UniS

University of Surrey

Department of Music and Sound Recording

School of Arts

An onset-guided spatial analyser for binaural audio

by Ben Supper

August 2005

Thesis submitted in fulfilment of the requirement
of the degree of Doctor of Philosophy

ABSTRACT

A novel system of computer algorithms is formulated to perform onset-guided source localisation using binaural stimuli. This system, called the *spatial analyser*, will analyse spatial attributes including source location. It is computationally efficient, compatible with streamed binaural data, and uses psychophysically-valid analysis techniques wherever possible. The main components of the system are a model of the human auditory periphery, an onset detector, a running localisation algorithm, and some logic to combine these.

The onset detector is designed specifically for spatial analysis, using a combination of linear regression and band-pass filtering techniques to produce a response that is sensitive to auditory onsets and robust to noise. It also features an implementation of the precedence effect.

To localise sounds, an efficient method is found for extracting interaural time difference cues using the interaural cross-correlation function. Instantaneous interaural time and intensity differences of the binaural signal are calculated and mapped to lateral angle using a database of interaural cues. A cross-weighting formula combines the interaural time and intensity data across frequency bands. Loudness weighting is then applied to every critical band to produce an output.

Spatial information is handled throughout the localisation algorithm in the form of lateral angle histograms. These are discrete functions, which specify localisation strength against lateral angle for any particular combination of cues.

In a series of validation experiments, the spatial analyser determines the direction of most sound sources to within 10° in a reverberant environment. For most sources, this performance is maintained even when a substantial amount of white noise is added to the audio as a confusing signal. The output data is also shown to be compatible with auditory source width extraction. With slight modifications, the spatial analyser can also approximate source distance.

CONTENTS

FIGURES AND TABLES

1 Introduction

2 Early reflections and spatial impression

3 Onset detection algorithm

4 Localisation algorithm

5 Investigation

6 Conclusion

EQUATIONS

2 Early reflections and spatial impression

3 Onset detection algorithm

4 Localisation algorithm

5 Investigation

ACKNOWLEDGEMENTS

many people have contributed to this project in many different ways . in particular, I thank the following people and institutions for their support . above all : pharos communications limited : for making this research project possible through their generous sponsorship . my parents : for helping me through twenty-one years of full-time education, thus unwittingly commissioning this thesis . audio engineering society educational foundation : for the substantial award that sustained me throughout 2003-2004 . digital music research network : for funding flights and accommodation, permitting me to present a poster at the 116th AES convention . tim brookes : for supervising me so patiently . francis rumsey : for his positive meta-supervision . michelle rosenthal : for proof-reading, companionship, support, and sanity . russell mason : for lending me his old copy of MATLAB 5 and acquiring a Cortex dummy head, also for encouragement, criticism, and occasionally for humour . slawek zielinski : for listening enthusiastically, talking through things, and proof-reading my papers . rafael kassier , tobias neher , hyun-kook lee , natanya ford , martin dewhirst : for keeping me company and letting me speak semi-articulately about my project for up to half an hour at a time . andy baker , ben beeson , nick fletcher , miriam iorwerth , sarah and steven pain , simon pearson , jessie taylor , rich watts , catherine lee : for reminding me that there was always something else to do . david griesinger , gilbert soulodre , chris travis : for helping to nudge this project along its eventual path . simon caton , jake gladman , adrian klemens , ross richardson , rich wheatley , matt derbyshire , mike reed : for being in bands with me that weren't quite organised enough to get anywhere, thus saving me from a moral dilemma . all shaped this thesis, and to all I extend my gratitude .

1 INTRODUCTION

This research project is intended to fulfil a need in the broadcasting industry for a reliable real-time visual indicator of the spatial attributes of audio signals. In a large broadcasting operation, it is not always possible to monitor the sound from every television and radio channel being transmitted, but it is usually possible for a single operator to survey many video screens at the same time. Furthermore, broadcasting suites often have limited floor space and high levels of background noise, so optimal listening conditions are seldom possible. This is true especially when dealing with surround formats, as these require precise placement of loudspeakers and a good deal of space around the installation. A visual display that represents the changing spatial sound attributes of an audio input would be useful in these circumstances, so that the spatial attributes of the programme material can be checked without the need for surround loudspeakers and a high-quality listening environment.

The invention of a visual display of spatial attributes would require considerable advances in the simulation of auditory perception, and the consolidation of many disparate bodies of research. As this project has progressed, it has become clear that the applications for this research are not limited to the realisation of a visual display for broadcasting: there may also be conceivable applications for a working spatial analysis algorithm in the field of architectural acoustics. As more becomes known about spatial hearing, more tools are available to designers of environments where good acoustics are crucial, such as lecture rooms, theatres, and concert halls.

1.1 Definition of 'spatial attributes'

In order to proceed meaningfully, it is important to ascertain what is meant in this thesis by 'spatial attributes'. A basic spatial attribute of a sound is its location relative to the listener. As with all physical concepts that we can perceive, the term 'source location' may refer to either the actual placement of a sound source, or its perceived location. When referring to source location, it must be clear which of these concepts is meant. This may be signalled either by preceding the term with the word 'actual' or 'perceived', or by using an unambiguous psychological term to refer to the perceived phenomenon. For example, the word 'location' can denote either the actual or perceived source position, but ‘localisation’ is used to refer an estimated source position elicited from a human listener or computer algorithm.

Localisation is often deemed to be the fundamental purpose of spatial hearing, because it is useful to an individual’s survival. For this reason, perceived source location will now be termed the *primary spatial attribute*. The other spatial properties of a sound, *secondary spatial attributes*, do not involve source location.

Source width is a convenient example of a secondary spatial attribute. The term ‘source width’ may refer to the physical concept of breadth, to its perception, or to both. For example, a grand piano or a car engine may have a large perceived width at close quarters because the actual width of the source is large — coherent energy is being radiated over a large physical surface. On the other hand, sound produced by a solo cello in an orchestra pit may under some circumstances appear wide, owing not to its size, but to the reverberant properties of the room. Although the test stimuli used within this project presume a link between actual and perceived secondary spatial attributes, it is the perceived spatial attributes that are considered important.

The dichotomy into primary and secondary attributes is necessary because the majority of spatial hearing research deals only with the problem of localisation. Early literature that investigated secondary attributes tended to treat their perception as a unidimensional phenomenon, using a term such as *auditory spaciousness* [Blauert and Lindemann 1986] or *spatial impression* [Barron and Marshall 1981; Blauert and Lindemann 1986]. These terms are still in use, but ‘spatial impression’ is now generally treated as an umbrella-term, to describe a collection of perceived phenomena [Rumsey 2002].

A large part of recent auditorium acoustics literature focuses on auditory or apparent source width, ASW. This term refers to a phenomenon that is perceived, but has no direct physical correlate. For example, it is not always possible to gain an indication of ASW by applying a measuring tape to a listening room and taking down the dimensions of the sound source and its environment. However, there are ways of approximating ASW by analysing recordings, and many analytical methods have been developed in recent years.

In general, ASW is associated with early room reflections. It is not the only example of a secondary spatial attribute, but it has attracted considerable attention in recent years, and along with apparent source distance, will serve well as an example of a secondary spatial attribute in this project.

1.2 Fundamental specifications

A processing algorithm that interfaces with a real-time visual display must satisfy three requirements. Firstly, it must be compatible with streamed data. Secondly, it must be basic enough to run in real time on a practical system. Thirdly, the delay between an audio signal entering the system and an output being generated must be as short as possible. Ideally, this would be less than 33ms, because this is the shortest video frame period used in broadcasting (33ms is approximately the reciprocal of the NTSC frame rate — 30 frames per second).

An additional set of requirements is imposed by the fact that the algorithm must mimic, to an extent, the response of the human auditory system in order to extract spatial information from it. A degree of correlation is required between what would be perceived by a listener and what is indicated by the system.

From the beginning it was decided that a binaural format would best suit this system. A large proportion of recent research on spatial scene analysis uses binaural source signals, and using the same format would make those findings directly compatible with this project. Also it would be sensible to use a format that is designed to be replayed straight into a listener's ears, as it already contains most of the cues of interest, and requires no extra processing before analysis. One further advantage of the binaural format is entirely pragmatic: binaural audio is simple to record because the microphone array is self-contained and, when using a modern dummy recording head, already calibrated. It is also easy to record, store and transfer binaural audio, because the majority of tape formats and computers are designed to deal with two audio channels. Furthermore, it is relatively simple to convolve a usable binaural signal from any audio format that is intended for loudspeaker reduction.

Binaural data is not without its problems. These are examined in detail in Section 4.2. These shortcomings reduce the ability of a listener (and therefore an analyser that is based on human audition) to discriminate between sounds coming from the front and rear hemispheres, and to discern the angle of elevation of a source. Many of the problems stem from the fact that a person listening to a recording cannot interact, even in a simple way, with the recorded scene. In nature, a listener's head and body are unconstrained, and so a feedback loop operates between the listener and the binaural data. When a stationary dummy recording head is replayed, this feedback loop is lost, and some of the more advanced localisation tasks become extremely difficult, if not impossible.

Some compromises are necessary. These are covered in more detail in Chapter 4, but they are important enough to state briefly here. The spatial analysis algorithm designed within this project will not attempt to discriminate between the front and rear hemispheres of audition, and will use the one-dimensional measure of lateral angle instead. Therefore, rear loudspeakers in a surround set-up will not be perceived explicitly as rear loudspeakers. This simplification also implies that any existing research which relates spatial impression to front-to-back energy ratio (for example, Morimoto [1997]) cannot be taken into account in this project.

1.3 Scope and aims of the thesis

The task of designing of a visual display of spatial attributes of sound can be divided into two stages. The first stage is to find a way to extract the fluctuating spatial attributes from the binaural data. Then a method must be developed to display these attributes. This project concentrates only on the first stage: the spatial analysis of binaural audio.

Thus, the aim of the thesis is to develop methods for extracting spatial information from a streamed binaural signal in a way that is psychoacoustically motivated, computationally efficient, and as precise as possible. The algorithms that perform this task will be referred to collectively as the *spatial analyser*.

The research question is how this spatial analyser may best be realised. A number of systems for extracting auditory features from streaming data have emerged over the last decade, most of which focus on timbral or rhythmic attributes, and relatively few of which are designed to process spatial information. The best of these spatial systems are able to monitor complex and arbitrary binaural signals, discern their component auditory events and sources, and estimate source positions. The principal aim of this project is to build such a localiser with its basis in known psychoacoustic mechanisms. This localiser can then be extended to extract secondary spatial attributes.

The techniques of this project are based on the extraction of source direction. This is because source direction is easy to measure, control, and compare when recording and analysing test stimuli, and has already been thoroughly researched. However, the auditory processing methods that are used in this project have been designed specifically to be extended to cover secondary spatial attributes. It will be shown that, with some minor modifications, the spatial analyser can be used to extract the attributes of source width and source distance automatically. It is not unreasonable to expect that other spatial attributes could also be obtained.

1.4 Thesis method and structure

The topology of the spatial analyser is based on the Zurek model [Zurek 1987], which is discussed in detail at the beginning of Chapter 2. This model, upon which Figure 1.1 is based, is a framework for spatial audio analysis. It specifies the interconnection of a running localisation algorithm, a separate auditory onset detector, and a method for suppressing the spatial processing of early room reflections. The output data is obtained from a fourth algorithm, which Zurek calls simply 'time averaging', but is called 'location gate' in Figure 1.1. This revised nomenclature is based on a paper by Griesinger [1997], and reflects the fact that the algorithm will need to be more sophisticated if secondary spatial attributes are to be extracted. For localisation, however, its purpose is identical: to make sense of the data that emerges.

As well as being a starting point for a computer implementation of spatial hearing, the Zurek model is also psychoacoustically valid. The neurophysics of spatial hearing is often divided into the same four categories. Thus, the chapter divisions within this thesis also encapsulate the separate components of the Zurek model.

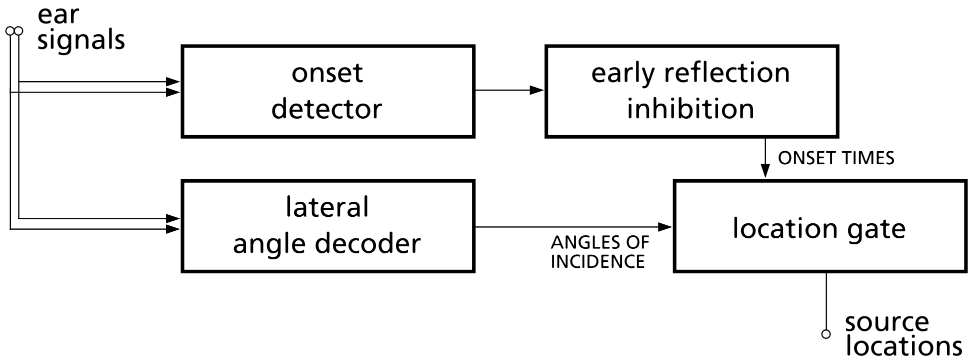


Figure 1.1. Processing structure based on the Zurek model [Zurek 1987].

Chapter 2 of this thesis reviews research into the various aspects of the precedence effect, and echo suppression in human audition. These phenomena are integral to spatial perception because they become active in the first milliseconds after onset, when room reflections are providing their most fundamental, unambiguous information about the dimensions and properties of the acoustic environment. In the first 100ms after a new sound source appears, many secondary spatial attributes are heard that are perceptually fused with the source, such as width and distance. To understand these aspects of spatial hearing therefore requires a firm understanding of echo suppression phenomena.

In **Chapter 3**, a new auditory onset detector is developed that is specifically matched to the requirements of the spatial analyser. Aspects of existing onset detection mechanisms are combined with new techniques in order to enhance the performance of the algorithm.

Chapter 4 describes the design of the running localisation algorithm, which extracts spatial information to complement the timing information from the onset detector. This algorithm is based on a combination of existing approaches, and has been optimised to run efficiently whilst sacrificing a minimum of spatial accuracy.

Using a variety of stimuli recorded under controlled conditions, **Chapter 5** investigates the performance of the onset detector and localisation algorithm. These are always tested together, although inferences can be made about the performance of each component. Specifically, the strengths and weaknesses of the system, and the causes behind them, can be examined. Some of these problems are inherent in binaural listening, and are encountered in human audition; others are specific artefacts of the algorithms used in this project.

Chapter 6 summarises the findings from this project, reviews the contributions that this thesis has made to the field of spatial hearing, and stipulates the improvements that may be made to the algorithms in future.

1.5 Summary

This thesis describes and tests an algorithm that aims to extract a number of spatial source attributes from arbitrary binaural data. While existing algorithms concentrate on source location, the aim of this project is to produce an analysis method that additionally enables the extraction of *secondary spatial attributes* — those that are not directly associated with localisation, such as the detection of auditory source width and apparent source distance.

The main application of this technology is the formulation of a new type of meter for use in broadcasting applications, but the knowledge obtained in creating this also has uses in the field of auditorium acoustics. The broadcasting application imposes some requirements on the algorithm. Firstly, it must be designed in such a way that it is compatible with streamed audio. Secondly, computational efficiency is an important priority, so that a real-time implementation is possible. Thirdly, the algorithms should be physiologically motivated wherever this is feasible.

Input data is required to be in binaural format. This format is helpful because it requires a physiological approach to be taken to data analysis. Furthermore, it is easy to acquire, and may be convolved from any loudspeaker format with little effort. However, the simplicity of binaural representation imposes limitations on the usefulness of the input data: front-back discrimination and angle-of-elevation detection cannot be performed reliably, and will not be attempted.

Zurek's model is the basis of the structure of this spatial analyser. This model requires separate components for auditory onset detection, running source localisation, an echo suppression model, and a system for making sense of the data that is generated by these components. The Zurek model's simple framework also provides a convenient chapter structure for this thesis.

2 EARLY REFLECTIONS AND SPATIAL IMPRESSION

The purpose of this chapter is to present a consolidated summary of research into the precedence effect and echo suppression phenomena in human audition. A hierarchy of mechanisms are responsible for the ways in which early lateral reflections are perceived, and these have a profound influence on spatial perception. Thus, the influence of the precedence effect permeates much of this thesis. This chapter will therefore form a basis for justifying the methods that are chosen and the theories that are advanced over the following chapters.

This chapter will not attempt to describe in detail every facet of the human echo suppression mechanism: certain parts of it cannot be implemented owing to the sophistication of the human auditory system, and certain aspects are under conscious control [Clifton and Freyman 1997].

