where β and σ represent some constants controling the shape of the function [Rickard, 2006]. Data sparseness can also be visually exposed using a scatterplot. A scatterplot is generally used in statistics to determine the association between two variables. The first variable is located along the horizontal axis and the second variable is located along the vertical axis. Figure 3.1 show the scatterplot of two linear mixtures of signals in time and frequency domains. In time domain, each point corresponds to a couple of sample from each mixture. In the frequency domain, each point corresponds to a couple of real complex from each mixture. We can clearly see lines and the direction of each line correspond to the mixing coefficients used for the mixture [Bofill and Zibulevsky, 2001]. Thus, sparse signals are more appropriate for source separation. Signal processing on sparse signals is also more efficient due to the lower number of significant coefficients to compute.

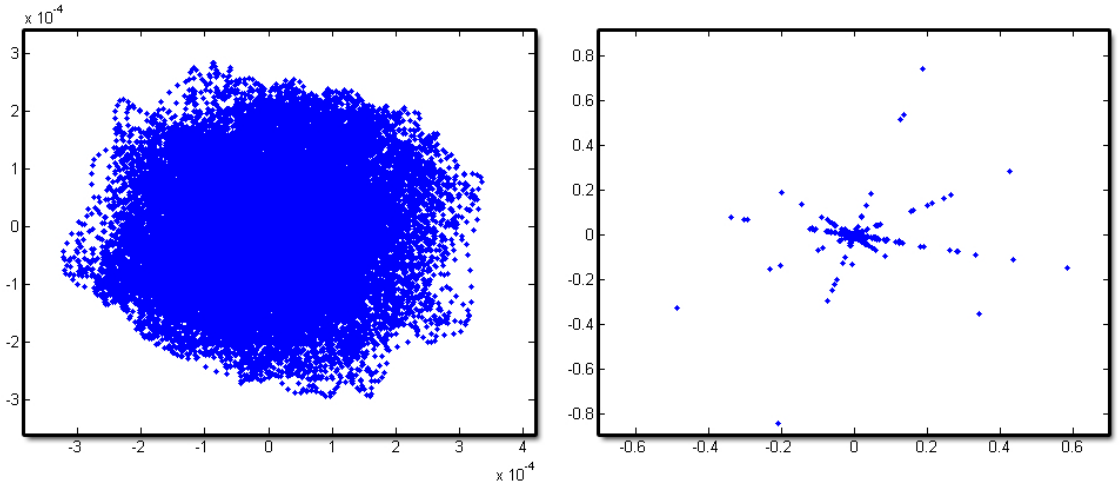

Figure 3.1 Data sparseness : (a) Scatterplot of two linear mixtures in the time domain. (b) Scatterplot of two linear mixtures in the frequency domain. Note : In the frequency domain, lines appear and their directions correspond to the mixing coefficients used for the mixture.

3.1.3 Sparse Representation

A sparse representation of a signal is usually constructed using a decomposition into a basis of elementary functions. The Fourier transform uses sinusoidal basis functions, the wavelet decomposition can use several, like Haar or Daubechies basis functions. However, the transforms do not provide a sparse representation for every kind of signal since the base does not take the signal into account. For instance, the Fourier transform is not adequate to represent the Dirac function. Mallat proposes a representation of the signal with an overcomplete basis named "Matching pursuit" [Mallat and Zhang, 1993]. This algorithm searches into a dictionary of redundant functions which provides the best match. The element of the dictionary is then subtracted out and the algorithm continues iteratively until the residual error is small. Later, Chen improved the signal decomposition to minimize the $L1$ norm of the coefficients [Chen et al., 1998]. However, functions in the dictionary are predetermined and do not take the signal into consideration.

The problem can be resolved with a probabilistic approach to ensure that coefficients are sparse and to minimize the reconstruction error. Finding a sparse representation can be formulated as :

$$x = As + \epsilon, \tag{3.6}$$

where x is the signal to decompose, A the matrix of basis functions (the dictionary) s the vector of coefficients and ϵ is the residual error with s the sparsest possible.

The decomposition is not unique. However, this could be resolved using a Bayesian approach [Olshausen and Field, 1996, Lewicki and Sejnowski, 2000] :

$$\hat{s} = \arg\max_s P(s|x, A) = \arg\max_s P(x|A, s)P(s), \tag{3.7}$$

where $\hat{s}$ is the optimal representation of x and $P(s|x, A)$ the posterior distribution.

Assuming that ϵ is Gaussian noise, the data likelihood is given by :

$$logP(x|A, s) \propto -\frac{1}{2\sigma^2}(x - As)^2, \tag{3.8}$$

where σ is the noise variance.

s is supposed to be statistically independent and sparse :

$$P(s) = \prod_i P(a_i), \tag{3.9}$$

$$P(a_i) = e^{-Q(a_i)}, \tag{3.10}$$

where Q is a non convex function to ensure sparseness of the coefficient.

Figure 3.2 provides a comparison between a sparse representation and a spectrogram of a voice sample. The dictionary used for the sparse representation is constructed with a 128 ERB-spaced gammatone filter.

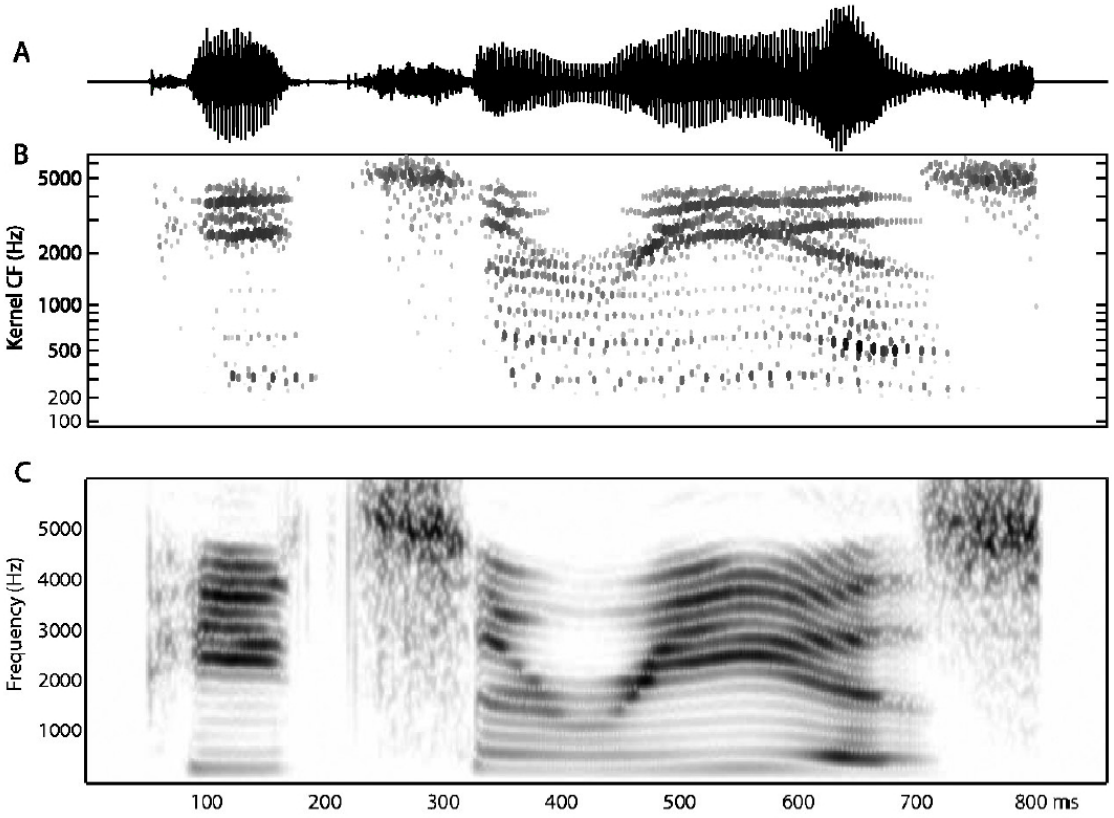

Figure 3.2 Sparse code representation [Smith and Lewicki, 2005]. (a) Signal in time domain, (b) Sparse representation using a gammatone based dictionary, (c) spectrogram of the signal.

3.2 Virtual Auditory Space Simulation

This section will review the auralization procedure. "Auralization" is a term defining the process of simulating a virtual auditory scene. Thus, this section will present the basic primitives used during rendering and effects generally involved in the simulation.

3.2.1 Auditory Scene Modeling

Point Sound Source

Point sound sources are traditionally used to model a virtual auditory scene. These sound sources represent objects localized in space that emit sound. In order to localize the sound source, localization cues must be simulated [Chowning, 1971]. Typically, interaural cues are employed (see 2.2.2). Since the source or the listener can move, the wavelength at the receiver differs from the emission. This phenomenon is called the "Doppler effect" and can be simulated using resampling. Different source characteristics can modify sound emission. For example, a sound source can be omni-directional or have a more complex emission pattern [Savioja et al., 1999]. Point sound source do not physically exist in reality. This model assumes that the region that emits the sound is small enough or is far enough from the listener to be represented as a point source.

Signal of the sound source can either be recorded or fully synthesized and should preferably be anechoic if additional effects have to be applied. For natural signals, the sound must be recorded in quiet environment with low reverberation and pre-equalized to fit the virtual emitting object. The signal is generally not compressed to allow manipulating it efficiently, or a hardware based decompression is used in some new game-console platforms. Pure synthetic sounds are generally created by physical modeling. They can be generated on the fly or pre-computed depending on memory or CPU requirements. Sound synthesis is well suited to virtual reality as a variety of sounds can be modeled and rendered depending on the parameters obtained from the simulation [Doel, 1998, Roads, 1996]. For instance, it is well suited to simulation of contact sounds [Knott et al., 2003] or other physical phenomena [Dobashi et al., 2003].

Complex sound sources

Complex sound sources are generally modeled as collections of multiple elementary sound sources which have to be simulated separately to represent the sound of a virtual object in the scene. For instance, to model the sound of a car correctly, it is necessary to simulate with different sound sources the tire noise, the resonances of the muffler and the engine harmonics as they do not produce the same sound, same position and the same directivity pattern.

Diffuse sound area is generally simulated by sampling the region with point sound source. The signals of each sound source have to be similar, but must be uncorrelated to avoid phasiness effects generated by the small distance between the sources. For better performance, previous papers [Sibbald, 2001, Tsingos et al., 2004] compute a dynamic level of detail of the group of sources according to the distance of the listener

(See 3.3.4).

3.2.2 Indirect Sound Scattering and Reverberation

Reverberation denotes sound coming indirectly to the listener. The environment of the scene disturbs the propagation wave through various phenomena such as reflection, refraction, diffraction or dissipation. These phenomena are often modeled using an impulse response representing the delay and amplitude of sounds arriving along different propagation paths. An impulse response is composed of the direct sound, early reflections and late reverberation. The early reflections depend on the geometry of the room and the position of source and listener. They reach the listener during a period, generally established, of up to 80ms [Roads, 1996]. Late reverberation is due to the volume of the room and is mostly independent of the position of source and receiver (see Figure 3.3)

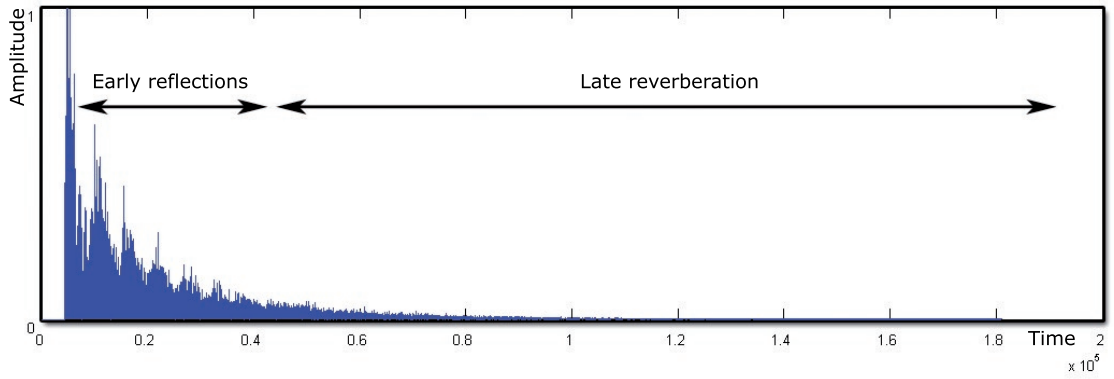

Figure 3.3 Example of room impulse response composed of the direct sound, early reflections and late reverberation.

To create a reverberant effect, the impulse response can be measured from a real room. Then, this filter is applied to the signal by convolution. The computational time is generally proportional to the length of the filter. The impulse response of a virtual room can be also calculated using geometrical acoustics algorithms given a scene description and all properties of the materials. Finally, it can be artificially created to fit a characteristic model of the reverberation. In this case, reverberation can be applied using fast recursive algorithms.

Measures of Room Reverberation

The general method to measure an impulse response is to generate a signal with high energy (e.g. a chirp), record it at the same time and determine the impulse

response by deconvolving the recorded signal with the emitted one. In this case, an average of successive measures is realized to improve the signal-to-noise ratio of the impulse response measure. Schroeder introduced a method using maximal length noise sequences as excitation signals [Schroeder, 1979]. Such sequences are periodic and their autocorrelation is a Dirac function. Thus, the correlation of the signal emitted with that recorded directly gives the impulse response, improving the signal to noise ratio of the impulse response. Moreover, Borish accelerates the algorithm using a transform of Hadamard [Borish and Angell, 1983]. A comparable method has also been introduced by Foster using Golay codes [Foster, 1986].

Geometrical Techniques for Reverberation

Geometrical techniques allow for computing various propagation paths by modeling the wave as a set of rectilinear rays. The reverberation is created by the contribution of all the paths to the listener through an impulse response representing the delay and amplitude of sounds arriving along different propagation paths.

The first geometric approach used ray tracing [Schroeder, 1962, Krokstad et al., 1968, Schroeder, 1970]. Each sound source of the scene generates a large number of sound rays in every direction. Rays travel until they reach the observer or can hit an element of scene. In this case, the material absorbs some energy of the sound before re-emiting it by reflection or diffraction. The listener is generally modeled as a sphere to limit aliasing and improve computational efficiency. The image-source technique is another popular technique to compute specular reflections [Allen and Berkley, 1979, Borish, 1984]. On the assumption that only specular reflections are considered, each path can be represented by a virtual sound source. Indeed, if the sound wave is reflected by a wall, it is possible to create a virtual source by mirroring the source, thus reducing the problem to direct sound sources (see Figure 3.4).

Geometric approaches fail at low frequencies when room resonance and diffraction are more prevalent. However, some extensions have been proposed to model edge diffraction based on the geometrical theory of diffraction and beamtracing [Tsingos et al., 2001]. Room acoustics modeling using a digital waveguide mesh was introduced by Savioja and provided another approach to room response simulation [Savioja et al., 1994]. The method is a time domain finite difference method. The algorithm solves the sound pressure propagation through a waveguide mesh for low frequencies and combines with a ray tracing technique for high frequencies. At each point of the grid, the sound pressure is resolved according to its neighborhood and to the last computed value.

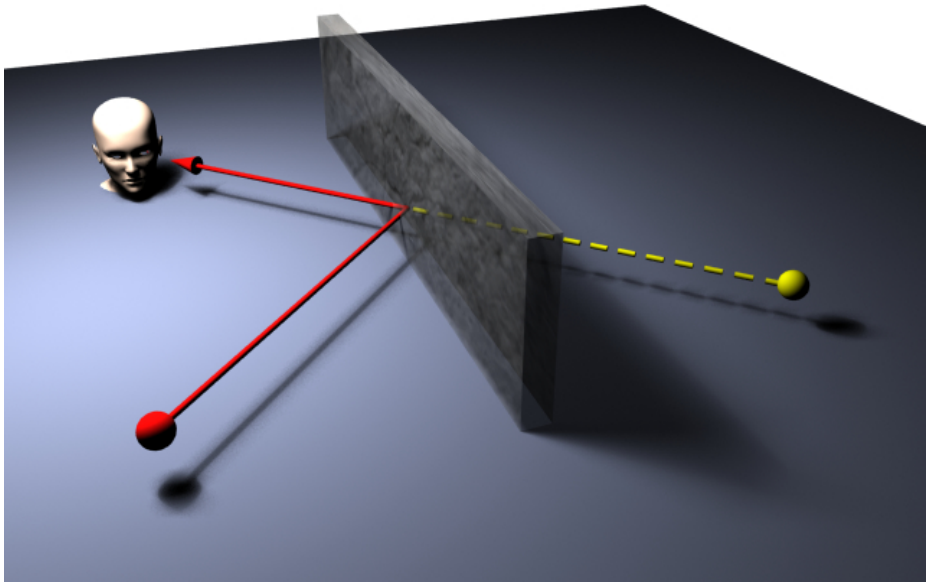

Figure 3.4 The reflection of the sound source (red sphere) solved with a source-image technique. A virtual source (yellow sphere) is created reducing the problem to that of a direct sound source.

Artificial Reverberation

The previous reverberation techniques are computationally expensive. Moreover, the listener might not perceive the full complexity of the reverberation. Besides, for some applications, a higher level control might be preferable to physical modeling. This can justify the usage of an artificial reverberation approach. An artificial reverberation is not based on the geometrical properties of the room, but rather it attempts to approximate the reverberation by adjusting the parameters perceptually.

The first artificial reverberation model was introduced by Schroeder using comb and all-pass filters to reduce the coloration of the sound [Schroeder, 1962] and extended by Moorer [Moorer, 1979]. Gerzon generalized the Schroeder all-pass to N dimensions using a unitary matrix [Gerzon, 1976]. He showed that if the matrix is unitary, then the energy is preserved. Stautner and Puckette proposed a general reverberation model called "feedback delay networks" [Stautner and Puckette, 1982] based on delay lines and a feedback matrix connecting the output to the input (see Figure 3.5). Jot proposed a practical reverberator design based on perceptual criteria resulting in an efficient reverberation technique [Jot, 1997].

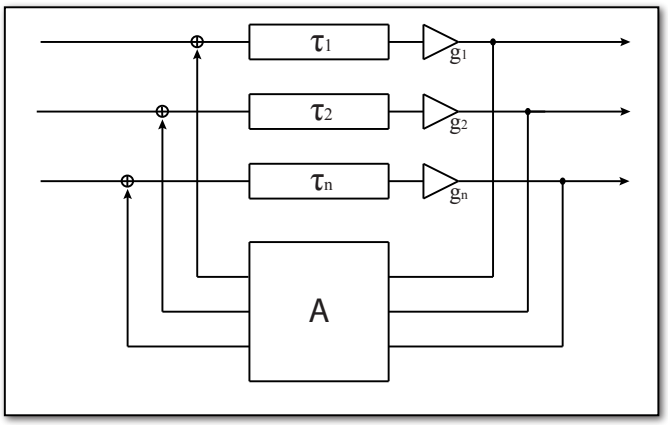

Figure 3.5 Feedback delay networks topology. A is the feedback matrix, τ_i are the delays and g_i are the gains.

3.3 Spatial Audio Reproduction

Once the simulation is completed, the computed sound field must be reproduced to the ears of the listener. The spatialization algorithm depends on the devices, such as loudspeaker or headphone, used for the reproduction. The challenge of sound spatialization is to provide a correct 3D image of the sound position.

3.3.1 Stereophonic Techniques and Multichannel Extensions

The simplest technique to spatially reproduce a sound is based on intensity panning [Blumlein, 1931, Bauer, 1961]. From the signal emanating by two equidistant loudspeakers and differing only by magnitude or sign, a phantom source can be perceived at the angular localization θ_I given by the stereophonic law of sines formulated in phasor form :

$$\frac{sin(\theta_I)}{sin(\theta_A)} = (S_l - S_r)/(S_l + S_r). \tag{3.11}$$

where θ_A is the half angle between the two speakers, and S_l, S_r is respectively the signal sent to the left and right speaker (see Figure 3.6).

In 1997, Pulkki introduced the "Vector Base Amplitude Panning" (VBAP) techniques that generalized the method of the intensity panning for N speakers [Pulkki, 1997]. This method requires all speakers to be nearly equidistant to the listener. The algorithm selects the three speakers nearest to the virtual source and computes the

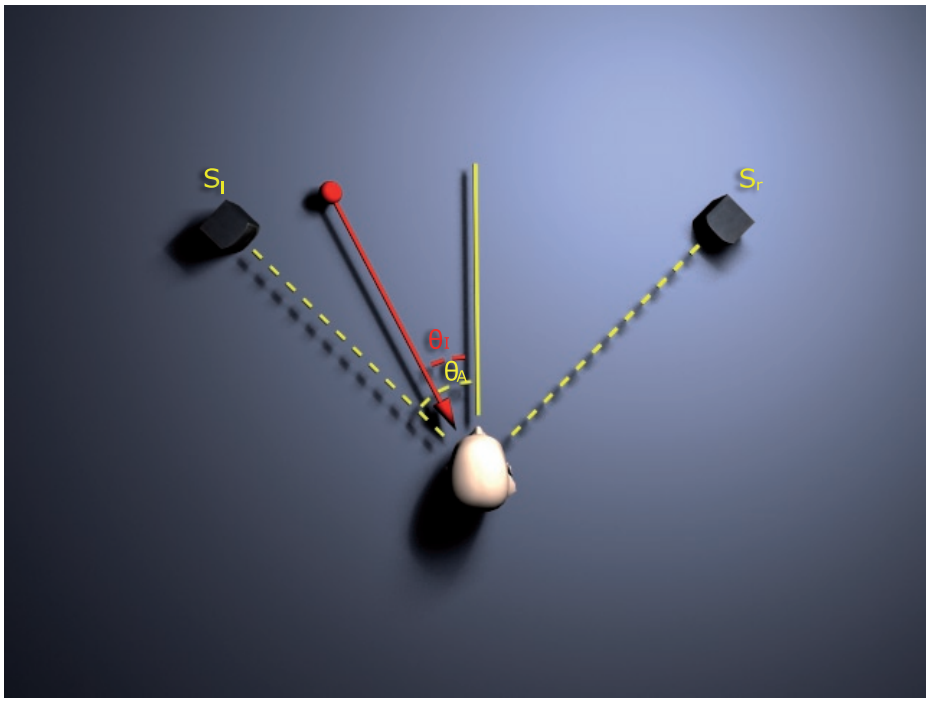

Figure 3.6 Stereophonic law of sines. A phantom source (red sphere) is created using an intensity panning between the two speakers.

gain coefficients for each speaker, solving the equation:

$$p = g_1 l_1 + g_2 l_2 + g_3 l_3 \tag{3.12}$$

where g_i is the gain to be applied on the speaker i, l_i is the vector from the listener to the speaker i and p is the vector from the listener to the virtual source (see Figure 3.7).

Virtual sound can also be reproduced with speakers using a 5.1 surround technique. This technique has been introduced in the cinematographic field and has been primarily designed to be used with a large screen. A specific speaker configuration is required : Two front speakers for the stereo, one front speaker for the dialog, two rear speakers for the ambiance and one subwoofer for enhancing the low frequencies. Many audio coding schemes such as Dolby Digital (AC3) are dedicated to this setup [Steinke, 1996].

A general drawback of these stereophony inspired techniques is to limit the listener to be positioned only in an area called "sweet spot" where the reproduction is valid.

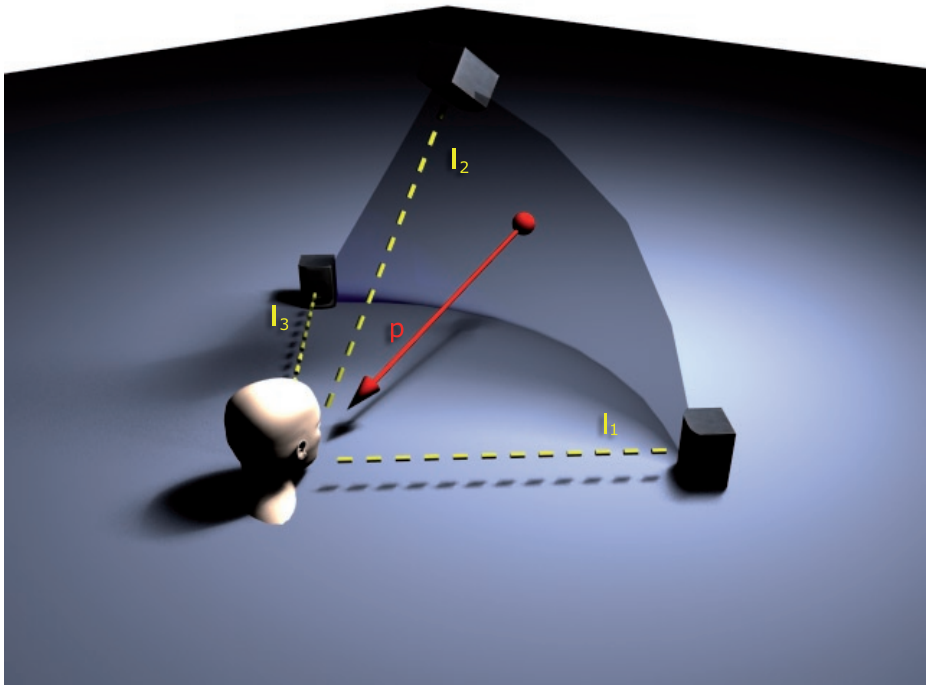

Figure 3.7 Vector Base Amplitude Panning. A phantom source (red sphere) is created using N speakers (N=3 in this example).

3.3.2 Binaural Rendering Techniques

Headphones are a common device employed for sound simulation. They takes advantage of the fixed position of the listener's ears relative to the speaker and deliver accurate sound.

The binaural rendering method is simple, requiring only a single filtering operation. It provides the spectral cues for localization which are essential for the perception of the elevation and rear hemisphere sound reproduction. (See 2.2.2).

The filter is defined individually from the characteristic of the head, torso and pinna and is called the Head Related Transfer Function. The common technique used to measure HRTFs is to place small microphone into the ears of the listener and to record a the associated impulse response from a large number of directions. The sound stimulus used in the acquisition process can be a chirp, a sine sweep or pseudo random noise (MLS). The filter is created by comparing the recorded signal with the emitted stimulus [Wightman and Kistler, 1989]. Another possible measurement option is to simulate the filter using Boundary Element Methods (BEM) simulation [Katz, 2001]. This method resolves the wave propagation equation with a 3D model of a head and deduces the corresponding HRTF filter. This technique is more reliable at low

frequency where the measurements yield inaccurate. A major problem is the variation between different individuals, which requires that HRTFs be measured individually, especially for children who do not have the same morphology as adults. Some works tend to provide appropriate HRTF by asking listeners to select from a catalogue of HRTFs [Seeber and Fastl, 2003].

Another drawback of using headphones for audio simulation is the sensation that the sounds move with the listener position. Head tracking provides a solution and is essential for virtual reality applications. Other benefits of head tracking during the simulation is that it resolves front/back ambiguities using head movements (see 2.2.2) and improves sound localization performance [Begault et al., 2001].

It is possible to listen to binaural recordings with loudspeakers. However, using loudspeakers for the reproduction setup requires another step due to the fact that the speakers are distant from the ears and the sound emitted from one speaker arrives to both ears. To listen to a binaural recording through speakers, a filter is applied to remove the crosstalk signals and the filtering from the head to the speaker. The technique is known as Transaural synthesis [Schroeder, 1975,Cooper and Bauck, 1989, Jot et al., 1995]. When emitting signals X_r and X_l from speakers, the binaural signals Y_r,Y_l perceived from the listener are convolved by a filter H (see Figure 3.8):

$$\begin{bmatrix} Y_l \\ Y_r \end{bmatrix} = \begin{bmatrix} H_{ll} & H_{rl} \\ H_{lr} & H_{rr} \end{bmatrix} \cdot \begin{bmatrix} X_l \\ X_r \end{bmatrix} \tag{3.13}$$

In order to render the correct binaural signal, the inverse matrix H^{-1} must be applied to the emitted signal X such that :

$$\begin{bmatrix} X_l \\ X_r \end{bmatrix} = \frac{\begin{bmatrix} H_{rr} & -H_{rl} \\ -H_{lr} & H_{ll} \end{bmatrix}}{(H_{ll} \cdot H_{rr}) - (H_{lr} \cdot H_{rl})} \cdot \begin{bmatrix} Y_l \\ Y_r \end{bmatrix} \tag{3.14}$$

3.3.3 Physically Based Rendering Techniques

The method presented in the previous section simulates auditory scenes based on collections of point sound sources. This method provides a free view point rendering of the scene interactively and is well adapted for 3D virtual reality applications and games. However, the rendering might not sound realistic due to approximations of the physical model. Alternatively, the surrounding environment can be recorded from a microphone, and spatially rendered using various techniques.

The oldest and simplest is binaural recording. With just two microphones placed inside the ears, the recording captures the filtering of the head (see 3.3.2) and constitutes a reference for the perceived sound at the recording location. The recording can

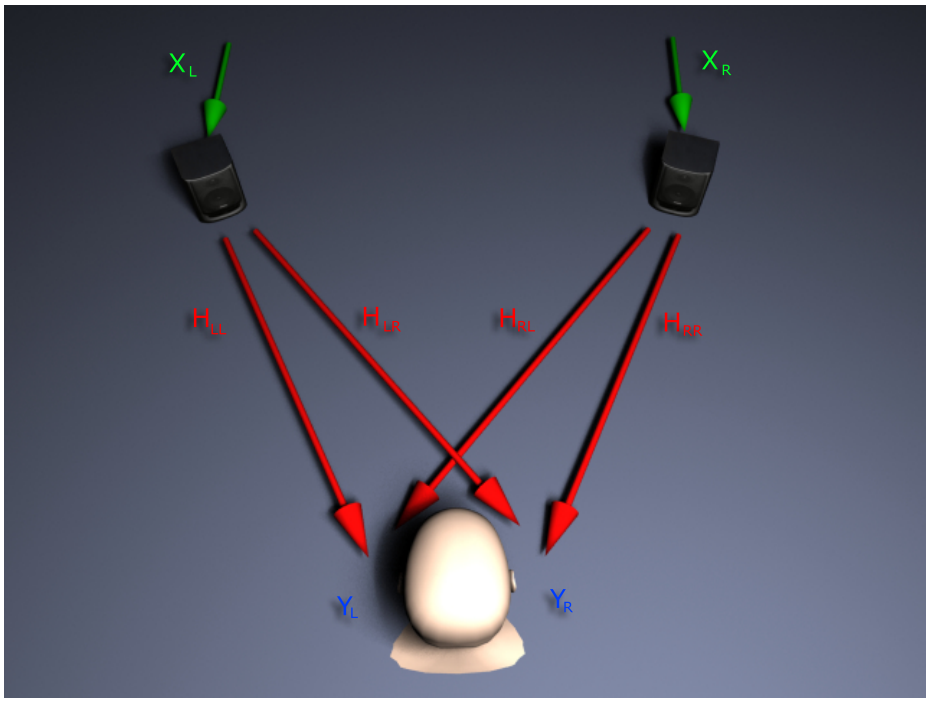

Figure 3.8 Transaural technique overview. the perceived signal Y_r,Y_l is a mixture of the signal X_l and X_r convolved by a filter H.

be carried out with a human head or an artificial head. The quality of localization increases when recording with the head of the actual listener. The only drawback of the method is the "static" capture of the scene. After the recording is made, there are no ways to control the displacement or rotation of the listener in the scene or to modify the scene itself.