Before going any further, it is worth defining the terms ‘precedence effect’ and ‘echo suppression’. These are often used interchangeably, but in their strictest senses refer to different things. The first mention of the ‘precedence effect’ is by Wallach et al. [1949], and it refers to the aspect of echo suppression that is investigated in their paper:

> If two sounds that are nearly alike follow each other in close sequence, they will be heard as one sound; and if an interval of at least 1ms (0.3 m length of path) separates them, the total sound will be heard coming from the location of the prior sound. This localization in terms of the earlier sound we shall term the “precedence effect.” [Wallach et al. 1949: 819]

This effect was observed to break down for impulsive stimuli when the inter-click interval was greater than 5ms, or after approximately 40ms when musical signals were used. While the precedence effect was operating, the delayed click could influence the perceived direction of the overall auditory event by a maximum of 7°.

As Hartmann states, the power of the automatic dereverberation that human listeners unconsciously perform demonstrates that other echo suppression mechanisms exist. While the influence of the precedence effect is fading 50ms after the direct sound arrives at the listener, the influence of other dereverberation phenomena can extend for several seconds after the onset of direct sound [Hartmann 1997]. These other mechanisms are commonly referred to as ‘echo suppression’ phenomena.

Unfortunately, the use of the term ‘echo suppression’ is now controversial for two reasons. Firstly, the word ‘echo’ was originally used to refer to all room reflections, but now implies an isolated, individually-audible reflection. The word is used in this sense by Haas [1951] and Blauert [1997: 224], and extended into the term ‘echo threshold’ to describe the upper limit of the precedence effect. Secondly, the word ‘suppression’ infers the existence of a masking effect, which would prevent the perception of reflections altogether. While the perceived intensity of the reflection may be reduced by the presence of the direct sound, it is often just the suppression of the spatial content of a reflection that is of interest. The extent of this suppression is never total. Even ‘suppressed’ spatial content may be perceived in a number of ways: for example, as a broadening of the source, or as a sense of environment-related spatial impression.

The following terminology is used in this chapter: ‘the precedence effect’ refers to the short-term effect explored by Wallach et al., Haas, and their successors. In the absence of a better umbrella-term, all the phenomena that rely on the selective inhibition of spatial information — including the precedence effect — will be referred to as ‘echo suppression’ phenomena. In turn, the precedence effect will be examined, along with evidence of its adaptability. The impact of the precedence effect on spatial perception will then be examined, followed by the more complicated echo suppression mechanisms. Finally, the Griesinger model will be investigated. This is the only current elaboration of Zurek’s model (see Chapter 1, Figure 1.1), besides the one in this thesis, that attempts to account for the perception of secondary spatial attributes.

2.1 The precedence effect

Haas approached the precedence effect from the point of view of sound reinforcement. His experiment investigates the thresholds in time and intensity for which a loudspeaker-generated ‘reflection’ of speech is perceived as equally loud as the direct sound, and the thresholds beyond which this reflection becomes disturbing to a listener [Haas 1951]. These results are summarised in Figure 2.1.

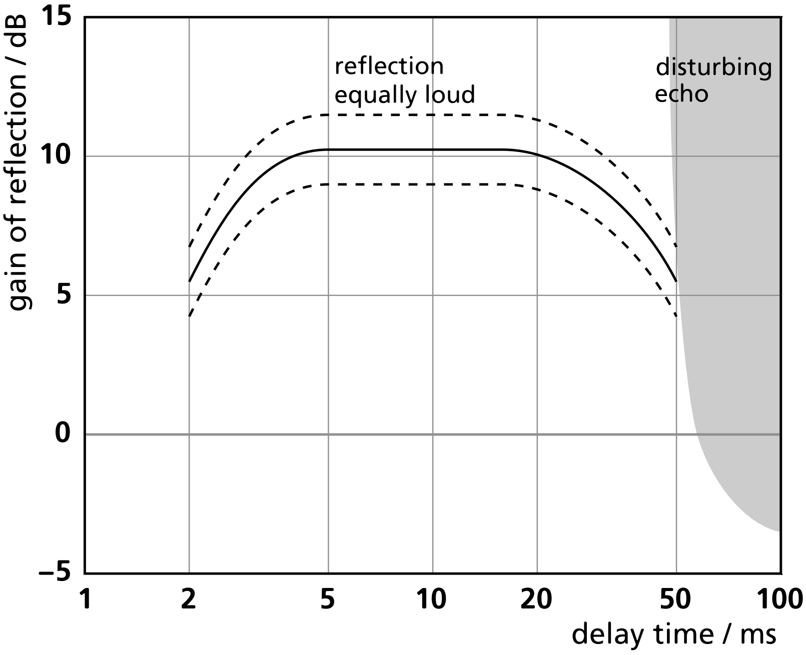


Figure 2.1. Summary of Haas’s data [Haas 1951], obtained from 15 observers. The solid line shows the mean response, and the dashed lines are the extent of the deviation. The ‘disturbing echo’ criteria is based upon 50% of subjects referring to the echo as disturbing. This was found to be a function also of rapidity of speech, reverberation time, and the timbre of the echo.

There are a number of differences in methodology between the experiments of Haas and those of Wallach et al. [1949]. Haas uses loudspeaker delivery as opposed to headphone delivery, and examines the masking-type suppression of a reflection rather than the extent of its spatial fusion with the direct sound. For the latter reason, the Haas effect is not the same as the precedence effect. Nevertheless, the two papers concur as far as the establishment of the echo threshold for musical stimuli.

As well as providing broader insights into the precedence effect, subsequent investigations have uncovered a number of additional complexities. A series of experiments conducted by Barron [1971] investigate the effects of a single lateral reflection on spatial impression. These are essentially investigations into the precedence effect, as the phenomena under investigation were the extent of image shift caused by the distracting reflection and the nature of the effect of the reflection on spatial impression. Data from Barron's experiments, summarised in Figure 2.2, essentially combine many of the conditions of Haas's experiments — loudspeaker listening conditions, programme-based test stimuli, and a single delayed artificial reflection — with the spatial echo suppression phenomenon studied by Wallach et al..

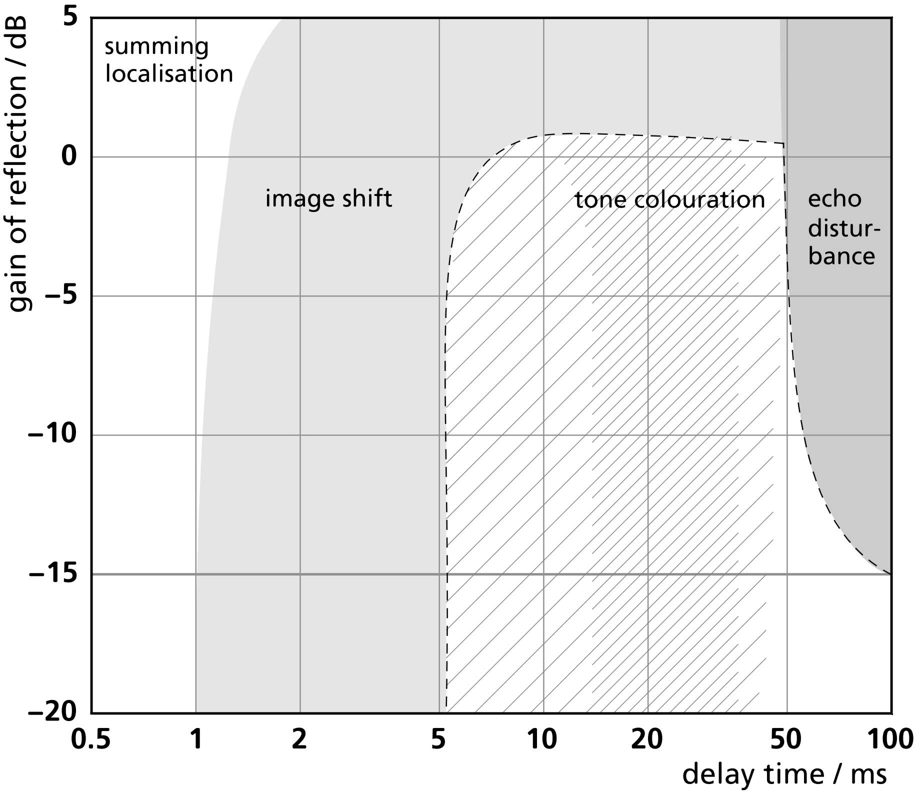


Figure 2.2. The effect of a single reflection on an orchestral music source. When the parameters of the reflection fall inside the dashed area, a sense of spatial impression results. The precedence effect causes image shift and spatial impression. Data based on a graph by Barron [1971]; summing localisation data based on information in Blauert [1997] and Hartmann [1997].

An important series of three experiments by Zurek examines the precedence effect. Like the experiments of Wallach et al, the effect is discovered to behave differently depending on the envelope of the stimulus.

The second of Zurek's experiments employs pairs of 1ms noise bursts, presented over headphones. In Zurek's methodology, the latter noise burst of every pair contains a small interaural time or amplitude difference. Three pairs of noise bursts are presented in each test interval, separated by 400ms, but the interaural cues are reversed on either the second or third noise-burst pair. The listener has to detect this, and is compelled to answer by pushing one of two buttons corresponding to the second or third pair. The 'correct' answer is then indicated by the equipment before the next test interval begins. In this way, the just-noticeable difference (JND) in each interaural cue can be determined against inter-burst delay.

The results of this experiment are reasonably uniform across listeners. They show the JND to be influenced strongly by the inter-burst interval, and concur with the click-based experiments of Wallach et al.. The JND is minimal, and thus spatial acuity is most sensitive, for an inter-burst delay of 500µs or less. This is what Blauert calls the period of *summing localisation*. The JND climbs sharply after this, and peaks 2–3ms after onset for both interaural cues. For a detection rate of 67%, the JND rises to around 250µs or 10dB above threshold. This is when the precedence effect is regarded to be maximal. The hearing system regains full spatial acuity when the inter-click interval reaches 10ms. These findings are very similar to those from the click experiments of Wallach et al.

Zurek's third experiment employed a more continuous stimulus: a 50ms burst of interaurally-coherent noise, in which is hidden a 5ms noise burst containing an interaural difference in either time or intensity. JNDs were discovered using the same hidden-reference, forced-choice paradigm. Again, the precedence effect was found to be maximal at 2–3ms. Although the JND fell monotonically after this, the loss of spatial acuity persisted throughout the duration of the noise. These findings are sketched in Figure 2.3.

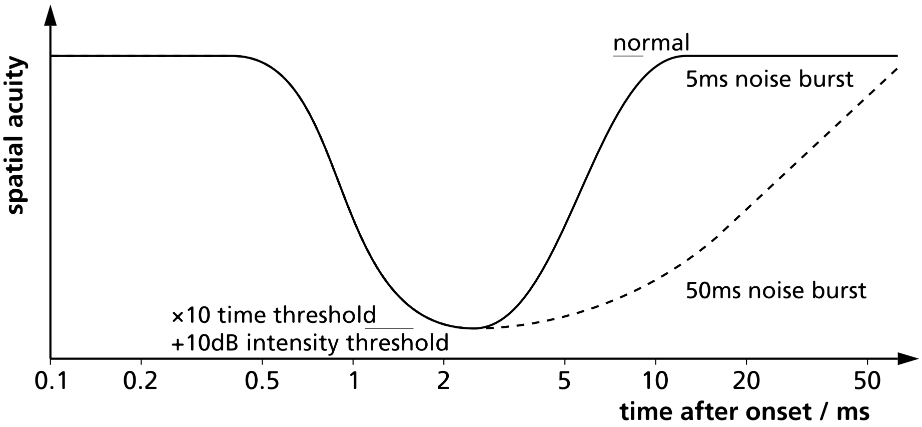


Figure 2.3. Sketch of the results of Zurek [1980]. 'Spatial acuity' is plotted on a relative scale. After 2ms, JNDs for interaural time difference are about ten times larger than their minimal value, whereas JNDs for interaural intensity differences are approximately 10dB higher than their minimal value (approximately 1dB).

2.1.1 Reflection angle and onset rate

Wallach et al., Haas, Barron, and Zurek's experiments form a comprehensive overview of the precedence effect. Together, they quantify the time constants involved in the precedence effect, and indicate the manner in which reflections of different times and intensities are perceived. They also demonstrate the extent to which sensitivity to spatial information is affected.

The exact manner in which the rate of onset and stimulus duration affect the behaviour of the precedence effect, and the extent of the image shift caused by early reflections, are the subject of some further investigations by Rakerd and Hartmann.

All the stimuli in the experiments of Hartmann [1983] and Rakerd and Hartmann [1985; 1986] are based on noise bursts of a minimum length of 50ms. The methodology throughout this series of experiments is based around the replay of the noise stimulus from one of eight closely-spaced loudspeakers arranged in an arc around the listener. Localisation performance could be studied by extracting the angular error between the actual sounding speaker and the loudspeaker chosen by the listener.

The first of these experiments [Hartmann 1983] was conducted in a hall with variable acoustics and ceiling height to investigate the distracting effect of early reflections. The second and third experiments clarified certain aspects of the findings by moving the experimental set-up to an anechoic chamber and using a rectangle of particle board to simulate one first-order room reflection at a time. Then the effect of onset rate on localisation accuracy was investigated by employing onset ramps of up to 100ms.

For the noise stimulus, Hartmann discovered that reverberation time had very little effect on localisation accuracy [Hartmann 1983]. Haas discovered that lengthening reverberation time slightly increases the echo threshold (the point at which an echo becomes 'disturbing') [Haas 1951]. However, Hartmann's experiment does not show a correlation between the echo suppression aspect of the precedence effect and reverberation time.

Localisation accuracy was enhanced significantly when the ceiling was lowered [Hartmann 1983]. This suggests that ceiling reflections, which come from a similar direction to the source, have a less confounding effect on source localisation than reflections from the side walls. This concurs with the findings of Barron and Marshall [1981] and Ando and Gottlob [1979], who find that ceiling reflections contribute some sense of spatial impression to the direct sound, but that lateral reflections contribute more. Increasing the amount of spatial impression or perceived breadth of the source also reduces interaural coherence, and therefore impairs localisability. This inverse relationship between source width and localisability is widely accepted.

Rakerd and Hartmann [1985] extended these investigations, and confirmed the earlier findings. When only one acoustic reflection is present, a reflection from the direction of the floor and ceiling confounds localisation accuracy, but markedly less than reflections from the side walls.

The level and rate of onset was found to be important to the working of the precedence effect, and the salience of the distracting reflections. In Rakerd and Hartmann [1986], an onset time of 100ms (equivalent to a rate of about 500dB per second in this experiment) still triggered the precedence effect, but the effect begins to fail for onsets longer than this. Rakerd and Hartmann also found that the upper onset time limit of the precedence effect is dependent on the delay between direct sound and reflection, with longer delays imposing longer onset time limits.

2.1.2 Lindemann's precedence effect model

Lindemann's simulation of the low-level precedence effect is one of few serious attempts to model the phenomenon, and it is therefore worth examining to consolidate the information presented so far. Lindemann's model is unusual as its principal aim is to model the human auditory system. Although a number other precedence effect algorithms exist, such as the ones formulated by Huang et al. [1997] for a robotic system, and by Schwartz et al. [2001] for speech localisation, the motivation behind the inclusion of echo suppression mechanisms in these systems is pragmatic. Localising transient sounds in rooms is almost impossible without a precedence effect algorithm, but modelling the psychophysics of the precedence effect is usually regarded as less important than producing a computationally efficient and effective algorithm.

Lindemann's precedence effect model is designed for incorporation into a running cross-correlation algorithm. In this algorithm, eighty identical stages run simultaneously and are linked together, each one testing for a different interaural time difference. (A block diagram of one stage of Lindemann's running cross-correlation algorithm is presented in Chapter 4, Figure 4.9.) Every stage runs an instance of the precedence effect model, which Lindemann calls 'dynamic inhibition'.

The amount of inhibition is controlled by a low-pass filtered version of the output of the stage. Thus, the more correlated a signal at the stage's characteristic interaural time difference, the stronger the inhibition applied. This inhibition attenuates two neighbouring stages, passing the influence of the precedence effect downstream. Practically, this cascading does not significantly affect the output of a correlation stage unless it is adjacent to, and downstream from, a maximally excited stage. However, it makes the exact working of the precedence effect algorithm heavily dependent on input signal.

The design of the filter is crucial to this simulation of the precedence effect. Each successive value depends on the previous value according to the following formula:

$$\Phi(n) = x(n-1) + \Phi(n-1)\, e^{-1.25\times10^{-3}} \big(1 - x(n-1)\big) \qquad (2.1)$$

where $x(n)$ is the current input signal value, where $0 \leq x \leq 1$, and $\Phi(n)$ is the inhibition constant. This formula is designed to work with an input sampling frequency of 80kHz. The resulting variation of $\Phi(n)$ against time can be seen in Figure 2.4.

Lindemann's model simulates only some aspects of the precedence effect. For a correlated transient sound, it mimics human response fairly well. The dynamic inhibition reaches a peak within 1–2ms after onset. For naturalistic signals, this inhibition will be almost complete ($\Phi(n) = 1$) so that further spatial information will be suppressed entirely. Haas's findings state that even suppressed echoes can be amplified so that they are perceived to be of equal loudness to the direct sound, and Wallach et al. and Zurek show that some spatial acuity remains during the operation of the precedence effect. All note a loss of sensitivity of approximately 10dB. Thus Lindemann's nearly-complete inhibition, even for highly-correlated signals, is not accurate. In the Lindemann model, recovery from transient excitation takes between 10 and 20ms, depending on the correlation of the transient. According to Zurek, this is entirely representative of the workings of the human auditory system.

However, Lindemann's model is not so compatible with continuous data. While Zurek's listening experiments show a prolonged recovery of spatial sensitivity when continuous and correlated signals are presented, the Lindemann model will continue to inhibit the signal maximally until the signal is removed or its cross-correlation declines. Thus, there is no 'echo threshold' in Lindemann's simulation of the precedence effect.

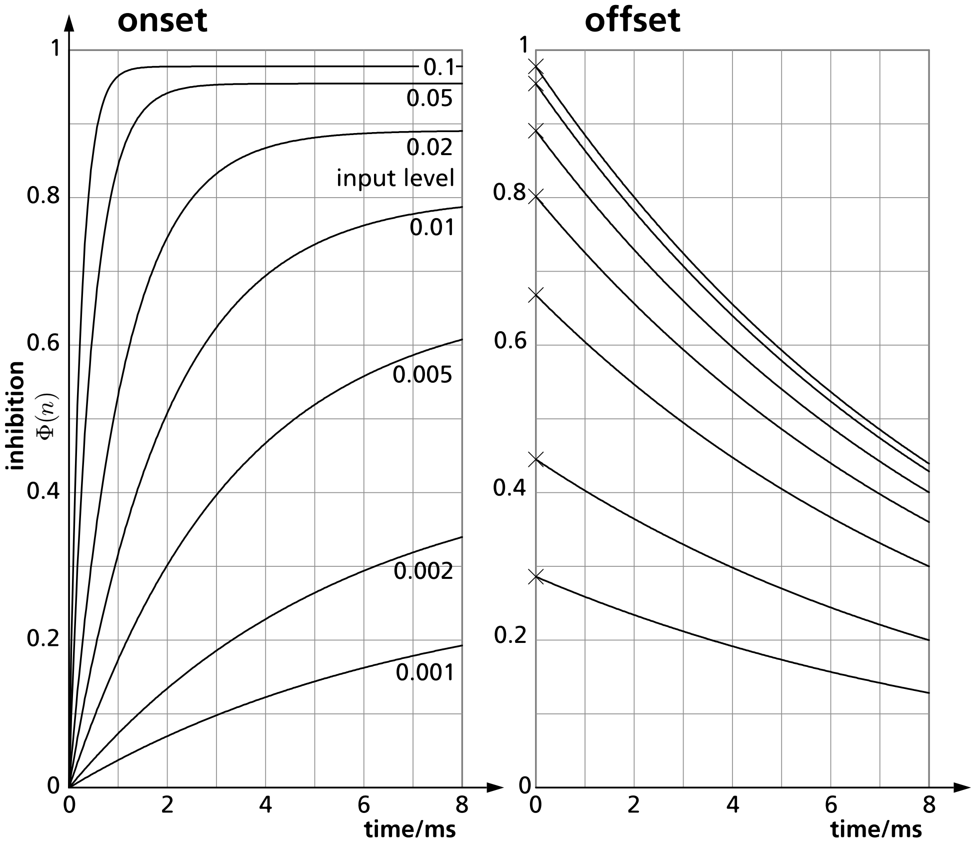


Figure 2.4. Behaviour of Lindemann's dynamic inhibition model for input signals of different constant levels [Lindemann 1986].
Left panel: onset characteristic: input signal applied after silence.
Right panel: offset characteristic, after a long input signal is removed.

Part of the reason for this incompatibility with continuous data is Lindemann's reliance on click-based experiments for listener data. There are two problems with listening experiments that use clicks as stimuli. The first is that they attempt to make general inferences from extremely short auditory events with unnaturally rapid onset and decay times. This disregards the characteristics of music, speech, and most everyday sounds, which have slow onset and offset characteristics, and periods of sustained activity. The behaviour of the precedence effect is highly dependent on the envelope of the sound source under investigation [Rakerd and Hartmann 1986; Mason and Rumsey 2001]. Unfortunately, this cannot be investigated using click-based experiments.

Furthermore, a growing body of research suggests that the precedence effect exhibits neural plasticity: it adapts over a few seconds upon exposure to a new listening situation [Clifton and Freyman 1997]. After a limited amount of exposure, listeners can suppress, either consciously or unconsciously, some strong, isolated reflections that would otherwise be audible. This adaption was manifested in Wallach et al.'s series of experiments [Wallach et al. 1949], in which different thresholds for image shift were recorded depending on whether the delay time between the first and second click was routinely being increased or decreased during the experiment. This training effect was later investigated by Saberi and Perrott [1990] who concluded that with sufficient training, the spatial echo suppression produced by the precedence effect can be almost entirely annulled in click experiments.

For these reasons, click experiments risk falsely simplifying the precedence effect. However, if the inter-click interval is kept fairly short, as it is in the series of experiments conducted by Clifton and Freyman [1997], the auditory system can adapt as it would to a continuous stimulus, and the effects of neural plasticity can be included in the experiment.

2.2 Measuring spatial impression

The precedence effect is central to the perception of spatial impression [Morimoto 2002]. A number of experiments that investigate the effect of lateral reflections on spatial perception have already been examined. It is clear from these experiments that all reflections that arrive from a direction away from the direct sound contribute a sense of spatial impression. Furthermore, lateral reflections contribute a greater sense of spatial impression than frontal reflections. Barron and Marshall [1981] represented these findings by creating the prototypical metric for spatial impression, called the *lateral energy fraction* (L_f), from which many measures are descended:

$$L_f = \sum_{t=5\text{ms}}^{80\text{ms}} r(t) \cos\phi \Big/ \sum_{t=0\text{ms}}^{80\text{ms}} r(t) \tag{2.2}$$

In this equation, $r(t)$ is the signal energy, t is the time after onset, and ϕ is the angle between a single reflection and the origin of the aural axis. (The aural axis passes through both of the listener's ears. Its origin is defined as the centre of the head.) In this form, the L_f must be calculated from architectural plans, rather than measured. The incident angles, arrival times, and relative levels of every reflection that reaches the ears within the first 80ms have to be determined and combined. The numerator excludes the direct sound and suppressed early reflections, and the denominator includes these. The cosine weighting in the numerator ensures a higher L_f for environments with strong lateral reflections. This lateral energy fraction correlates well with the spatial impression data elicited from Barron and Marshall's listening subjects.

Although Kleiner specifies a way of measuring the L_f using two omnidirectional microphones and digital signal processing [Kleiner 1989] this system is rather complicated, and too similar to the now-popular interaural cross-correlation function (IACCF) to have become widely adopted. Unlike the IACCF, it is also too abstract to be psychologically plausible.

Subsequent investigations rely on a similar measurement that is more readily obtainable: the early lateral fraction, or LF_E. This is obtained from a two-microphone recording of an impulse response:

$$LF_E = \sum_{t=5\mathrm{ms}}^{80\mathrm{ms}} p_8^2(t) \Bigg/ \sum_{t=0\mathrm{ms}}^{80\mathrm{ms}} p_o^2(t) \tag{2.3}$$

$p_8(t)$ is the waveform recorded by a figure-of-eight microphone with its null pointed towards the source; $p_o(t)$ is the waveform recorded at a coincident omnidirectional microphone, balanced to match the maximal response of the figure-of-eight. This version of the lateral fraction appears to originate with Bradley [1994]. In the equations cited by Bradley and by Hidaka et al. [1995], the numerator has been adjusted to consider signals from $t = 0$. It is therefore assumed that the impulse is positioned centrally, and that no reflections will arrive within the first few milliseconds.

A more advanced measure of the spatial impression of concert halls can be generated directly from a binaural impulse response by the interaural cross-correlation function (IACCF). The following formula is based on the standard cross-correlation algorithm [BS EN ISO 3382:2000]:

$$IACCF_{T_1}^{T_2} = \max \left| \frac{\int_{T_1}^{T_2} p_l(t-\tau)p_r(t+\tau)dt}{\sqrt{\int_{T_1}^{T_2} p_l^2 dt \int_{T_1}^{T_2} p_r^2 dt}} \right|_{\tau=-500\mu\mathrm{s}}^{\tau=500\mu\mathrm{s}} \tag{2.4}$$

$IACCF_{T_1}^{T_2}$ represents the interaural cross-correlation function of the signal between the times T_1 and T_2. $p_l(t)$ and $p_r(t)$ are the sound waveforms of the left and right ears. The use of the $IACCF_0^{80}$, as a measure of spatial impression caused by early reflections, is reviewed by Bradley [1994] and Hidaka et al. [1995].

There is a clear inverse correlation between the IACCF and the lateral fraction formula. This is because the maximum value of the function will register the direct, frontal sound, close to $\tau = 0$. Lateral reflections will increase the signal energy within the IACCF and hence increase the denominator of Equation 2.4, but will not change the numerator when $\tau = 0$, so the IACCF will decline. Frontal reflections, however, will reinforce both the numerator and the denominator. Bradley discovered a strong correlation (R > 0.8) between 1 – $IACCF_0^{80}$ and L_f for hall average responses, over the three octaves of the audio frequency spectrum centred on 250Hz, 500Hz, and 1kHz. However, the correlation is smaller for other frequency bands.

There have been a number of refinements to these formulae. Barron and Marshall [1981] suggest that spatial impression increases with sound pressure level, and many other researchers concur. A recording can be made to appear more spacious simply by adding gain. It is possible that this effect could be an epiphenomenon of the human auditory system, although this assertion has not been proven. However, increasing listening level does four things, all of which will increase perceived spaciousness:

- Critical bands widen as sound pressure level increases, owing to nonlinearity of travel of the basilar membrane [Ren 2002]. Thus a loud signal will stimulate cochlear hair cells whose characteristic frequency range would normally fall outside the range of the stimulus.
- A number of reflections will become audible that were previously below the threshold of perception.
- As long as the noise floor of the recording is substantially below the threshold of perception, adding gain to a recording increases the onset rate, in terms of dB/s. This changes the way in which the precedence effect works (see Section 2.1.1).
- If the microphone self-noise is above the threshold of perception, increasing the gain will cause it to become more audible. This noise will be decorrelated across the binaural channels, and will therefore be perceived as acoustically wide.

Results of an experiment by Blauert et al. [1986], which state that perceived spatial impression varies in proportion to the bandwidth of the reflections, may also be influenced mainly by associated changes in reflection energy.

There is one further important development of the *IACCF*, which relates to frequency selectivity. Hidaka et al. [1995] demonstrated that a newly-devised measure, the $IACC_{E3}$, correlates closely with a subjective ranking of concert hall quality. The $IACC_{E3}$ is an interaural cross-correlation function of the first 80ms of a binaural room impulse response, taken for three central octave bands (500Hz, 1000Hz, 2000Hz) only. This finding concurs with listening experiments conducted by Schroeder et al. [1974], which demonstrate a strong link between interaural coherence and hall preference.

2.2.1 Late-arriving sound energy

In most natural listening situations, reflected sound energy that arrives at the listener more than 50ms after the direct sound will have travelled a far greater distance than the direct sound. It will have been attenuated by every surface that has reflected it, so most late reflections are far quieter than the direct sound. Moreover, the statistical density of reflections increases with time after onset. After 50–100ms, the reflection density becomes very high, and reverberant energy within a room starts to approximate diffuse conditions.

For these reasons, reflections arriving after the echo threshold are seldom perceived as disturbing echoes. As they arrive too late to be fused perceptually with the direct sound, they become perceptually associated with the listening environment, and are heard as supporting reflections. Measurements of early- and late-reflected energy are therefore divided by referring to the early-arriving reflection class of measurements as ASW (auditory source width) and the late-arriving class as LEV (listener envelopment) parameters.

The two phenomena, and the terms ASW and LEV, are introduced in a paper by Bradley and Soulodre [Bradley and Soulodre 1995]. However, the different properties of early and late reflections have been appreciated for far longer. This explains the choice of an 80ms a cut-off period for reflection analysis in Barron and Marshall's L_f formula (Equation 2.2).

In concert hall impulse response analysis, 80ms has served almost universally as the time constant that separates early reflections from late reflections. It is a convenient compromise between psychophysical parameters and acoustical ones, mediating between the 50ms time constant that Haas attributes to the wearing off of the precedence effect, and the 50–100ms time period during which, in a medium-sized hall, the earliest third- and fourth-order reflections arrive at the listener. Around this time, the reflection pattern becomes appreciably denser and more diffuse (Figure 2.5).

Attempts to refine the 80ms time constant have arrived at largely similar values. Soulodre et al. [2003] conducted a series of experiments using an anechoic recording of Handel's 'Water Music'. The closest correlation with elicited LEV scores for a number of artificial sound fields is provided by an integration time that is a function of frequency. This value is 140ms for the octave band centred at 63Hz, and diminishes to 60ms for octave bands of 1kHz and greater. The fixed integration time that produces results that correlate best with the subjective LEV data is 105ms.

This experiment does have an important limitation: it is based on one performance of one type of music. Early research into spatial impression by Ando [1977] shows that for a single reflection, subjective spatial characteristics depend on the autocorrelation function of the test signal, and are therefore a function of its tempo.

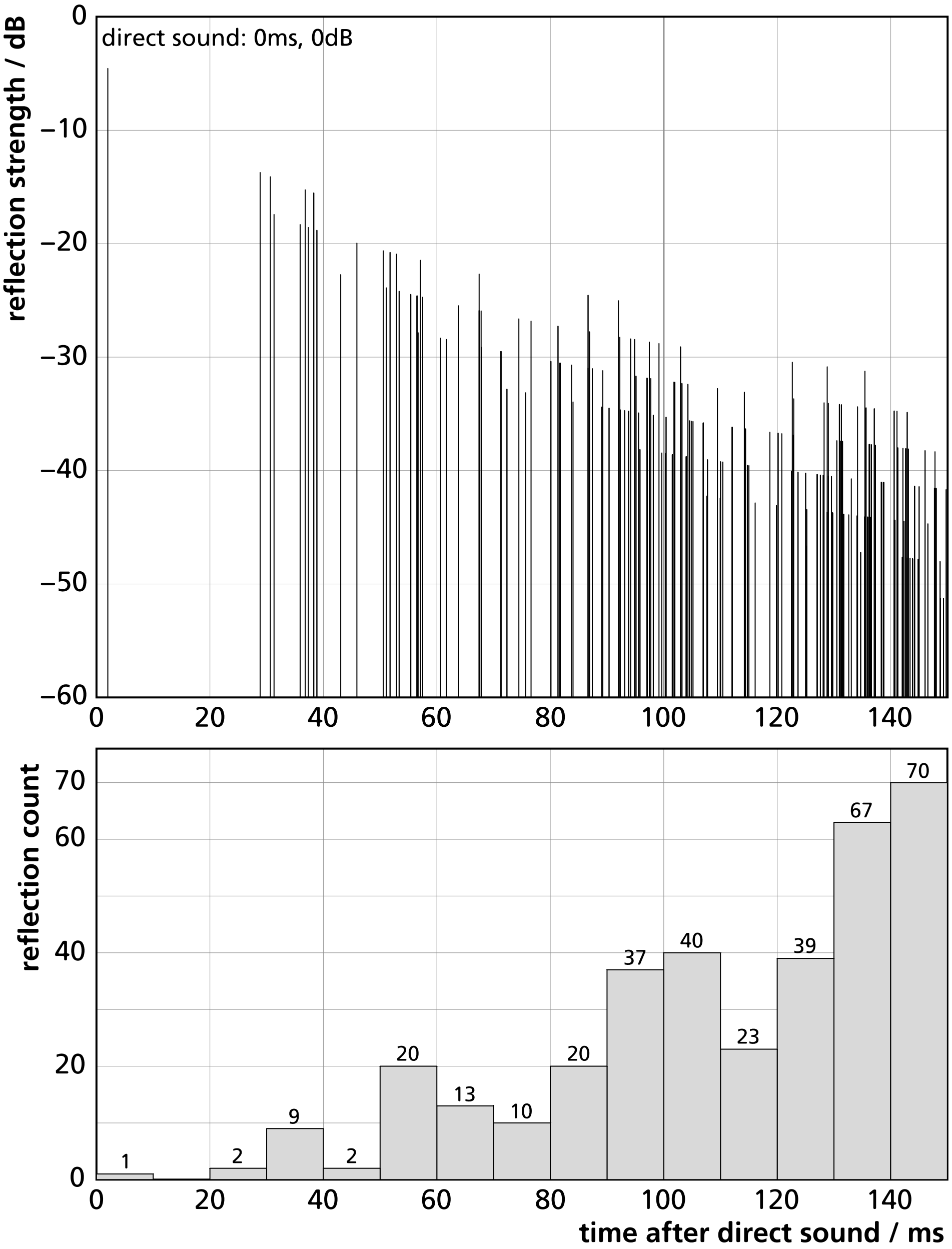


Figure 2.5. The earliest 349 acoustic reflections (up to 150ms) of an omnidirectional source in a medium-sized rectangular hall, approximating the dimensions of Studio 1 (see Chapter 5).