In 1931, Blumlein introduce a coincident microphone technique leading to new possibility in sound reproduction [Blumlein, 1931]. During the recording, two bidirectional microphones were placed in the same location but with a different orientation. Thus, the sound arrives with a different intensity and the same delay. The playback is made using the difference, creating stereo images directly from the real scene. Similar techniques appeared with other microphones and orientations. At the same time, non-coincident microphone techniques appeared, capturing sound with different time delay with two microphones (A-B stereo). These techniques were later extended to multichannel surround (e.g. Decca-tree [Everest, 1998]).

In the early 1970s, M. Gerzon introduced a surround sound technology called "Ambisonics" [Gerzon, 1985]. With the assumptions that at a point in space, the soundfield can be defined by an omnidirectional pressure and three difference pressures in X, Y, Z direction, a microphone with four capsules positioned in a tetrahedron is

used (see Figure 3.9).

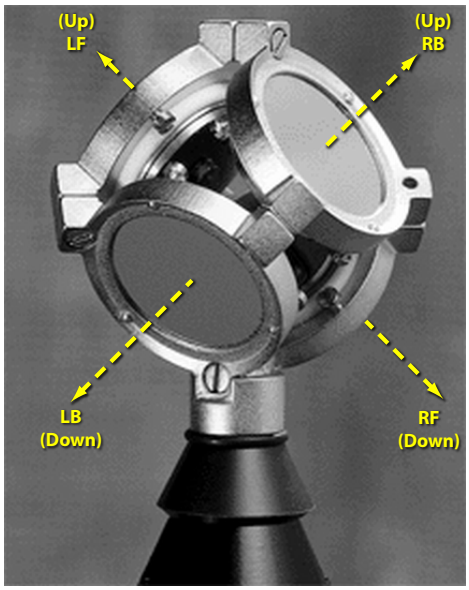

Figure 3.9 Four-capsule microphone positioned in a tetrahedron. The soundfield can be defined by an omnidirectional pressure and three difference pressures in X, Y, Z direction(©Soundfield).

The signals are created by a linear combination of the four capsules to provide the "B-format". The B-format has four components: W is the omnidirectionnal pressure signal, X the front-to-back directional information, Y the side-to-side directional information and Z the up-down directional information. Thus, the WXYZ component can be deduced from the microphone capsules using the following formula [Craven and Gerzon, 1977] :

$$\begin{cases} W = LF + RB + RF + LB \\ X = LF - RB + RF - LB \\ Y = LF - RB - RF + LB \\ Z = LF + RB - RF - LB \end{cases}$$

where LF is the left front capsule, RB the right back, RF the right front and LB is the left back.

These components also correspond to a decomposition onto a first-order spherical harmonics basis. Encoding the soundfield in such functions induces a spatial low pass filter in the localization (see Figure 3.10). Recently, Laborie extended the principle to acquire high order spherical harmonics using a microphone array [Laborie et al.,

2003]. With theses methods, spatial localization was improved in comparison to previous methods and rotation of the soundfield can be achieved. However, free listener walkthrough still can not be done with such approaches.

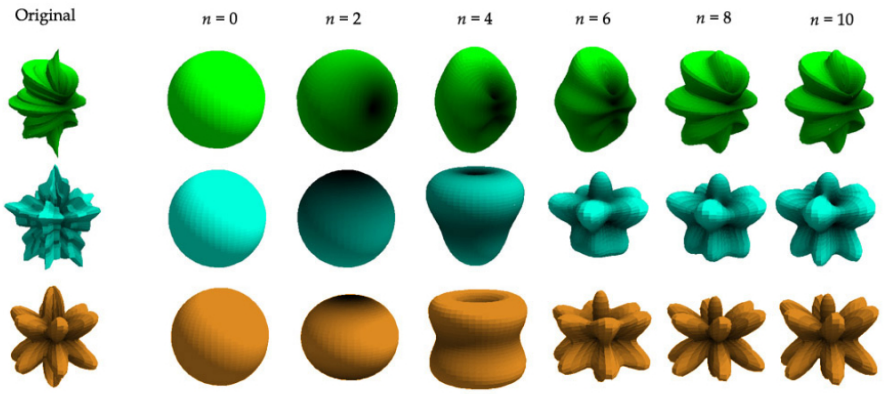

Figure 3.10 Example of three functions with their approximation using spherical harmonics functions of order N [Green, 2003].

Wavefield-synthesis (WFS) is based on the Huygens principle. This technique consists in reconstructing the acoustic wavefront. From a loudspeaker array, it is possible to recreate the wavefront at any point with appropriately weighted and delayed signals. Inside the loudspeaker array, the synthesized wavefront is identical to the measure, providing a large listening zone without sweet spot. The recording is done by close microphones which acquire the direct sound field of each source. High order ambisonics can also be decoded in WFS systems by matrixing [Daniel, 2003]. Even in this case it is not possible to create a virtual walkthrough application.

3.3.4 Perceptual Optimizations for Spatial Audio Rendering

This section will review the existing speed-up techniques for spatial audio rendering. Spatialization is a per-sound source processing and implies an incrementally large overhead as the number of source grows. However, our perception is not able to distinguish all the detail of sound spatialization. Thereby, spatial optimization can be done to reduce the processing cost of the sound simulation. Different types of spatial reduction approaches exists and this section will introduce some fixed and dynamic clustering approaches.

Fixed Clustering

The first approach to reduce the spatialization was introduced by Herder [Herder, 1999]. He proposed to cluster sound sources in a directional cone structure. The cone is selected according to the source location and the source is clustered if the perceptual error between the representative and all the sources continues to be small. The cluster representatives are used for rendering. This approach lead to several problems discussed by the author : perceptual artefacts might occur during source switching and sound spatialization errors might occur through averaging object attributes.
Recently, Jot proposes a method to reduce the binaural synthesis cost by spatially interpolating the HRTF cues through a set of fixed spatial functions [Jot et al., 2006].

Dynamic Per-Object Clustering

The clustering proposed by Sibbald is an object based method [Sibbald, 2001]. Sound sources related to an object or an area are grouped according to their distance to the listener. In near field, secondary sound sources are created and dynamically uncorrelated in order to improved the spatial sensation. In far field, sources are clustered together, accelerating the spatial rendering. The drawback of the method is that the clustering is evaluated on a per-object basis and does not consider all the elements of the scene. Moreover, the metric used for the clustering is based only on the distance, which could be improved.

Dynamic Global Clustering

Tsingos et al introduced a dynamic source clustering method based on both the scene and the signal [Tsingos et al., 2004]. The dynamic clustering is derived from the Hochbaum-Shmoys heuristic. The distance metric from the cluster position C and the source position S uses a criterion of combined loudness, distance and angle:

$$Dist(C,S) = L * (\beta \log_{10}(\|C\| \,/\, \|S\|) + \lambda(\frac{1}{2}(1 - C \cdot S) \tag{3.15}$$

where L is the loudness of the current frame of the source signal S, λ and β are weighting coefficients.

Thus, this distance metric creates longer clusters when sources are far from the listener, close in angle and quiet. The signal of the cluster representative is constructed from the signal of the clustered sound sources and its position used to spatialize the cluster using a dedicated algorithm according to the reproduction setup (see Figure 3.11). In order to render fewer sources, the algorithm also uses an on-line masking evaluation performed at each time frame to remove inaudible sound sources.

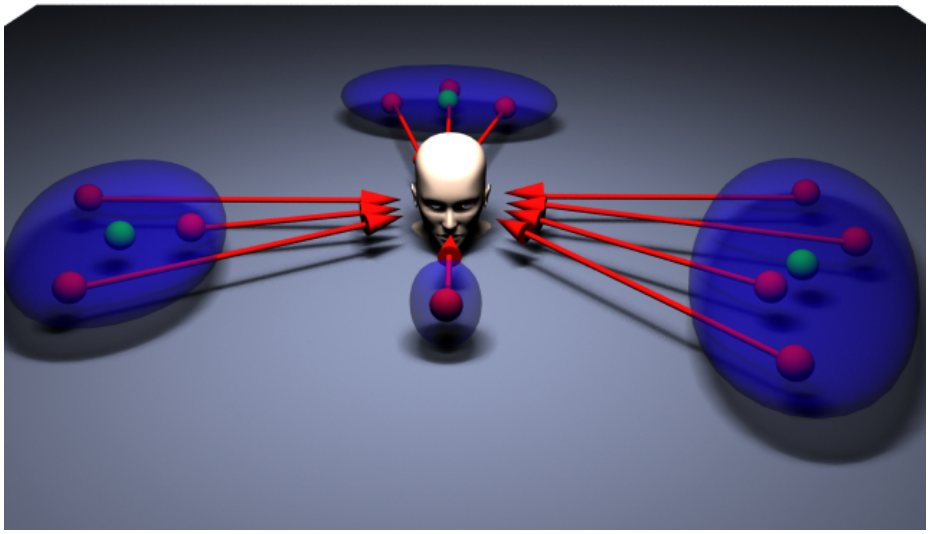

Figure 3.11 Dynamic clustering of point sound sources : When the cluster is composed of more than one source, an impostor is defined to replace all sources of the cluster (green sphere).

In this first part, we have introduced the necessary notions to understand the following chapters. We have seen that the auditory system analyzes the incoming signal through auditory filters and that a number of auditory phenomena depend on this analysis. We have seen that signals can be described with different representations and the sparseness of these representations can lead to intuitively separate a signal mixture. Moreover, sparse data have fewer significant coefficients leading to faster processing. Finally, we have presented an overview of sound rendering techniques. Point sound sources are generally used to model an auditory scene. We will now investigate hardware parallelism in order to process massive numbers of such sources efficiently.

Part II

Efficient Sound Rendering

Chapter 4

Massively Parallel Processing for Audio Rendering : A Case Study on GPUs.

Audio processing applications are among the most compute-intensive and often rely on additional DSP resources for real-time performance. However, programmable audio DSPs are in general only available to product developers. Professional audio boards with multiple DSPs usually support specific effects and products, while consumer "game-audio" hardware still only implements fixed-function pipelines which evolve at a rather slow pace.

The widespread availability and increasing processing power of programmable graphics hardware (GPUs) could offer an alternative solution. GPU features, such as multiply-accumulate instructions or multiple SIMD execution units, are similar to those of most DSPs [Eyre and Bier, 2000]. Moreover, their high-level programmability with floating point support and easy access to development kits turns them into attractive co-processors for non-graphics applications. Besides, 3D audio rendering applications require a significant number of geometric calculations, which are a perfect fit for the GPU. Our feasibility study investigates the use of GPUs for efficient audio processing.

4.1 GPU Architecture

Graphics hardware has a specific dataflow computational model. Its architecture is originally dedicated to manipulate 3D primitives like points, lines or polygons, perform some graphics operations and render the result on the screen. Primitives follow a sequence of operations before processed to the screen. Figure 4.1 shows the

essential steps of the pipeline. Basically, the application transmits data vertices of the primitives to the vertex processor. The vertex is a structure containing 3D and texture coordinates, color and normal vector. The vertex processor applies any mathematical transform to each vertex including transform from world space to projection space.

In the second step, vertices are assembled following the geometric primitives information. In this step, culling is computed to discard invisible polygons according to the normal of the polygon and the view direction. Next, clipping to the view frustum is applied before the rasterization. The view frustum is a set of plane which defined the field of view of the camera.

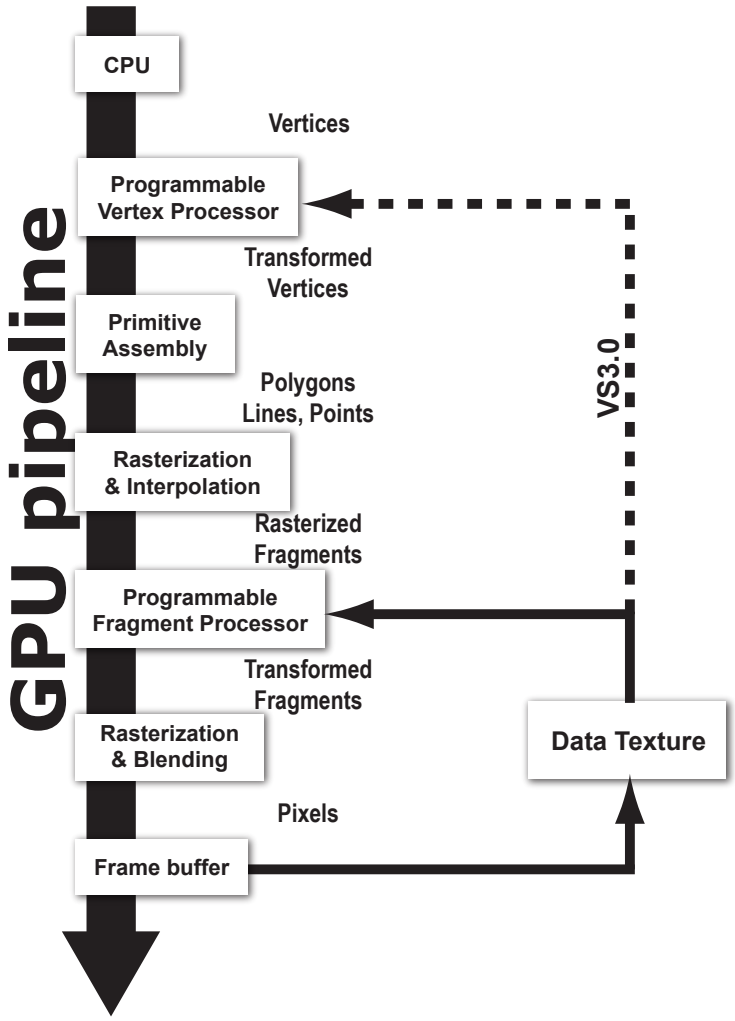

Figure 4.1 GPU pipeline. The vertex processor and the fragment processor are totally programmable.

The third step of the pipeline rasterizes the transformed primitives. Rasterization determines which fragment of the screen buffer is covered by a primitive. The term "fragment" is employed instead of pixel to make the difference between the resulting pixel and the one with other characteristics than color and containing possible operations to be done to get the result. The characteristics of the fragment, like color or textures coordinates, are interpolated between the transformed vertices of the primitive. The interpolation can be "nearest", meaning truncated from the nearest pixel, linear or bi-linear but any other interpolations can be applied via the programming features of more recent GPUs.

The next step is the most important one. Indeed, this stage processes the operations of the fragment. Many math operations can be computed to obtain the final pixel. The aim of the last step is to perform additional tests before writing the final fragment value (color) to the frame buffer. Tests includes depth testing, used to remove hidden pixel and alpha test, for compositing purposes.

The second and the fourth step can be overridden to define user programs. It is possible to program the GPU with high level C-like languages (e.g., CG, GLSL, HLSL). The power of the GPU is provided by its data parallel architecture. Each fragment programs is processed in parallel. G70 Nvidia card contains 24 pixels pipeline and the G80 contains 128 stream processors which are automatically attributed to vertex of fragment processing using massive multithread handling. The communication from the GPU towards the CPU is slow, even with the new PCI express bus. The fastest way to transfer resulting data back to fragment processor is to do multiple passes by rendering the frame buffer to a texture. The vertex shader 3.0 model allows transfer of data to the vertex processor. However, this communication is not really fast because all pixels have to be rendered to the frame buffer beforehand.

4.2 GPU-Accelerated Audio Rendering

4.2.1 Storing Data on the GPU

The GPU memory model is targetted to 3D graphics and is different from the CPU memory model. Taking it into consideration leads to better performance. The GPU memory was designed to work with images. It is thus highly optimized for the Red-Green-Blue plus Alpha (RGBA) data type. This structure is usually packed in floating point data. To fit this model, we decompose our signals to four frequency bands, in a perceptual scale and pack them in the RGBA structure (see Figure 4.2). As a result, the four frequency bands are stored in an interleaved manner. Each pixel of the texture represents a sample of the signal.

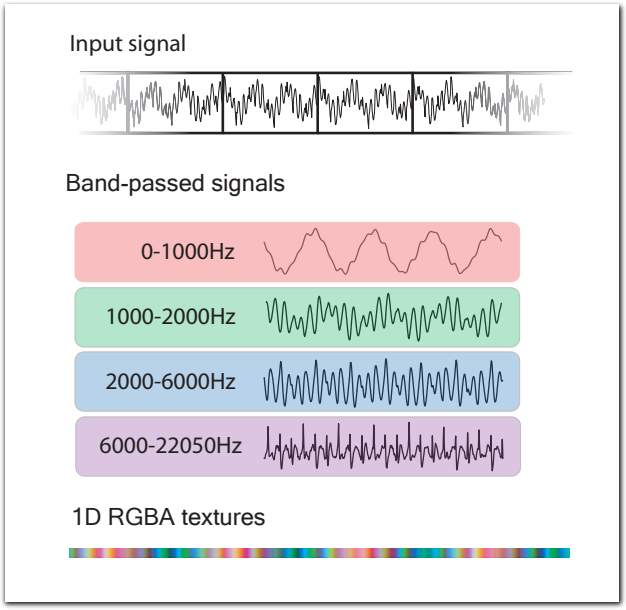

Figure 4.2 Audio data structure. (a) The incoming signal is sliced into frames. (b) The signal is decomposed into four frequency subbands. (c) the four subbands are stored in 1D RGBA textures.

4.2.2 Audio Processing

We consider a combination of two simple operations commonly used for 3D audio rendering: variable delay-line and filtering [Begault, 1994, Funkhouser et al., 2002]. The signal of each sound source is first delayed by the propagation time of the sound wave. This involves resampling the signal at non-integer index values thus automatically accounting for Doppler shifting. The signal is then filtered to simulate the effects of source and listener directivity functions, occlusions and propagation through the medium. We resample the signals using linear interpolation between the two closest samples. On the GPU this is achieved through texture resampling. Filtering is implemented using a simple 4-band equalizer. Assuming that input signals are band-pass filtered in a pre-processing step, the equalization is efficiently implemented as a 4-component dot product which is issued as a single GPU instruction.

Binaural stereo rendering requires applying this pipeline twice, using a direction-dependent delay and equalization for each ear, derived from head-related transfer

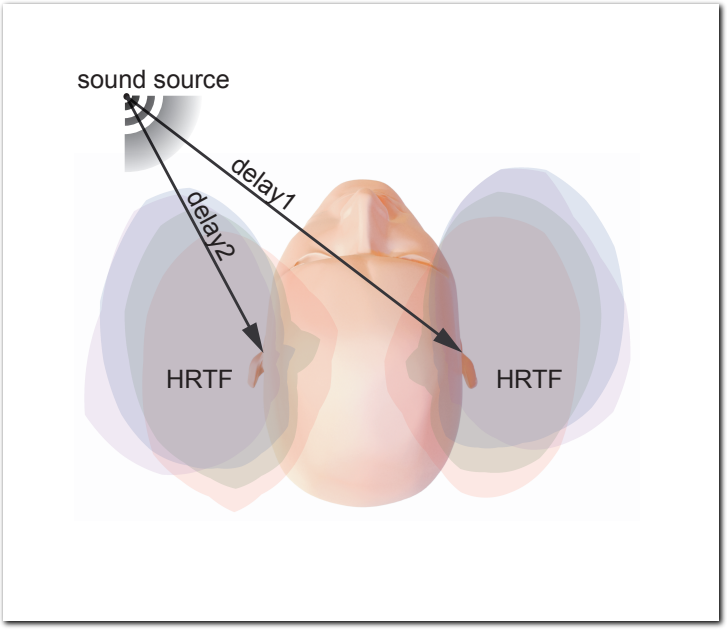

Figure 4.3 Audio processing involved in the GPU simulation. Each sound source is delayed by the propagation time and filtered to account for the distance attenuation and head-related transfer functions (HRTFs).

functions (HRTFs) [Begault, 1994]. The HRTF data is represented as an azimuth-elevation texture array (see Figure 4.4) where the RGBA component holds a gain value for the corresponding frequency band. Similar audio processing can be used to generate dynamic sub-mixes of multiple sound signals prior to spatial audio rendering (e.g., the perceptual audio rendering of [Tsingos et al., 2004]).

4.2.3 Results

We compared an optimized SSE (Intel's Streaming SIMD Extensions) assembly code running on a *Pentium 4 3GHz* processor and an equivalent *Cg/OpenGL* implementation running on a *Nvidia GeForce FX 5950 Ultra* graphics board on AGP 8X and a *Nvidia Quadro FX4500* on PCI express. Audio was processed at 44.1 KHz using 1024-sample long frames. All processing was 32-bit floating point.

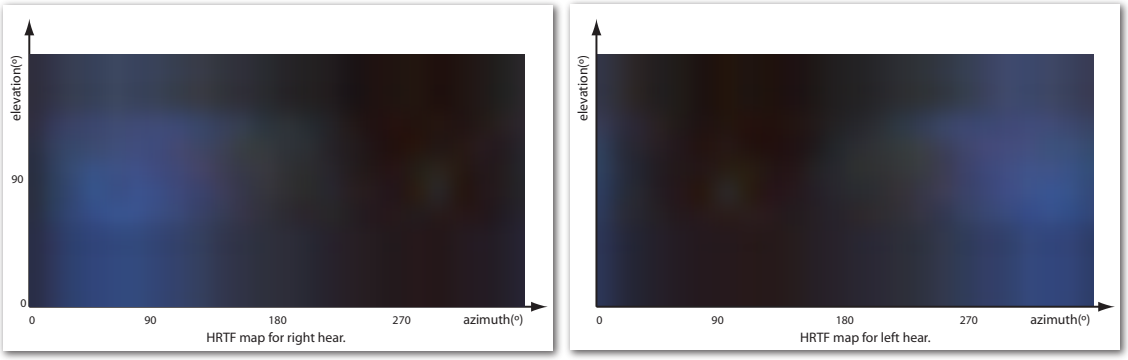

Figure 4.4 Azimuth-elevation HRTF map for the left (a) and the right ear (b). The intensity color of the RGBA component correspond to the attenuation for each frequency component generated from measured FIR data from the LISTEN HRTF database.

The GPU implementation, on a Nvidia Quadro FX4500, can perform binaural processing of up to 2250 sound sources in real time while the SSE version renders 700 sound sources in one time-frame (≈22.5 ms). However, resampling floating-point textures requires two texture fetches and a linear interpolation in the fragment shader. If floating-point texture resampling was available in hardware, GPU performance would increase. We have simulated this functionality on our GPU using a single texture-fetch and achieved real-time performance for up to 3100 sources. With the up-coming G80 processor, floating-point textures resampling is supported but not available at the time of this study.

For mono processing, the Quadro FX treats up to 6150 (1 texture fetch)/ 4580 (2 fetches and linear interp.) sources, while the CPU handles 1400 in the same amount of time. On average the GPU implementation using the Quadro FX4500 was about three times faster than the SSE implementation and it would become 50% faster if floating-point texture resampling was supported in hardware. The latest graphics architectures would significantly improve GPU performance due to their increased number of pipelines and their faster RAMDAC.

The huge pixel throughput of the GPU can also be used to improve audio rendering quality without reducing frame-size by recomputing rendering parameters (source-to-listener distance, equalization gains, etc.)on a per-sample rather than per-frame basis. This can be seen as an audio equivalent of per-pixel vs. per-vertex lighting in graphics. By storing directivity functions in cube-maps and recomputing propagation delays and distances for each sample, our GPU implementation can still render up to 180 sources in the same time-frame. However, more complex texture-addressing

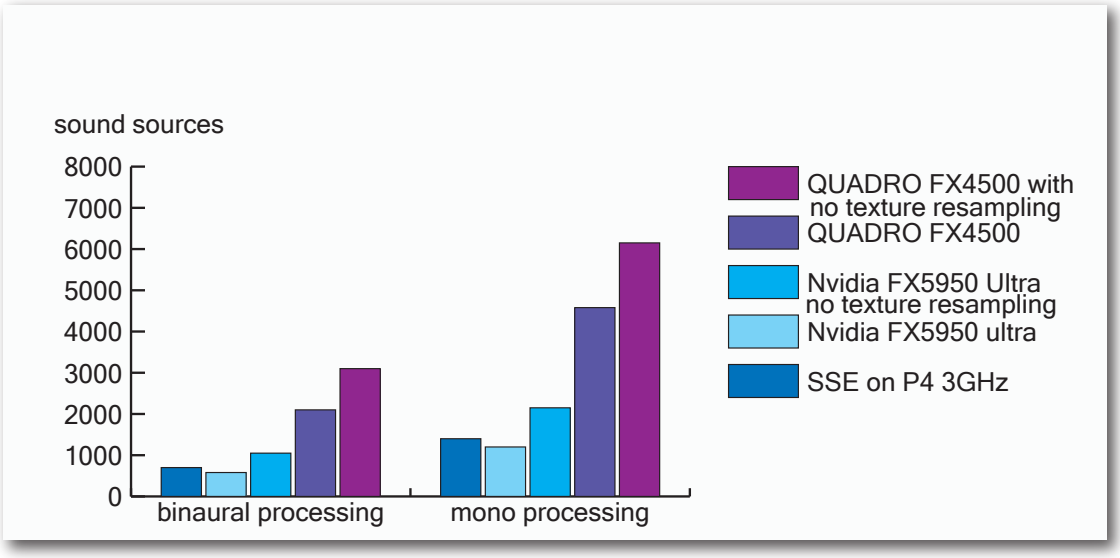

Figure 4.5 Performance tests for audio rendering on the CPU and GPU.

calculations are needed in the fragment program due to limited texture size. By replacing such complex texture addressing with a single texture-fetch, we also estimated that direct support for large 1D textures would increase performance by at least a factor of 2. Novel G80 processors now support this functionality.

Running audio effects on the GPU frees-up CPU time for other tasks and can even be combined with graphics rendering with little impact on display performances for moderately graphics-demanding applications. Example movie files including GPU-generated audio and graphics are available[1]. Both audio and graphics were generated in real-time with the GPU.

4.3 Discussion and Conclusion

The vertex and fragment processor of the graphics hardware can be fully programmed with assembly-like languages but is not really suitable for programming complex shaders. Similar in spirit to Renderman, a shading language used by Pixar for image rendering, high level C-like languages have been introduced to program the GPUs: "CG" from Nvidia, "OpenGL Shading Language (GLSL)" from 3DLabs in conjunction with OpenGL ARB and "High Level Shading Languages (HLSL)" from Microsoft. They give the most up-to-date functionality, but graphics notions are required to pro-

[1] http://www-sop.inria.fr/reves/projects/GPUAudio/

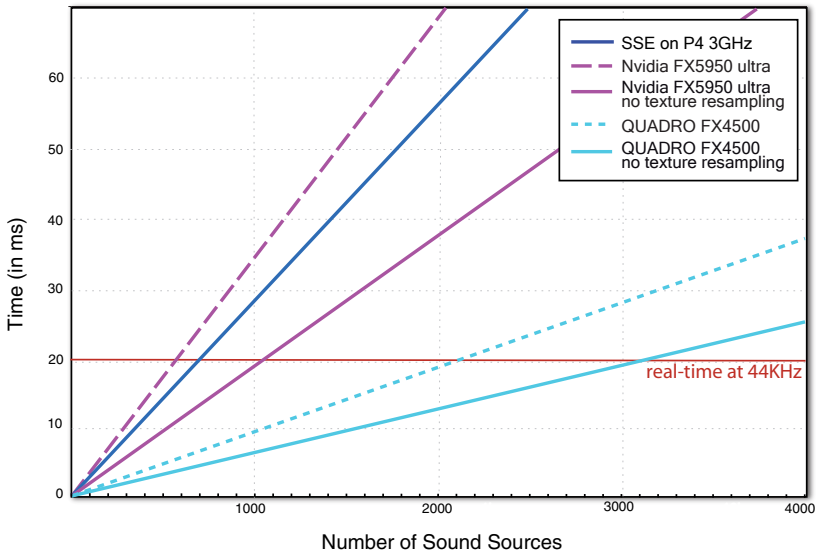

Figure 4.6 Performance for binaural audio rendering on the CPU and GPU.

gram them. They are also dedicated to particular platforms API (such as OpenGL or DirectX) and graphics processor vendor. Recently, with the success of the graphics hardware as a general purpose processor unit, high level languages have emerged, proposing a friendly programming approach and hiding graphics primitives and 3D manipulation instructions.

Nvidia and ATI/AMD have recently introduced their new general purpose C-like language to get optimal performance using their latest hardware and provided total abstraction of the graphics pipeline[2].

Microsoft introduced their new general purpose language for programming graphics-processor[3].

Academic research has also introduced new alternative languages which have evolved into commercial applications. They provide development environments for programming general purpose processor including the CPU, the GPU, or the Cell processor. They offer a good compromise because they are multi-platform and independent from the hardware[4].

Finally, the parallel stream architecture introduced by the GPUs tends to abstract from the graphics and evolve towards general purpose application. Due to this

[2]http://www.nvidia.com/object/cuda.html, http://ati.amd.com/companyinfo/researcher/documents.html
[3]http://research.microsoft.com/research/downloads/
[4]http://www.rapidmind.net, http://www.peakstreaminc.com

architecture, the GPUs provide better performance than the CPU.

The GPUs performance has increased dramatically over the last three years compared to CPUs [Owens et al., 2007]. While our first experiments, in 2004, suggested that GPUs can be used for 3D audio processing with similar or increased performance compared to optimized software implementations running on top-of-the-line CPUs, the latest GPUs clearly outperform CPUs by a factor of at least 3, and, thereby, are a perfect alternative for audio processing. Moreover, the GPUs surpass CPUs for a number of other tasks, including Fast Fourier Transform, a tool widely used for audio processing [Buck et al., 2004] and [Govindaraju et al., 2006]. Figure 4.7 shows a performance comparison of the 1D Fast Fourier Transform, the CPU implementation is based on the Intel Math Kernel library and the GPU implementation based on the GPUFFTW library [Govindaraju et al., 2006].

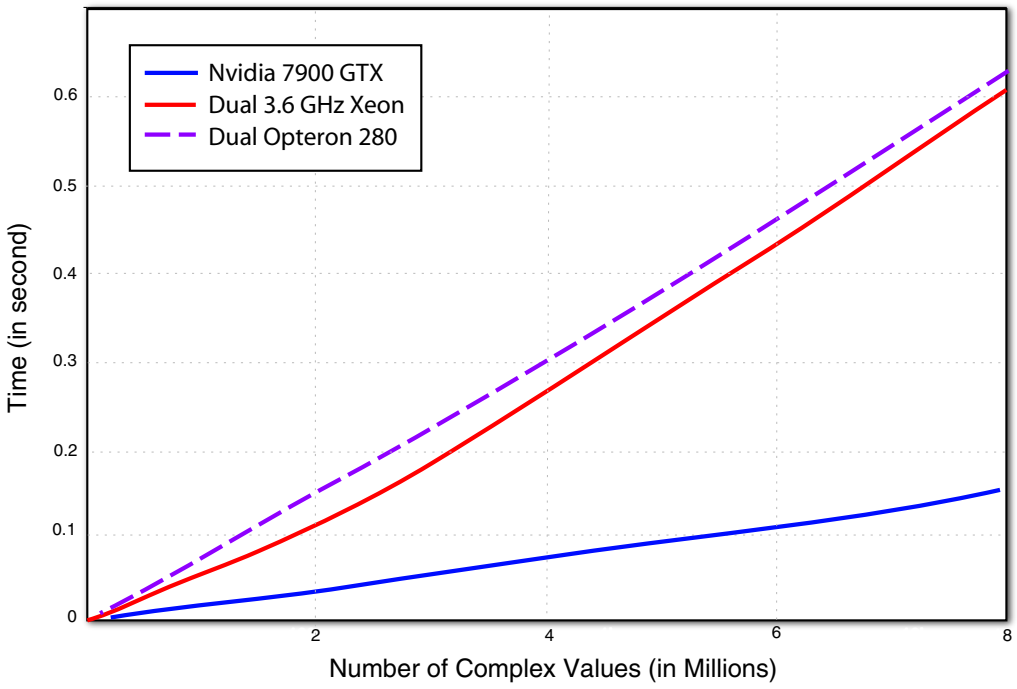

Figure 4.7 Comparison of 1D Fast Fourier Transform on CPU and GPU [Govindaraju et al., 2006].

In our first study, we had detected several shortcomings which prevent efficient use of GPUs for mainstream audio processing applications. Due to limitations in texture-access modes and texture-size, long 1D textures could not be easily indexed

and floating-point textures resampling was not supported. The latest G80 processor now overcomes these limitations.

However, other algorithms such as infinite impulse response (recursive) filtering cannot be implemented efficiently since past values are usually unavailable when rendering a given pixel in fragment programs. As suggested in [Buck et al., 2004], including persistent registers to accumulate results across fragments would solve this problem.

On a broader scale, our results demonstrate that stream-processing architectures are appropriate for audio rendering applications and that game-audio hardware, borrowing from graphics architectures and shading languages, may benefit from including programmable "voice shaders", enabling per-sample processing, prior to their main "effects" processor.

In this chapter, we have focused on rendering a complex auditory scene by exploiting parallel processing. A real-time 3D audio simulation has been presented as an application example including time-delay resampling, distance attenuation and HRTF equalization. The results showed that the GPU architecture is well adapted to handling audio processing. However, although we obtain a significant improvement in performance through the power of the GPU, the algorithm is still directly dependent on the number of sources. In the following chapter, we will examine the possibility of reducing the amount of data to be processed using human perception.

Chapter 5

Perceptual Progressive Rendering

Many applications ranging from video games to virtual reality or visualization and sonification require processing large number of audio signals in real-time. For instance, modern video games must render a large numbers of 3D sound sources using some form of spatial audio processing. Furthermore, each source's audio signal may itself be generated as a mixture of a number of sub-signals (e.g., a car-noise is a composite of engine and tire/surface noise) driven from real-time simulated physical parameters. The number of audio signals to process may often exceed hardware capabilities. Priority schemes which select the sounds to process according to a preset importance value are a common way of using hardware more efficiently, for instance by managing the limited number of hardware channels on a dedicated sound card. Usually, this value is determined by the sound designer at production time and might further be modulated by additional effects at run-time, such as attenuation of the sound due to distance or occlusion.

This chapter is focused on the problem of automatically prioritizing audio signals according to an importance metric, in order to selectively process these signals. Such a metric can then be used to tune the processing "bit-rate" in order to fit a given computational budget: for instance, allocating a budget of arithmetic operations to a complex signal processing task (e.g., a combination of mixing, filtering, etc.) involving a large number of source signals.
Figure 5.1 shows a basic example application where four speech signals have been prioritized according to a loudness metric and a mix has been generated simply by playing back the single most-important signal per processing frame (highlighted in yellow). Figure 5.3 demonstrates the same principle applied to several tracks of a song rendered with a variable subpart of the original data. This could typically be used for hardware voice management in video games.

This chapter presents a comparative study of several metrics that can be used to prioritize signals for selective real-time processing of audio signals. In section 5.1,

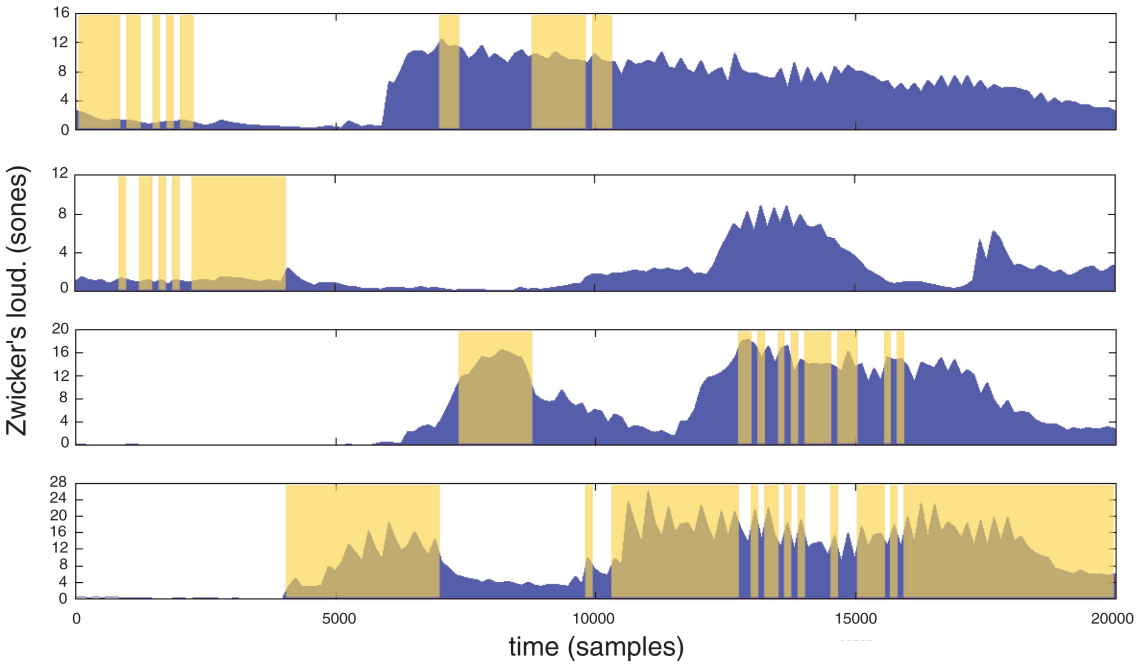

Figure 5.1 Four speech signals prioritized according to a loudness metric computed over successive short time-frames. The single most important frame across time is highlighted in yellow.

we start by reviewing previous work related to scalable and progressive audio processing. A coarse-grain selective processing algorithm is described in section 5.2. In particular, several metrics that can be used to prioritize the audio signals and selectively allocate the required operations are discussed in section 5.2.1. Our selective processing algorithm, which does not require a specific representation for the audio data, is demonstrated in the context of a time-domain pipeline comprising mixing and simple filtering operations in section 5.2.2. Results of a pilot subjective study are then presented which support the applicability of our technique in section 5.3. We finally discuss our approach and sketch other possible applications of our prioritization scheme before concluding.

5.1 Related Work

While parametric, progressive and scalable codecs are a key research topic in the audio coding community [Vercoe et al., 1998, Purnhagen, 1999, Herre, 2002], few attempts to date have been made to design a scalable or selective approach for real-time signal

processing.

Fouad et al. [Fouad et al., 1997] propose a level-of-detail rendering approach for spatialized audio where the sound samples are progressively generated based on a perceptual metric in order to respect a computing time budget. When the budgetted time is elapsed, missing samples are interpolated from the calculated ones. They tag the signals according to their overall energy. However, such a scheme will fail at capturing large variations of energy through time within the signal itself.

Wand and Straßer [Wand, 2004] proposed a multi-resolution approach to 3D audio rendering. At each frame of their simulation, they use an importance sampling strategy to randomly select a sub-set of all sound sources to render at each processing frames. However, their importance sampling strategy does not account for the variations in signal intensity which might be much more significant (factors of 10 or more can be easily observed on speech signals for instance) than variations in the control parameters such as distance attenuation, etc. which usually vary smoothly and slowly through time (except for very near-field sources).

In our previous work [Tsingos et al., 2004] we proposed a framework for 3D audio rendering of complex virtual environments in which sound sources are first sorted by an *importance* metric, in our case the *loudness level* of the sound signals. The importance metric is efficiently updated in real-time using pre-computed descriptors stored with small chunks of the input audio signals. Hence, variations within the signals are accounted for. The priority metric was used to determine inaudible sources in the environment due to auditory masking and group sound sources together for spatialization.

Our previous algorithm consists of computing the total power level of the scene during a time frame. Then, we render all sound sources until the auditory masking created by the rendered sources is sufficient to hide the remaining sources.

We evaluate masking in a conservative manner by first sorting the sources by decreasing order according to their normalized loudness L_t^k and progressively inserting them into the current mix until they mask the remaining ones. We start by computing the total power level of our scene

$$P_{tot} = \sum_k P_t^k(f) \tag{5.1}$$

At each frame, we maintain the sum of the power of all sources to be added to the mix, P_{toGo}, which is initially equal to P_{tot}. We then progressively add sources to the mix, maintaining the current tonality T_{mix}, masking threshold M_{mix}, as well as the current power P_{mix} of the mix (see section 2.1.3). We assume that sound power adds up which is a crude approximation but works reasonably well with real-world signals, which are typically noisy and uncorrelated. To perform the perceptual

culling, we apply the following algorithm, where ATH is the absolute threshold of hearing (corresponding to 2 phons) [Moore et al., 1997]:

$M_{mix} = -200$
$P_{mix} = 0$
$T = 0$
$P_{toGo} = P_{tot}$
while $(dB(P_{toGo}) > dB(P_{mix}) - M_{mix})$ and $(P_{toGo} > ATH)$ do
    add source S_k to the mix
    P_{toGo} -= P^k
    P_{mix} += P^k
    T += $P^k * T^k$
    $T_{mix} = T/P_{mix}$
    $M_{mix} = (14.5 + Bark(f_{max})) * T_{mix} + 5.5 * (1 - T_{mix})$
    k++
end

Similar to prior audio coding work [Painter and Spanias, 1997], we estimate the masking threshold, $M_{mix}(f)$ as :

$$M_{mix}(f) = (14.5 + Bark(f_{max})) * T_{mix}(f) + 5.5 * (1 - T_{mix}(f))(dB), \qquad (5.2)$$

where $Bark(f_{max})$ is the value of the maximum frequency in each frequency-band f expressed in Bark scale.
The Bark scale is a mapping of the frequencies in Hertz to Bark numbers, corresponding to the 25 critical bands of hearing [Zwicker and Fastl, 1999]. In our case we have for our four bands : Bark(500) = 5,Bark(2000) = 18,Bark(8000) = 24,Bark(22050) = 25. The masking threshold represents the limit below which a maskee is going to be masked by the considered signal. To better account for binaural masking phenomena, we evaluate masking for left and right ears and assume the culling process is over when the remaining power at both ears is below the masking threshold of the current mix. Since we always maintain an overall estimate for the power of the entire scene, our culling algorithm behaves well even in the case of a scene composed of many low-power sources. This is the case for instance with image-sources resulting from sound reflections. A naive algorithm might have culled all sources while their combination is actually audible. This chapter extends this approach by comparing several priority metrics and their subjective effect on selective processing of audio signals even for cases where removed sub-parts of the signals are above masking threshold.

Where the previous method does not provide any control on the amount of processing, this chapter will propose a level of detail approach, using a speed versus quality trade-off.

Other scalable approaches based, for instance, on modal synthesis, have also been proposed for real-time modal synthesis of multiple contact sounds in virtual environments [Lagrange and Marchand, 2001, Doel et al., 2002, Doel et al., 2004]. Similar parametric audio representations [Vercoe et al., 1998, Purnhagen, 1999] also allow for scalable audio processing (e.g. pitch shifting or time-stretching, frequency content alteration, etc.) at limited additional processing cost, since processing would only concern a limited number of parameters rather than the full PCM audio data. However, this approach might imply real-time coding and decoding of the sound representations. As parametric representations are not widely standardized and commonly used in interactive applications, available standard hardware decoders do not usually give access to the coded representation in a convenient form for the user to further manipulate. Eventhough processing in coded domain might be achieved through modified software implementation of standard audio codecs (e.g. MPEG-1 layer 3, MPEG-2 AAC) [Touimi, 2000], the overhead due to partial decoding would probably be overwhelming for a real-time application handling many signals.

5.2 Selective Audio Processing

We propose a coarse-grain selective audio processing framework which can be separated in two steps : 1) we assign a priority to each frame of the input signals and 2) we select the frames to process by decreasing priority order until our pre-specified budget is reached. Remaining frames are simply discarded from the final result. Both steps are applied at each processing frame to produce a frame of processed output signal. The following sections detail both steps.

5.2.1 Priority Metrics

In our approach, as well as others we described in section 5.1, processing management is driven by a given importance metric. The choice of this metric is then a crucial step: the audibility of the artifacts introduced by any processing optimizations will depend on its quality.

Loudness seems a good candidate since it has been shown to be closely related to masking phenomena [Zwicker, 1984, Baumgarte, 1997]. Using loudness as an importance metric might hence allow important maskers to be processed first. But one can imagine that weighting may be more efficiently performed on the basis of more cognitive aspects. For instance, in the context of a collision avoidance experimental

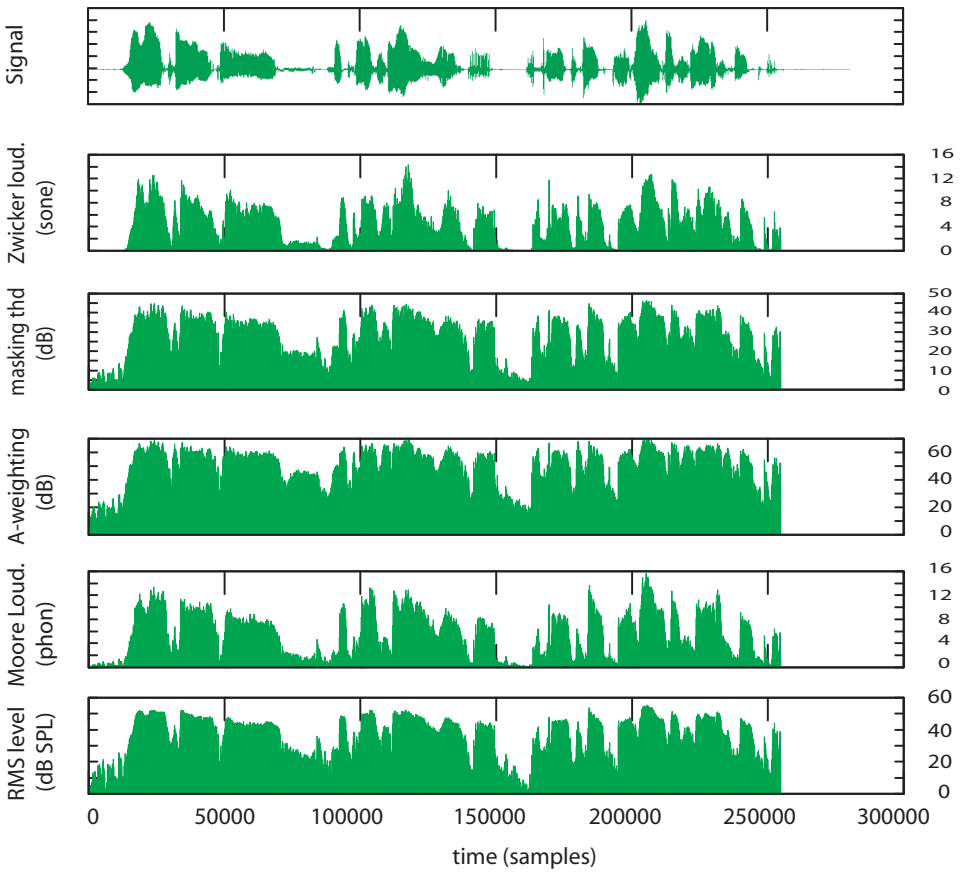

Figure 5.2 Several priority metrics calculated for an example speech signal using 3 ms-long frames.

setup, Robert Graham [Graham, 1999] noticed that faster braking reaction times were measured when drivers were warned by car horns sounds, even if they were less loud than other tested sounds. There is a vast literature aiming at building psychoacoustic relationships between acoustic parameters of a sound and its so-called *urgency* (see Stanton and Edworthy [Stanton and Edworthy, 1999] for an overview). Derivations of these urgency metrics may form a more cognitive-founded importance metric.
As a starting point, this chapter examines the ability of several level-related metrics to optimize audio processing. In particular, we evaluated the following importance metrics:

1. RMS level, expressed in dB SPL,

2. *A*-weighted level, expressed in dB [Acoustics-FAQ, 1997],

3. Moore, Glasberg & Baer's loudness level [Moore et al., 1997], expressed in *phons*, calculated assuming a stimulus is a band-limited noise.

4. Zwicker's loudness [Zwicker et al., 1991], expressed in *sones*,

5. "Masking level" model defined as the level of the source minus a masking-threshold offset, expressed in dB, predicted from the *tonality* index of the signal [Painter and Spanias, 1997, Brandenburg, 1992]. The tonality index is typically derived from a *spectral flatness measure* and indicates the tonal or noisy nature of the signal (see section 2.1.3).

Each metric is evaluated for small processing frames along our test signals, typically every 3 to 23 ms (i.e., 128 to 1024 samples at 44.1kHz). Results were not significantly different for the various frame sizes. Smaller frames give better time-resolution and can result in more optimal interleaving of the signals during the processing step. However, frames which are too short can result in highly degraded audio information since interleaved signals will no longer be recognizable, a problem closely related to the illusion of continuity [McAdams et al., 1998]. Figure 5.2 shows a comparison of several loudness metrics evaluated on a fragment of speech data.

Table 5.1 shows the average rank correlation obtained with various metrics on a three different mixtures speech, ambient and music signals. Rank correlation measures how correlated the orderings obtained with the various metrics are. As can be seen in this table, results appear to be dependent on the type of signals. For speech and ambient sounds, metrics are correlated although not strongly. For the musical mixture, results are more pronounced showing stronger correlation between Zwicker's and Moore's loudness models and very low correlation between loudness models and all the others.

5.2.2 Selective Processing Algorithm

Our budget allocation algorithm is designed for real-time streaming applications. Hence, it has to be efficient and has to find a solution locally at each processing frame. To do this, the importance of each frame of the signal is evaluated and until our computational budget is reached, the algorithm selects which sub-parts of the signals should be processed, by decreasing priority value, using a greedy approach. An example is shown in Figure 5.1. The result is thus constructed as an interlaced mixture of the most important frames in all signals. To avoid artefacts during the reconstruction step, an overlap-add method (3ms frames with 10% overlap) was used. Another example is shown in Figure 5.3. Selected frames for different budgets are

speech	Zwicker loud.	mask. thr.	Moore loud.	RMS level	A-weight.
Zwicker loud.	1	0.37	0.57	0.40	0.35
mask thr.	0.37	1	0.54	0.73	0.56
Moore loud.	0.57	0.54	1	0.54	0.36
RMS level	0.40	0.73	0.54	1	0.54
A-weight.	0.35	0.56	0.36	0.54	1
ambient	Zwicker loud.	mask. thr.	Moore loud.	RMS level	A-weight.
Zwicker loud.	1	0.40	0.44	0.42	0.37
mask thr.	0.40	1	0.48	0.51	0.35
Moore loud.	0.44	0.48	1	0.47	0.37
RMS level	0.42	0.51	0.47	1	0.33
A-weight.	0.37	0.35	0.37	0.33	1
music	Zwicker loud.	mask. thr.	Moore loud.	RMS level	A-weight.
Zwicker loud.	1	0.05	0.42	0.04	0.03
mask thr.	0.05	1	0.04	0.42	0.40
Moore loud.	0.42	0.04	1	0.02	0
RMS level	0.04	0.42	0.02	1	0.36
A-weight.	0.03	0.40	0	0.36	1

Table 5.1 Rank correlation matrices for three test mixtures of speech, ambient and musical signals. Rank correlation was calculated using Spearman's formula [Howell, 1992] and averaged over all frames of the mixture.

highlighted. As can be seen in the figure, our approach directly takes care of any sparseness in the mix by removing input frames below audibility threshold from the final mix. This might already result in a significant gain. For the various mixtures we used (ambient sounds, music and speech), we estimated that 0.7% to 33% of the input frames could be trivially removed (0.7% for ambient sounds, 24.5% for music and 33% for speech).

To improve the frequency resolution of our approach, we can further evaluate the priority metric for a number of sub-bands of the signals. In our experiments, we used four sub-bands corresponding to 0-500 Hz, 500-2000 Hz, 2000-8000 Hz, 8000-22000 Hz and treated each sub-band as if it were an additional input sound signal to prioritize. This would be typically useful for applications performing some kind of sub-band correction of the audio signal (e.g., equalizers). Necessary band-pass filtering can then be performed only on the selected sub-parts of the signal.

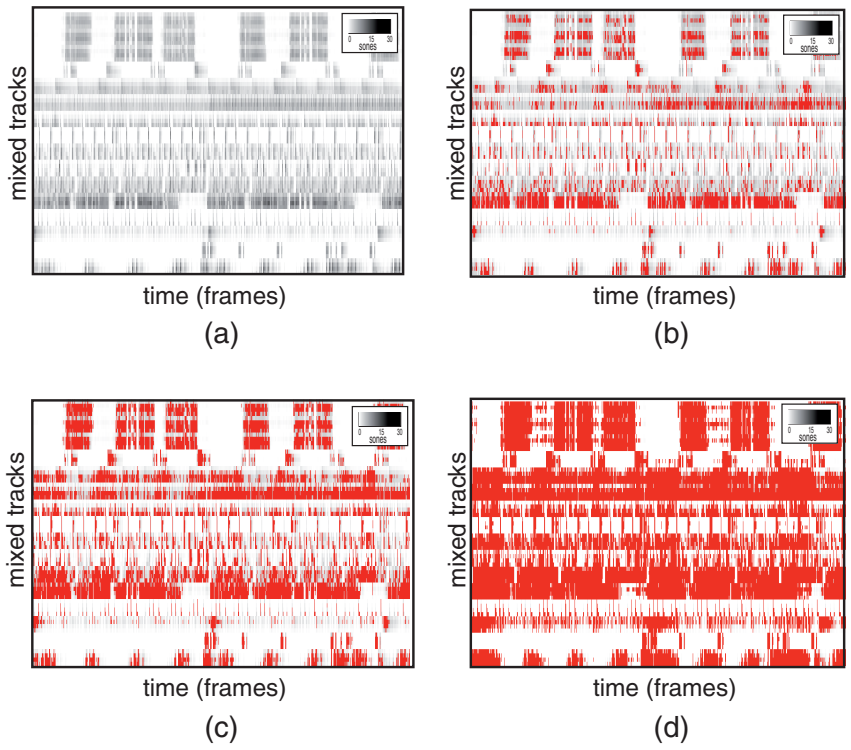

Figure 5.3 (a) Loudness values (using Zwicker's loudness model) through time for the 17 tracks of a musical mix. Each track was selectively filtered and processed into 4 frequency sub-bands resulting in 68 signals to prioritize. (b) Priority map showing the first 12.5% most important frames highlighted in red. (c) First 25% most important frames and (d) first 50% most important frames.

5.2.3 Integration in a Real-Time Processing Framework

Although most of the level-related priority metrics we used cannot be directly evaluated in real-time for large numbers of audio streams, they can be efficiently computed from additional descriptors stored with the audio data, in a manner similar to [Tsingos et al., 2004]. Loudness information, in particular, can be retrieved from pre-computed loudness tables, energy levels and tonality indices stored with each corresponding frame of input audio data. This information may also be stored for several sub-bands of the signal. Such an approach allows us to further modulate the importance value in real-time depending on various other effects affecting the signal during the simulation. Contrary to parametric coding this information does not aim

at reconstructing the final signal directly so it can remain quite compact (typically about 1 to 4 Kb/sec of input audio data) and can be interleaved with standard PCM audio data for streaming or kept resident in memory for random access while the PCM data is streamed on-demand.

Our coarse-grain selective algorithm integrates well within standard time-domain audio processing pipelines. We tested it with a 3D audio processing application for virtual reality. In this case, the signals of each virtual sound source undergo filtering and resampling operations to simulate propagation effects (atmospheric scattering, occlusion, Doppler shift, etc.) and binaural hearing (e.g., HRTF filtering) before being combined to produce the final mix. We implemented a scalable 3D audio processing pipeline implementing these effects using time-domain resampling and attenuation over several sub-bands of each source signal, computed using second-order biquad filters. Using our selective processing pipeline, we can process the signals using a budgetted number of operations resulting in a performance gain directly proportional to the selected budget. As sources of decreasing priority are processed, an alternate solution is to simplify the operations (for instance, using linear resampling instead of better quality spline-based resampling) rather than maintain high-quality processing for all selected frames and simply drop low priority frames. Example movie files demonstrating the approach are available at:
http://www-sop.inria.fr/reves/projects/scalableAudio/.

5.3 Pilot Subjective Evaluation

In order to evaluate subjective differences between the various metrics we ran a short pilot evaluation study described in the following sections.

5.3.1 Experimental Conditions

Subjects: 18 subjects (10 women and 8 men, 19 to 48 years old) volunteered as listeners. All reported normal hearing. Most of them were computer scientists, very few with any experience in acoustics or music practice. None of them was familiar with audio coding techniques, nor listened heavily to coded audio.
Stimuli: Three mixtures of various *types of signals* were generated: 1) a multi-track musical mix, 2) male and female Greek, French and Polish speech and 3) ambient sounds. The mixtures were created respectively from 17, 6 and 4 recordings separated in four sub-bands, resulting in 68, 24 and 16 signals to prioritize. Mixtures were generated at three *resolutions*, selecting the most important frames according to our priority metrics, using only 50%, 25% and 12.5% of the input signal data. Five different priority *metrics* (see section 5.2.1) were tested. A total of 45 stimuli (3

types of signals * 3 resolutions * 5 metrics) were hence created. All signals were at CD quality (44.1 kHz sampling rate and 16 bits quantization)[1].

Apparatus: We ran the test on a laptop computer using an in-house test program (see Figure 5.4). It was conducted using headphone presentation in a quiet office room. *Sennheiser HD600* headphones were used (diotic listening), calibrated to a reference listening level at the eardrum. The sounds were stored on the computer's hard drive and played through the SigmaTel C-Major integrated sound-board. They were played back at a comfortable level.

Procedure: The subjects were given written instructions explaining the task. They were asked to rate the 45 resulting output mixtures relative to the corresponding reference mix. We used the ITU-R[2] recommended *triple stimulus, double blind with hidden reference* technique, previously used for quality assessment of low bit-rate audio codecs [Grewin, 1993]. Subjects were presented with three stimuli, R, A and B, corresponding to the reference, the test stimulus and a hidden reference stimulus[3]. Test stimuli were presented to each subject in a different (random) order. The hidden reference was randomly assigned to button A or B. Our test program automatically kept track in an output log file of the presentation order and the ratings given respectively to the stimulus and the hidden reference signal for each test.

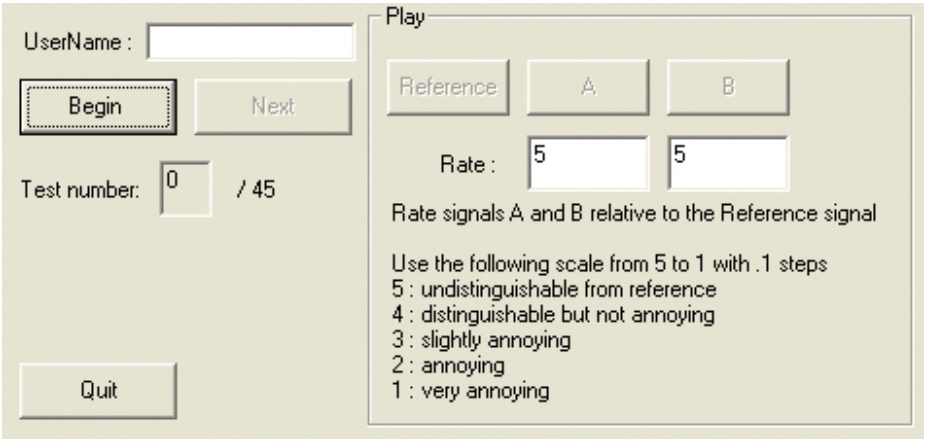

Figure 5.4 Snapshot of the interface designed for our listening tests.

Subjects could switch between the three stimuli at any time during the test by pressing the corresponding buttons on the interface (see Figure 5.4). They were asked to rate differences between each test stimuli (A and B) and the Reference from

[1]The stimuli used for the tests can be found at:
`http://www-sop.inria.fr/reves/projects/scalableAudio/`

[2]International Telecommunication Union

[3] *i.e.,* the subjects did not know which of A or B was the actual test or the reference.

"imperceptible" to "very annoying", using a scale ranging from 5.0 to 1.0 (with one decimal) [ITU-R, 1994].
After the test, subjects were invited, during a semi-guided interview, to describe the differences that they heard between the processed and the original sounds.

5.3.2 Analysis

Correlations between the subjects: All subjects raw judgments were significantly correlated ($p < 0.01$) except for one who was removed from further analysis. After removing this subject, the correlation coefficients ranged from 0.38 to 0.92.
Analysis of variance: A three-way analysis of variance was performed over the judgments (repeated design). Results are given in Table 5.2. The experimental factors affecting the judgments are: $\mathcal{S}$: subjects, $\mathcal{R}$: resolution , $\mathcal{M}$: metric and $\mathcal{T}$: type of signals. All principal effects are significant (resolution: F(2,32)=195.0, p corrected

Source	df	Sum of squares	Mean squares	F-value	p cor.
$\mathcal{S}$	16	139.8	8.7		
$\mathcal{R}$	2	967.7	483.2	195.0	0.000(**)
$\mathcal{S}$*$\mathcal{R}$	32	79.4	2.5		
$\mathcal{M}$	4	23.4	5.8	16.3	0.001(**)
$\mathcal{S}$*$\mathcal{M}$	64	23.0	0.3		
$\mathcal{R}$*$\mathcal{M}$	8	5.6	0.7	1.9	0.191 (ns)
$\mathcal{S}$*$\mathcal{R}$*$\mathcal{M}$	128	48.2	0.4		
$\mathcal{T}$	2	112.7	56.3	41.1	0.000 (**)
$\mathcal{S}$*$\mathcal{T}$	32	43.8	1.4		
$\mathcal{R}$*$\mathcal{T}$	4	27.5	6.9	6.8	0.0191(*)
$\mathcal{S}$*$\mathcal{R}$*$\mathcal{T}$	64	64.7	1.0		
$\mathcal{M}$*$\mathcal{T}$	8	24.5	3.0	8.6	0.010(**)
$\mathcal{S}$*$\mathcal{M}$*$\mathcal{T}$	128	45.5	0.4		
$\mathcal{R}$*$\mathcal{M}$ *$\mathcal{T}$	16	17.2	1.07	2.8	0.115(ns)
$\mathcal{S}$*$\mathcal{R}$*$\mathcal{M}$*$\mathcal{T}$	256	99.1	0.4		
Total		1722.0	2.2		

df: degree of freedom
p cor.: corrected probability (conservative F-test)
* p<0.05; ** p<0.01; ns: not significant

Table 5.2 Anova table for the subjective evaluation

< 0.01; metric: F(4,64)=16.3, p corrected < 0.01; type of signal: F(2,32)=41.1, p

corrected < 0.00). Only the interactions between between *metric* and *type of signals* is significant at the lower threshold ($F(8,128)=8.6$ p corrected < 0.01). The principal effects of the experimental factors are depicted on Figure 5.5 (vertical bars represent the standard deviation).