The lower graph is a count of the discrete reflections shown in the upper graph. Sudden increases in the density of reflections can clearly be seen after 50ms, and again after 90ms.

The dimensions of the simulated room are 15×16×7.6m, with the source positioned 4m from the target. The 60dB reverberation time of the space has been set to one second, using the Sabine formula.

2.2.2 Beyond the impulse response

The ASW and LEV measures presented so far are based on the analysis of room impulse responses. An important advantage of impulse response techniques that makes them applicable to room acoustics is their simplicity. An impulse response, unlike a test recording of a musical or vocal source, can be processed meaningfully with very little effort. Conversely, even basic automatic spatial analysis of musical or oratory sources cannot be attempted without a reasonably complex scene analysis model.

The principal drawback of impulse response analysis is that the methods developed for this field cannot be used for, or easily adapted to, any kind of signal that is not also impulsive. The IACCF formula, for example, is of limited use against a complex stream of programme material, as it will fluctuate considerably depending on the activity of the sound source. A second disadvantage of impulse responses is the obverse of this situation: that human listeners cannot process the spatial information they contain. The unnaturalness of these signals means that the impulse response has little meaning to the human auditory system, and the room impulse will need to be convolved with anechoic programme material before human validation of a listening environment or a new impulse analysis technique is possible.

To compound these complications, it is clear from listening experiments that the spatial impression produced by an instrument or an ensemble in a hall is a function of properties of the instrument or ensemble [Mason and Rumsey 2001] and the speed of music played [Ando 1977] or the speech recited [Haas 1951], as much as it is a function of the hall's acoustic properties and the listening position. The subtleties of spatial impression are not immediately apparent from an impulse response. Instead, interpretation of impulse response data requires the ready availability of several other room parameters and some considerable interpretation.

To answer these problems, at least two dynamic measures of spatial impression have been created, along with one theoretical paradigm for extracting spatial information. The two measures are Mason's interaural cross-correlation fluctuation function (IACCFF) [Mason 2002] and Griesinger's diffuse-field transfer function (DFT) [Griesinger 1998]. These are superficially similar in approach. The IACCFF and the DFT both analyse short-term fluctuations in interaural time differences at low frequencies, and follow the prototypical flowchart shown in Figure 2.6. The important differences between them are tabulated in Table 2.1.

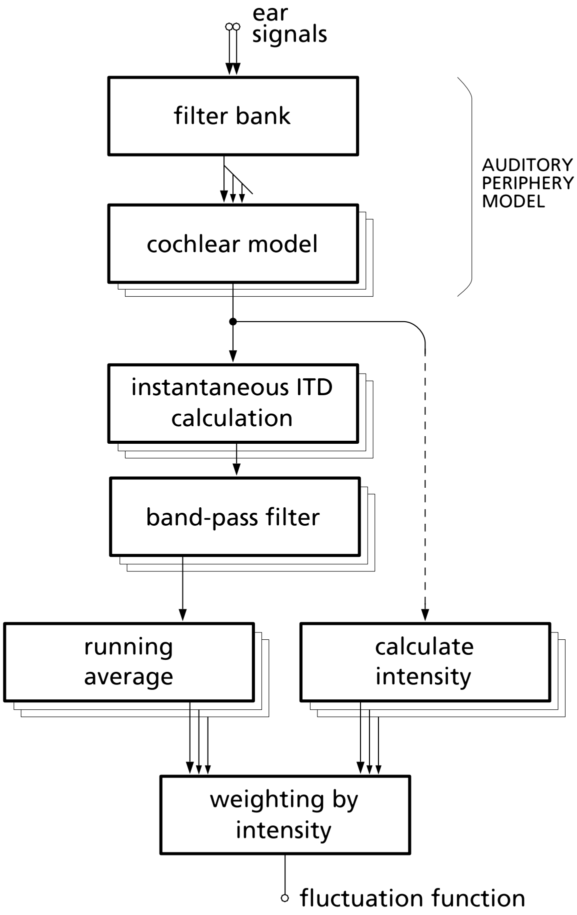


Figure 2.6. Flowchart of the fluctuation functions devised by Griesinger [1998] and Mason [2002].

	IACCFF [Mason 2002]	DFT [Griesinger 1998]
Intended input signal	arbitrary	white noise
ITD calculation method	cross-correlation	zero-crossing based
Upper frequency limit	2500Hz	1360Hz
ITD band-pass filter range	10–125Hz	3–17Hz

Table 2.1. Comparison of the IACCFF and DFT functions.

As Mason and Rumsey demonstrate in their comparison of DFT and IACCF measurements, the limited sophistication of these models restrict their applications. A deal of interpretation is needed with both measurements, as their values and characteristics vary depending on the pitch, amplitude envelope and harmonic content of the source signal, as well as other dynamic aspects such as vibrato [Mason and Rumsey 2001].

2.3 Griesinger's model for spatial analysis

In order to implement the precedence effect, a special topology must be adopted that divides the spatial analyser into four components. This suggests the topology of Zurek's model [Zurek 1987], upon which the project's spatial analyser is based (see Section 1.4). David Griesinger extends the Zurek model into a theoretical system that would listen to an arbitrary binaural signal, accommodate spatial cues from early and late reflections, and hence derive measures of early and late secondary spatial attributes, in addition to source location. This is shown in Figure 2.7.

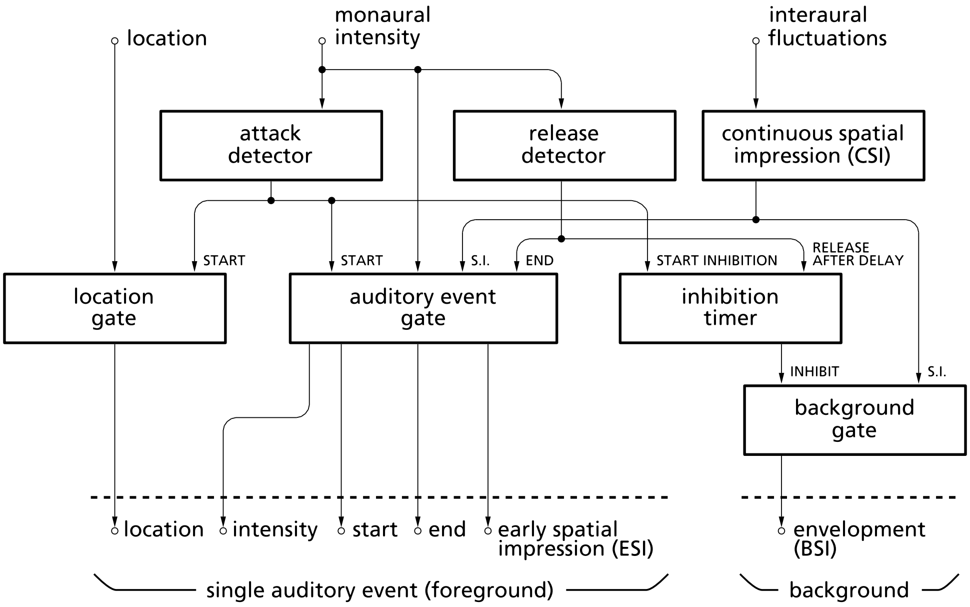


Figure 2.7. Griesinger's model. This includes basic scene analysis features to extract secondary spatial attributes. Adapted from [Griesinger 1997; 1999].

The importance of onset and offset detection (*attack* and *release*) in the workings of Griesinger's model extends Zurek's model, which requires only an onset detector. Onset detection is the first significant problem to be tackled in this thesis. The need to complement this algorithm with an offset detector becomes apparent when the extraction of secondary spatial attributes is attempted (see Section 5.4.2).

The second important extension to Zurek's model is the addition of *background spatial impression* (BSI). Griesinger uses this term in the same way many authors use the term 'listener envelopment'. According to Griesinger, reverberant energy that occurs less than 100ms after release of an auditory event is masked by the human auditory system. This time constant agrees very closely with the research conducted by Soulodre et al. [2003], into listener envelopment perception. When processing the room impulse response for listener envelopment, they found that a minimum cut-off time of 105ms produced a metric that most closely correlated with their listeners' data.

Griesinger's schema is comprehensive in terms of what it defines, and also in its stipulation of the way in which its constituent processes interrelate to produce an automatic spatial analyser. This system would be broadly aware of different auditory objects, and capable of extracting foreground (ASW) and background (LEV) spatial information. Nevertheless, this system is a framework. Griesinger has not ventured details of the workings of any of the high-level processes in the diagram, every one of which presents considerable research problems and design challenges.

Some amendments to the structure of Griesinger's model are proposed within this thesis. Details of these can be found in Section 6.7.1.

2.4 Faller and Merimaa's model for spatial analysis

Faller and Merimaa's localisation algorithm [2004] provides an alternative to Zurek's structure (see Section 1.4, Figure 1.1). By using interaural coherence as their only onset cue, Faller and Merimaa eliminate the need for a dedicated intensity-based onset detector and precedence effect model.

The interaural coherence (IC) function described in Faller and Merimaa's paper is identical to the normalised *IACCF* shown in Equation 2.4. Theoretically, this function peaks whenever the direct sound from a source dominates other sources and room reflections, and it is therefore a good substitute for level-based onset detection.

This approach is attractive because it is computationally simple. The localisation algorithm computes a running interaural cross-correlation in order to extract interaural time differences. All that is required to turn this in to an onset cue is to normalise the peak value of this cross-correlation according to the instantaneous input level.

However, the same difficulties of implementation are encountered when using this approach as with any onset detection algorithm. The first is that the contrast of IC data under typical listening conditions is not as pronounced as it is in most idealised, simulated environments (see, for example, the peak truth value data in Figures 5.3, 5.6, and 5.9, which relate to interaural coherence). In the experiments in Chapter 5, it will become clear that it is no small challenge to set a threshold value for onset detection that applies even to a good majority of musical signals and listening conditions. If the localisation algorithm is to be applied to generic signals, Faller and Merimaa suggest that the threshold must adapt slowly over time, and that this could account for precedence effect phenomena. The modulation of this onset threshold is the only way to incorporate the precedence effect in Faller and Merimaa's model, and therefore it would need to be implemented with great care.

The second significant problem with Faller and Merimaa's model is that it depends entirely on interaural coherence measurements. Therefore it cannot account for onset detection under monaural listening conditions or amplitude-panned headphone listening, where the IC will be either constant (in the presence of a signal) or undefined (for silence). Although it is plausible that the human auditory system employs interaural coherence as an onset cue, it must also apply a complementary level-based method.

In spite of these shortcomings, Faller and Merimaa prove their approach to be effective under many conditions. This research has been published fairly recently, and further investigation would be necessary to determine whether the IC is worth incorporating into the spatial analyser.

2.5 Summary

The precedence effect is a complicated neural inhibition phenomenon, in which approximately the first millisecond of sound that arrives at a listener dominates over later-arriving sound energy. Both the perceived loudness of the later information and the spatial content it contains are inhibited by approximately 10dB. Although later-arriving energy can exert a small influence on the perceived location of the sound, the two are perceived as a single, spatially-fused auditory object. The precedence effect hence makes the auditory localisation system robust to disturbance by early room reflections.

The period during which the precedence effect acts on a stimulus depends chiefly on its amplitude envelope. For clicks, the influence of the precedence effect is maximal between 2 and 3ms after onset, and has ceased by 10ms after onset. For more natural stumuli, which tends to be continuous or slowly-decaying, the influence of the precedence effect is again maximal between 2–3ms after onset, but can extend about 50ms into the auditory event.

A number of other factors govern the strength of the precedence effect. The slower the rate of onset, the less pronounced the inhibition will be. Some research suggests that the precedence effect plays little or no part in sounds whose rates of onset are less than around 400dB/ms.

To some extent, human listeners will also adapt consciously or unconsciously to an environment in order to exploit the advantages of echo suppression, and are able to desensitise themselves to certain types of strong isolated reflection after a short period of exposure.

Lindemann's model of the precedence effect [Lindemann 1986] has been analysed. Theoretically, it would work well with click-based stimuli, but is not versatile enough to respond accurately to continuous, musical, or speech stimuli.

Owing to its profound effect on the perception of early reflections, the precedence effect greatly influences the perception of spatial attributes: particularly auditory source width (ASW). This attribute has received considerable attention in recent years, and a number of ways of measuring source width objectively have been proposed.

Lateral (side wall) reflections decorrelate a frontally-positioned source more than frontal reflections. Therefore, lateral reflections decrease the localisability of the direct sound and thus increase its perceived breath more than frontal reflections. Three early measures of ASW, the L_f, LF_E, and IACCF, incorporated this observation. They process room impulse responses, weighting lateral reflections higher than frontal reflections, and express this as a proportion of total unweighted sound energy. 80ms is usually used as an empirical cut-off point for early reflection integration in these quotients.

Sound energy arriving after approximately 80ms is not spatially fused with the direct sound. Instead, it is perceived as supporting reverberation. This gives rise to the term listener envelopment (LEV) to describe late-arriving signal energy. Attempts to quantify LEV currently depend on the analysis of post-80ms reflections in binaural room impulse responses.

At least two individual attempts have been made to depart from impulse response analysis, which would allow an environment to be analysed using more continuous stimuli, and natural sounds such as musical instruments, ensembles, or speech. Griesinger's diffuse-field transfer function (DFT) is designed to process fluctuations in interaural time difference so that continuous noise can be used to gain an impression of auditory source width. Mason's interaural cross-correlation fluctuation function (IACCFF) operates in a similar way, but has been tested on different instrumental stimuli. Output data from the IACCFF is heavily dependent on the amplitude envelope and pitch of the source material used, so knowledge of the source stimulus is required to interpret it.

Griesinger [1999] proposed a series of extensions to the Zurek model (see Chapter 1.4) to enable extraction of secondary spatial attributes. Griesinger's model includes plausible mechanisms for the extraction of ASW, LEV, and a number of other secondary spatial attributes. Like the Zurek model, however, it is only a framework for future development, and none of this has been realised. Thus, Griesinger's model cannot presently be verified, either formally or informally, as a valid approach to spatial attribute extraction.

Faller and Merimaa [2004] describe an alternative to the Zurek model that employs the interaural cross-correlation function as its sole onset cue. This framework could be expanded to form a model of the precedence effect, and is demonstrated to work well under a number of listening conditions. It could therefore be investigated as a complementary onset cue for the spatial analyser. However, it would be difficult to adapt this algorithm for generic listening conditions, and it would not work with monaural or amplitude-panned signals.

3 ONSET DETECTION ALGORITHM

***Note**: Parts of the onset detection algorithm described in this chapter have been documented previously in a journal paper [Supper et al. 2005].*

This chapter details the design of the onset detector. The presence of this component of the spatial analyser has been motivated by the previous two chapters. To discern when an auditory onset occurs is to know when the direct sound dominates reflected energy in a sound field. This allows the source to be localised as unambiguously as possible, and sense to be made of the spatial data.