The bottom graph in this figure clearly shows the effect of the resolution on the average judgments: when 50% of the data are kept, average judgments lay between 4 and 5. This indicates that subjects were almost unable to catch any difference between the processed sound and the original sound. For 25% resolution, average judgments fall between 3 and 4, and slide down to less than 2 for a resolution of 12.5%. The top graph in the figure indicates that musical signals were, on average, ranked higher than the other type of signals (subjects freely mention during post-experimental interviews that differences were harder to notice for musical sounds). This indicates that the alterations of the signal produced by the algorithm are less perceptible for musical sounds. The middle graph in the figure represents the effect of the metric on the average judgments. Results were not quite as pronounced, but a first conclusion is that the A-weighting metric leads to the most audible difference between the processed and original sounds (average judgments are weaker). Further understanding is obtained by studying the significant interaction between metric and type of signal, depicted in Figure 5.6.

The patterns of effects for the metrics are qualitatively identical for both musical and speech signals: A-weighting leads to the worst results, RMS level and Zwicker's loudness model results in better judgments, Moore's loudness yields to weaker judgments. On the other hand, for ambient sounds, Zwicker's loudness model results in weaker results, whereas Moore's loudness model leads the processed sounds to be better evaluated with reference to the original ones. Although, our evaluation was aimed at a totally different purpose, our results appear to be in agreement with the recent paper by Skovenborg and Nielsen [Skovenborg and Nielsen, 2004]. However, we could not test their two new loudness models, which seem to perform best. This would be an interesting future study to conduct.

5.4 Discussion

Several conclusions can be drawn from this study. First of all, when only 50 % of the original data are used, subjects are almost unable to hear any difference between processed and original mixtures. When only 25 % of the sounds are kept, average judgments lay between 3 and 4 (respectively "slightly annoying" and "perceptible but not annoying"). This indicates that our algorithm can reduce the required number of operations by more than 50 % without dramatically distorting the resulting mixtures

(see Figure 5.7).
Another conclusion is that the judgments seem to be strongly influenced by the type of signal. However, as this variable also integrates several other effects (numbers of signals in the mix, sparseness of the mix, energy distribution in the mix, etc.) it requires further testing.

Nevertheless, differences between original and processed sounds were more difficult to detect for musical sounds. Two hypothesis can be formulated to explain this phenomenon: first of all, due to their nature, musical sounds are more sparse than other sounds. Energy peaks occur at regular rhythmic patterns, and there might be a significant amount of low energy frames between these rhythmic accents. In our example, we estimated the sparseness of our musical mix to be about 25%, which would make it well suited to our algorithm. However, the speech mixture was found to be much sparser that the ambient mixture (33% vs. less than 1%) although the results for these two cases were rather similar.

Another hypothesis is that the metrics were in general better suited for musical sounds.
Zwicker's loudness model is more suitable for speech and musical sounds, while Moore's loudness model is more suitable for ambient sounds. This is consistent with our implementation of Moore's loudness model for noisy signals (it can be reasonably assumed that ambient sounds are noisier than musical sounds).
These conclusions were confirmed during the subject interviews. Many subjects reported that they used different criteria for the different types of signals. For speech signals, they reported to produce favorable judgments as long as the intelligibility was preserved, although most of the mixture was foreign language to them. For musical sounds, they did not hear any difference until the sounds were drastically distorted. Finally, for ambient sounds, they seem to have performed some kind of "spectral listening"; a typical remark being: "I tried to notice if there was more or less bass/treble". Hence, we can conclude that no metric seems to perform best in all cases but, rather, that the importance metric has to be adapted to the type of signal.

5.5 Conclusions

We have presented an approach for coarse-grain selective processing of audio signals. Several level-related metrics that can be used to drive the selection process were compared, showing significant difference between the various metrics in terms of the ordering induced on the signals. A pilot subjective evaluation study suggests that A-weighting does not perform as well as the other metrics at prioritizing the sound signals. While RMS level appears as a good compromise, other metrics, loudness in particular, can yield better results depending on the type of signals. Our selective

processing approach integrates well within standard audio processing pipelines and can be used to reduce the necessary operations by 50% being almost transparent and 75% with reasonable impact on the perceived quality.

As future work, we would like to explore extensions to finer-grain processing by combining our selection scheme with parametric audio coders or alternate representations for audio signals. Although the ordering produced by the metrics might be the same, the resulting priority remains different (especially due to dynamic compression performed by the various metrics) and its influence might grow as the selection can be performed at finer levels.

We believe that proposing and evaluating more sophisticated priority metrics is of primary interest for a wide range of applications including memory/resource management (e.g. 3D hardware voices, streaming from main storage space), real-time masking evaluation [Tsingos et al., 2004], on-the-fly multi-track mixing [Pachet and Delerue, 2000], dynamic coding and transmission of spatial audio content and more generally for computational auditory scene analysis.

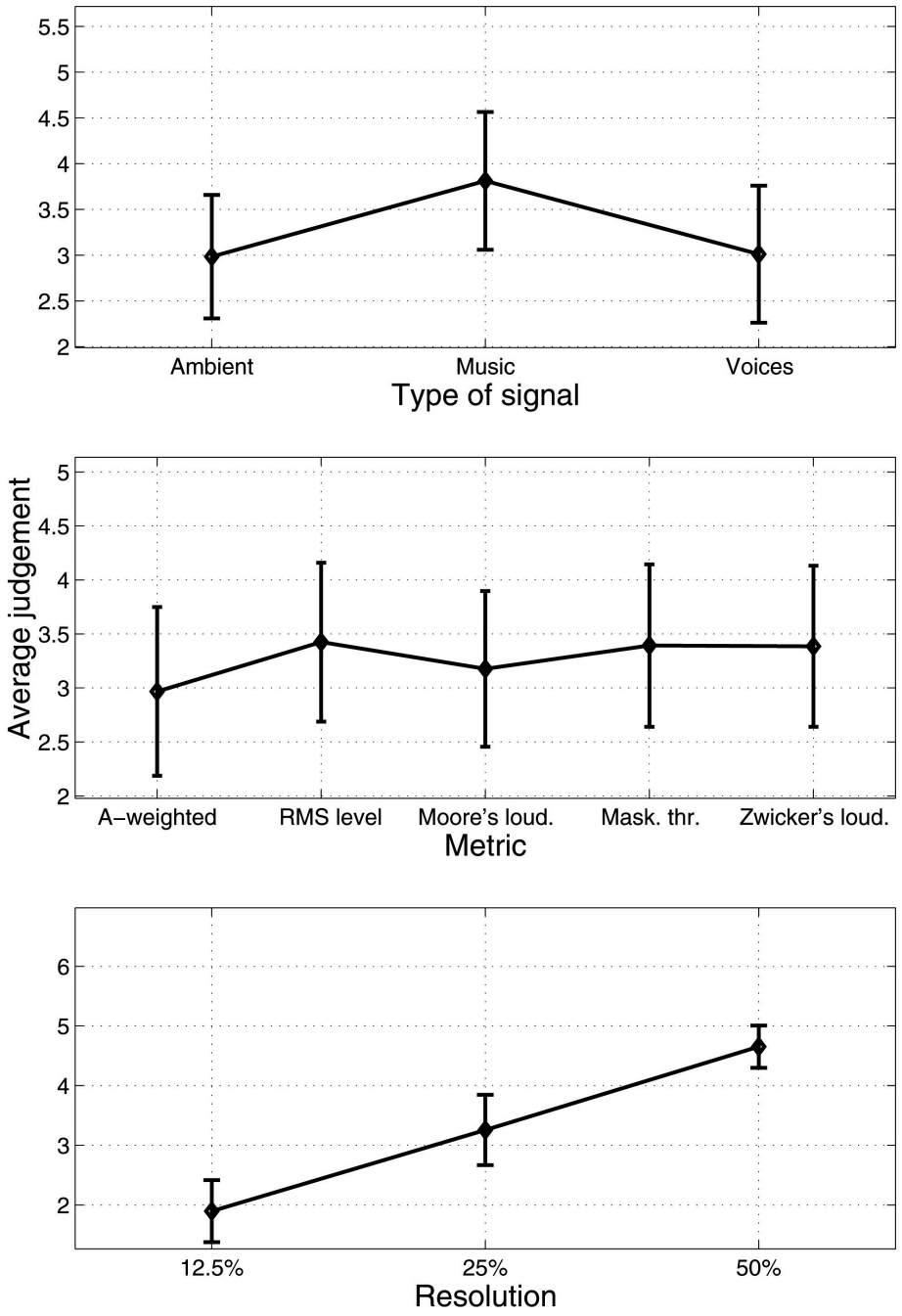

Figure 5.5 Principal effects of the experimental factors. Vertical bars represent standard deviation.

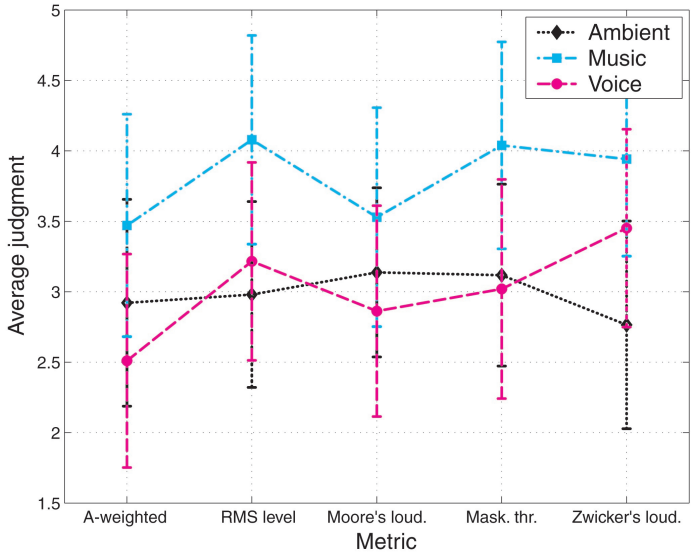

Figure 5.6 Interactions between the effects of the metric and the type of signal on the average judgments.

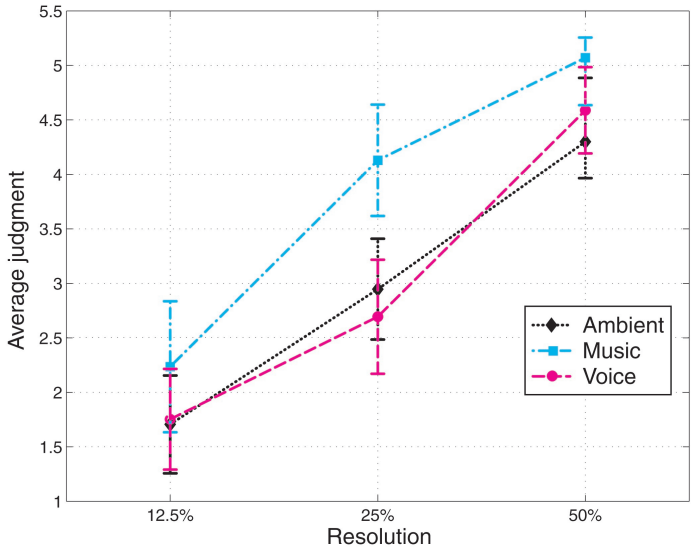

Figure 5.7 Averaged judgments for the three test mixtures and for three levels of detail. When only 50% of the input audio data was used, the resulting mixture was highly rated regardless of the stimuli.

In this first part, we have introduced new methods to accelerate audio rendering. Two complementary approaches have been discribed. In the first chapter, the audio rendering was efficiently performed exploiting parallel computing resources. In the second chapter, audio rendering was optimized using properties of human perception. We studied various priority metrics that can be used to progressively select sub-parts of a number of audio signals for real-time processing. We conducted both an objective and a pilot subjective evaluation study aimed at evaluating which metric would perform best at reconstructing mixtures of various types using only a budget amount of original audio data. RMS level as a metric offers a good compromise for all cases. Finally, our results show that significant sub-parts of the original audio data can be omitted in most cases, without noticeable degradation in the generated mixtures, which validates the usability of our selective processing approach for real-time applications.

Part III

Authoring and Re-Rendering from Field Recordings

Chapter 6

Segmenting and Re-Rendering Field-Recordings

While hardware capabilities allow for real-time rendering of increasingly complex environments, authoring realistic virtual audio-visual worlds is still a challenging task. This is particularly true for interactive spatial auditory scenes for which few content creation tools are available.

Current models for authoring interactive 3D-audio scenes often assume that sound is emitted by a set of monophonic point sources for which a signal has to be individually generated. In the general case, source signals cannot be completely synthesized from physics-based models and must be individually recorded, which requires enormous time and resources. Although this approach gives the user the freedom to control each source and freely navigate throughout the auditory scene, the overall result remains an approximation due to the complexity of real-world sources, limitations of microphone pick-up patterns and limitations of the simulated sound propagation models.

On the opposite end of the spectrum, spatial sound recordings which encode directional components of the sound-field can be directly used to capture live auditory environments as a whole [Malham and Myatt, 1995, Soundfield, 2007]. They produce lifelike results but offer little control, if any, at the playback end. In particular, they are acquired from a single location in space, which makes them insufficient for walk-through applications or rendering of large near-field sources. In practice, for virtual reality applications, their use is mostly limited to the rendering of an overall ambiance. Besides, since no explicit position information is directly available for the sound sources, it is difficult to tightly couple such spatial recordings with matching visuals.

This chapter presents a novel analysis-synthesis approach which bridges the two previous strategies. Our method builds a higher-level spatial description of the audi-

tory scene from a set of field recordings. By analyzing how different frequency components of the recordings reach the various microphones through time, it extracts both spatial information and audio content for the most significant sound events present in the acquired environment. This spatial mapping of the auditory scene can then be used for post-processing and re-rendering the original recordings. Re-rendering is achieved through a frequency-dependent warping of the recordings, based on the estimated positions of several frequency subbands of the signal. Our approach makes positional information about the sound sources directly available for generic 3D-audio processing and integration with 2D or 3D visual content. It also provides a compact encoding of complex live auditory environments and captures complex propagation and reverberation effects which would be very difficult to render with the same level of realism using traditional virtual acoustics simulations.

Our work complements image-based modeling and rendering approaches in computer graphics [Chen and Williams, 1993,Horry et al., 1997,Buehler et al., 2001,Aliaga and Carlbom, 2001]. Moreover, similar to the *matting* and *compositing* techniques widely used in visual effects production [Porter and Duff, 1984], we show that the various auditory components segmented out by our approach can be pasted together to create novel and spatially consistent soundscapes. For instance, foreground sounds can be integrated in a different background ambiance.

Our technique opens many interesting possibilities for interactive 3D applications such as games, virtual/augmented reality or off-line post-production. We demonstrate its applicability to a variety of situations using different microphone setups.

6.1 Related Work

Our approach builds upon prior work in several domains including spatial audio acquisition and reproduction, structure extraction from audio recordings and blind source separation. A fundamental difference between the approaches is whether they attempt to capture the spatial structure of the wavefield through mathematical or physical models or attempt to perform a higher-level auditory scene analysis to retrieve the various, perceptually meaningful, sub-components of the scene and their 3D location. The following sections give a short overview of the background most relevant to our problem.

6.1.1 Spatial Sound-Field Acquisition and Reproduction

Processing and compositing live multi-track recordings is of course a widely used method in motion-picture audio production [Yewdall, 2003]. For instance, recording a scene from different angles with different microphones allows the sound editor to

render different audio perspectives, as required by the visual action. Thus, producing synchronized sound-effects for films requires carefully planned microphone placement so that the resulting audio track perfectly matches the visual action. This is especially true since the required audio material might be recorded at different times and places, before, during and after the actual shooting of the action on stage. Usually, simultaneous monaural or stereophonic recordings of the scene are composited by hand by the sound designer or editor to yield the desired track, limiting this approach to off-line post-production. Surround recording setups (e.g., *Surround Decca Trees*) [Streicher, 2003,Everest, 1998], which historically evolved from stereo recording, can also be used for acquiring a sound-field suitable for reproduction in typical cinema-like setups (e.g., 5.1-surround). However, such recordings can only be played-back directly and do not support spatial post-editing.

Other approaches, more physically and mathematically grounded, decompose the wavefield incident on the recording location on a basis of spatial harmonic functions such as spherical/cylindrical harmonics (e.g., *Ambisonics*) [Gerzon, 1985,Malham and Myatt, 1995, Daniel et al., 1998, Leese, 1998, Merimaa, 2002] or generalized Fourier-Bessel functions [Laborie et al., 2003]. Such representations can be further manipulated and decoded over a variety of listening setups. For instance, they can be easily rotated in 3D space to follow the listener's head orientation and have been successfully used in immersive virtual reality applications. They also allow for beamforming applications, where sounds emanating from any specified direction can be further isolated and manipulated. However, these techniques are practical mostly for low order decompositions (order 2 already requiring 9 audio channels) and, in return, suffer from limited directional accuracy [Jot et al., 1999]. Most of them also require specific microphones [Abhayapala and Ward, 2002, Meyer and Elko, 2004, Soundfield, 2007,Laborie et al., 2004] which are not widely available and whose bandwidth usually drops when the spatial resolution increases. Hence, higher-order microphones do not usually deliver production-grade audio quality, perhaps with the exception of *Trinnov's SRP* system [Laborie et al., 2004] (www.trinnov.com) which uses regular studio microphones but is dedicated to 5.1-surround reproduction. Finally, a common limitation of these approaches is that they use coincident recordings which are not suited to rendering walkthroughs in larger environments. Closely related to the previous approach is wave-field synthesis/holophony [Berkhout et al., 1993, Boone et al., 1995]. Holophony uses the Fresnel-Kirchoff integral representation to sample the sound-field inside a region of space. Holophony could be used to acquire live environments but would require a large number of microphones to avoid aliasing problems, which would jeopardize proper localization of the reproduced sources. In practice, this approach can only capture a live auditory scene through small acoustic "windows". In contrast, while not providing a physically-accurate reconstruction of the soundfield, our approach can provide stable localization cues regardless of the frequency and number

of microphones.

Finally, some authors, inspired from work in computer graphics and vision, proposed a dense sampling and interpolation of the *plenacoustic function* [Ajdler and Vetterli, 2002, Do, 2004] in the manner of *lumigraphs* [Gortler et al., 1996, Levoy and Hanrahan, 1996, Aliaga and Carlbom, 2001]. However, these approaches remain mostly theoretical due to the required spatial density of recordings. Such interpolation approaches have also been applied to measurement and rendering of reverberation filters [Pellegrini, 1999, Horbach et al., 1999]. Our approach follows the idea of acquiring the plenacoustic function using only a sparse sampling and then warping between these samples interactively, e.g., during a walkthrough. In this sense, it could be seen as an "unstructured plenacoustic rendering".

6.1.2 High-Level Auditory Scene Analysis

A second large family of approaches aims at identifying and manipulating the components of the sound-field at a higher-level by performing auditory scene analysis [Bregman, 1990]. This usually involves extracting spatial information about the sound sources and segmenting out their respective content.

Spatial Feature Extraction and Reproduction

Some approaches extract spatial features such as binaural cues (interaural time-difference, interaural level difference, interaural correlation) in several frequency subbands of stereo or surround recordings. A major application of these techniques is efficient multi-channel audio compression [Baumgarte and Faller, 2003, Faller and Baumgarte, 2003] by applying the previously extracted binaural cues to a monophonic down-mix of the original content. However, extracting binaural cues from recordings requires an implicit knowledge of the reproduction system.

Similar principles have also been applied to flexible rendering of directional reverberation effects [Mungamuru and Aarabi, 2004] and analysis of room responses [Merimaa, 2002] by extracting direction of arrival information from coincident or near-coincident microphone arrays [Pulkki, 2006].

This chapter generalizes these approaches to multi-channel field recordings using arbitrary microphone setups and no *a priori* knowledge of the reproduction system. We propose a direct extraction of the 3D position of the sound sources rather than binaural cues or direction of arrival.

Blind Source Separation

Another large area of related research is blind source separation (BSS) which aims at separating the various sources from one or several mixtures under various mixing models [Vincent et al., 2003, O'Grady et al., 2005]. Most recent BSS approaches rely on a sparse signal representation in some space of basis functions which minimizes the probability that a high-energy coefficient at any time-instant belongs to more than one source [Rickard, 2006]. Some work has shown that such sparse coding does exists at the cortex level for sensory coding [Lewicki, 2002]. Several techniques have been proposed such as independent component analysis (ICA) [Comon, 1994] or the more recent *DUET* technique [Jourjine et al., 2000, Yilmaz and Rickard, 2004] which can extract several sources from a stereophonic signal by building an inter-channel delay/amplitude histogram in Fourier frequency domain. In this aspect, it closely resembles the aforementioned binaural cue coding approach. However, most BSS approaches do not separate sources based on spatial cues, but directly solve for the different source signals assuming *a priori* mixing models which are often simple. Our context would be very challenging for such techniques which might require knowing the number of sources to extract in advance, or need more sensors than sources in order to explicitly separate the desired signals. In practice, most auditory BSS techniques are devoted to separation of speech signals for telecommunication applications but other audio applications include up-mixing from stereo to 5.1 surround formats [Avendano, 2003].

In this work, however, our primary goal is not to finely segment each source present in the recorded mixtures but rather to extract enough spatial information so that we can modify and re-render the acquired environment while preserving most of its original content. Closer in spirit, the DUET technique has also been used for audio interpolation [Radke and Rickard, 2002]. Using a pair of closely spaced microphones, the authors apply DUET to re-render the scene at arbitrary locations along the line passing through the microphones. The present work extends this approach to arbitrary microphone arrays and re-rendering at any 3D location in space.

6.2 Overview

We present a novel acquisition and 3D-audio rendering pipeline for modeling and processing realistic virtual auditory environments from real-world recordings.

We propose to record a real-world soundscape using arbitrarily placed omnidirectional microphones in order to get a good acoustic sampling from a variety of locations within the environment. Contrary to most related approaches, we use widely-spaced microphone arrays. Any studio microphones can be used for this purpose, which

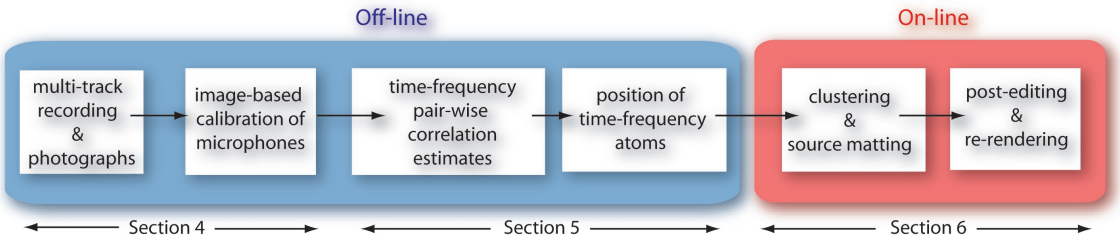

Figure 6.1 Overview of our pipeline. In an off-line phase, we first analyze multi-track recordings of a real-world environment to extract the location of various frequency subcomponents through time. At run-time, we aggregate these estimates into a target number of clustered sound sources for which we reconstruct a corresponding signal. These sources can then be freely post-edited and re-rendered.

makes the approach well suited to production environments. We also propose an image-based calibration strategy making the approach practical for field applications. The obtained set of recordings is analyzed in an off-line pre-processing step in order to segment various auditory components and associate them with the position in space from which they were emitted. To compute this spatial mapping, we split the signal into short time-frames and a set of frequency subbands. We then use classical time-difference of arrival estimation techniques between all pairs of microphones to retrieve a position for each subband at each time-frame. We evaluate the performance of existing approaches in our context and present an improved hierarchical source localization technique from the obtained time-differences.

This high-level representation allows for flexible and efficient on-line re-rendering of the acquired scene, independent of the reproduction system. At run-time during an interactive simulation, we use the previously computed spatial mapping to properly warp the original recordings when the virtual listener moves throughout the environment. With an additional clustering step, we recombine frequency subbands emitted from neighboring locations and segment spatially-consistent sound events. This allows us to select and post-edit subsets of the acquired auditory environment. Finally the location of the clusters is used for spatial audio reproduction within standard 3D-audio APIs.

Figure 6.1 shows an overview of our pipeline. Sections 6.3, 6.4 and 6.5 describe our acquisition and spatial analysis phase in more detail. Section 6.6 presents the on-line spatial audio resynthesis based on the previously obtained spatial mapping of the auditory scene. Finally, Section 6.7 describes several applications of our approach to realistic rendering, post-editing and compositing of real-world soundscapes.

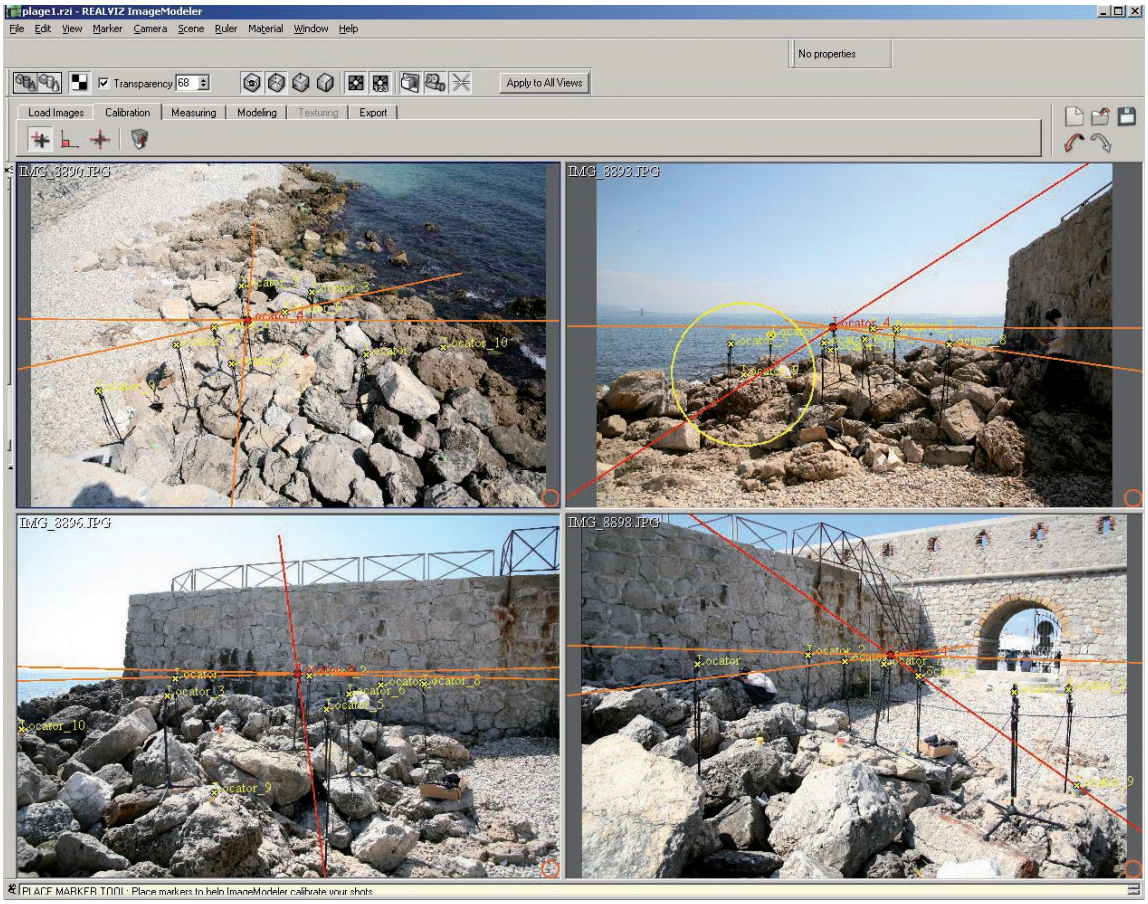

Figure 6.2 We retrieve the position of the microphones from several photographs of the setup using a commercial image-based modeling tool. In this picture, we show four views of a recording setup, position of the markers and the triangulation process yielding the locations of the microphone capsules.

6.3 Recording Setup and Calibration

We acquire real-world soundscapes using a number of omnidirectional microphones and a multi-channel recording interface connected to a laptop computer. In our examples, we used up to 8 identical *AudioTechnica AT3032* microphones and a *Presonus Firepod* firewire interface running on batteries. The microphones can be arbitrarily positioned in the environment. Section 6.7 shows various possible setups. To produce the best results, the microphones should be placed so as to provide a compromise between the signal-to-noise ratio of the significant sources and spatial coverage.

In order to extract correct spatial information from the recordings, it is necessary

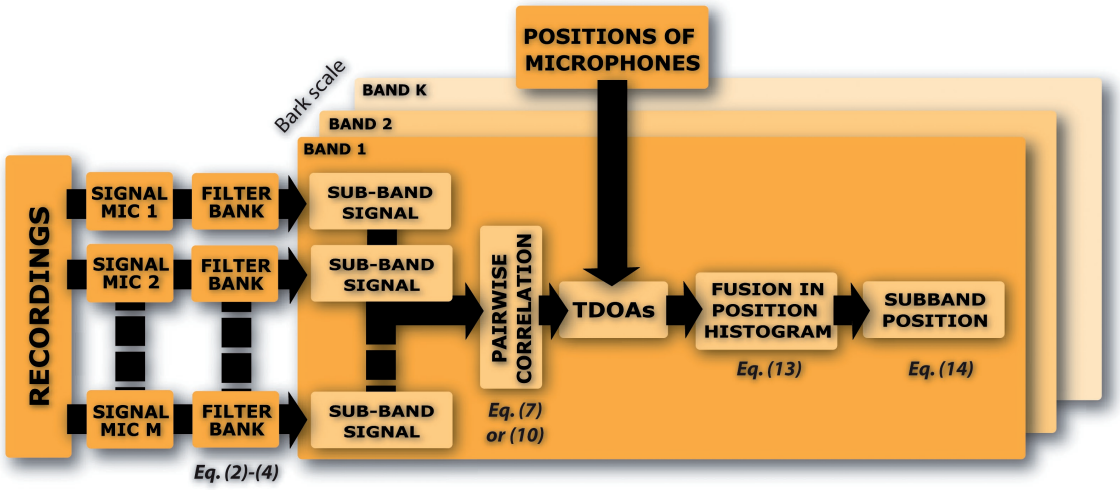

Figure 6.3 Overview of the analysis algorithm used to construct a spatial mapping for the acquired soundscapes.

to first retrieve the 3D locations of the microphones. Maximum-likelihood autocalibration methods could be used based on the existence of pre-defined source signals in the scene [Moses et al., 2002], for which the time-of-arrival (TOA) to each microphone has to be determined. However, it is not always possible to introduce calibration signals at a proper level in the environment. Hence, in noisy environments obtaining the required TOAs might be difficult, if not impossible. Rather, we use an image-based technique from photographs which ensures fast and convenient acquisition on location, not requiring any physical measurements or homing device. Moreover, since it is not based on acoustic measurements, it is not subject to background noise and is likely to produce better results. We use REALVIZ *ImageModeler* (www.realviz.com) to extract the 3D locations from a small set of photographs (4 to 8 in our test examples) taken from several angles, but any standard algorithm can be applied for this step [Faugeras, 1993]. To facilitate the process we place colored markers (tape or balls of modeling clay) on the microphones, as close as possible to the actual location of the capsule, and on the microphone stands. Additional markers can also be placed throughout the environment to obtain more input data for calibration. The only constraint is to provide a number of non-coplanar calibration points to avoid degenerate cases in the process. In our test examples, the accuracy of the obtained microphone locations was of the order of one centimeter. Image-based calibration of the recording setup is a key aspect of our approach since it allows for treating complex field recording situations such as the one depicted in Figure 6.2 where microphones stands are placed on large irregular rocks on a seashore.