The following sections propose a new definition of *auditory onset* that is compatible with the spatial analysis task. The requirement for a new onset detection algorithm can then be justified. A specification will then be formulated for this algorithm, and its design and implementation described.

3.1 Definition of auditory onset

An *auditory onset* is usually defined as the span of time during which a new auditory event begins, and in doing so exhibits a rapid and significant increase in sound energy [Bello and Sandler 2003]. When the purpose of an onset detector is to assist with a non-spatial machine listening task, such as note transcription or tempo detection, this definition is satisfactory. However, the definition is not adequate for an onset detector used in spatial analysis, for the reasons described in Section 3.2.

The term *auditory onset* has therefore been extended here to cope with the nature of spatial scene analysis, in which the sound from the onset of an auditory event, which is mostly reflection-free, allows the source to be localised. An *auditory onset* thus defines any region of time during which directly-arriving sound dominates over reflected energy, so that reliable localisation information can be extracted from the auditory stream. This redefinition includes as auditory onsets those quickly-rising attack portions that are covered by the more usual definition. Under the extended definition, however, a single auditory event that contains more than one rapid rise in level can possess more than one auditory onset. Furthermore, if a source is close to the listener, any steady-state portion of its waveform will furnish correct localisation cues, and will therefore constitute part of its onset.

Thus an auditory onset may have a greater duration under its newer definition, even if only the beginning of this region is used to trigger spatial processing. The new definition excludes note onsets which are not strong enough for the direct sound to dominate over the reflected energy, because these would cause subsequent spatial processing to produce erroneous results.

Extending the meaning of ‘auditory onset’ has counter-intuitive ramifications for particular stimuli. For example, under anechoic conditions, the entire waveform is now classed as an auditory onset. This may cause problems when non-spatial properties are considered, but for localisation, all portions of an anechoic waveform contain usable directional cues, so any portion will suffice for source localisation. Some confusion involving the redefinition may also occur when considering the auditory onset of a distant stimulus with a slow attack time. There may be no time during which direct sound dominates substantially over the reverberant field, so the auditory event may not possess an auditory onset at all. This can upset the spatial processing of some stimuli. However, most sounds in nature are not entirely continuous, and fluctuate in amplitude enough for reflection-free localisation to be attempted even if they build up slowly.

Furthermore, human listeners can be deceived when a sound field changes without an onset occurring. This effect is exemplified by the Franssen illusion [Franssen 1960]. To produce the Franssen illusion, a sine tone is played through a loudspeaker 30° to the left of a listening subject in a reverberant room. This tone is faded out immediately as a complementary tone is faded in to a second loudspeaker 30° to the right of the listener. The complementary tone is then faded to silence. The listener locates the entire event at the left loudspeaker only. An engineering, rather than physiological, approach has been taken in the development of the algorithm detailed in this paper, but it has been designed to provide a perceptually representative output. For the purposes of this system, then, if a human listener is misled by such changes, it is also acceptable for an artificial listener to be misled.

3.2 The requirement for the new auditory onset detection algorithm

As described in Chapter 2, the first few milliseconds of incident sound are largely free from reflections from walls and other physical objects. Many 'real-world' circumstances, and almost all musical or oratory performances, involve sounds with rapid rising edges and some continuous component heard at a distance in a room. Under these circumstances, acoustic reflections build up so quickly that the usual interaural time and level difference cues used for direction finding cease to be reliable after tens of milliseconds. The human auditory system uses its echo suppression mechanisms to inhibit spatial processing of later-arriving sound energy (see Section 2.1). The onset detector must include, or be informed by, an algorithm that imitates the faster, lower-level aspects of echo suppression.

For detecting secondary spatial attributes, auditory onsets must be detected relatively infrequently. Interaural fluctuations more than 100ms after onset have been demonstrated to be important to the spatial perception of a sound source (see Chapter 2). If an onset is inferred indiscriminately from every rapid increase in amplitude, there will seldom be an occasion in many real sound fields where no onset occurs for 100ms, and this will have a deleterious effect on the quality of data available to the spatial feature extraction processes. Also, the more sensitive the onset detector, the greater the likelihood that individual specular reflections will trigger the detector falsely. The sensitivity of this detection algorithm must therefore be controlled so that misleading detections are minimised. No other auditory onset detection algorithm can be found for which infrequent onset detection is an explicit specification.

The optimal threshold for level- and rate-based onset detection varies widely depending upon the stimulus used. Creating a detector that retrieves onsets at an acceptable rate, and does so in a perceptually valid way, is therefore not just a matter of desensitising existing algorithms. Piano notes in particular have chaotic, steeply-fluctuating decay curves, and problems may also be encountered in cases where the source or the listener is positioned near a large reflective surface, so that some reflections are particularly early and strong. The approach described in this paper employs a number of simple techniques that allow its sensitivity to be altered without affecting the reliability of detection. The new algorithm is not immune to false onsets and missed onsets, but owing to its relative insensitivity to short-term isolated signal fluctuations, mislocation owing to false detections occurs infrequently.

3.3 Design specification and strategy

The onset detector is designed for an application that requires it to work in real time. Therefore its algorithm must be able to handle streamed audio. Although the prototype onset detector works only with stored files, the necessity for a stream-compatible algorithm has been taken into account.

This algorithm cannot ‘look ahead’ at forthcoming auditory data more than 2ms away. A look-ahead method has been implemented in practice by allowing a short delay between input and output.

For the most part, the onset detector described in this chapter deliberately employs techniques that are computationally simple and physiologically plausible. There is an important exception to this physiological validity: the dynamic range compression imposed by the inner ear is absent in this algorithm. Instead, the signal that reaches the processing stage bears a closer resemblance to the input signal. This approach has enabled the onset detector to be built using standard signal processing techniques. Although some perceptual accuracy may have been sacrificed by taking this approach, it has imparted two worthwhile advantages. Firstly, familiar processes can be applied to the signal with familiar results. Secondly, avoiding non-linear processes renders the absolute input level to the system largely unimportant: the input does not need to be calibrated and scaled with respect to a fixed sound pressure level.

The ability of the human auditory system to localise sound, to separate two spatially disparate sound sources, and to hear subtleties within them, is better under binaural conditions than it is under monaural conditions. When one ear is damaged or obstructed, localisation becomes more difficult and masking thresholds are raised [Moore 2000]. Unfortunately, no model of this binaural interaction has been universally accepted. For pragmatic reasons, the onset detector defined here follows the example set by almost every other algorithm, and employs separate monaural processors for each ear. The outputs from these processors are combined to increase the algorithm’s sensitivity over monaural conditions.

There is a wealth of published research that concerns auditory onset detection, and the strategies that have been attempted over the years are varied. However, a quantity of this research describes algorithms that are not psychoacoustically inspired, and hence are not directly applicable to this research. Furthermore, there is no previous published work that relates to the detection of auditory onsets that are relevant to spatial scene analysis, as defined in the preceding section. However, a number of existing onset detectors are relevant to the development of this algorithm, and the similarities and differences between these processes and the one described here is included in Section 3.4.

3.4 System details

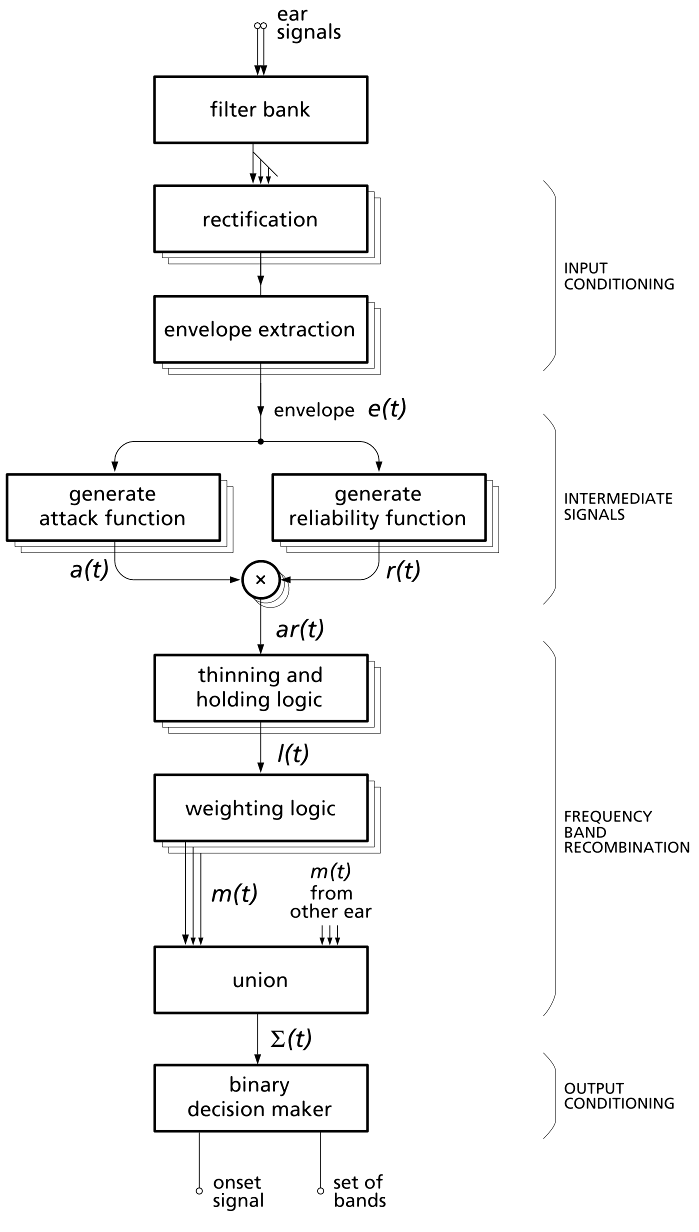


Figure 3.1. Overview of the onset detection algorithm.

Figure 3.1 shows an overview of the monaural onset detector algorithm. Its input data are 24 filtered versions of each ear signal: this gives 48 audio-rate signals. It outputs a two-valued function that equals unity whenever an onset is detected and is zero at all other times, together with the set of the filter bands that contributed most significantly to each onset decision. Only those frequency bands with substantial rising signal level, and therefore the largest proportion of direct sound to reverberant sound and noise, need to be considered in the spatial processing. The detector is split into four sections, each of which will be considered separately.

Input conditioning reduces the data rate of the input signal, in order to speed up processing and to filter out information that is irrelevant to onset detection. This generates the envelope function, $e(t)$. Two independent functions, the *intermediate signals*, are calculated from $e(t)$. One of these, $a(t)$, produces temporally accurate, high-contrast information pertaining to onsets, but is also sensitive to false onsets. The other function, $r(t)$, has a lower contrast, is slower to rise and fall, and therefore is less sensitive to slow onsets. $r(t)$ is also constructed to be immune to disruptions caused by background noise. These signals are expressed in the [0 1] domain to preserve their continuous nature and to allow standard fuzzy logic processes to be applied to them. They are combined by piecewise multiplication to create $ar(t)$. This is a high-contrast function that is insensitive to false onsets.

Forty-eight versions of $ar(t)$ are calculated across the two channels and 24 frequency bands. These are then combined and summed into one function of time, $\Sigma(t)$. Three processes, the *frequency band recombination stage*, generate this signal. The final onset decision is based entirely upon $\Sigma(t)$.

Lower frequency bands have longer impulse responses that spread energy over a longer period of time than higher bands. This is partly compensated by using higher bandwidths at lower frequencies, and partly by the first of the recombination functions, ‘thinning and holding’. This removes weaker fluctuations and aligns near-coincidences of $ar(t)$ peaks across frequency bands. The second function cross-weights the output signals in favour of coincident peaks that occur across adjacent bands. All bands are summed in the third stage to derive $\Sigma(t)$.

Finally, a simple binary decision maker determines when the fluctuations in this sum signal constitute an onset. When an onset is detected, those bands whose individual contributions to the sum exceed a secondary threshold are flagged as contributing towards it. These procedures are covered in detail in the following sections.

3.4.1 Filter bank and rectification

The filter bank used in this project is based on Slaney's efficient implementation of a gammatone filter bank [Slaney 1993]. Low and high cutoff frequencies for each of these filters were taken from Gaik's cross-correlation model [Gaik 1993: 100]. These parameters are reproduced in Table 3.1 and Figure 3.2. Each channel of the incoming binaural signal is thus divided into 24 critical bands. Most of the frequency bands are around a quarter of an octave wide (Q = 5.79). The lower frequency bands are wider to model the bandwidths of critical bands. This widening also reduces the length of impulse responses, preventing the disruption of timing information that would otherwise occur.

It was necessary to add two high-pass filters to the Slaney filter bank. These block DC and extreme LF content that may otherwise leak through the lowest two bands and cause level-detection problems after rectification. One filter is cascaded to the Band 1 filter output, and the other to the Band 2 filter output. First-order Butterworth filters have been chosen, with cutoff frequencies of 18Hz.

Band	f_l	f_h	Centre freq.	Bandwidth	Q	Band	f_l	f_h	Centre freq.	Bandwidth	Q
1	**20**	**100**	60	80	0.75	13	**1720**	**2000**	1860	280	6.64
2	**100**	**200**	150	100	1.50	14	**2000**	**2320**	2160	320	6.75
3	**200**	**300**	250	100	2.50	15	**2320**	**2700**	2510	380	6.61
4	**300**	**400**	350	100	3.50	16	**2700**	**3150**	2925	450	6.50
5	**400**	**510**	455	110	4.14	17	**3150**	**3700**	3425	550	6.23
6	**510**	**630**	570	120	4.75	18	**3700**	**4400**	4050	700	5.79
7	**630**	**770**	700	140	5.00	19	**4400**	**5300**	4850	900	5.39
8	**770**	**920**	845	150	5.63	20	**5300**	**6400**	5850	1100	5.32
9	**920**	**1080**	1000	160	6.25	21	**6400**	**7700**	7050	1300	5.42
10	**1080**	**1270**	1175	190	6.18	22	**7700**	**9500**	8600	1800	4.78
11	**1270**	**1480**	1375	210	6.55	23	**9500**	**12000**	10750	2500	4.30
12	**1480**	**1720**	1600	240	6.67	24	**12000**	**15500**	13750	3500	3.93

Table 3.1. Lower and upper frequencies for each band of the filter bank, adapted from Gaik's cross-correlation model [Gaik 1993]. The centre frequency, bandwidth, and quality factor (Q) of each band are also included.

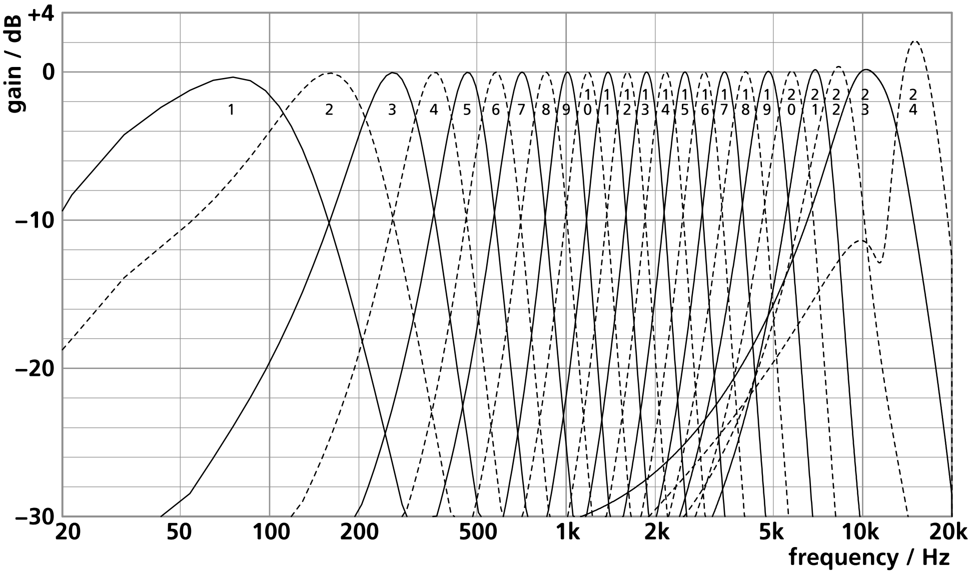


Figure 3.2. Filter bank amplitude responses. Even-numbered bands are shown as dashed lines for clarity.