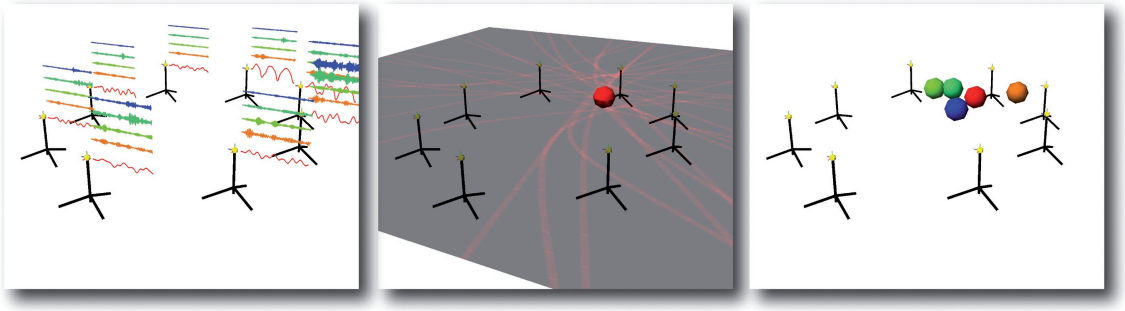

Figure 6.4 Illustration of the construction of the global spatial mapping for the captured sound-field. (a) At each time-frame, we split the signals recorded by each microphone into the same set of frequency subbands. (b) Based on time-difference of arrival estimation between all pairs of recordings, we sample all corresponding hyperbolic loci to obtain a position estimate for the considered subband. (c) Position estimates for all subbands at the considered time-frame (shown as colored spheres).

6.4 Propagation model and assumptions for source matting

From the M recorded signals, our final goal is to localize and re-render a number J of *representative sources* which offer a good perceptual reconstruction of the original soundscape captured by the microphone array. Our approach is based on two main assumptions.

First, we consider that the recorded sources can be represented as point emitters and assume an ideal anechoic propagation model. In this case, the mixture $x_m(t)$ of N sources $s_1(t), .., s_n(t)$ recorded by the m^{th} microphone can be expressed as:

$$x_m(t) = \sum_{n=1}^{N} a_{mn}(t) s_n(t - \delta_{mn}(t)), \tag{6.1}$$

where parameters $a_{mn}(t)$ and $\delta_{mn}(t)$ are the attenuation coefficients and time-delays associated with the n^{th} source and the m^{th} microphone.

Second, since our environments contain more than one active source simultaneously, we consider K frequency subbands, $K >= J$, as the basic components we wish to position in space at each time-frame (Figure 6.4 (a)). We choose to use non-overlapping frequency subbands uniformly defined on a Bark scale [Moore et al., 1997] to provide a more psycho-acoustically relevant subdivision of the audible spectrum (in our examples, we experimented with 1 to 32 subbands).

In frequency domain, the signal x_m filtered in the k^{th} Bark band can be expressed at each time-frame as:

$$Y_{km}(z) = W_k(z) \sum_{t=1}^{T} x_m(t) e^{-j(2\pi z t/T)} = W_k(z) X_m(z), \tag{6.2}$$

where

$$W_k(f) = \begin{cases} 1 & \frac{25k}{K} < Bark(f) < \frac{25(k+1)}{K} \\ 0 & \text{otherwise} \end{cases} \tag{6.3}$$

$$Bark(f) = 13atan(\frac{0.76f}{1000}) + 3.5atan(\frac{f^2}{7500^2}), \tag{6.4}$$

$f = z/Zf_s$ is the frequency in Hertz, f_s is the sampling rate and $X_m(z)$ is the $2Z$-point Fourier transform of $x_m(t)$. We typically record our live signals using 24-bit quantization and $f_s = 44.1KHz$. The subband signals are computed using $Z = 512$ with a Hann window and 50% overlap before storing them back into time-domain for later use.

At each time-frame, we construct a new representation for the captured soundfield at an arbitrary listening point as:

$$\hat{x}(t) \approx \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{\alpha}_{km}^{j} y_{km}(t + \hat{\delta}_{km}), \forall m \tag{6.5}$$

where $y_{km}(t)$ is the inverse Fourier transform of $Y_{km}(z)$, $\hat{\alpha}_{km}^{j}$ and $\hat{\delta}_{km}$ are correction terms for attenuation and time-delay derived from the estimated positions of the different subbands. The term $\hat{\alpha}_{km}^{j}$ also includes a matting coefficient representing how much energy within each frequency subband should belong to each representative source. In this sense, it shares some similarity with the *time-frequency masking* approach of [Yilmaz and Rickard, 2004].

The obtained representation can be made to match the acquired environment if $K >= N$ and if, following a sparse coding hypothesis, we further assume that the contents of each frequency subband belong to a single source at each time-frame. This hypothesis is usually referred to as *W-disjoint orthogonality* [Yilmaz and Rickard, 2004] and given N sources $S_1, .., S_N$ in Fourier domain, it can be expressed as:

$$S_i(z) S_j(z) = 0 \quad \forall i \neq j \tag{6.6}$$

When the two previous conditions are not satisfied, the representative sources will correspond to a mixture of the original sources and Equ. 6.5 will lead to a less accurate approximation.

6.5 Spatial Mapping of the Auditory Scene

In this step of our pipeline, we analyze the recordings in order to produce a high-level representation of the captured soundscape. This high-level representation is a mapping, global to the scene, between different frequency subbands of the recordings and positions in space from which they were emitted (Figure 6.4).

Following our previous assumptions, we consider each frequency subband as a unique point source for which a single position has to be determined. The localization of a sound source from a set of audio recordings, using a single-propagation-path model, is a well studied problem with major applications in robotics, people tracking and sensing, teleconferencing (e.g, automatic camera steering) and defense. Approaches rely either on time-difference of arrival (TDOA) estimates [Aarabi, 2003, Knapp and Carter, 1976, Huang et al., 2000], high-resolution spectral estimation (e.g., MUSIC) [Schmid, 1986, Krim and Viberg, 1996] or steered response power using a beamforming strategy [DiBiase et al., 2001, Chen et al., 2003, Mungamuru and Aarabi, 2004]. In our case, the use of freely positioned microphones, which may be widely spaced, prevents from using a beamforming strategy. Besides, such an approach would only lead to direction of arrival information and not a 3D position (unless several beamforming arrays were used simultaneously). In our context, we chose to use a TDOA strategy to determine the location of the various auditory events. Since we do not know the directivity of the sound sources nor the response of the microphones, localization based on level difference cannot be applied.

Figure 6.3 details the various stage of our source localization pipeline.

6.5.1 Time-Frequency Correlation Analysis

Analysis of the recordings is done on a frame by frame basis using short time-windows (typically 20ms long or 1024 samples at 44.1 kHz). For a given source position and a given pair of microphones, the propagation delay from the source to the microphones generates a measurable time-difference of arrival. The set of points which generate the same TDOA defines an hyperboloid surface in 3D (or an hyperbola in 2D) whose foci are the locations of the two microphones (Figure 6.4 (b)).

In our case, we estimate the TDOAs, $\hat{\tau}_{mn}$, between pairs of microphones $< m, n >$ in each frequency subband k using standard generalized cross-correlation (GCC) techniques in the frequency domain [Knapp and Carter, 1976, Rabinkin et al., 1996, Chen et al., 2006]:

$$\hat{\tau}_{mn} = \arg\max_{\tau} GCC_{mn}(\tau), \tag{6.7}$$

where the GCC function is defined as:

$$GCC_{nm}(\tau) = \sum_{z=1}^{Z} \psi_{nm}(z) E\left\{Y_{kn}(z) Y_{km}^{*}(z)\right\} e^{j(2\pi\tau z/Z)}. \tag{6.8}$$

Y_{kn} and Y_{km} are the $2Z$-point Fourier transforms of the subband signals (see Eq. 6.2), $E\left\{Y_{kn}(z) Y_{km}^{*}(z)\right\}$ is the cross spectrum, $\tau <$ window size and $.^{*}$ denotes the complex conjugate operator.

For the weighting function, ψ, we use the PHAT weighting which was shown to give better results in reverberant environments [Chen et al., 2006]:

$$\psi_{mn}(z) = \frac{1}{|Y_n(z) Y_m^{*}(z)|} \tag{6.9}$$

Note that phase differences computed directly on the Fourier transforms, e.g. as used in the DUET technique [Jourjine et al., 2000, Yilmaz and Rickard, 2004], cannot be applied in our framework since our microphones are widely spaced.

We also experimented with an alternative approach based on the average magnitude difference function (AMDF) [Merimaa, 2002, Chen et al., 2005]. The TDOAs are then given as:

$$\hat{\tau}_{nm} = \arg\min_{\tau} AMDF_{nm}(\tau), \tag{6.10}$$

where the AMDF function is defined as:

$$AMDF_{nm}(\tau) = \frac{1}{Z} \sum_{z=1}^{Z} |y_{kn}(\tau) - y_{km}(k+\tau)| \tag{6.11}$$

We compute the cross-correlation using vectors of 8192 samples (185 ms at 44.1KHz). For each time-frame, we search the highest correlation peaks (or lowest AMDF values) between pairs of recordings in the time-window defined by the spacing between the corresponding couple of microphones. The corresponding time-delay is then chosen as the TDOA between the two microphones for the considered time-frame.

In terms of efficiency, the complexity of AMDF-based TDOA estimation (roughly $O(n^2)$ in the number n of time-domain samples) makes it unpractical for large time-delays. In our test-cases, running on a *Pentium4 Xeon* 3.2GHz processor, AMDF-based TDOA estimations required about 47 s. per subband for one second of input audio data (using 8 recordings, i.e., 28 possible pairs of microphones). In comparison, GCC-based TDOA estimations require only 0.83 s. per subband for each second of recording.

As can be seen in Figure 6.7, the two approaches resulted in comparable subband localization performance and we found both approaches to perform reasonably well in

all our test cases. In more reverberant environments, an alternative approach could be the adaptive eigenvalue decomposition [Huang et al., 2000]. From a perceptual point-of-view, listening to virtual re-renderings, we found that the AMDF-based approach lead to reduced artifacts, which seems to indicate that subband locations are more perceptually valid in this case. However, validation of this aspect would require a more thorough perceptual study.

6.5.2 Position Estimation

From the TDOA estimates, several techniques can be used to estimate the location of the actual sound source. For instance, it can be calculated in a least-square sense by solving a system of equations [Huang et al., 2000] or by aggregating all estimates into a probability distribution function [Rui and Florencio, 2003, Aarabi, 2003]. Solving for possible positions in a least-square sense lead to large errors in our case, mainly due to the presence of multiple sources, several local maxima for each frequency subband resulting in an averaged localization. Rather, we choose the latter solution and compute a histogram corresponding to the probability distribution function by sampling it on a spatial grid (Figure 6.5) whose size is defined according to the extent of the auditory environment we want to capture (in our various examples, the grid covered areas ranging from 25 to 400 m^2). We then pick the maximum value in the histogram to obtain the position of the subband.

For each cell in the grid, we sum a weighted contribution of the distance function $D_{ij}(\mathbf{x})$ to the hyperboloid defined by the TDOA for each pair of microphones $< i, j >$:

$$D_{ij}(\mathbf{x}) = |(||M_i - \mathbf{x}|| - ||M_j - \mathbf{x}||) - DDOA_{ij}|, \tag{6.12}$$

where M_i resp. M_j is the position of microphone i resp. j, $\mathbf{x}$ is the center of the cell and $DDOA_{ij} = TDOA_{ij}/c$ is the signed distance-difference obtained from the calculated TDOA (in seconds) and the speed of sound c.

The final histogram value in each cell is then obtained as :

$$\begin{aligned} H(\mathbf{x}) = \textstyle\sum_{ij} \Big[\frac{e^{(\gamma(1-D_{ij}(\mathbf{x})))}}{e^{\gamma}} (1 - DDOA_{ij}/||M_i - M_j||) \\ \text{if } D_{ij}(\mathbf{x}) < 1,\ 0 \text{ otherwise } \Big] . \end{aligned} \tag{6.13}$$

The exponentially decreasing function controls the "width" of the hyperboloid and provides a tradeoff between localization accuracy and robustness to noise in the TDOA estimates. In our examples, we use $\gamma = 4$. The second weighting term reduces the contribution of large TDOAs relative to the spacing between the pair of microphones. Such large TDOAs lead to "flat" ellipsoids contributing to a large number of neighboring cells in the histogram and resulting into less accurate position estimates [Ajdler et al., 2004].

The histogram is re-computed for each subband at each time-frame based on the corresponding TDOA estimates. The location of the k_{th} subband is finally chosen as the center point of the cell having the maximum value in the probability histogram (Figure 6.4 (c)):

$$B_k = \arg\max_{\mathbf{x}} H(\mathbf{x}) \tag{6.14}$$

In the case where most of the sound sources and microphones are located at similar height in a near planar configuration, the histogram can be computed on a 2D grid. This yields faster results at the expense of some error in localization. A naive calculation of the histogram at each time-frame (for a single frequency band and 8 microphones, i.e., 28 possible hyperboloids) on a 128×128 grid requires 20 milliseconds on a *Pentium4 Xeon* 3.2GHz processor. An identical calculation in 3D requires 2.3 s. on a $128 \times 128 \times 128$ grid. To avoid this extra computation time, we implemented a hierarchical evaluation using a quadtree or octree decomposition [Kalman, 1960]. We recursively test only a few candidate locations (typically 16 to 64), uniformly distributed in each cell, before subdividing the cell in which the maximum of all estimates is found. Our hierarchical localization process supports real-time performance requiring only 5 ms to locate a subband in a $512 \times 512 \times 512$ 3D grid. In terms of accuracy, it was found to be comparable to the direct, non-hierarchical, evaluation at maximum resolution in our test examples.

6.5.3 Indoor Validation Study

To validate our approach, we conducted a test-study using 8 microphones inside a 7m×3.5m×2.5m room with a short reverberation time (about 0.3 sec. at 1KHz). We recorded three people speaking while standing at locations specified by colored markers. Figure 6.6 depicts the corresponding setup. We first evaluated the localization accuracy for all subbands by constructing spatial energy maps of the recordings. As can be seen in Figure 6.7, our approach properly localizes the corresponding sources. In this case, the energy corresponds to the signal captured by a microphone located at the center of the room.

Figure 6.10 shows localization error over all subbands by reference to the three possible positions for the sources. Since we do not know *a priori* which subband belongs to which source, the error is simply computed, for each subband, as the minimum distance between the reconstructed location of the subband and each possible source position. Our approach achieves a maximum accuracy of one centimeter and, on average, the localization accuracy is of the order of 10 centimeters. Maximum errors are of the order of a few meters. However, listening tests exhibit no strong artefacts showing that such errors are likely to occur for frequency subbands containing very little energy. Figure 6.10 also shows the energy of one of the captured

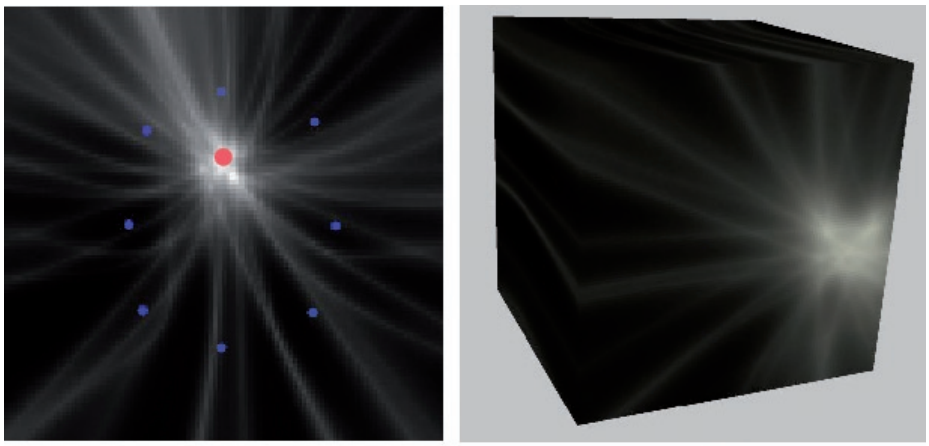

Figure 6.5 (a) A 2D probability histogram for source location obtained by sampling a weighted sum of hyperbolas corresponding to the time-difference of arrival to all microphone pairs (shown in blue). We pick the maximum value (in red) in the histogram as the location of the frequency band at each frame. (b) A cut through a 3D histogram of the same situation obtained by sampling hyperboloid surfaces on a 3D grid.

signals. As can be expected, the overall localization error is also correlated with the energy of the signal.

We also performed informal comparisons between reference binaural recordings and a spatial audio rendering using the obtained locations, as described in the next section. Corresponding audio files can be found at:
`http://www-sop.inria.fr/reves/projects/audioMatting`.

They exhibit good correspondence between the original situation and our renderings showing that we properly assign the subbands to the correct source locations at each time-frame.

6.6 3D-Audio Resynthesis

The final stage of our approach is the spatial audio resynthesis. During a real-time simulation, the previously pre-computed subband positions can be used for re-rendering the acquired sound-field while changing the position of the sources and listener. A key aspect of our approach is to provide a spatial description of a real-world auditory scene in a manner independent of the auditory reproduction system. The scene can thus be re-rendered by standard 3D-audio APIs: in some of our test examples, we used *DirectSound 3D* accelerated by a *CreativeLabs Audigy2 NX* sound-

Figure 6.6 Indoor validation setup using 8 microphones. The 3 markers (see blue, yellow, green arrows) on the ground correspond to the location of the recorded speech signals.

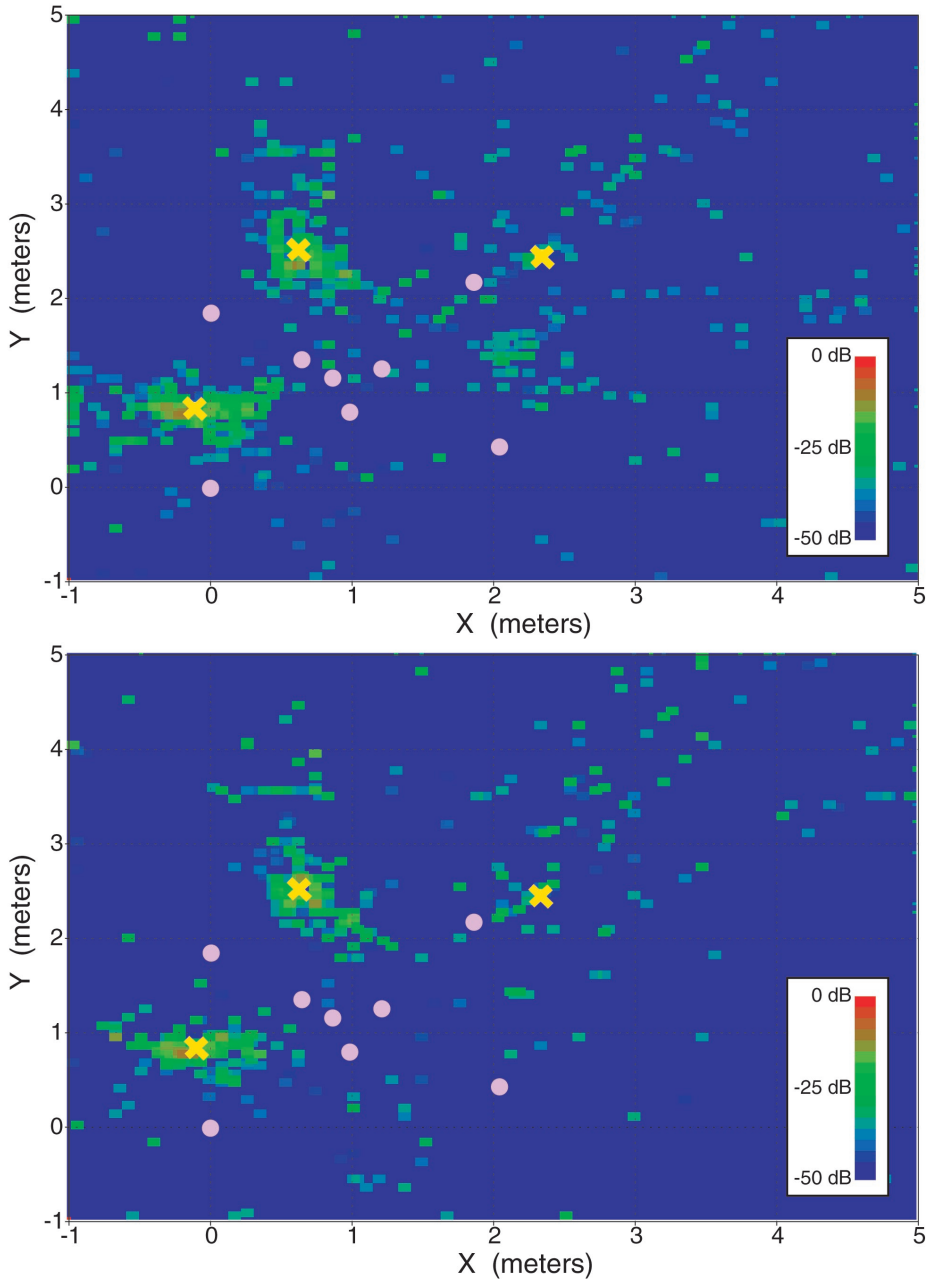

Figure 6.7 Energy localization map for a 28s.-long audio sequence featuring 3 speakers inside a room (indicated by the three yellow crosses). Light-purple dots show the location of the 8 microphones. The top map is computed using AMDF-based TDOA estimation while the bottom map is computed using GCC-PHAT. Both maps were computed using 8 subbands and corresponding energy is integrated over the entire duration of the sequence.

card and also implemented our own software binaural renderer, using head-related transfer function (HRTF) data from the LISTEN HRTF database[1].

Inspired by binaural-cue coding [Faller and Baumgarte, 2003], our re-rendering algorithm can be decomposed in two steps, that we detail in the following sections:

- First, as the virtual listener moves throughout the environment, we construct a *warped monophonic signal* based on the original recording of the microphone closest to the current listening position.
- Second, this warped signal is spatially enhanced using 3D-audio processing based on the location of the different frequency subbands.

These two steps are carried out over small time-frames (of the same size as in the analysis stage). To avoid artefacts we use a 10% overlap to cross-fade successive synthesis frames.

6.6.1 Warping the Original Recordings

For re-rendering, a monophonic signal best matching the current location of the virtual listener relative to the various sources must be synthesized from the original recordings.

At each time-frame, we first locate the microphone closest to the location of the virtual listener. To ensure that we remain as faithful as possible to the original recording, we use the signal captured by this microphone as our reference signal $R(t)$.

We then split this signal into the same frequency subbands used during the off-line analysis stage. Each subband is then warped to the virtual listener location according to the pre-computed spatial mapping at the considered synthesis time-frame (Figure 6.8).

This warping involves correcting the propagation delay and attenuation of the reference signal for the new listening position, according to our propagation model (see Eq.6.1). Assuming an inverse distance attenuation for point emitters, the warped signal $R'_i(t)$ in subband i is thus given as:

$$R'_i(t) = r_1^i / r_2^i R_i(t + (\delta_1^i - \delta_2^i)), \tag{6.15}$$

where r_1^i,δ_1^i are respectively the distance and propagation delay from the considered time-frequency atom to the reference microphone and r_2^i,δ_2^i are the distance and propagation delay to the new listening position.

[1] `http://recherche.ircam.fr/equipes/salles/listen/`

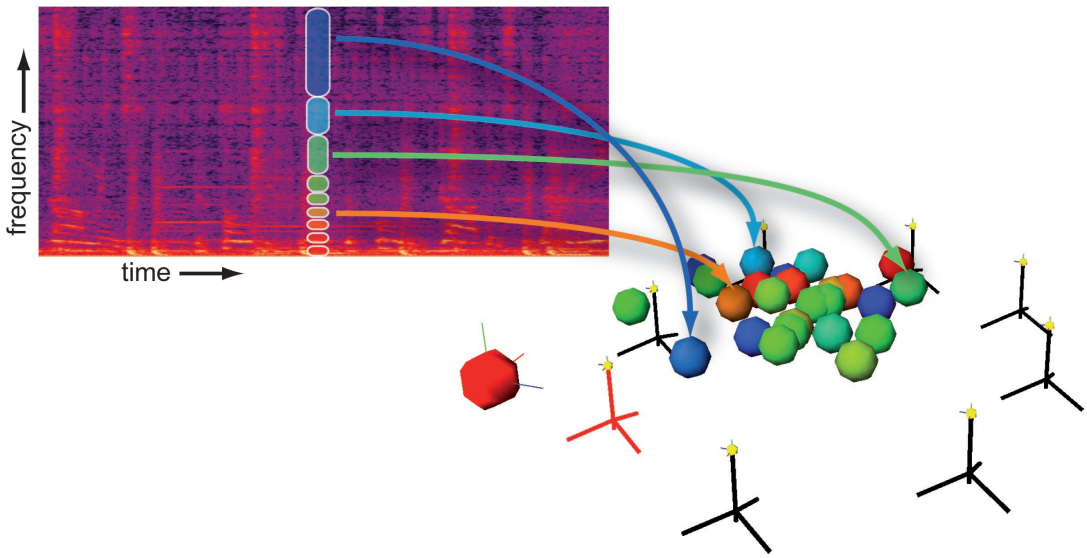

Figure 6.8 In the resynthesis phase, the frequency components of the signal captured by the microphone closest to the location of the virtual listener (shown in red) is warped according to the spatial mapping pre-computed in the off-line stage.

6.6.2 Clustering for 3D-audio Rendering and Source Matting

To spatially enhance the previously obtained warped signals, we run an additional clustering step to aggregate subbands which might be located at nearby positions using the technique of [Tsingos et al., 2004]. The clustering allows to build groups of subbands which can be rendered from a single representative location and might actually belong to the same physical source in the original recordings. Thus, our final rendering stage spatializes N representative point sources corresponding to the N generated clusters, which can vary between 1 and the total number of subbands. To improve the temporal coherence of the approach we use an additional Kalman filtering step on the resulting cluster locations [Kalman, 1960].

With each cluster we associate a weighted sum of all warped signals in each subband which depends on the Euclidean distance between the location of the subband B_i and the location of the cluster representative C_k. This defines matting coefficients α_k, similar to alpha-channels in graphics [Porter and Duff, 1984]:

$$\alpha(C_k, B_i) = \frac{1.0/(\epsilon + ||C_k - B_i||)}{\sum_i \alpha(C_k, B_i)}. \tag{6.16}$$

In our examples, we used $\epsilon = 0.1$. Note that in order to preserve the energy distribu-

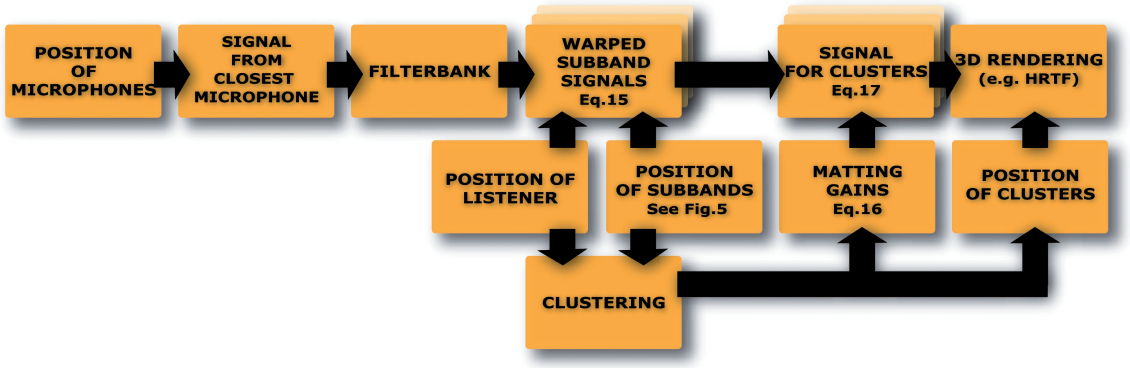

Figure 6.9 Overview of the synthesis algorithm used to re-render the acquired soundscape based on the previously obtained subband positions.

tion, these coefficients are normalized in each frequency subband.

These matting coefficients control the blending of all subbands rendered at each cluster location and help smooth the effects of localization errors. They also ensure a smoother reconstruction when sources are modified or moved around in the re-rendering phase.

The signal for each cluster $S_k(t)$ is finally constructed as a sum of all warped subband signals $R'_i(t)$, as described in the previous section, weighted by the matting coefficients $\alpha(C_k, B_i)$:

$$S_k(t) = \sum_i \alpha(C_k, B_i) R'_i(t). \tag{6.17}$$

The representative location of each cluster is used to apply the desired 3D-audio processing (e.g., HRTFs) without *a priori* knowledge of the reproduction setup.

Figure 6.9 summarizes the complete re-rendering algorithm.

6.7 Applications and Results

Our technique opens many interesting application areas for interactive 3D applications, such as games or virtual/augmented reality, and off-line audio-visual post-production. Several example renderings demonstrating our approach can be found at the following URL:
`http://www-sop.inria.fr/reves/projects/audioMatting`.