The inner ear is approximated by full-wave rectification followed by low-pass filtering. The latter is performed by a second-order Butterworth filter with a cut-off frequency of 1100Hz. It is more usual, particularly in onset detection systems that include inner hair cell transduction models (such as [Smith 2001] and [Martin 1995b]), to use half-wave rectification. Although full-wave rectification does not correspond to any neurophysical model, it produces an output signal with a higher level, lower ripple, and the same rise time as half-wave rectification. Full-wave rectification is therefore preferable for signal processing. The Butterworth filters simulate the refractoriness of inner hair cells. In the cochlea, this refractoriness causes the break-up of neural phase-locking at mid-frequencies. It is important to account for this phenomenon in other areas of the spatial analysis algorithm. Although Hafter and Carrier [1972: 1852] attribute phase-locking breakdown to frequencies of over 5kHz, lower frequencies are often used in simulations because the wavelength of sound at 1100Hz begins to match the dimensions of a listener's head. Thus, the human auditory system's reliance on interaural time differences begins to decrease above this frequency. This approach to the selection of cut-off frequency has also been employed in the periphery models of Blauert and Cobben [1978] and Lindemann [1986], who employed a first-order low-pass filter with a cut-off frequency of 800Hz.

The cochlear model used in this algorithm is fairly basic compared with many of the cochlear models that are available, such as the Meddis model [Meddis et al. 1990] used in Martin's localisation system [Martin 1995b], and many of the inner hair cell transduction models reviewed by Mountain and Hubbard [1996]. However, the aim of simulating the action of the inner ear must be balanced with the need to extract data with a minimum of processing. The advantage of the approach used here is that it preserves the linear scale of the input data. The justification for this decision was presented in Section 3.3.

3.4.2 Envelope extraction

The envelope extraction removes data that is redundant to onset detection. For ease of implementation, it employs two single-pole IIR filters. One filter is set to a cut-off frequency of 90Hz, and the other to 150Hz. Equations 3.1 and 3.2 are used to calculate the IIR filter coefficients. Equations 3.3 and 3.4 perform the filtering and rectification:

$$R_f = \cos\frac{2\pi f}{f_S} \tag{3.1}$$

$$k_f = (2 - R_f) - \sqrt{(R_f - 1)(R_f - 3)} \tag{3.2}$$

$$y(T) = k_{90}\, y(T-1) + (1 - k_{90})\ |x(T)| \tag{3.3}$$

$$e(T) = k_{150}\, e(T-1) + (1 - k_{150})\, y(T) \tag{3.4}$$

f_S is the sampling frequency of the system (equal to 44.1kHz in this system); k_f is the operating constant for a single-pole IIR filter of cut-off frequency *f*, and R_f is used in the calculation of this constant. $x(T)$ is a band-pass filtered audio signal; $y(T)$ passes audio between the two IIR filters; $e(T)$ is the output amplitude envelope. T is a discrete time variable, so that (T - 1) refers to the sampling interval that precedes T.

The two IIR filters remove the higher frequencies that convey fine signal detail, as this is not useful for onset detection. Filtering also limits the waveform's rise time. The filter constants themselves were chosen conservatively. They respond quickly compared with the mechanism responsible for integrating loudness in the human hearing system, which (to a simple approximation) can be attributed an integration time of around 80ms [Scharf 1978]. They are also steep enough to attenuate the rectified 1kHz component of an input signal by almost 40dB whilst preserving the signal envelope. Consequently, $e(T)$ may be decimated simply by removing samples.

When $x(T)$ has a sampling frequency of 44.1kHz, $e(T)$ is decimated by 1:18 to 2.45kHz. This decimation ratio is chosen to match the speed of the localisation algorithm. The choice is governed by the use of convenient multiples to increase the speed of the algorithm, and the minimum rate at which the localisation algorithm must run in order to be effective. (These considerations are covered in detail in Chapter 4.) As the discrete time variable T refers to the original sampling frequency, the variable t will used to refer to the new sampling frequency. The decimated $e(T)$ is referred to as $e(t)$.

3.4.3 Intermediate signals

Two signals are generated from the envelope signal $e(t)$. The processes that form these are flowcharted in Figure 3.3. One signal, referred to as $a(t)$, is calculated from the envelope signal's rate of change. The other signal, $r(t)$, is calculated by dividing $e(t)$ by a non-linearly filtered version of itself — termed a *follower signal.* This signal ascends slowly and descends quickly. A declining $e(t)$ will thus be assigned a low or zero $r(t)$, and an increasing $e(t)$ will be assigned a high $r(t)$. The follower signal is limited so that its level cannot fall below a fixed noise floor. Therefore noise from the recording microphones and the environment, and rounding errors within the filter bank, cannot normally influence the output. Thus $r(t)$ pertains to the reliability of the input data.

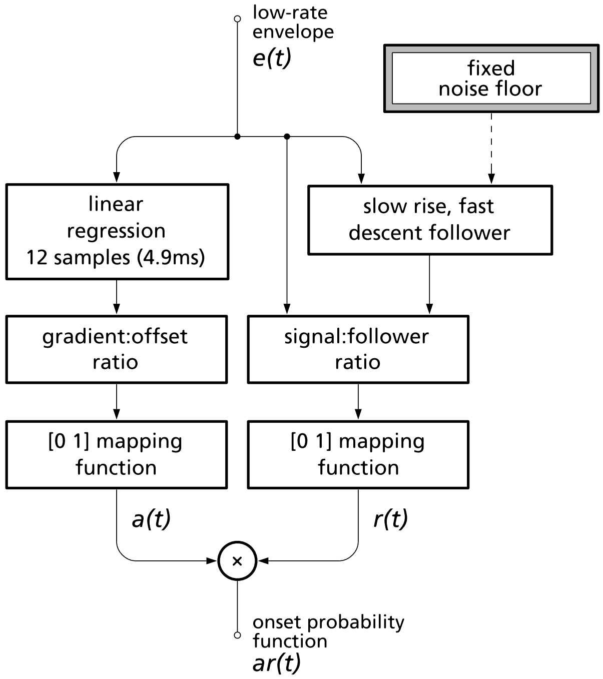


Figure 3.3. Generating the two intermediate signals.

To generate $a(t)$, a standard linear regression model is applied to twelve surrounding values (4.9ms) of $e(t)$, from $e(t$ - $6)$ to $e(t+5)$. This provides an equation describing the line of best fit as $e(t) = mt + c$. The formulae used to calculate m and c in a twelve-point linear regression are presented in equations 3.5 to 3.7:

$$\bar{e} = \sum_{n=1}^{12} e(n) \Big/ 12 \tag{3.5}$$

$$m = \frac{\sum_{n=1}^{12} ne(n)}{143} - \frac{6\bar{e}}{11} \tag{3.6}$$

$$c = \bar{e} - 6.5m \tag{3.7}$$

where n is an integer between 1 and 12, representing a position in the array $e(t$ - $6 \ldots t+5)$.

Linear regression has become a familiar tool in onset detection, because it emphasises increasing or decreasing trends in a signal at the expense of short-term fluctuations. Linear regression is also used in the detection algorithms of Martin [1995b] and Dixon [2001], but these algorithms are intended for signal categorisation, so the sampling windows employed are an order of magnitude greater than the 4.9ms used here. The alternative method of increasing contrast between onsets and noise would be to employ another low-pass filter. This would impose longer delays and would also flatten gradients.

In order to convert the $e(t) = mt + c$ fit into a form that works for exponentially increasing and decreasing signals, m is divided by the offset c to produce the gradient-to-offset ratio. This value is invariant for a signal that changes exponentially, and will not change if this signal is amplified or attenuated. Klapuri's onset detection process [Klapuri 1999] applies a similar process to the extracted envelope, without using a regression model. Klapuri points out that the operation of dividing intensity increase by absolute intensity has a logarithmic equivalent:

$$\frac{\Delta I(t)}{I(t)} \equiv \frac{d}{dt} \log I(t) \tag{3.8}$$

According to Klapuri, this operation detects onsets earlier than a linear derivative would, and performs better with complicated signals. A cosinusoidal function is applied to map the output data to the [0 1] range:

$$a(t) = \begin{cases} 0 & : \alpha(t) \leq \alpha_0 \\ -\frac{1}{2}\left(\cos\frac{\pi\,(\alpha(t)-\alpha_0)}{\alpha_1-\alpha_0} + 1\right) & : \alpha_0 < \alpha(t) < \alpha_1 \\ 1 & : \alpha(t) \geq \alpha_1 \end{cases} \tag{3.9}$$

where $\alpha(t)$ is the gradient-to-offset ratio, and a_0 and a_1 are lower and upper limits of the transition region. In this implementation, $\alpha_0 = 60/f_s$ and $\alpha_1 = 300/f_s$ are chosen as a result of trial and error (see Section 5.3.1 for some further consideration of this choice). The cosinusoidal function was chosen because it appears that a transition curve with shallow slopes near the top and bottom of its range would best suit the distribution of $\alpha(t)$ data. A cosinusoid is the simplest function that possesses this property.

$r(t)$ is generated by mapping a function, $\rho(t)$, into the [0 1] domain. $\rho(t)$ is obtained by dividing the envelope signal $e(t)$ by a follower signal, which will be termed $v(t)$. The equations used to derive $v(t)$, and hence to calculate $\rho(t)$, are shown below:

$$z_Q = 10^{-6}/Q \tag{3.10}$$

$$g_T = 2^{1000/Tf_s} \tag{3.11}$$

$$v(t) = \begin{cases} g_{20}\, v(t-1) & : \quad e(t) > v(t-1) \\ \frac{v(t-1)+e(t)}{2} & : z_Q < e(t) \leq v(t-1) \\ z_Q & : z_Q \geq e(t) \leq v(t-1) \end{cases} \tag{3.12}$$

$$\rho(t) = e(t)/v(t) \tag{3.13}$$

The only absolute level threshold employed within the onset detector is the fixed noise floor, z_Q . This is calculated for each frequency band in inverse proportion to its quality factor Q, so that a pink noise floor is assumed. The constant g_{20} causes $v(t)$ to increase at a rate of 6dB per 20ms. f_s is 2.45kHz: the sampling frequency of $e(t)$.

When $v(t)$ descends, it uses a moving-average filter which combines each input sample with the last output sample in a 1:1 ratio. This produces a descent rate of approximately 13dB/ms, and counteracts a problem that is otherwise observed when the input signal approaches zero for one sample. Subsequent recovery from such minima creates a series of false positive results. The signal-to-follower ratio is mapped into the [0 1] range using a power function:

$$r(t) = \begin{cases} 0 & : \quad \rho(t) \leq \rho_0 \\ \left(\frac{\rho(t)-\rho_0}{\rho_1-\rho_0}\right)^{1.8} & : \quad \rho_0 < \rho(t) < \rho_1 \\ 1 & : \quad \rho(t) \geq \rho_1 \end{cases} \tag{3.14}$$

$\rho(t)$ is the signal-to-follower ratio, and ρ_0 and ρ_1 are lower and upper limits of the transition region. This implementation uses empirically-chosen values, $\rho_0 = 1.40$ and $\rho_1 = 2.25$. It would now be useful to consider an example input signal, and to generate the intermediate signals $a(t)$ and $r(t)$ from this. Figure 3.4 shows the envelope signals extracted from the right ear signal of a binaural recording of a grand piano. This excerpt is a short section of a fugue, played in a medium-sized recording studio with a reverberation time of 1.2s. The range and nature of this excerpt is demonstrated by the score in Figure 3.5. The piano was positioned 40 degrees right of, and 5 metres away from, a binaural dummy recording head.

It may be noted in Figure 3.4 that some signal level is registered even in the highest frequency bands — those bands which a piano barely excites. This occurs because these plots are normalised, so the activity can be attributed to low-amplitude leakage from out-of-band signals. The lowest three bands are showing mostly noise. It is for this reason that $r(t)$ employs the fixed noise floor.

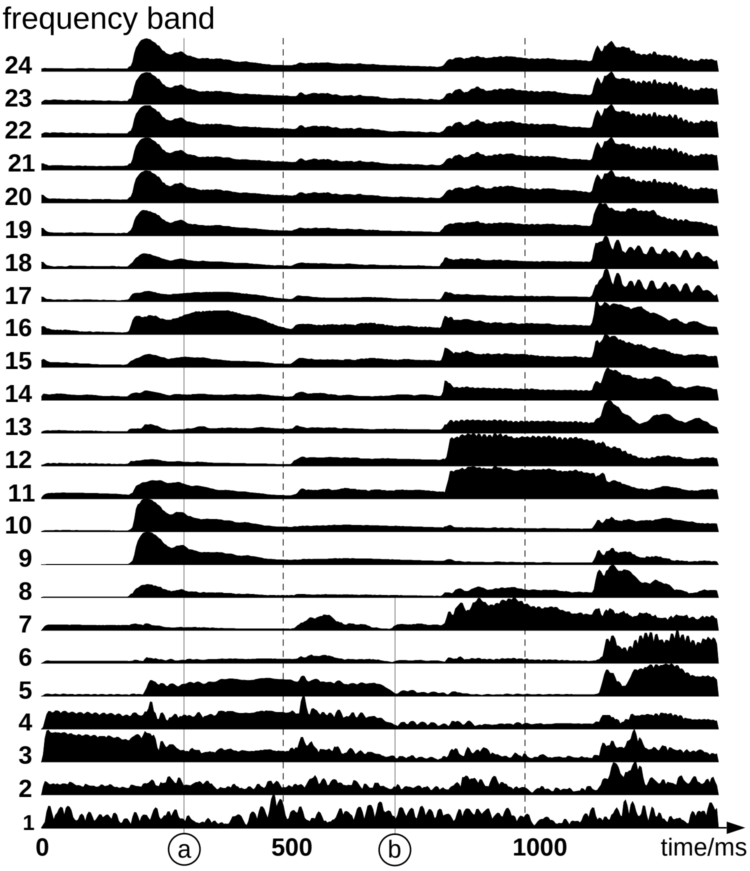
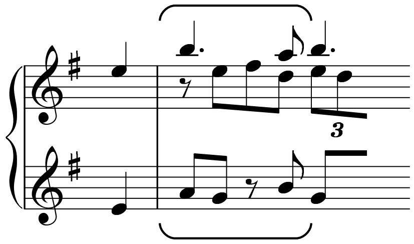


Figure 3.4. Extracted envelope signals for four notes of a piano fugue, right ear. The waveforms have been filled to improve visibility. At the beginning of the graph, the preceding note can be seen decaying. Point a) indicates the attack of the second note of a chord: this separate attack is inaudible unless the audio is slowed down. Point b) marks an area where a note 'swells' suddenly across many lower frequency bands. These are commonplace characteristics of piano waveforms.



Excerpt from 'Fugue' in *Tombeau de Couperin*, Maurice Ravel.

Figure 3.5. Extract from the score played in Figure 3.4, showing its pitch range and number of notes. The brackets show the extent of the excerpt. Eight notes are played. Because some of these notes sound simultaneously, only four separate auditory events are heard.

Figure 3.6 shows the $a(t)$ and $r(t)$ signals generated by the piano stimulus. $a(t)$ and $r(t)$ both possess onset-enhancing properties. Generally, $a(t)$ produces a high-contrast signal with a fast response, but is upset by chaotic signals. $r(t)$ is more stable in this respect, but its value characteristically remains high for much longer, and then falls relatively slowly. Presenting these two signals in the [0 1] domain allows them to be manipulated and combined easily with one another. All that is required to combine them is a fuzzy AND operation, in which the two signals are multiplied together. The resulting signal, $ar(t)$, is shown in Figure 3.7.