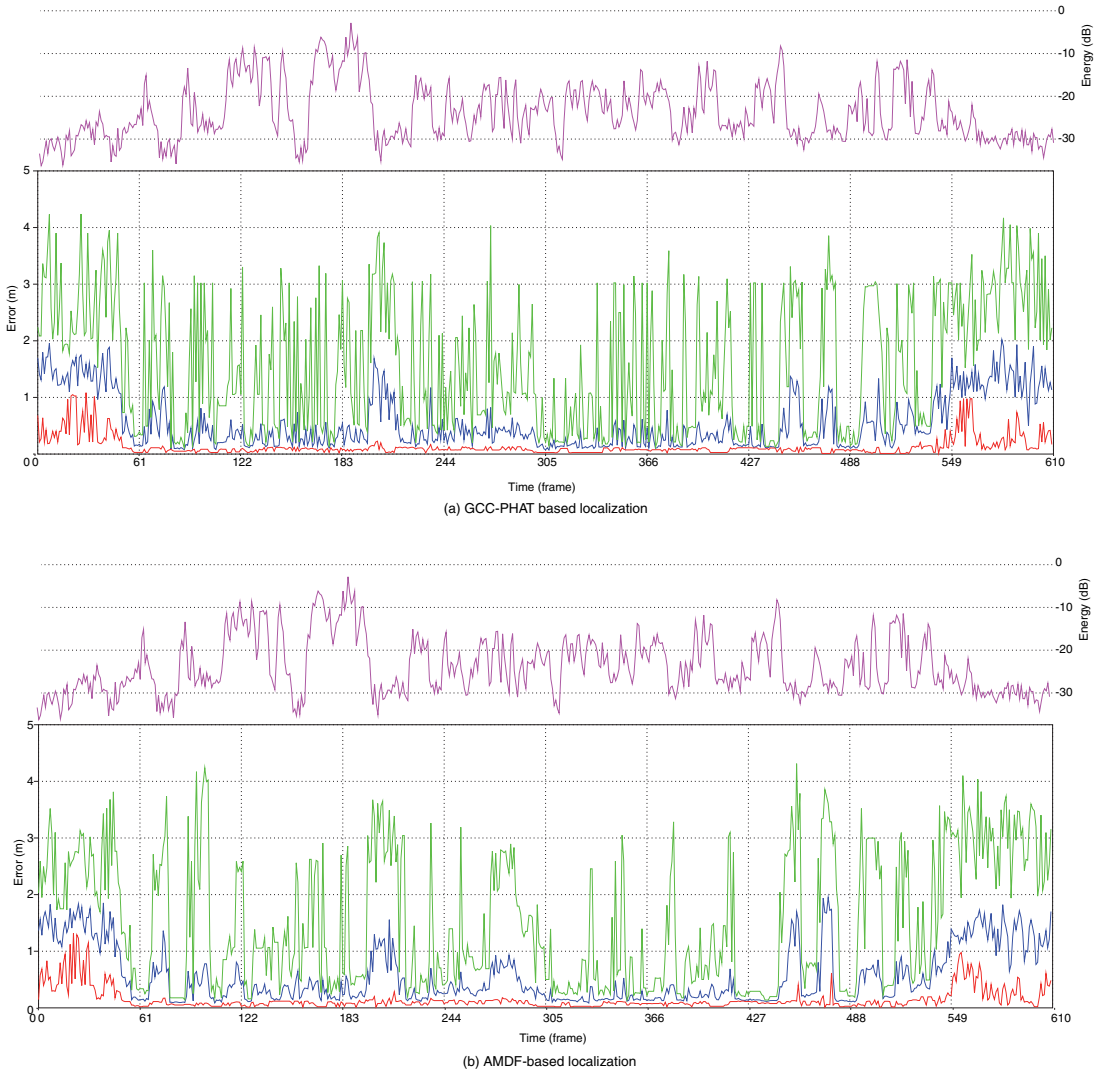

Figure 6.10 Localization error for the same audio sequence as in Figure 6.7. computed over 8 subbands. Averaged error over all subbands is displayed in blue, maximum error in green and minimum error in red. The top (magenta) curve represents the energy for one of the input recordings and shows its correlation with the localization error (clearly larger when the energy drops out).

6.7.1 Modeling Complex Sound Sources

Our approach can be used to render extended sound sources (or small soundscapes) which might be difficult to model using individual point sources because of their complex acoustic behavior. For instance, we recorded a real-world sound scene involving a car which is an extended vibrating sound radiator. Depending on the point of view around the scene, the sound changes significantly due to the relative position of the various mechanical elements (engine, exhaust, etc.) and the effects of sound propagation around the body of the car. This makes an approach using multiple recordings very interesting in order to realistically capture these effects. Unlike other techniques, such as *Ambisonics O-format* [Malham, 2001], our approach captures the position of the various sounding components and not only their directional aspect. In the accompanying examples, we demonstrate a re-rendering with a moving listening point of a car scenario acquired using 8 microphones surrounding the action (Figure 6.11). In this case, we used 4 clusters for re-rendering. Note in the accompanying video available on-line, the realistic distance and propagation effects captured by the recordings, for instance on the door slams. Figure 6.12 shows a corresponding energy map clearly showing the low frequency exhaust noise localized at the rear of the car and the music from the on-board stereo audible through the driver's open window. Engine noise was localized more diffusely mainly due to interference with the music.

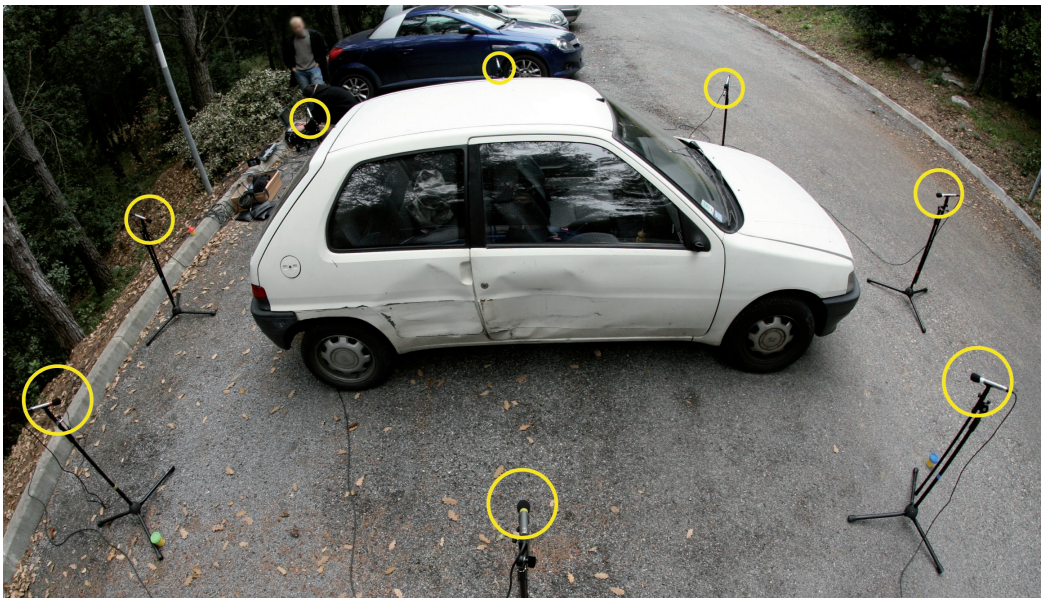

Figure 6.11 We capture an auditory environment featuring a complex sound source (car engine/exhaust, passengers talking, door slams and on-board stereo system) using 8 microphones surrounding the action.

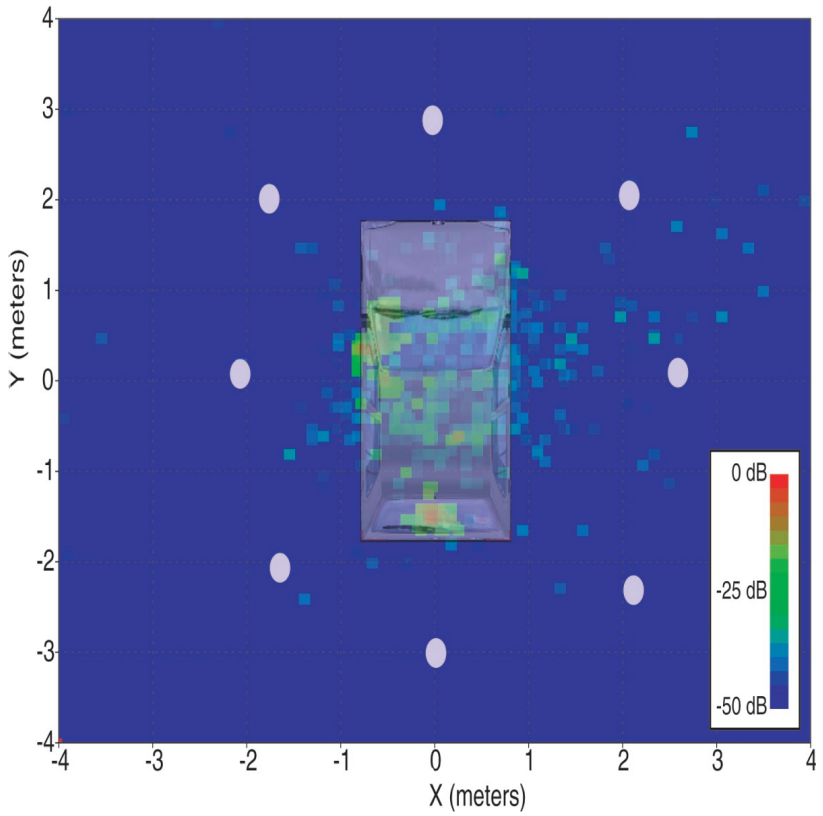

Figure 6.12 Energy localization map for a 15 sec.-long recording of our car scenario featuring engine/exhaust sounds and music (on the on-board stereo system and audible through the open driver-window). Positions were computed over 8 subbands using GCC-PHAT-based TDOA estimation. Energy is integrated over the entire duration of the input audio sequence.

6.7.2 Spatial Recording and View-Interpolation

Following binaural cue coding principles, our approach can be used to efficiently generate high-resolution surround recordings from monophonic signals. To illustrate this application we used 8 omnidirectional microphones located in a circle-like configuration about 1.2 meters in diameter (Figure 6.13) to record three persons talking and the surrounding ambiance (fountain, birds, etc.). Then, our pre-processing was applied to extract the location of the sources. For re-rendering, the monophonic signal of a single microphone was used and respatialized as described in Section 6.6.1, using 4 clusters (Figure 6.15). Please refer to the accompanying video provided on the web site to evaluate the result.

Another advantage of our approach is to allow for re-rendering an acquired audi-

Figure 6.13 Microphone setup used to record the fountain example. In this case the microphones are placed at the center of the action.

tory environment from various listening points. To demonstrate this approach on a larger environment, we recorded two moving speakers in a wide area (about 15×5 meters) using the microphone configuration shown Figure 6.14 (gray dot). The recording also features several background sounds such as traffic and road-work noises. Figure 6.14 shows a corresponding spatial energy map. The two intersecting trajectories of the moving speakers are clearly visible.

Applying our approach, we are able to re-render this auditory scene from any arbitrary viewpoint. Although the rendering is based only on the *monophonic* signal of the microphone closest to the virtual listener at each time-frame, the extracted spatial mapping allows for convincing reproduction of the motion of the sources. Note in the example video provided on the accompanying web-site how we properly capture front-to-back and left-to-right motion for the two moving speakers.

6.7.3 Spatial Audio Compositing and Post-Editing

Finally, our approach allows for post-editing the acquired auditory environments and compositing several recordings.

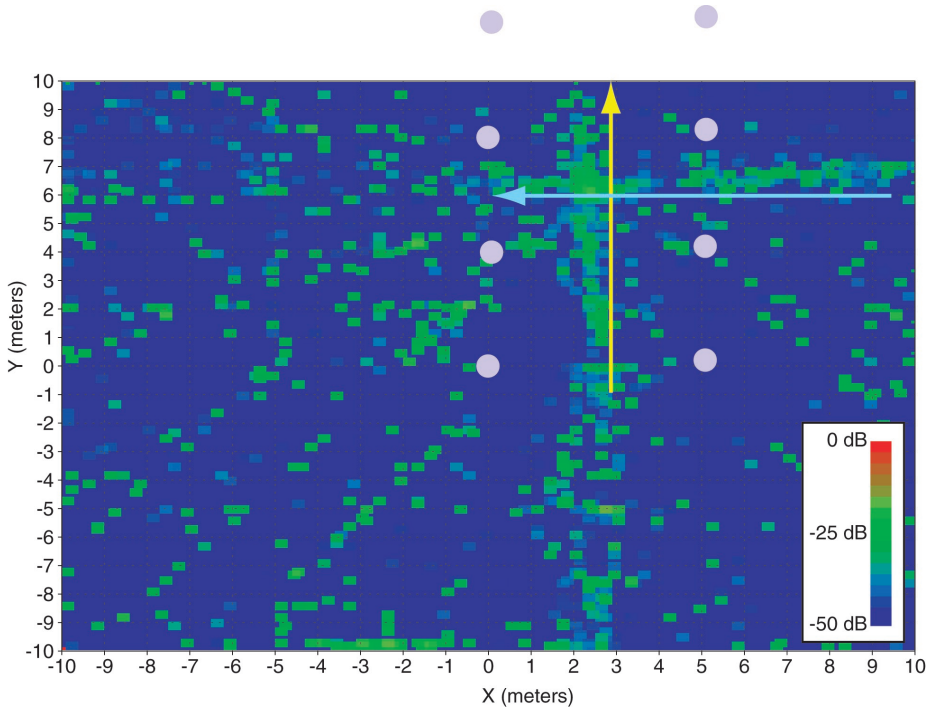

Figure 6.14 Energy map for a recording of our moving speaker scenario. The arrows depict the trajectory of the two speakers. Energy is integrated over the entire duration of the input audio sequence. Note how the two intersecting trajectories are clearly reconstructed.

Source Re-Localization and Modification

Using our technique, we can selectively choose and modify various elements of the original recordings. For instance, we can select any spatial area in the scene and simply relocate all clusters included in the selected region. We demonstrate an example interactive interface for spatial modification where the user first defines a selection area then a destination location. All clusters entering the selection area are translated to the destination location using the translation vector defined by the center of the selection box and the target location. In the accompanying video, we show two instances of source re-localization where we first select a speaker on the left-hand side of the listener and move him to the right-hand side. In a second example, we select the fountain at the rear-left of the listener and move it to the front-right (Figure 6.15).

Compositing

Since our recording setups are spatially calibrated, we can integrate several environments into a single composite rendering which preserves the relative size and positioning of the various sound sources. For instance, it can be used to integrate a close-miked sound situation into a different background ambiance. We demonstrate an example of sound-field compositing by inserting our previous car example (Figure 6.11) into the scene with the two moving speakers in a wide area. The resulting composite environment is rendered with 8 clusters and the 16 recordings of the two original soundscapes. Future work might include merging the representations in order to limit the number of composite recordings (for instance by "re-projecting" the recordings of one environment into the recording setup of the other and mixing the resulting signals).

Figure 6.15 An example interface for source re-localization. In this example we select the area corresponding to the fountain (in purple) and translate it to a new location (shown as a yellow cross). The listener is depicted as a large red sphere, the microphone array as small yellow spheres and the blue spheres show cluster locations.

Real/Virtual Integration

Our approach permits spatially consistent compositing of virtual sources within real-world recordings. We can also integrate virtual objects, such as walls, and make them interact with the original recordings. For instance, by performing real-time ray-casting between the listener and the location of the frequency subbands, we can add occlusion effects due to a virtual obstacle using a model similar to [Tsingos et al., 1998]. Please refer to the accompanying examples at the previously mentioned URL for a demonstration. Of course, perfect integration would also require correcting for the reverberation effects between the different environments to composite. Currently, we experimented only in environments with limited reverberation but blind extraction of reverberation parameters [Baskind and Warusfel, 2002] and blind deconvolution are complementary areas of future research in order to better composite real and virtual soundfields.

6.8 Discussion

Although it is based on a simple mixing model and assumes W-disjoint orthogonality for the sources, we were able to apply our approach to real-world recording scenarios. While not production-grade yet, our results seem promising for a number of interactive and off-line applications.

While we tested it for both indoor and outdoor recordings, our approach is currently only applicable to environments with limited reverberation. Long reverberations will have a strong impact on our localization process since existing cross-correlation approaches are not very robust to interfering sound reflections. Other solutions based on blind channel identification in a reverberant context could lead to improved results [Chen et al., 2006].

Errors in localization of the frequency subbands can result in noticeable artefacts especially when moving very close to a source. These errors can come from several factors in our examples particularly low signal-to-noise ratio for the source to localize, blurring from sound reflections, correlation of two different signals in the case of widely spaced microphones or several sources being present in a single frequency subband. As a result, several overlapping sources are often fused at the location of the louder source. While the assumption of W-disjoint orthogonality has been proven to be suitable for speech signals [Rickard and Yilmaz, 2002], it is more questionable for more general scenarios. It will only be acceptable if this source can perceptually mask the others. However, recent approaches for efficient audio rendering have shown that masking between sources is significant [Tsingos et al., 2004], which might explain why our approach can give satisfying results quite beyond the validity domain of the

underlying models. Alternate decompositions [Mallat and Zhang, 1993, Lewicki and Sejnowski, 2000] could also lead to sparser representations and better results within the same framework.

The signal-to-noise ratio of the different sound sources is also directly linked to the quality of the result when moving very close to the source since our warping is likely to amplify the signal of the original recording in this case.

We are working on several improvements to alleviate remaining limitations of the system and improve the rendering quality:

Currently, we do not interpolate between recordings but select the signal of the microphone closest to the listener location for subsequent warping and re-rendering. This provides a correct solution for the case of omnidirectional anechoic point sources. In more general situations, discontinuities might still appear when switching from one microphone to the next. This can be caused, for instance, by the presence of a sound source with a strong directionality. A solution to this problem would be to warp the few microphones closest to the listener and blend the result at the expense of a higher computing cost. Note that naive blending between microphone signals before warping would introduce unwanted interferences, very noticeable in the case of widely-spaced microphones. Another option would be to experiment with morphing techniques [Slaney et al., 1996] as an alternative to our position-based warping. We could also use different microphones for each frequency subband, for instance choosing the microphone closer to the location of each subband rather than the one closest to the listener. This would increase the signal-to-noise ratio for each source and could be useful to approximate a close-miking situation in order to edit or modify the reverberation effects for instance.

The number of bands also influences the quality of the result. More bands are likely to increase the spatial separation but since our correlation estimates are significantly noisy, it might also make artefacts more audible. In our case, we obtained better sounding results using a limited number of subbands (typically 8 to 16). Following the work of Faller et al. [Baumgarte and Faller, 2003, Faller and Baumgarte, 2003, Faller and Merimaa, 2005], we could also keep track of the inter-correlation between recordings in order to precisely localize only the frames with high correlation. Frames with low correlation could be rendered as "diffuse", forming a background ambiance which cannot be as precisely located [Merimaa and Pulkki, 2004]. This could be seen as explicitly taking background noise or spatially extended sound sources into account in our mixing model instead of considering only perfect anechoic point sources. We started to experiment with an explicit separation of background noise using noise-removal techniques [Ephraim and Malah, 1984]. The obtained foreground component can then be processed using our approach while the background-noise component can be rendered separately at a lower spatial resolution. Example renderings available on the web site demonstrate improved quality in complex situations such as a seashore

recording.

Sound source clustering and matting also strongly depends on the correlation and position estimates for the subbands. An alternative solution would be to first separate a number of sources using independent component analysis (ICA) techniques and then run TDOA estimation on the resulting signals [Saruwatari et al., 2003, Huang et al., 2005]. However, while ICA might improve separation of some sources, it might still lead to signals where sources originating from different locations are combined.

Another issue is the microphone setup used for the recordings. Any number of microphones can be used for localization starting from two (which would only give directional information). If more microphones are used, the additional TDOA estimates will increase the robustness of the localization process. From our experience, closely spaced microphones will essentially return directional information while microphone setups surrounding the scene will give good localization accuracy. Microphones uniformly spaced in the scene provide a good compromise between signal-to-noise ratio and sampling of the spatial variations of the sound-field. We also experimented with cardioid microphone recordings and obtained good results in our car example. However, for larger environments, correlation estimates are likely to become noisier due to the increase in separation between different recordings, making them difficult to correlate. Moreover, it would make interpolating between recordings more difficult in the general case. Our preferred solution was thus to use a set of identical omnidirectional microphones. However, it should be possible to use different sets of microphones for localization and re-rendering which opens other interesting possibilities for content creation, for instance by generating consistent 3D-audio flythroughs while changing the focus point on the scene using directional microphones.

Finally, our approach currently requires an off-line step which prevents it from being used for real-time analysis. Being able to compute cross-correlations in real-time for all pairs of microphones and all subbands would make the approach usable for broadcast applications.

6.9 Conclusions

We presented an approach to record, edit and re-render real-world auditory situations. Contrary to most related approaches, we acquire the sound-field using an unconstrained, widely-spaced, microphone array which we spatially calibrate using photographs. Our approach pre-computes a spatial mapping between different frequency subbands of the acquired live recordings and the location in space from which they were emitted. We evaluated standard TDOA-based techniques and proposed a novel hierarchical localization approach. At run-time, we can apply this mapping to the frequency subbands of the microphone closest to the virtual listener in order to

resynthesize a consistent 3D sound-field, including complex propagation effects which would be difficult to simulate. An additional clustering step allows for aggregating subbands originating from neighboring locations in order to segment individual sound sources or small groups of sound sources which can then be edited or moved around. To our knowledge, such level of editing was impossible to achieve using previous state-of-the-art techniques and could lead to novel authoring tools for 3D-audio scenes.

We believe our approach opens many novel perspectives for interactive spatial audio rendering or off-line post-production environments, for example to complement image based rendering techniques or free-viewpoint video. Moreover, it provides a compact encoding of the spatial sound-field, which is independent of the reproduction system. In the near future, we plan to run more formal perceptual tests in order to compare our results to binaural or high-order *Ambisonics* recordings in the case of fixed-viewpoint scenarios and to evaluate its quality using various reproduction systems. From a psychophysical point of view, this work suggests that real-world sound scenes can be efficiently encoded using limited spatial information.

Other promising areas of future work would be to exploit perceptual localization results to improve localization estimation [Wilson and Darell, 2006] and apply our analysis-synthesis strategy to the real-time generation of spatialized audio textures [Lu et al., 2004]. Finally, making the calibration and analysis step interactive would allow the approach to be used in broadcasting applications (e.g., 3D TV).

We presented a novel approach to real-time spatial rendering of realistic auditory environments and sound sources recorded live, in the field. Our approach automatically extracts a high-level representation from the recording which is compatible with the point-source model used for 3D audio rendering. Using the high-level representation thus obtained, we can edit and re-render the acquired auditory scene over a variety of listening setups. In particular, we can move or alter the different sound sources and arbitrarily choose the listening position. We can also composite elements of different scenes together in a spatially consistent way. We demonstrated a wide range of possible applications for games, virtual and augmented reality and audio-visual post-production. However, the current algorithm assumes that the sound is emitted by a point source, which is not true for every case. In the last chapter, we will present a new approach to solve this problem based on a foreground and background decomposition.

Chapter 7

Improved Background and Foreground Classification and Perceptual Evaluation

In the previous chapter, we developed a novel analysis-synthesis approach from field recording. Inspired by spatial audio coding [Faller and Baumgarte, 2003, Baumgarte and Faller, 2003, Breebaart et al., 2005, Pulkki and Faller, 2006, Goodwin and Jot, 2006] and blind source separation [Yilmaz and Rickard, 2004, Vincent et al., 2003, Radke and Rickard, 2002], our method builds a higher-level spatial description of the auditory scene from a small set of monophonic recordings. This description can then be used for real-time post-processing and re-rendering of the original recordings, for instance by smoothly varying the listening point inside the environment and editing/moving sound sources. Contrary to previous spatial audio coding work, the recordings are made from widely-spaced locations and sample both content and spatial information for the sound sources present in the scene. Our approach is also mostly dedicated to live recordings since it reconstructs estimates of the 3D locations of the sound sources from physical propagation delays. This information might not be available in studio recordings which rely on non-physical panning strategies. However, in the case of live field recordings, this approach suffers from several limitations. First, the underlying hypothesis of time-frequency sparseness for the acquired signals is often not true in practice, especially in the presence of significant background noise [Rickard, 2006]. This results in noisy position estimates and low quality signal reconstruction when virtually moving throughout the environment. Second, our approach uses a limited number of frequency subbands, acting as representative point sources, to model the auditory environment at each time-frame. While point sources might be appropriate to render well-localized events, background ambiance and extended sources (e.g., the sea on a seashore) cannot be convincingly reproduced

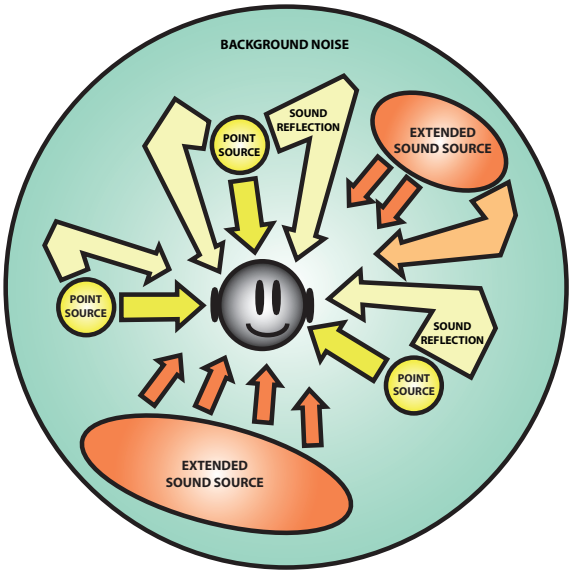

Figure 7.1 Typical components of a real-world auditory scene. In this chapter, we propose to explicitly separate *foreground*, non-stationary and well localized, sound events from *background* components that are more stationary and spatially diffuse.

using this model (Figure 7.1).

In this chapter, we propose a solution to these shortcomings based on an *a priori* segmentation of foreground sound events and background ambiance which we describe in Section 7.1.1. We also present an improved re-rendering solution specifically adapted to these two components which preserves the independence from the reproduction setup. In particular, we propose to render the foreground sound events using a set of separate point sources while the background component is encoded using a smoother low-order spherical harmonics representation. Details can be found in Sections 7.1.2 and 7.1.3.

Section 7.2 describes the results of a pilot perceptual evaluation study aimed at assessing the quality of our approach relative to reference binaural and B-format recordings in the case of fixed-listening-point scenarios.

Finally, our approach introduces additional authoring capabilities by allowing separate manipulation of each component, which we briefly outline in Section 7.3 before concluding.

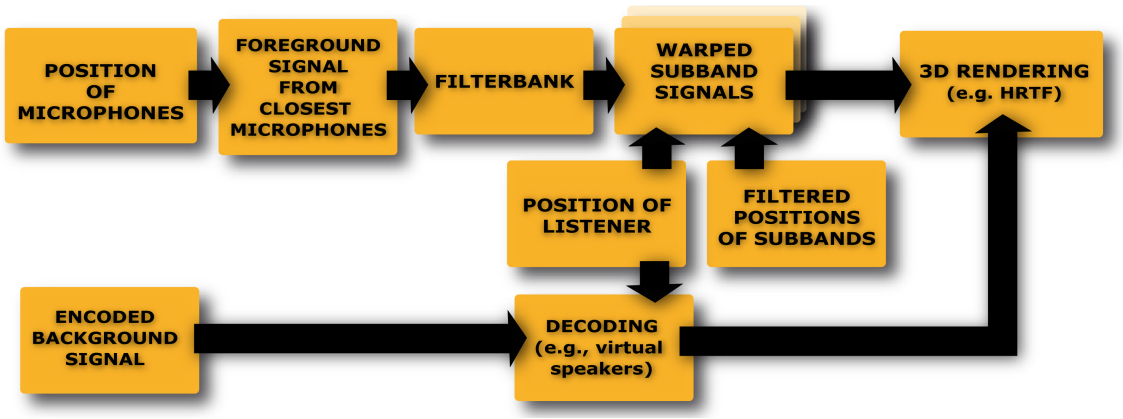

Figure 7.2 Overview of our re-synthesis pipeline. Foreground sound events are rendered as point sources while background sounds are encoded using a low-order spherical harmonics decomposition.

7.1 Improved Analysis and Re-Synthesis

This section addresses a set of possible improvements to our previous technique. They are based on an *a priori* segmentation of background and foreground components leading to a two-layer model, similar in spirit to the pairwise/non-directional and direct/diffuse decompositions used in some spatial audio coding approaches [Pulkki and Faller, 2006, Goodwin and Jot, 2006, Merimaa and Pulkki, 2004, Breebaart et al., 2005]. However, since we are warping the direct component when re-rendering from different listening points, switching at each time frame between localized/diffuse models on a per-subband basis would introduce audible artefacts in our case. We chose to perform a finer-grain segmentation of the input recordings as a pre-processing step which does not rely on position estimates. Such an approach was already reported to improve results for blind source separation problems [Choi, 2003]. We also propose re-rendering strategies tailored to each component.

7.1.1 Background/Foreground Segmentation

Previous work, such as performed by Avendano, proposed an ambiance extraction from stereo signals using the assumption that the left and right ambience signals are correlated. We chose to segment stationary background noise from non-stationary sound events using the technique by Ephraim and Malah [Ephraim and Malah, 1984], originally developed for denoising of speech signals. This approach assumes that the distributions of Fourier coefficients for signal and noise are statistically independent zero-mean Gaussian random variables. Under this assumption, the spectral amplitude

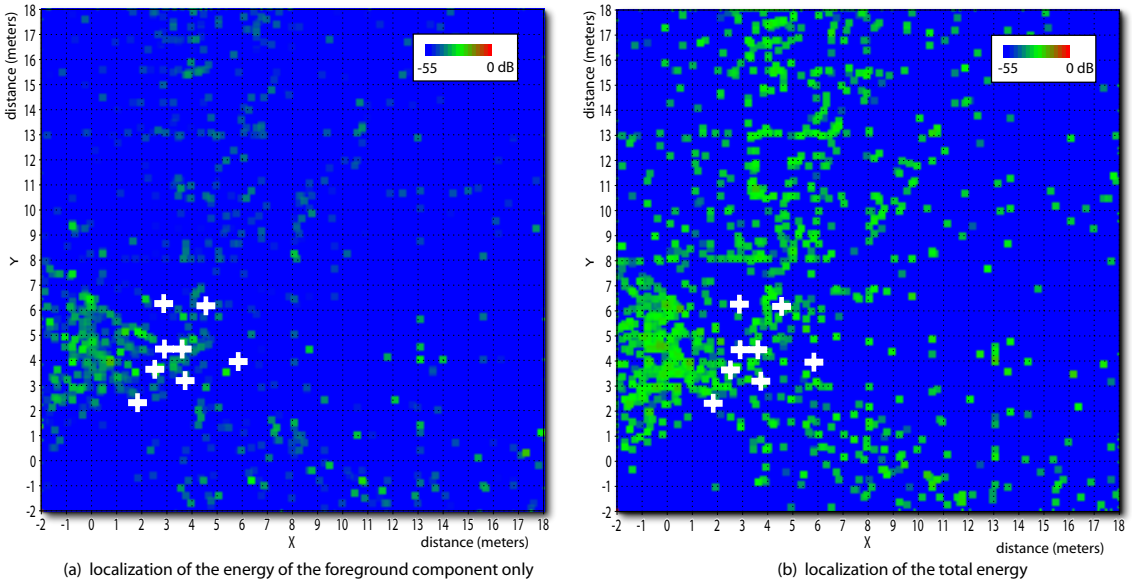

Figure 7.3 Comparison between energy localization in the seashore example of Section 7.3 for (a) the foreground component only and (b) the complete recording. The figure shows the reconstructed location of all subbands integrated through the entire duration of the sequence. White crosses indicate the locations of the microphones used for recording.

of the denoised signal is estimated using a minimum mean-square error criterion. The background noise signal is then simply obtained by subtracting the denoised signal from the original. We found the algorithm to perform quite well. While not perfect, it leads to a foreground component with limited musical noise. In most cases, this noise is masked when re-combined with the background component at re-rendering time. The extracted foreground component, containing non-stationary sounds is also better suited to our underlying assumption of time-frequency sparseness than the original recordings (see Figure 7.3). However, several foreground sound sources might still overlap in time-frequency. Background and foreground segmentation is performed independently on the signals from all microphones.

7.1.2 Background "Panorama" Generation

The separated foreground and background components are both processed using the analysis pipeline described in Chapter 6 (see also Figure 7.1). However, in the case of the background component, we obtain noisier position estimates since this component will generally correspond to background noise and sources with low signal-to-noise

ratios. In order to produce a smooth spatial background texture, we use the obtained positions to encode the corresponding subband signals on a 1^{st}-order spherical harmonic basis. No warping is applied to the background component in this case (Figure 7.2).

As our signals are real-valued, we encode them with real spherical harmonics defined as:

$$y_l^m(\theta,\phi) = \begin{cases} \sqrt{2}K_l^m cos(m\phi)P_l^m(cos\theta) & m > 0 \\ \sqrt{2}K_l^m cos(-m\phi)P_l^{-m}(cos\theta) & m < 0 \\ K_l^0 P_l^0(cos\theta) & m = 0 \end{cases} \tag{7.1}$$

where l is the order, $m \in [-l;+l]$, P is the associated Legendre polynomial and K is a scaling factor defined as:

$$K_l^m = \sqrt{\frac{(2l+1)}{4\pi}\frac{(l-|m|)!}{(l+|m|)!}}. \tag{7.2}$$

For each subband signal, we compute the minimum and maximum elevation and azimuth of the obtained positions over the entire duration of the recording. Then, we uniformly expand the background signal in this area. We choose the background signal to encode from the monophonic recording closest to the center of the acquired scene. Accordingly, the background texture is encoded relative to a fixed reference point, for instance the central point of the scene.

This background panorama can thus be encoded in a pre-processing stage so that only the decoding is performed at run-time, e.g., when freely navigating in the recordings. Several decoding options are available depending on the desired reproduction setup [Jot et al., 1999].

7.1.3 Improved Foreground Re-Synthesis

At re-rendering time, we perform a warping of the original foreground recordings in order to generate a signal as consistent as possible with the desired virtual listening position (Figure 7.2). Assuming an inverse distance attenuation for point emitters, the warped signal $R'_i(t)$ in subband i is given as:

$$R'_i(t) = r_1^i / r_2^i R_i(t + (\delta_1^i - \delta_2^i)), \tag{7.3}$$

where r_1^i,δ_1^i are respectively the distance and propagation delay from the considered time-frequency atom to the reference microphone and r_2^i,δ_2^i are the distance and propagation delay to the desired listening position.

This warping heavily relies on the fact that we consider the subband signals to be re-emitted by anechoic point sources. In real-world environments this model is

challenged, due to the strong directionality of some sound sources. As a result, discontinuities can appear when the virtual listener is moving around if the signal from a single reference microphone is used (e.g., the one closest to the desired virtual position). To avoid such problems and roughly compensate for the limitations of our anechoic point source model, we propose to continuously warp the signals of the two microphones closest to the desired virtual listening position and blend them together to generate a smoothly varying monophonic signal. Blending can be simply controlled by the relative distance of the virtual listener to these two reference microphones. Note that blending the signals prior to warping would introduce comb filtering effects that can be very noticeable when the microphones are widely spaced. To further improve the re-rendering quality of the foreground component, we also smooth our position estimates for the subbands using Kalman filtering [Kalman, 1960]. This prevents large and fast position changes and limits possible "wobbling" effects due to jittery subband positions.

7.2 Pilot Subjective Evaluation

In order to evaluate the quality of a spatial audio reproduction system based on our approach, we compared it to binaural and B-format recordings in the context of various scenarios with fixed listening points.

7.2.1 Test Stimuli and Procedure

We recorded test scenarios in two different environments: indoors in a moderately reverberant room (RT60 $\approx$ 0.3 sec. at 1KHz) and outdoors (see Figure 7.4). For each scenario, we used 8 monophonic recordings made with *AudioTechnica 3032* omnidirectional microphones to run our localization and re-rendering approach. A pair of *Sennheiser MKE-2 gold* microphones was placed inside the ears of a subject to capture reference binaural recordings and we also acquired a B-format version of the scenes using a *Soundfield ST250* microphone. Eventually, four recordings (one indoors, three outdoors), each about 50 sec. long, were chosen for quality testing.

We used 8 non-overlapping subbands uniformly distributed on a Bark scale to run our spatial analysis. Then, a binaural rendering from a point of view similar to the binaural and B-format recordings was generated from the monophonic input of the closest omnidirectional microphone and the time-varying locations obtained for the subbands. The signal of the same microphone was used to generate both a binaural rendering of the foreground events and the 1^{st}-order spherical harmonic background decoded over headphones using a *virtual loudspeakers* technique. In both cases, we used head related transfer functions (HRTFs) of the *LISTEN* database

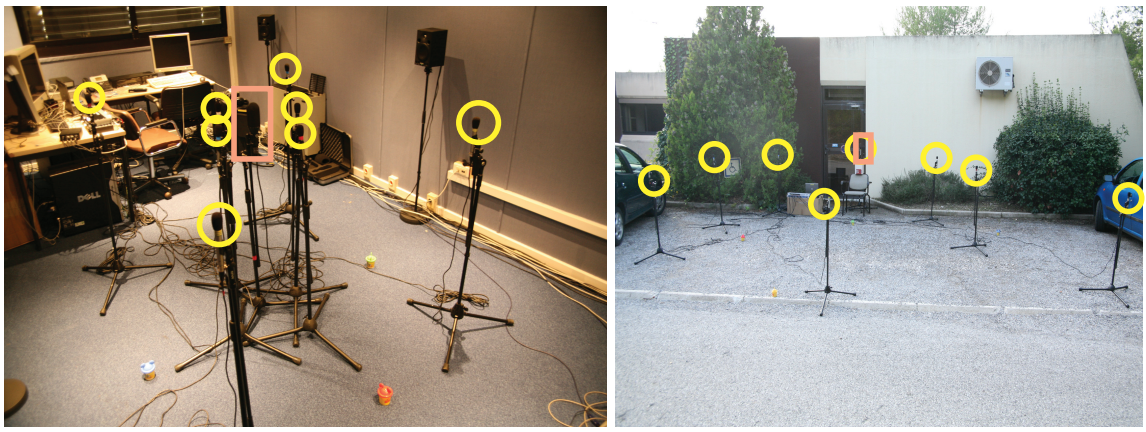

Figure 7.4 Example recording setups. We used 8 omnidirectional microphones (circled in yellow) to capture the auditory scene as well as a *Soundfield* microphone (highlighted with a light red square) to simultaneously record a B-format version. A binaural recording using microphones placed in the ears of a subject provided a reference recording in each test case.

(http://recherche.ircam.fr/equipes/salles/listen/) for re-rendering. We also generated a re-rendering without explicit background/foreground segmentation considering the original recording to be entirely foreground. B-format recordings were also converted to binaural using a similar virtual loudspeaker approach.

We used a protocol derived from *Multiple Stimuli with Hidden Reference and Anchors* procedure (MUSHRA, ITU-R BS.1534) [Stoll and Kozamernik, 2000, Union, 2003, ITU-R, 2003] to evaluate each scenario, using four tests stimuli (binaural reference, B-format, our approach with foreground only, our approach with background/foreground segmentation) and a hidden reference. We also provided one of our 8 monophonic recordings and the omnidirectional (W) component of the B-format recordings as anchors, resulting in a total of 7 signals to compare. Corresponding test stimuli are available at the following URL: http://www-sop.inria.fr/reves/projects/aes30. Test stimuli were presented over *Sennheiser HD600* headphones. Monaural anchor signals were presented at both ears.

Five subjects, aged 23 to 40 and reporting normal hearing, volunteered for this evaluation. They were asked to primarily focus on the spatial aspects of the sounds, paying particular attention to the position of the sources. Since the recordings were made with different microphones, we asked them to avoid specific judgments comparing the general timbre of the recordings. However, the subjects were instructed to keep track of any artefact compromising the quality of the reproduction. Their comments were gathered during a short post-screening interview. Subjects were in-

structed to rank the signals on a continuous [0,100] quality scale and give the highest possible score to the signal closest to the reference. They were also instructed to give the lowest possible score to the signal with the worst *spatial degradation* relative to the reference.

7.2.2 Results

Figures 7.5 and 7.6 summarize the results of this study. The subjects were able to identify the hidden reference and it received a maximal score in all test cases. In most cases, our approach was rated higher than B-format recordings in terms of quality of spatial reproduction. This is particularly true for the foreground-only approach which does not smooth the spatial cues and obtains a very high score. However, the subjects reported artefacts due to subbands whose localization varies rapidly through time, which limits the applicability of the approach in noisier environments. Our approach including background/foreground separation leads to smoother spatial cues since the low order background signal may mask the foreground signal. Hence, it was rated only slightly better than the B-format recordings. Subjects did not report specific artefacts with this approach, showing an improved signal quality. As could be expected, the monophonic anchors received the lowest scores. However, we can note that in some of our test cases, they received scores very close to the B-format reproduction. This is probably due to the low spatial resolution of B-format but could also arise from a non-optimal HRTF-based decoding.

Looking at the various test-cases in more detail, Figure 7.6 highlights a significantly different behavior for the indoor scenario (TEST#3). In this case, very little background sound was present, hence our approach based on background and foreground separation did not lead to any improvement and, in fact, resulted in a degraded spatial impression. The B-format reproduction, however, obtained significantly better scores in this case, probably due to the favorable configuration of the three speakers (one in front, one to the left, and one to the right).

7.2.3 Discussion

In terms of audio quality, feedback from the subjects of the tests shows that our improved algorithm outperforms the previous foreground-only solution. This is of course due to the smoothly varying background and more robust foreground estimates. However, our proposed approach appears less convincing in terms of localization accuracy. Significant parts of the foreground sounds can still be present in the background component and will be spatialized using a different strategy. The resulting blend tends to blur out the localization cues leading to a poorer spatial impression. Improving the quality of the segmentation would probably lead to better results. Another possibility

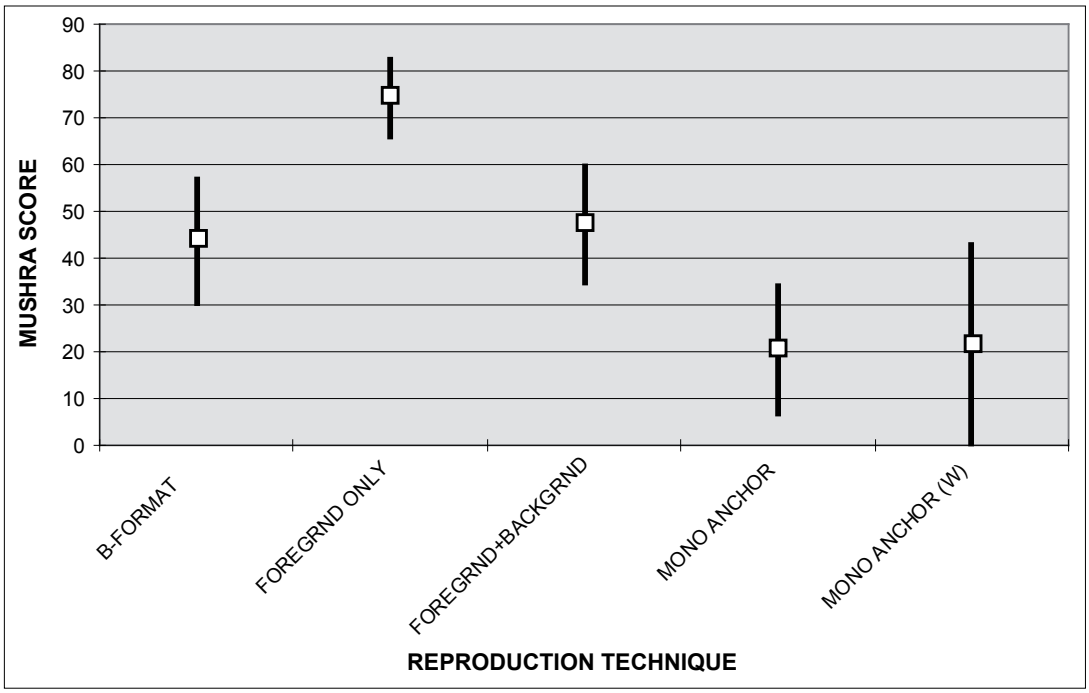

Figure 7.5 Average MUSHRA scores and 95% confidence intervals for all subjects and all scenarios.

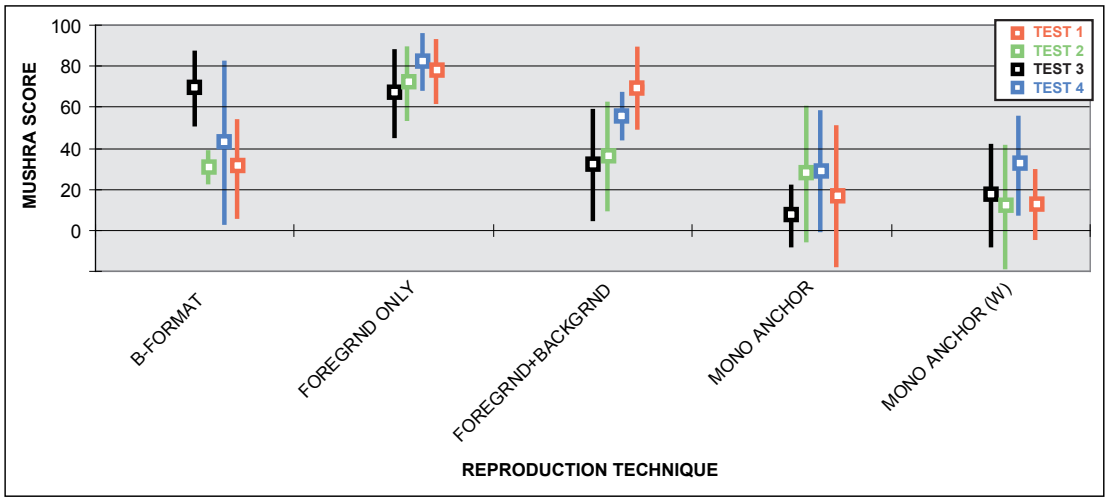

Figure 7.6 Average MUSHRA scores and 95% confidence intervals for all subjects in each of our 4 test scenarios.

Figure 7.7 Recording setup used for the seashore recordings.

would be to use energy and not only time-differences of arrival to extract possible localization information for the background component. We used a small number of frequency subbands in our tests which can challenge our time-frequency orthogonality assumption resulting in noisier position estimates for the foreground component. However, we obtained less convincing results with an increased number of frequency subbands due to less accurate correlation estimates for narrower subbands signals. We do not currently model sources "at infinity", which may appear in the background but also in the foreground component. Our position estimation can return erroneous position estimates in this case due to the limited extent of our position histogram. This could also explain the perceived degradation of spatial cues compared to the reference. Explicit detection of far-field sources is a component we are planning to add in the near future. Finally, non-individualized HRTF processing could also be a major cause of spatial degradation. Running the test with head-tracking and individualized HRTFs might lead to improved results.

7.3 Applications

Our approach can lead to spatial audio coding applications for live audio footage in a way similar to [Pulkki and Faller, 2006, Goodwin and Jot, 2006, Merimaa and Pulkki, 2004, Breebaart et al., 2005], but it also offers novel decoding/authoring capabilities

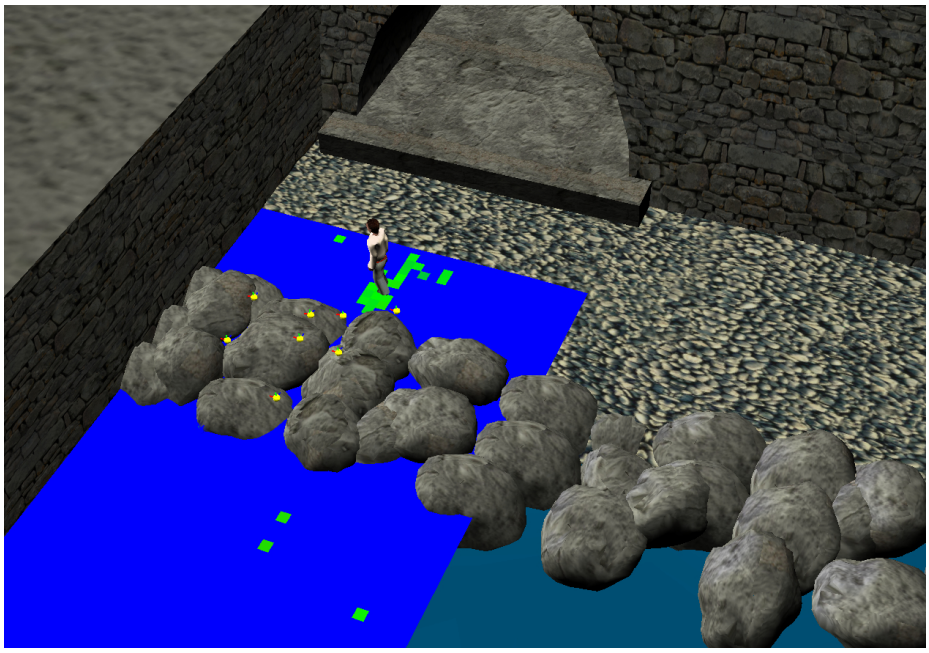

Figure 7.8 Example virtual reconstruction of a seashore with walking pedestrian. Yellow spheres correspond to the locations of the microphones used for recording.

not available with previous techniques such as *free-viewpoint* walkthroughs. Figure 7.8 illustrates the virtual reconstruction of a seashore scene with a pedestrian walking on a pebble beach recorded with the setup shown in Figure 7.7. A spatial energy map is overlaid, highlighting the location of foreground time-frequency atoms. Note how the position of the footsteps sounds is well reconstructed by our approach. The sound of sea waves hitting the rocks on the shore is mostly captured by the background component (see also Figure 7.3). Please, visit the web pages mentioned in Sections 7.2 for example audio files and videos.

Spatial Re-Synthesis with Free-Moving Listener

Our approach allows for a "free-viewpoint" spatial audio rendering of the acquired soundscapes. As the virtual listener moves throughout the scene, the foreground component is rendered using a collection of point sources corresponding to each time-frequency atom, as described in section 7.1.3. The background component is simply rotated based on the current orientation of the listener in order to provide a consistent rendering. Our representation encodes spatial cues in world space and can thus be rendered on a variety of reproduction setups (headphones, multichannel, *etc.*).

Background/Foreground Editing

Our two-layer model allows for independent control of the background and foreground components. Their overall level can be adjusted globally or locally, for instance to attenuate foreground sounds with local virtual occluders while preserving the background. The foreground events can also be copied and pasted over a new background ambiance.

Re-Rendering with Various Microphones

Finally, the microphones used for the analysis process can be different from the ones used for re-rendering. For instance, it is possible to use any directional microphone to obtain a combined effect of spatial rendering and focussing on a specific source.

7.4 Conclusion

We presented an approach to convert field recordings into a structured representation suitable for generic 3D audio processing and integration with 2D or 3D visual content. It applies both to outdoor environments or indoor environments with limited reverberation, provides a compact encoding of the spatial auditory cues and captures propagation and reverberation effects which would be very difficult to render with the same level of realism using traditional virtual acoustics simulations.

Perceptual comparisons with reference binaural and B-format recordings showed that our approach outperforms B-format recordings and can get close to reference binaural recordings when all time-frequency atoms are rendered as foreground point sources. However, artefacts due to background noise lead to reduced signal quality. An alternative solution was proposed based on the explicit segmentation of stationary "background noise" and non-stationary "foreground events". While the signal quality is significantly improved when re-rendering, spatial cues were perceived to be degraded, probably due to non-optimal background separation. In the future, we would like to improve on our background/foreground segmentation approach, possibly based on auditory *saliency* models [Kayser et al., 2005] or taking advantage of the signals from all microphones. Alternative sparse representations of the signals [Lewicki and Sejnowski, 2000, Mallat and Zhang, 1993] could also be explored in order to improve our approach. Further comparisons to other sound-field acquisition techniques, for instance based on high-order spherical harmonic encoding [Abhayapala and Ward, 2002, Meyer and Elko, 2004], Fourier-Bessel decomposition [Laborie et al., 2003, Laborie et al., 2004] or directional audio coding [Pulkki, 2006, Pulkki and Faller, 2006] would also be of primary interest to evaluate the quality vs. flexibility/applicability

tradeoffs of the various approaches. We believe our approach opens many novel perspectives for interactive spatial audio rendering or off-line post-production environments, for example to complement image based rendering techniques or free-viewpoint video.

In this chapter, we have presented an approach to automatically extract and re-render a structured auditory scene from field recordings obtained with a small set of microphones, freely positioned in the environment. From the recordings and the calibrated position of the microphones, the 3D location of various auditory events can be estimated together with their corresponding content. This structured description is reproduction-setup independent. We proposed solutions to classify foreground, well-localized sounds and more diffuse background ambiance and we adapted our rendering strategy accordingly. We showed that warping the original recordings during playback allows for simulating smooth changes in the listening point or position of sources. We also presented comparisons with reference binaural and B-format recordings showing that our approach achieves good spatial rendering while remaining independent of the reproduction setup and offering extended authoring capabilities.

Chapter 8

Conclusion

In this thesis, we were interested in the problems related to virtual sound rendering of complex scenes, for instance containing many sound sources. Two difficulties prevent the interactive rendering of such scenes. The signal processing involved in the simulation is beyond current CPU capabilities and the process of authoring such complex scenes is difficult and tedious. In this thesis, we proposed solutions to both problems.

8.1 Summary of Contributions

In order to perform the massive number of audio processing operations required by 3D audio rendering, we have proposed to leverage the novel parallel architecture provided by the graphics hardware(GPU). Although the GPU is designed for graphics applications, its flexibility and its data parallel processing architecture yields a good alternative solution that clearly outperforms current CPUs. Moreover, GPU performance has increased dramatically in the last three years in comparison to CPUs and GPUs now tend to become truly multi-purpose processors. Our studies showed that this architecture is well designed for audio processing tasks. In the future such architectures are likely to become *de-facto* standards and we believe that soundcards may benefit from including the same type of architecture and programmability.

In order to simplify the auditory scene and provide a progressive approach to audio rendering, we proposed a new algorithm which exploits the properties of human hearing, such as auditory masking and illusory continuity. The proposed method provides a scalable approach by progressively processing the important components of a scene. This algorithm is well suited to any kind of application which includes signal processing and in particular for 3D sound rendering. It provides a speed versus quality trade-off well adapted to real-time applications, yielding a rendering solution that can be tuned to any computer. Within this method, we have introduced an emergence

criterion and we have assessed a few importance metrics. A subjective evaluation validates our algorithm, and shows that the audio operations can be reduced by 50 % without degradation in perceived quality. This approach can be also used in compression algorithms or transmission of audio streams over networks. The solutions proposed in this thesis allow for rendering thousands of sound sources on a variety of platforms from laptops to top-of-the-line workstations

In the second part of this thesis, we have introduced a method to automatically create virtual auditory scenes from recordings of a real scene. Moreover, this technique complements current methods for spatial sound recording. Our approach extracts a higher level description of the scene, using live recordings from several microphones without any constraint on their location. We reconstruct a spatial scene representation based on the location of the emitted signals and their extracted frequency content. This method can be used for outdoor scenes, as well as in low-reverberation indoor scenes. We believe our approach can offer new perspectives for post-production environments. It avoids the problem of capturing each source individually while offering a similar level of interaction with the scene. Furthermore, the resulting scene representation provides a compact encoding of the spatial soundfield which is independent of the restitution system. This work also suggests that real-life sound scenes can be efficiently encoded using limited spatial information.

We extended our previous assumptions used for segmenting the scene to diffuse sound sources. The proposed method separates the signal into foreground and background components under the hypothesis that the background component is stationary. We reconstruct the audio scene from any point-of-view by spatializing the foreground component using our previous approach. The background component is encoded with low order spherical harmonics in order to provide a spatially diffuse rendering. A subjective evaluation of the quality of the spatial reconstruction, comparing our method with other spatial recording techniques such as binaural and B-format has been performed. It showed that our approach typically outperforms B-format recording and can get close to reference binaural recording. Our approach allows for interactive auralization of complex real-world auditory scenes while maintaining re-rendering flexibility.

8.2 Future Research and Applications

This dissertation opens many promising directions for future work.

We have seen that the GPU is well suited to 3D audio processing. Unfortunately, we could not evaluate our algorithms on the latest processors G80 which solve the few remaining problems highlighted by our study and would certainly improve the performance. It would be interesting to test other audio processing algorithms such

as "Infinite Impulse Response" (IIR) or "Finite Impulse Response" (FIR) filtering. Indeed, these algorithms are the basic tools to include reverberation effects in virtual reality applications. In this case, reverberation parameters could be computed directly by analyzing the 3D scene geometry through the GPU. A performance evaluation with other DSP processors could be also interesting to perform.

The scalability of the algorithm presented in Chapter 5 could be improved by adding a finer grain selection or by using another importance metric such as a time varying loudness model. Moreover, the illusory continuity was not explicitly used in the emergence metric, and may improve the result. An interesting extension will be to use this algorithm with other signal representations such as those obtained by sparse coding [Lewicki, 2002]. In this case, our elementary grains will be the sparse atoms. The algorithm presented can also be used in other applications, to reduce bus/network traffic or code only important parts of the signal.

The second part of the thesis could also be improved in several ways:

- First, we assumed a W-Disjoint hypothesis in the frequency domain for source separation but the Fourier domain is in general not sparse enough for this assumption to hold in complex real-world scenarios. Working in another sparser domain could improve the source separation.

- Second, for every frame, we searched for a source location regardless of the quality of the estimation. We could instead keep track of good estimates during a lapse of time and interpolate between them when no satisfying estimation is found. In this case, the difficulty is to find a "goodness measure" for our estimates.

- Third, in its current state, the algorithm uses a fixed number of bands. An alternative strategy would be to use dynamic splitting strategies for every frame.

- Fourth, the localization of the microphone is derived from photographs. It would be better to find the positions of microphones using a time-delay of arrival approach to be able to directly calibrate on-site. This could be useful for broadcasting applications.

- Finally, in the segmented background component, there are still some foreground parts of the signal. A possible improvement would be to select the part of the background signal where the foreground signal magnitude is high and to fully replace the selected background which contains some foreground signal by texture synthesis strategies similar to methods in audio restoration.

In conclusion, we believe that this thesis has achieved the goals set out in the introduction, that is to advance the speed of processing for complex auditory scenes,

and to improve the authoring process. We believe that the successful results and the directions for future work described above illustrate the strong potential of this research.

Bibliography

[Aarabi, 2003] Aarabi, P. (2003). The fusion of distributed microphone arrays for sound localization. *EURASIP Journal on Applied Signal Processing*, 2003(4):338–347.

[Aarts, 1992] Aarts, R. M. (1992). A comparison of some loudness measures for loudspeaker listening tests. *Journal of the Audio Engineering Society*, 40(3):142–146.

[Abhayapala and Ward, 2002] Abhayapala, T. D. and Ward, D. B. (2002). Theory and design of high order sound field microphones using spherical microphone array. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*.

[Acoustics-FAQ, 1997] Acoustics-FAQ (1997). A-weighting formula, section 8.1. http://www.faqs.org/faqs/physics-faq/acoustics/.

[Ajdler et al., 2004] Ajdler, T., Kozintsev, I., Lienhart, R., and Vetterli, M. (2004). Acoustic source localization in distributed sensor networks. *Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA*, 2:1328–1332.

[Ajdler and Vetterli, 2002] Ajdler, T. and Vetterli, M. (2002). The plenacoustic function and its sampling. *Proc. of the 1st BeneluxWorkshop on Model-based processing and coding of audio (MPCA2002), Leuven, Belgium*.

[Aliaga and Carlbom, 2001] Aliaga, D. G. and Carlbom, I. (2001). Plenoptic stitching: a scalable method for reconstructing 3d interactive walk throughs. *In SIGGRAPH 01: Proceedings of the 28th annual conference on Computer graphics and interactive techniques, ACM Press, New York, NY, USA*, pages 443–450.

[Allen and Berkley, 1979] Allen, J. B. and Berkley, D. A. (1979). Image method for efficiently simulating small-room acoustics. *Journal of the Audio Engineering Society*, 65 (4):943–950.

[Ashmead et al., 1990] Ashmead, D. H., LeRoy, D., and Odom, R. (1990). Perception of relative distances of nearby sound sources. *Perception and Psychophysics*, 47(4):326–331.

[Avendano, 2003] Avendano, C. (2003). Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppresssion and re-panning applications. *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA2003), New Paltz, NY, USA*.

[Bashford et al., 1988] Bashford, J. A., Meyers, J. M. D., Brubaker, B. S., and Warren, R. M. (1988). Illusory continuity of interrupted speech: Speech rate determines durational limits. *Journal of the Acoustical Society of America*, 84(5):1635–1638.

[Baskind and Warusfel, 2002] Baskind, A. and Warusfel, O. (2002). Methods for blind computational estimation of perceptual attributes of room acoustics. In *proceedings of the AES 22nd Intl. Conf. on virtual, synthetic and entertainment audio, Espoo, Finland*.

[Batteau, 1967] Batteau, D. W. (1967). The role of the pinna in human localization. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 168(1011):158–180.

[Bauer, 1961] Bauer, B. B. (1961). Phasor analysis of some stereophonic phenomena. *Journal of the Acoustical Society of America*, 33(11):1536–1539.

[Baumgarte, 1997] Baumgarte, F. (1997). A physiological ear model for auditory masking applicable to perceptual coding. In *Proc. of 103rd Convention of the Audio Engineering Society, New York, USA*.

[Baumgarte and Faller, 2003] Baumgarte, F. and Faller, C. (2003). Binaural cue coding - part i: Psychoacoustic fundamentals and design principles. *IEEE Transaction on Speech and Audio Processing*, 11(6).

[Begault, 1994] Begault, D. R. (1994). 3d sound for virtual reality and multimedia. *Academic Press Professional*.