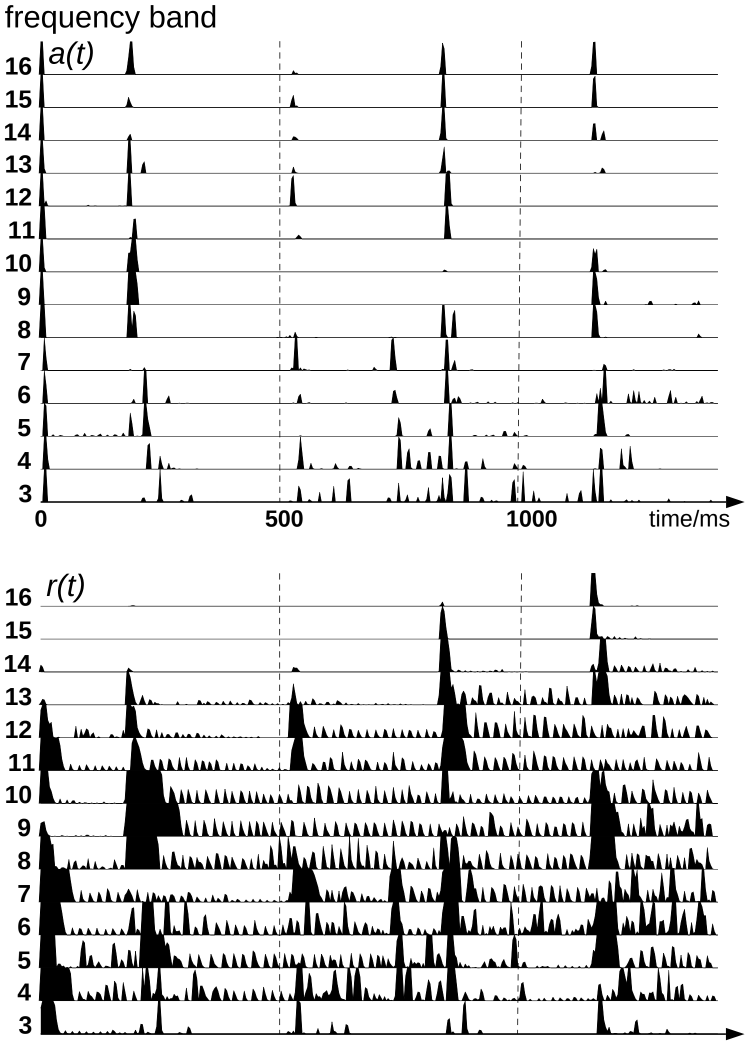


Figure 3.6. $a(t)$ **and** $r(t)$ **for the right ear channel of the piano extract. All signals are in the [0 1] domain. Only significant frequency bands are shown. The different properties of the two signals can be seen clearly.**

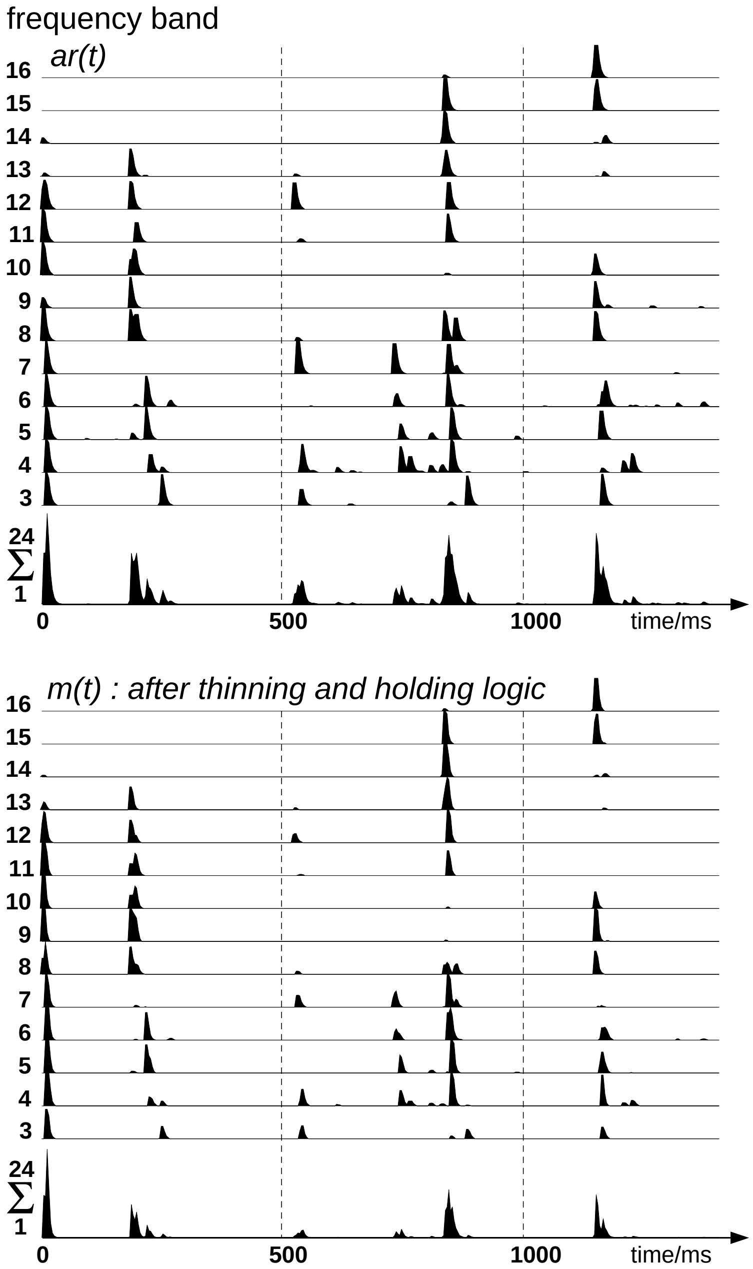


Figure 3.7. Recombination stages applied to the piano excerpt, before [top] and after [bottom] thinning, holding, and cross-weighting logic. The bottom set of graphs have been clipped to fit [0 1]. Sum signals appear below each graph set. This figure shows the right ear signals.

3.4.4 Frequency band recombination

$ar(t)$ is generated for every frequency band of both binaural channels, so that forty-eight $ar(t)$ signals are produced in total. To combine these into a single output, each $ar(t)$ is processed to thin out closely-spaced groups of peaks, and to extend its signal peaks. The thinning-and-holding operation is built by combining hold-and-decay envelope generators. The formulae that control these envelope generators are shown in the following equations:

$$s_T = 0.1^{1000/Tf_s} \tag{3.15}$$

$$\tau(0) = 0 \tag{3.16a}$$

$$\tau(t) = \tau(t-1)+1 \tag{3.16b}$$

$$H(h,d,t) = \begin{cases} 1 & : \tau(t) \le hf_s/1000 \\ s_d & : \tau(t) > hf_s/1000 \end{cases} \tag{3.17}$$

To generate a hold-and-decay envelope with hold time h and decay time d milliseconds, a starting value is multiplied iteratively by each emerging value of $H(h,d,t)$. The decay time is the interval between the hold phase ending and the envelope decaying to 10% of its initial value. Each hold-and-decay generator also has a reset condition. Whenever this condition is satisfied, the counting variable $\tau(t)$ will be reset to zero. The thinning and holding logic is computed by three parallel hold-and-decay envelope generators:

$$c_m(t) = \begin{cases} 1.3\,ar(t) & : ar(t) > c_m(t-1) \\ H_M(10,4,t)\,c_m(t-1) & : ar(t) \le c_m(t-1) \end{cases} \tag{3.18}$$

$$c_s(t) = \begin{cases} 0.8\,ar(t) & : ar(t) > c_s(t-1) \\ H_S(50,20,t)\,c_s(t-1) & : ar(t) \le c_s(t-1) \end{cases} \tag{3.19}$$

$$l(t) = \begin{cases} 1 & : ar(t) > c_m(t-1) \\ H_O(10,4,t)\,l(t-1) & : ar(t) \le c_m(t-1) \end{cases} \tag{3.20}$$

$$H_M(10,4,t) \quad \text{reset} \; : \; ar(t) > c_s(t-1) \tag{3.21}$$

$$H_S(50,20,t) \quad \text{reset} \; : \; ar(t) > c_s(t-1) \tag{3.22}$$

$$H_O(10,4,t) \quad \text{reset} \; : \; ar(t) > c_m(t-1) \tag{3.23}$$

The input signal $ar(t)$ is thus compared with two thresholds, c_m and c_s , which are controlled by two hold-and-decay envelope generators. Whenever

$ar(t)$ exceeds the larger threshold c_m , an output spike is generated in $l(t)$ using the third envelope generator, and the two thresholds are recalculated. Whenever $ar(t)$ exceeds the secondary threshold c_s , the envelope generators H_M and H_S are reset, and c_s is re-computed. $H_S(50{,}20{,}t)$ takes 70ms from triggering to decaying to 10%, and this governs the speed of re-triggering. The choice of this time constant is based upon the 50ms release time for continuous stimuli in the precedence effect (see Section 2.1), and the need to prevent re-triggering for at least 100ms after a detected auditory event (see Section 3.2).

Significant onsets will generate activity across several frequency bands. Less important fluctuations, for example those caused by the beating of one room mode with another, are characterised by a narrow bandwidth, and will activate only a small number of frequency bands. Therefore a cross-weighting process is applied to $l(t)$ that takes into account the values of neighbouring frequency bands. This procedure is best described by a general formula. Let $l_j(t)$ refer to the function $l(t)$ that corresponds to the jth frequency band. $l_{\text{j-1}}(t)$ thus refers to the function $l(t)$ of the lower neighbouring frequency band, and $l_{\text{j+1}}(t)$ to the higher neighbouring frequency band. The output of the cross-multiplication network, $m_j(t)$, is given by:

$$\begin{aligned} m_j(t) &= 0.3\, l_j(t) \\ &+ 0.6\left(l_{j-1}(t)l_j(t) + l_j(t)l_{j+1}(t)\right) \\ &+ 0.9\left(l_{j-1}(t)l_{j+1}(t)\right) \end{aligned} \tag{3.24}$$

This is a weighted sum with a theoretical maximum of 2.4. Although $m(t)$ is thus capable of exceeding [0 1], for practical signals this rarely happens. When it does, the signal can be clipped without sacrificing useful information. Zeros are substituted for $l_0(t)$ and $l_{25}(t)$. The effect of this processing stage, and of the thinning and holding logic, can be seen in Figure 3.7.

3.4.5 Output conditioning

Having tried a number of ways of geometrically combining the left- and right-ear signals, the following root-sum-square strategy was found to be a good compromise between high contrast and monaural sensitivity:

$$\Sigma(t) = \sum_{j=1}^{24} \sqrt{m_{L\,j}^2(t) + m_{R\,j}^2(t)} \tag{3.25}$$

where $m_{Lj}(t)$ and $m_{Rj}(t)$ are the left and right ear $m_j(t)$signals, and $\Sigma(t)$ is the sum signal. $\Sigma(t)$ is shown in Figure 3.8.

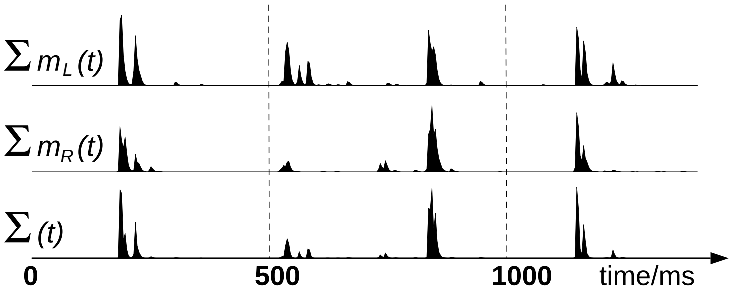


Figure 3.8. Sum of left and right ear $m(t)$ signals for the piano excerpt, and the root-sum-square resultant, $\Sigma(t)$. These traces have been scaled individually, and therefore are not mutually to scale. The large transient at zero time that occurs in previous figures has been removed for clarity.

It is now necessary to convert the fluctuating function into a binary one. There are at least two established ways of performing this task. The fuzzy logic approach involves filtering values according to a fixed threshold. Any function value that exceeds this threshold generates a positive output; otherwise, the output is zero. This approach is used in Klapuri's onset detector [Klapuri 1999], where information on sound intensity before and after onset candidates is available. However, this approach is not compatible with streaming signals for two reasons. Firstly, $\Sigma(t)$ does not just spike at an onset: it takes time to rise and decay. Therefore a threshold detector has to remove the stream of contiguous positive outputs that the function generates. $\Sigma(t)$ also reflects the presence of some small fluctuations that would mislead the spatial analyser if they were detected as onsets. Setting an absolute threshold regardless of the nature and content of $\Sigma(t)$ would cause the binary decision maker to miss onsets in some signals and to detect onsets falsely in others.

The methods of Smith [2001] and Marolt et al. [2002] use integrate-and-fire neurons to generate the binary output. The activity of an integrate-and-fire neuron is governed by the following formula (adapted from [Marolt et al. 2002]:

$$\frac{dO}{dt} = I - \gamma O \qquad (3.26)$$

in which O is the output activity of the neuron, and I is the input to it. Thus γ

symbolises the ‘leakiness’ of the running integration process. When O exceeds a certain threshold, the neuron fires, O is reset to zero, and a period of either total or relative insensitivity to the input follows, depending on the sophistication of the model.

The operation of a leaky running integrator is comparable to a band-pass filter, since an integrator is a low-pass filter, and leakiness constitutes low-frequency attenuation. Decision tasks in some onset detectors explicitly use band-pass filters: for example, they are included in an early model by Smith [1994] and also an algorithm by Schwartz et al. [1999]. They appear just as frequently in a more implicit form: as an envelope extraction (low-pass filter) procedure followed by delay-programmed inhibition, for example in Mellinger [1991] and Palomäki et al. [2004].

Schwartz et al. [1999] use adaptive filters which converge to band-pass filters. In the method used within this project, the band-pass and refractory properties of integrate-and-fire neurons have inspired an alternative method, where $\Sigma(t)$ is compared continually with the sum of two thresholds: a fixed threshold, h_f, and a variable threshold, $h_v(t)$:

$$o(t) = \begin{cases} 1 & : \quad \Sigma(t) \geq h_f + h_v(t) \\ 0 & : \quad \Sigma(t) < h_f + h_v(t) \end{cases} \tag{3.27}$$

A value of $o(t) = 1$ signifies an onset. $h_f = 3.0$ for all examples in this thesis. Levels of this magnitude appear frequently in most musical signals, but the level is set high enough to prevent triggering by most chaotic signal components. The variable threshold is altered depending on preceding values of $o(t)$, $\Sigma(t)$, $h_v(t)$, and a hold-and-decay function (as described by Equation 3.17), $H(25,30,t)$:

$$h_v(t) = \begin{cases} 4\Sigma(t-1) & : \quad o(t-1) = 1 \\ s_{15}h_v(t-1) + (1-s_{15})\Sigma(t-1) & : \quad o(t-1) = 0,\ \Sigma(t-1) \geq h_v(t-1) \\ H(25,30,t-1)h_v(t-1) & : \quad \Sigma(t-1) < h_v(t-1) \end{cases} \tag{3.28}$$

$$H(25,30,t) \quad \text{reset} \ : \quad \Sigma(t-1) > h_v(t-1) \tag{3.29}$$

s_{15} is defined in Equation 3.15. It is used here to form an IIR filter that interpolates between the variable threshold $h_v(t)$ and the input $\Sigma(t)$, with an integration time of 15ms.

The piano excerpt is passed through this final stage of the onset detection process in Figure 3.9. Two onset spikes are generated during the first 10ms of the excerpt. The remaining four are spaced approximately 350ms apart, and coincide with the beginnings of each group of notes.

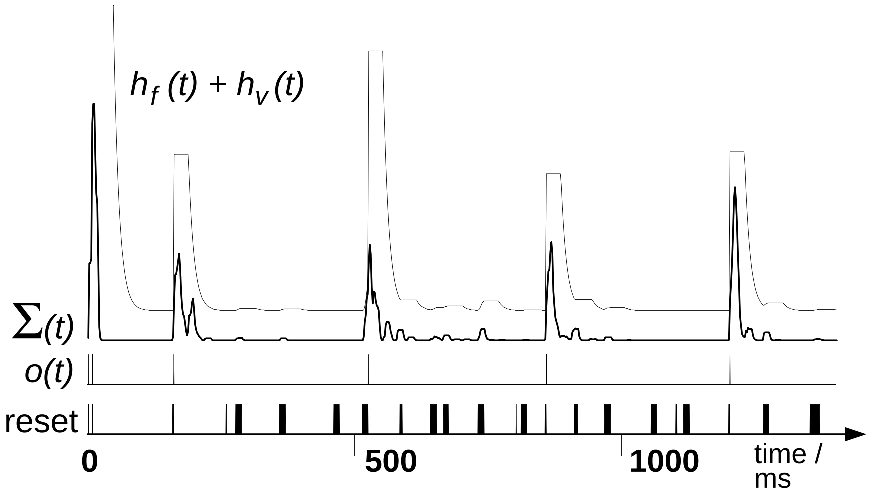


Figure 3.9. Performance of the binary decision maker, showing the temporal locations of the six detected onsets. Two onsets occur at the beginning of the excerpt, and four more mark the beginnings of each of the four groups of notes. Also shown are the 'reset' signal — this is positive whenever the hold-and-decay is restarted — and the sum of the fixed and moving thresholds.