[Begault et al., 2001] Begault, D. R., Wenzel, E. M., and Anderson, M. R. (2001). Direct comparison of the impact of head tracking, reverberation and individualized head-related transfer functions on the spatial perception of a virtual speech source. *Journal of the Audio Engineering Society*, 49:904–916.

[Berkhout et al., 1993] Berkhout, A., de Vries, D., and Vogel, P. (1993). Acoustic control by wave field synthesis. *Journal of the Acoustical Society of America*, 93:2764–2778.

[Blumlein, 1931] Blumlein, A. D. (1931). British patent specification 394325. *Reprinted in Stereophonic Techniques, Journal of the Audio Engineering Society, NY, 1986.*

[Bofill and Zibulevsky, 2001] Bofill, P. and Zibulevsky, M. (2001). Underdetermined blind sourc separation using sparse representations. *Signal Processing*, 81:2353–2362.

[Boone et al., 1995] Boone, M., Verheijen, E., and van Tol, P. (1995). Spatial sound-field reproduction by wave-field synthesis. *Journal of the Audio Engineering Society*, 43:1003–1011.

[Borish, 1984] Borish, J. (1984). Extension of the image model to arbitrary polyhedra. *Journal of the Acoustical Society of America*, 75 (6):1827–1836.

[Borish and Angell, 1983] Borish, J. and Angell, J. (1983). An efficient algorithm for measuring the impulse response using pseudorandom noise. *Journal of the Audio Engineering Society*, 31:478–488.

[Brandenburg, 1992] Brandenburg, K. (1992). mp3 and aac explained. In *AES 17th International Conference on High-Quality Audio Coding.*

[Breebaart et al., 2005] Breebaart, J., Herre, J., Faller, C., Rödén, J., Myburg, F., Disch, S., Purnhagen, H., Hotho, G., Neusinger, M., Kjörling, K., and Oomen, W. (2005). MPEG spatial audio coding/MPEG surround: Overview and current status. *Proc. 119th AES Convention, New York, USA. Preprint 6599.*

[Bregman, 1990] Bregman, A. S. (1990). *Auditory Scene Analysis, The perceptual organization of sound.* The MIT Press.

[Buck et al., 2004] Buck, I., Foley, T., Horn, D., Sugerman, J., and Hanrahan., P. (2004). Brook for gpus: Stream computing on graphics hardware. *ACM Transactions on Graphics, Proceedings of SIGGRAPH 2004.*

[Buehler et al., 2001] Buehler, C., Bosse, M., McMillan, L., Gortler, S., and Cohen, M. (2001). Unstructured lumigraph rendering. *Proc. of ACM SIGGRAPH.*

[Chen et al., 2005] Chen, J., Benesty, J., and Huang, Y. (2005). Performance of gcc and amdf based time-delay estimation in practical reverberant environments. *EURASIP Journal on Applied Signal Processing*, pages 25–36.

[Chen et al., 2006] Chen, J., Benesty, J., and Huang, Y. A. (2006). Time delay estimation in room acoustic environments: An overview. *EURASIP Journal on Applied Signal Processing*, 2006:Article ID 26503.

[Chen et al., 2003] Chen, J. C., Yao, K., and Hudson, R. E. (2003). Acoustic source localization and beamforming: Theory and practice. *EURASIP Journal on Applied Signal Processing*, pages 359–370.

[Chen and Williams, 1993] Chen, S. E. and Williams, L. (1993). View interpolation for image synthesis. *Computer Graphics*, 27(Annual Conference Series):279–288.

[Chen et al., 1998] Chen, S. S., Donoho, D. L., and Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61.

[Choi, 2003] Choi, C. (2003). Real-time binaural blind source separation. *Proc. of the 4th Intl. Symp. on Independant Component Analysis and Blind Source Separation (ICA2003), Nara, Japan, april.*

[Chowning, 1971] Chowning, J. M. (1971). The simulation of moving sound sources. *Journal of the Audio Engineering Society*, 19(2-6).

[Coleman, 1963] Coleman, P. (1963). An analysis of cues to auditory depth perception in free space. *Psychological Bulletin*, 60:302–315.

[Comon, 1994] Comon, P. (1994). Independent component analysis: A newconcept. *Signal Processing*, 36:287 – 314.

[Cooper and Bauck, 1989] Cooper, D. and Bauck, J. (1989). Prospects for transaural recording. *Journal of the Audio Engineering Society*, 37(1/2):3–19.

[Craven and Gerzon, 1977] Craven, P. G. and Gerzon, M. A. (1977). Coincident microphone simulation covering three dimensional space and yielding various directional outputs. *Official Gazette, United States Patent Office, brevet no 404277.*

[Daniel, 2003] Daniel, J. (2003). Introducing distance coding filters and a viable, new ambisonic format. *AES 23rd International Conference.*

[Daniel et al., 1998] Daniel, J., Rault, J.-B., and Polack, J.-D. (1998). Ambisonic encoding of other audio formats for multiple listening conditions. In *105th AES convention, preprint 4795.*

[DiBiase et al., 2001] DiBiase, J. H., Silverman, H. F., and Brandstein, M. S. (2001). Robust localization in reverberant rooms. *in Microphone arrays: signal processing techniques and applications, M. Brandstein and D. Ward, Eds. Berlin: Springer-Verlag*, pages 157–180.

[Do, 2004] Do, M. N. (2004). Toward sound-based synthesis: the far-field case. *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, Canada*.

[Dobashi et al., 2003] Dobashi, Y., Yamamoto, T., and Nishita, T. (2003). Real-time rendering of aerodynamic sound using sound textures based on computational fluid dynamics. *ACM Transactions on Graphics 22(3), Proceedings of SIGGRAPH*.

[Doel, 1998] Doel, K. (1998). *Sound Synthesis for Virtual Reality and Computer Games*. PhD thesis, University of British Columbia.

[Doel et al., 2004] Doel, K., Knott, D., and Pai, D. K. (2004). Interactive simulation of complex audio-visual scenes,. *Presence: Teleoperators and Virtual Environments*, 13(1).

[Doel et al., 2002] Doel, K., Pai, D. K., Adam, T., Kortchmar, L., and Pichora-Fuller, K. (2002). Measurements of perceptual quality of contact sound models. In *Proceedings of the International Conference on Auditory Display (ICAD 2002), Kyoto, Japan*, pages 345–349.

[Ephraim and Malah, 1984] Ephraim, Y. and Malah, D. (1984). Speech enhancement using a- minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. on Acoustics, Speech and Signal*, ASSP-32(6):1109–1121.

[Everest, 1998] Everest, R. S. F. A. (1998). *The new stereo soundbook*. Audio Engineering Associate, Pasadena (CA), USA, 2nd edition.

[Eyre and Bier, 2000] Eyre, J. and Bier, J. (2000). The evolution of dsp processors. *IEEE Signal Processing Magazine*.

[Faller and Baumgarte, 2003] Faller, C. and Baumgarte, F. (2003). Binaural cue coding - part II: Schemes and applications. *IEEE Transaction on Speech and Audio Processing*, 11(6).

[Faller and Merimaa, 2005] Faller, C. and Merimaa, J. (2005). Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *J. of the Acoustical Society of America*, 116(5):3075–3089.

[Faugeras, 1993] Faugeras, O. (1993). *Three-dimensional computer vision: a geometric viewpoint*. MIT Press.

[Fletcher and Munson, 1933] Fletcher, H. and Munson, W. A. (1933). Loudness, its definition, measurement and calculation. *Journal of the Acoustical Society of America*, 5 (2):82–108.

[Foster, 1986] Foster, S. (1986). Impulse response measurement using Golay codes. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '86)*, 11:929– 932.

[Fouad et al., 1997] Fouad, H., Hahn, J., and Ballas, J. (1997). Perceptually based scheduling algorithms for real-time synthesis of complex sonic environments. In *proceedings of the 1997 International Conference on Auditory Display (ICAD 97), Palo Alto, USA,*.

[Funkhouser et al., 2002] Funkhouser, T., Jot, J.-M., and Tsingos, N. (2002). Sounds good to me ! computational sound for graphics, vr and interactive systems. *SIGGRAPH 2002 course 45*.

[Gamble, 1909] Gamble, E. (1909). Intensity as a criterion in estimating the distance of sounds. *The Philosophical Review*, 16(416-426).

[Gardner and Gardner, 1973] Gardner, M. B. and Gardner, R. S. (1973). Problem of localization in the medial plane: effect of pinnae cavity occlusion. *Journal of the Acoustical Society of America*, 53:400–408.

[Gerzon, 1976] Gerzon, M. A. (1976). Unitary (energy preserving) multichannel networks with feedback. *Electronics Letters*, 12(11):278–279.

[Gerzon, 1985] Gerzon, M. A. (1985). Ambisonics in multichannel broadcasting and video. *Journal of the Audio Engineering Society*, 33(11):859–871.

[Glasberg and Moore, 2002] Glasberg, B. R. and Moore, B. C. J. (2002). A model of loudness applicable to time-varying sounds. *Journal of the Audio Engineering Society*, 50:331–342.

[Goodwin and Jot, 2006] Goodwin, M. and Jot, J.-M. (2006). Analysis and synthesis for universal spatial audio coding. *In 121th AES Convention, San Francisco, USA. Preprint 6874*.

[Gortler et al., 1996] Gortler, S., Grzeszczuk, R., Szeliski, R., and Cohen, M. F. (1996). The lumigraph. *In SIGGRAPH'96: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, ACM Press, New York, NY, USA*, pages 43–54.

[Govindaraju et al., 2006] Govindaraju, N. K., Larsen, S., Gray, J., and Manocha, D. (2006). A memory model for scientific algorithms on graphics processors. *UNC Tech. Report*.

[Graham, 1999] Graham, R. (1999). Use of auditory icons as emergency warnings: evaluation within a vehicle collision avoidance application. *Ergonomics*, 42(9):1233–1248.

[Green, 2003] Green, R. (2003). Spherical harmonic lighting: The gritty details. In *Game Developpers Conference*.

[Grewin, 1993] Grewin, C. (1993). Methods for quality assessment of low bit-rate audio codecs. In *proceedings of the 12th AES conference*, pages 97–107.

[Herder, 1999] Herder, J. (1999). Optimization of sound spatialization resource management through clustering. *The Journal of Three Dimensional Images, 3D-Forum Society*, 13(3):59–65.

[Herre, 2002] Herre, J. (2002). Audio coding - an all-round entertainment technology. In *Audio Engineering Society 22nd International Conference on Virtual, Synthetic and Entertainment Audio (AES'22), Espoo, Finland*, pages 139–148.

[Horbach et al., 1999] Horbach, U., Karamustafaoglu, A., Pellegrini, R., Mackensen, P., and GüntherTheile (1999). Design and applications of a data-based auralization system for surround sound. *106th Convention of the Audio Engineering Society, preprint 4976*.

[Horry et al., 1997] Horry, Y., Anjyo, K.-I., and Arai, K. (1997). Tour into the picture: using a spidery mesh interface to make animation from a single image. *In SIGGRAPH 97: Proceedings of the 24th annual conference on Computer graphics and interactive techniques, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA*, pages 225–232.

[Howell, 1992] Howell, D. C. (1992). Statistical methods for psychology. *PWSKent*.

[Huang et al., 2005] Huang, G., Yang, L., and He, Z. (2005). Multiple acoustic sources location based on blind source separation. *Proc. of the First International Conference on Natural Computation (ICNC'05)*.

[Huang et al., 2000] Huang, Y., Benesty, J., and Elko, G. (2000). Microphone arrays for video camera steering. *Acoustic Signal Processing for Telecommunications*.

[ITU-R, 1994] ITU-R (1994). Methods for subjective assessment of small impairments in audio systems including multichannel sound systems. *ITU-R BS 1116*.

[ITU-R, 2003] ITU-R (2001-2003). Method for the subjective assessment of intermediate quality level of coding systems. *Recommendation ITU-R BS.1534-1*.

[Johnston, 1988] Johnston, J. (1988). Estimation of perceptual entropy using noise masking criteria. *International Conference on Acoustics, Speech, and Signal Processing*, 5:2524–2527.

[Jot, 1997] Jot, J.-M. (1997). Efficient models for reverberation and distance rendering in computer music and virtual audio reality. *Proc. 1997 International Computer Music Conference*, pages 236–243.

[Jot et al., 1999] Jot, J.-M., Larcher, V., and Pernaux, J.-M. (1999). A comparative study of 3d audio encoding and rendering techniques. *Proceedings of the AES 16th international conference, Spatial sound reproduction, Rovaniemi, Finland*.

[Jot et al., 1995] Jot, J.-M., Larcher, V., and Warusfel, O. (1995). Digital signal processing issues in the context of binaural and transaural stereophony. *98th Convention of the Audio Engineering Society Preprint 3980*.

[Jot et al., 2006] Jot, J.-M., Walsh, M., and Philp, A. (2006). Binaural simulation of complex acoustic scenes for interactive audio. *121th Convention of the Audio Engineering Society Preprint 6950*.

[Jourjine et al., 2000] Jourjine, A., Rickard, S., and Yilmaz, Ö. (2000). Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'00), Istanbul, Turkey*.

[Kalman, 1960] Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transaction of the ASME-Journal of Basic Engineering 82 (Series D)*, pages 35–45.

[Katz, 2001] Katz, B. F. G. (2001). Boundary element method calculation of individual head-related transfer function. i. rigid model calculation. *Journal of the Acoustical Society of America*, 110(5):2440–2448.

[Kayser et al., 2005] Kayser, C., Petkov, C., Lippert, M., and Logothetis, N. (2005). Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology*, 15:1943–1947.

[Kleiner et al., 1993] Kleiner, M., Dalenbäk, B., and Svensson, P. (1993). Auralization: an overview. *Journal of the Audio Engineering Society*, 41(11):861–875.

[Knapp and Carter, 1976] Knapp, C. and Carter, G. C. (1976). The generalized correlation method for estimation of time delay. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 24(4):320–327.

[Knott et al., 2003] Knott, D., van den Doel, K., and Pai, D. K. (2003). Particle system collision detection using graphics hardware. *SIGGRAPH 03 Conference, Sketches and Applications,*.

[Krim and Viberg, 1996] Krim, H. and Viberg, M. (1996). Two decades of array signal processing research. *IEEE Signal Processing Magazine*, 13(4):67–94.

[Krokstad et al., 1968] Krokstad, A., Strøm, S., and Srøsdal., S. (1968). Calculating the acoustical room response by use of a ray tracing technique. *Journal of Sound Vibration*, 8(1):118–125.

[Laborie et al., 2003] Laborie, A., Bruno, R., and Montoya, S. (2003). A new comprehensive approach of surround sound recording. *Proc. 114th convention of the Audio Engineering Society, preprint 5717*.

[Laborie et al., 2004] Laborie, A., Bruno, R., and Montoya, S. (2004). High spatial resolution multi-channel recording. *Proc. 116th convention of the Audio Engineering Society, preprint 6116*.

[Lagrange and Marchand, 2001] Lagrange, M. and Marchand, S. (2001). Real-time additive synthesis of sound by taking advantage of psychoacoustics. *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01), Limerick, Ireland*.

[Leese, 1998] Leese, M. J. (1998). Ambisonic surround sound faq (version 2.8) http://members.tripod.com/martin_leese/ambisonic/.

[Levoy and Hanrahan, 1996] Levoy, M. and Hanrahan, P. (1996). Light field rendering. *In SIGGRAPH '96: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, ACM Press, New York, NY, USA*, pages 31–42.

[Lewicki, 2002] Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neurosceince*, 5(4).

[Lewicki and Sejnowski, 2000] Lewicki, M. S. and Sejnowski, T. J. (2000). Learning overcomplete representations. *Neural Computation*, 12(2):337–365.

[Lu et al., 2004] Lu, L., Wenyin, L., and Zhang, H.-J. (2004). Audio textures: Theory and applications. *IEEE Transactions on Speech and Audio Processing*, 12(2):156–167.

[Malham, 2001] Malham, D. G. (2001). Spherical harmonic coding of sound objects - the ambisonic 'o' format. *Proc. of the 19th AES Conference, Surround Sound, Techniques, Technology and Perception, Schloss Elmau, Germany*.

[Malham and Myatt, 1995] Malham, D. G. and Myatt, A. (1995). 3-d sound spatialization using ambisonic techniques. *Computer Music Journal*, 19(4):58–70.

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

[McAdams et al., 1998] McAdams, S., Botte, M.-C., and Drake, C. (1998). Auditory continuity and loudness computation. *Journal of the Acoustical Society of America*, 103(3):1580–1591.

[Merimaa, 2002] Merimaa, J. (2002). Applications of a 3d microphone array. *Proc. of 112th Convention of the Audio Engineering Society, preprint 5501.*

[Merimaa and Pulkki, 2004] Merimaa, J. and Pulkki, V. (2004). Spatial impulse response rendering. *Proc. of the 7th Intl. Conf. on Digital Audio Effects (DAFX'04), Naples, Italy.*

[Mershon and King, 1975] Mershon, D. H. and King, L. E. (1975). Intensity and reverberation as factors in auditory perception of egocentric distance. *Perception and Psychophysics*, 18:409–415.

[Meyer and Elko, 2004] Meyer, J. and Elko, G. (2004). Spherical microphone arrays for 3d sound recording. *chap. 2 in Audio Signal Processing for next-generation multimedia communication systems, Eds. Yiteng (Arden) Huang and Jacob Benesty, Bosten, Kluwer Academic Publisher.*

[Mills, 1958] Mills, A. W. (1958). On the minimum audible angle. *Journal of the Acoustical Society of America*, 30 (4):237–246.

[Moore, 1997] Moore, B. C. J. (1997). *An Introduction to the Psychology of Hearing.* Academic Press, London, 4th edition.

[Moore and Glasberg, 1996] Moore, B. C. J. and Glasberg, B. (1996). A revision of Zwicker's loudness model. *Acustica - Acta Acustica*, 82:335–345.

[Moore et al., 1997] Moore, B. C. J., Glasberg, B., and Baer, T. (1997). A model for the prediction of thresholds, loudness and partial loudness. *Journal of the Audio Engineering Society*, 45(4):224–240. Software available at http://hearing.psychol.cam.ac.uk/Demos/demos.html.

[Moorer, 1979] Moorer, J. A. (1979). About this reverberation business. *Computer Music Journal*, 3(2):13–28.

[Moses et al., 2002] Moses, R. L., Krishnamurthy, D., and Patterson, R. (2002). An auto-calibration method for unattended ground sensors. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, 3:2941– 2944.

[Mungamuru and Aarabi, 2004] Mungamuru, B. and Aarabi, P. (2004). Enhanced sound localization. *IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics*, 34(3).

[O'Grady et al., 2005] O'Grady, P. D., Pearlmutter, B. A., and Rickard, S. T. (2005). Survey of sparse and non-sparse methods in source separation. *Intl. Journal on Imaging Systems and Technology (IJIST), special issue on Blind source separation and deconvolution in imaging and image processing*.

[Olshausen and Field, 1996] Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607 – 609.

[Owens et al., 2007] Owens, J. D., Luebke, D., Govindaraju, N., Harris, M., Krüger, J., Lefohn, A. E., and Purcell, T. J. (2007). A survey of general-purpose computation on graphics hardware. *Computer Graphics Forum*, 26.

[Pachet and Delerue, 2000] Pachet, F. and Delerue, O. (2000). On-the-fly multi track mixing. In *Proceedings of the 109th Audio Engineering Society Convention*.

[Painter and Spanias, 1997] Painter, T. and Spanias, A. (1997). A review of algorithms for perceptual coding of digital audio signals. In *Proceedings of the International Conference on Digital Signal Processing*, pages 179–205.

[Patterson et al., 1992] Patterson, R., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand, M. (1992). Complex sounds and auditory images. *Auditory physiology and perception, (Eds.) Y Cazals, L. Demany, K.Horner, Pergamon, Oxford*, pages 429–446.

[Pellegrini, 1999] Pellegrini, R. S. (1999). Comparison of data and model-based simulation algorithms for auditory virtual environments. *106th Convention of the Audio Engineering Society, preprint 4953*.

[Porter and Duff, 1984] Porter, T. and Duff, T. (1984). Compositing digital images. *Proceedings of ACM SIGGRAPH*, pages 253–259.

[Pulkki, 1997] Pulkki, V. (1997). Virtual source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, 45(6):456–466.

[Pulkki, 2006] Pulkki, V. (2006). Directional audio coding in spatial sound reproduction and stereo upmixing. *Proc. of the AES 28th Int. Conf, Pitea, Sweden*.

[Pulkki and Faller, 2006] Pulkki, V. and Faller, C. (2006). Directional audio coding: Filterbank and STFT-based design. *In 120th AES Convention, Paris, France, Preprint 6658*.

[Purnhagen, 1999] Purnhagen, H. (1999). Advances in parametric audio coding. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'99), New-Paltz, NY*.

[Rabinkin et al., 1996] Rabinkin, D., Renomeron, R., French, J., and Flanagan, J. (1996). Estimation of wavefront arrival delay using the cross-power spectrum phase technique. *132th meeting of the Acoustical Society of America, Honolulu*.

[Radke and Rickard, 2002] Radke, R. and Rickard, S. (2002). Audio interpolation. *In the Audio Engineering Society 22nd International Conference on Virtual, Synthetic and Entertainment Audio (AES'22), Espoo, Finland*, pages 51–57.

[Raleigh and Strutt, 1907] Raleigh, L. and Strutt, J. W. (1907). Our perception of sound direction. *Philosophical Magazine*, 13:214–232.

[Rickard, 2006] Rickard, S. (2006). Sparse sources are separated sources. *Proceedings of the 16th Annual European. Signal Processing Conference, Florence, Italy,*.

[Rickard and Yilmaz, 2002] Rickard, S. and Yilmaz, O. (2002). On the approximate W-disjoint orthogonality of speech. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*.

[Roads, 1996] Roads, C. (1996). *The Computer Music Tutorial*. MIT Press, 5 edition.

[Roffler and Butler, 1968] Roffler, S. K. and Butler, R. A. (1968). Factors that influence the localization of sounds in the equatorial plane. *Journal of the Acoustical Society of America*, 43:1255–1259.

[Rui and Florencio, 2003] Rui, Y. and Florencio, D. (2003). New direct approaches to robust sound source localization. *Intl. Conf. on Multimedia and Expo (ICME)*.

[Saruwatari et al., 2003] Saruwatari, H., Kurita, S., Takeda, K., Itakura, F., Nishikawa, T., and Shikano, K. (2003). Blind source separation combining independent component analysis and beamforming. *EURASIP Journal on Applied Signal Processing*, 11:1135–1146.

[Savioja et al., 1999] Savioja, L., Huopaniemi, J., Lokki, T., and Väänänen, R. (1999). Creating interactive virtual acoustic environments. *Journal of Audio Eng. Soc*, 47:675–705.

[Savioja et al., 1994] Savioja, L., Rinne, T., and T.Takala (1994). Simulation of room acoustics with a 3-D finite difference mesh. *Proceedings of the International Computer Music Conference*, pages 463–466.

[Schmid, 1986] Schmid, R. O. (1986). Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, AP-34(3):276–280.

[Schroeder, 1975] Schroeder, M. (1975). Models of hearing. *Proc. IEEE*, 63:1332–1350.

[Schroeder, 1962] Schroeder, M. R. (1962). Natural sounding artificial reverberation. *Journal of the Audio Engineering Society*, 10(3):219–223.

[Schroeder, 1970] Schroeder, M. R. (1970). Digital simulation of sound transmission in reverberant spaces. *Journal of the Acoustical Society of America*, 47(2A):424–431.

[Schroeder, 1979] Schroeder, M. R. (1979). Integrated-impulse method measuring sound decay without using impulses. *Journal of the Acoustical Society of America*, 66(2):497–500.

[Seeber and Fastl, 2003] Seeber, B. and Fastl, H. (2003). Subjective selection of non-individual head-related transfer functions. *Proceedings of the International Conference on Auditory Display (ICAD 2003)*, pages 259–262.

[Shaw, 1966] Shaw, E. A. G. (1966). Earcanal pressure generated by a free sound field. *Journal of the Acoustical Society of America*, 39 (3):465–470.

[Sibbald, 2001] Sibbald, A. (2001). Sensaura whitepapers, zoomfx for 3D-sound. http://www.sensaura.com/whitepapers/.

[Skovenborg and Nielsen, 2004] Skovenborg, E. and Nielsen, S. (2004). Evaluation of different loudness models with music and speech material. In *Proc. of 117th Convention of the Audio Engineering Society, San Francisco.*

[Slaney et al., 1996] Slaney, M., Covell, M., and Lassiter, B. (1996). Automatic audio morphing. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96).*

[Smith and Lewicki, 2005] Smith, E. and Lewicki, M. S. (2005). Efficient coding of time-relative structure using spikes. *Neural Computation*, 17(1):19–45.

[Soundfield, 2007] Soundfield (2007). http://www.soundfield.com.

[Stanton and Edworthy, 1999] Stanton, N. A. and Edworthy, J. (1999). *Human Factors in Auditory Warnings*. Ashgate Publishing Ltd.

[Stautner and Puckette, 1982] Stautner, J. and Puckette, M. (1982). Designing multichannel reverberators. *Computer Music Journal*, 6(1).

[Steinke, 1996] Steinke, G. (1996). Surround sound - the new phase. an overview. *100th Convention of the Audio Engineering Society Preprint 4286*.

[Stevens, 1936] Stevens, S. S. (1936). A scale for the measurement of a psychological magnitude: loudness. *Psychological Review*, 43:405–416.

[Stevens, 1975] Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects*. John Wiley and Sons.

[Stoll and Kozamernik, 2000] Stoll, G. and Kozamernik, F. (2000). EBU subjective listening tests on internet audio codecs. *EBU TECHNICAL REVIEW, European Broadcast Union (EBU)*.

[Streicher, 2003] Streicher, R. (2003). The decca tree - it's not just for stereo anymore. *Mix magazine August 2003 issue*.

[Touimi, 2000] Touimi, A. B. (2000). A generic framework for filtering in subband-domain. In *Proc. of the Ninth DSP Workshop (DSP2000), Hunt, Texas*.

[Traunmüller, 1990] Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *Journal of the Acoustical Society of America*, 88:97–100.

[Tsingos et al., 1998] Tsingos, N., Gascuel, and J.-D. (1998). Fast rendering of sound occlusion and diffraction effects for virtual acoustic environments. *Proc. 104th Audio Engineering Society Convention, preprint 4699*.

[Tsingos et al., 2001] Tsingos, N., Funkhouser, T., Ngan, A., and Carlbom, I. (2001). Modeling acoustics in virtual environments using the uniform theory of diffraction. *SIGGRAPH 2001, Computer Graphics Proceedings*, pages 545–552.

[Tsingos et al., 2004] Tsingos, N., Gallo, E., and Drettakis, G. (2004). Perceptual audio rendering of complex virtual environments. *ACM Transactions on Graphics (SIGGRAPH Conference Proceedings)*, 23(3).

[Union, 2003] Union, E. B. (2003). Subjective listening tests on low-bitrate audio codecs. Technical report, Technical report 3296, European Broadcast Union (EBU), Projet Group B/AIM.

[Vercoe et al., 1998] Vercoe, B. L., Gardner, W. G., and Scheirer, E. D. (1998). Structured audio: Creation, transmission, and rendering of parametric sound representations. *Proceedings of IEEE*, 86:922–939.

[Vincent et al., 2003] Vincent, E., Févotte, C., Gribonval, R., Rodet, X., Carpentier, É. L., Benaroya, L., Rödel, A., and Bimbot, F. (2003). A tentative typology of audio source separation tasks. *Proc. of the 4th Intl. Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), Nara, Japan.*

[Wallach, 1940] Wallach, H. (1940). The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology*, 27:339–368.

[Wallach et al., 1949] Wallach, H., Newman, E. B., and Rosenzweig, M. R. (1949). The precedence effect in sound localization. *Journal of Psychology*, 62(3):315–336.

[Wand, 2004] Wand, W. S. M. (2004). Multi-resolution sound rendering. In *Symp. Point-Based Graphics.*

[Warren et al., 1988] Warren, R. M., Wrightson, J. M., and Puretz, J. (1988). Illusory continuity of tonal and infratonal periodic sounds. *Journal of the Acoustical Society of America*, 84(4):1338–1342.

[Wenzel et al., 1993] Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *Journal of the Acoustical Society of America*, 94(1):111–123.

[Wiener and Ross, 1946] Wiener, F. M. and Ross, D. A. (1946). The pressure distribution in the auditory canal in a progressive sound field. *Journal of the Acoustical Society of America*, 18 (2):401–408.

[Wightman and Kistler, 1989] Wightman, F. L. and Kistler, D. J. (1989). Headphone simulation of free-field listening. i: Stimulus synthesis. *Journal of the Acoustical Society of America*, 85 (2):858–867.

[Wightman and Kistler, 1992] Wightman, F. L. and Kistler, D. J. (1992). The dominant role of low-frequency interaural time differences in sound localization. *Journal of the Acoustical Society of America*, 91(3):1648–1661.

[Wightman and Kistler, 1997] Wightman, F. L. and Kistler, D. J. (1997). Monaural sound localization revisited. *Journal of the Acoustical Society of America*, 101 (2):1050–1063.

[Wilson and Darell, 2006] Wilson, K. and Darell, T. (2006). Learning a precedence effect-like weighting function for the generalized cross-correlation framework. *IEEE Journal of speech and audio processing. Special issue on statistical and perceptual audio processing.*

[Yewdall, 2003] Yewdall, D. L. (2003). *Practical Art of Motion Picture Sound*. Focal Press, 2nd edition.

[Yilmaz and Rickard, 2004] Yilmaz, Ö. and Rickard, S. (2004). Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847.

[Young, 1931] Young, P. T. (1931). The role of head movements in auditory localization. *Journal of Experimental Psychology*, 14:95–124.

[Zwicker, 1960] Zwicker, E. (1960). Ein verfahren zur berechnung der lautstärke. *Acustica*, 10:304–308.

[Zwicker, 1977] Zwicker, E. (1977). Procedure for calculating loudness of temporally variable sounds. *Journal of the Acoustical Society of America*, 62 (3):675–681.

[Zwicker, 1984] Zwicker, E. (1984). Dependence of post-masking on masker duration and its relation to temporal effects in loudness. *Journal of the Acoustical Society of America*, 75(1):219–223.

[Zwicker and E.Terhardt, 1980] Zwicker, E. and E.Terhardt (1980). Analytical expressions for critical band rate and critical bandwidth as a function of frequency. *The Journal of the Acoustical Society of America*, 68 (5):1523–1525.

[Zwicker and Fastl, 1999] Zwicker, E. and Fastl, H. (1999). *Psychoacoustics: Facts and Models*. Springer, 2 edition.

[Zwicker et al., 1991] Zwicker, E., Fastl, H., Widmann, U., Kurakata, K., Kuwano, S., and Namba, S. (1991). Program for calculating loudness according to DIN 45631 (ISO 532B). *Journal of the Acoustical Society of Japan*, 12:39–42.

[Zwicker et al., 1957] Zwicker, E., Flottorp, G., and Stevens, S. S. (1957). Critical band width in loudness summation. *Journal of the Acoustical Society of America*, 29 (5):548–557.