3.5 Summary

This chapter has described an onset detector designed to work with the spatial analyser. The special requirements of this detector are that it must process streamed data, and that it must be broadly compatible with the spatial analysis task. Onsets must be detected substantially more than 100ms apart in typical reverberant music and speech material, preferably at instants in the binaural signal where a human listener would consciously recognise onsets. Each onset must also be detected very quickly, while the sound pressure level at the ears is still rising sharply. Many models of onset detection already exist, but none are closely compatible with the spatial processing task. This has required a novel approach to be taken to onset detection.

The onset detection algorithm combines a number of techniques from existing detectors to provide a simple algorithm that is both sensitive to genuine onsets, and robust to false triggering from chaotic signals and strong early reflections. Techniques applied in this algorithm include simple envelope extraction that preserves the linear scale of the input data, and the use of linear regression to find general signal trends without resorting to low-pass filters. Filtering the signal further would introduce delays through phase shifting. Band-pass filters are used throughout, usually implicitly by comparing low-pass filtered signals with other, more heavily-filtered signals. Band-pass onset detection techniques are found explicitly in the onset detectors of Smith [1994] and Schwartz et al. [1999]. Band-pass filters are also comparable to the integrate-and-fire neuron approaches favoured in more recent papers by Smith [2001] and Marolt et al. [2002]. The functionality of the low-level precedence effect is implemented in this algorithm within the thinning-and-holding process and the binary decision mechanism. For several milliseconds after the onset detector fires, re-triggering is suppressed.

Of the many techniques employed in the development of this algorithm, most are inspired by fuzzy logic rather than neurophysical theory. Since there is little information about the way in the brain detects auditory onsets, the approach has been almost entirely data-driven.

The onset detector is tested in Chapter 5 as a part of the localisation system. A large proportion of the constants used within this project have been refined by a process of trial and error — it will be shown that the algorithm generally performs well, but some of these constants may not be optimal. Specifically, the fixed values chosen for the binary decision maker are not suited to slowly-changing signals. Fortunately, it has been demonstrated that only these values, rather than the algorithm itself, need to be changed to fix this problem. Specific conclusions about the performance of the onset detector will be drawn from the experiments in Chapter 5.

4 LOCALISATION ALGORITHM

This chapter describes the localisation algorithm that converts the binaural stream into a running map of lateral angle against time. This conversion is achieved by extracting interaural time and intensity differences (ITDs and IIDs) from the binaural stream, and comparing these differences with a number of look-up tables constructed from psychoacoustic data. An overview of this process is shown in Figure 4.1.

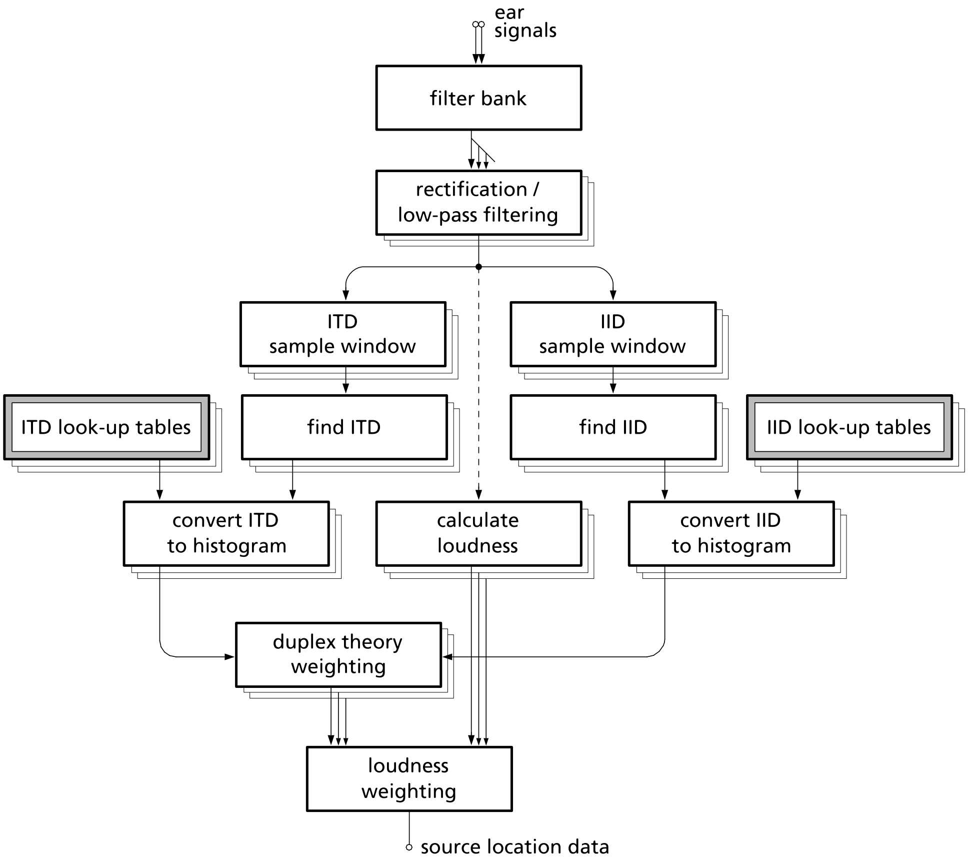


Figure 4.1. Overview of the localisation algorithm.

This flowchart shows the *analytical algorithm.* This chapter also describes another two algorithms that are not used during run-time, but are essential to the operation of the analytical algorithm. These are the *generative algorithms* which create the ITD and IID look-up tables.

There are a few similarities between the generative and analytical algorithms, and these will be explained as they are encountered. The two classes of algorithms will otherwise be described separately, starting with a review of existing analytical algorithms for extracting and decoding ITD and IID, and a detailed description of those procedures chosen for this project. Explaining the analytical algorithm before the generative algorithms allows the purpose of the look-up tables to be understood completely, so that the reasons behind the complexity of the generative algorithms are clear. A number of techniques will be described that are used to speed up the analysis of the binaural data.

Before any description is attempted, a polar co-ordinate system will need to be introduced to allow unambiguous representation of the auditory space. It is also important to discuss some of the limitations of binaural data, as these shortcomings are responsible for some of the decisions that have been made in the design of the algorithm.

4.1 Introducing the co-ordinate system

Two polar co-ordinate systems are used in this thesis. They are both expressed in terms of an angle on the horizontal plane. The second polar dimension, that of elevation, is not used in this project, because elevated sources are not considered explicitly. All possible sources are generalised to the diffuse-field, unelevated condition.

The most common system used in this thesis describes the lateral angle at which a source is positioned or localised. It is shown in Figure 4.2a. The second co-ordinate system, shown in Figure 4.2b, is used to describe monaural recordings. In both conventions, zero degrees refers to the situation in which the source is placed in the direction that the recording head or listener is facing.

In the first convention, +90° refers to 90° right of the listener, and –90° refers to 90° left. In the second [monaural] convention, 90° refers to the situation in which the ear is oriented towards the source, and 270° refers to the situation where it is facing away.

A polarity sign is always prepended to angles in the first convention, with the exception of 0° and 180°. No sign needs to be prepended in the monaural convention. In this project, samples are taken on the lateral angle domain around the head. This is always performed using a sampling interval of one degree.

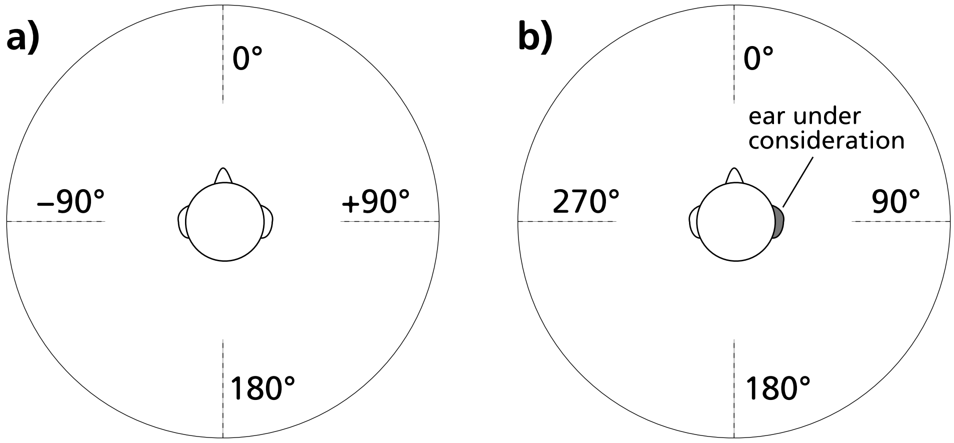


Figure 4.2.
a) The co-ordinate system that describes sound source placement and localisation.
b) The co-ordinate system that describes monaural recordings and situations. In the monaural convention, 90° is always oriented towards the ear under consideration, whether it is the left or the right ear.

4.2 The use of binaural data

It would be ideal to produce a algorithm that performs three-dimensional localisation as effectively as a human listener. Unfortunately, there are shortcomings in the binaural format that complicate this task, and additional compromises that must be made when generalising the localisation algorithm to work with any binaural recording.

A major problem is the *cone of confusion* phenomenon [Blauert 1997: 179]. A ‘cone of confusion’ is a locus on which the difference between the distances to the ears does not change. This locus approximates the surface of a baseless cone whose apex falls between the two ears on the interaural axis. At every point on this cone, interaural time differences are identical. Discrimination between locations that lie on the same cone of confusion using only interaural time differences is therefore impossible.

Beyond a few metres’ distance, inverse-pressure law effects cease to exert much influence on interaural intensity differences. If it is assumed that the head is approximately spherical, the cone of confusion will also hold for IIDs at distance.

If a more sophisticated model is applied, in which the head is non-spherical and pinna and torso reflections are taken into account, the relationship between ITD, IID, and listening position becomes more complicated. At low frequencies, the spherical model is a good approximation, because the listener's head has no detail that is significant at a wavelength scale. At high frequencies, though, reflections from the listener's head, pinnae and torso create spectral notches and peaks. These nonuniformities add extra IID cues that can help to pinpoint a single direction on the cone of confusion.

General trends can be spotted in the angular variation of IID cues among a group of listeners. Many of the irregularities are caused, for example, by head shadowing and concha resonances that tend not to vary widely from listener to listener. However, there are a number of peaks and notches in the frequency spectrum whose characteristics can vary substantially between listeners [Møller et al. 1995]. Considerable variation is also encountered in standardised dummy recording heads [Møller et al. 1999]. An example of this variation is the standing wave created in the ear canal. The ear's response to frequency components above approximately 3kHz is greatly influenced by very small interpersonal variations in the structure of the eardrum and ear canal (for example, [Stinson 1990]). Some of the differences are substantial enough to impair localisation performance significantly when a listener hears a recording made using another listener's ears. Localisation precision is further impaired if a listener is asked to locate sources in a recording made with an artificial head [Minnaar et al. 2001].

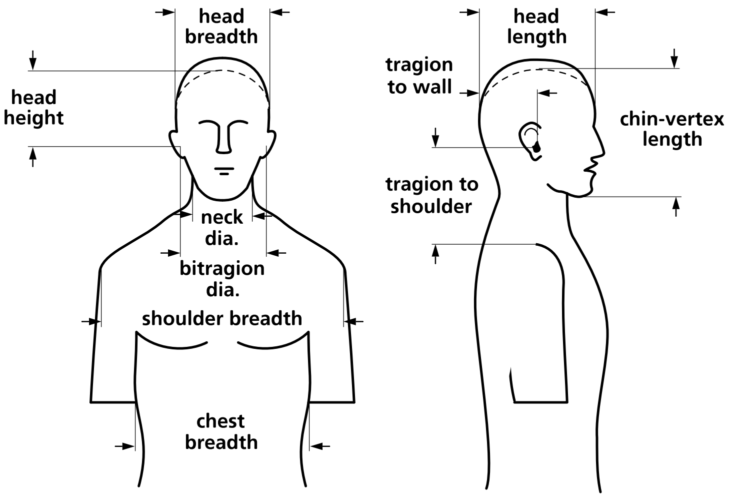
To transcend the cone of confusion, one of two things must be done. The simplest solution is to generalise the ITD and IID look-up tables to a particular model of dummy head, and accept that there will be errors when using binaural signals recorded using other models of head. Alternatively, a learning algorithm can be incorporated so that the model trains itself as it listens. This approach has been investigated for some years by Karjalainen and his research group, for example [Palomäki et al. 1999]. Both techniques involve making assumptions about the source stimulus, so they will work only with familiar listening material. The first technique is the simplest, but is prone to errors if data from a non-specific recording head is used. The second approach would be the most psychologically accurate because human listeners must adapt their own head-related cues as they grow, and subjects have demonstrated variously that they can re-train easily to artificial head data [Minnaar et al. 2001]. It would, however, require a considerable library of example data to remap the look-up tables with precision.

Even if difficulty prohibits venturing as far as full three-dimensional location, it might still be possible to process head-related impulse responses to enable 360° localisation in two Cartesian dimensions. There are a number of algorithms that attempt this with some success, albeit under anechoic conditions. For example, some results from an algorithm by Kunz and Bodden presented by Bodden [1998], and a set of results from Backman and Karjalainen [1993], both describe successful planar 360° localisation under anechoic conditions. However, there is no evidence of either system being fully evaluated, and their implementation details cannot be found.

The use of an ITD and IID database from one head, and analysis material from another, is tested and supported by the experimental research of Begault et al. [Begault et al. 2001]. Their listening experiment suggests that the localisation errors involved when non-individualised data is used are small, as long as front-back confusions are not counted as localisation errors. However, the localisation accuracy of their nine listening subjects falls considerably if front-back confusions are considered as localisation errors. Specifically, when the binaural simulation is non-interactive, the reliability of localisation judgements falls below 50%. One can then assume that the listeners are guessing the hemisphere in which the source is placed.

It can be concluded that human listeners make frequent front-back reversal errors when listening to binaural recordings. This finding is corroborated by the listening experiments of Møller et al. [1996] and Horbach et al. [1999], both of which also demonstrate that as long as interaction with the sound field is not permitted, front-back reversals are hardly less frequent whether a sphere, a dummy head, or another listener's ears are being used to supply ITD and IID cues. Both noted improvements in front-back localisation when the listener's own ears were used; however, these improvements were small. The greatest improvement in front-back discrimination accuracy in all cases occurred when a head-tracked simulation was used in place of the static binaural recording.

No convincing data exists to demonstrate that a computerised binaural analyser can localise sources any better than a human listener, without having access to additional cues. Therefore, to avoid cone-of-confusion-based complications, no attempt is made in this project to transcend the cone of confusion. This means that the database obtained from the KEMAR dummy head data set [Gardner and Martin 1994] is universally applicable, as long as a small amount of error can be tolerated. The recordings for analysis were made using a Cortex Instruments MK2 dummy head. A comparison of the dimensions of Cortex and KEMAR heads, shown in Figure 4.3, shows that they are almost identical. This is further evidence that ITD and IID discrepancies owing to head-shadowing and ear spacing should be very similar.



Dimension	KEMAR	IEC 959
Head breadth	152	
Head height	125	
Bitragion diameter	143	
Neck diameter	112	113
Shoulder breadth	440	
Chest breadth	282	
Head length	188	191
Tragion to wall	96.5	97
Chin-vertex length	224	
Tragion to shoulder	175	

Figure 4.3. Dimensions of the KEMAR head, and IEC 959 parameters, to which the Cortex Instruments MK2 head is designed. All measurements are in millimetres.
KEMAR parameters are taken from Burkhard and Sachs [1975]; IEC 959 parameters from Wojcik and Cardinal [1999].

Localisation in this project is achieved by specifying a source position as a lateral angle in the horizontal plane. Thus a source placed at +110° will be localised successfully at +70°, and while source location angles less than –90° and greater than +90° are valid in nature, they will not be considered by the algorithm. Elevated sources will be localised at the unelevated point on their cones of confusion. To extend this project and investigate the performance of the localisation algorithm under flat 360° conditions would form an interesting subject for further work.

4.3 The analytical algorithm

This section describes the processes behind the flowchart shown in Figure 4.1, at the beginning of this chapter. As there are many accepted ways of extracting ITD and IID information, and also many ways to make sense of them, the elaboration of each part of the algorithm will be interplayed with a review of relevant literature.

The output of the analytical algorithm is arranged, in common with many other existing localisation algorithms, three-dimensionally, as a map of truth values versus lateral angle and time. 'Truth value' is a fuzzy logic term, referring to a value in the [0 1] range that represents the degree of membership of a set. Hence, in this case, every truth value represents the degree to which a corresponding lateral angle belongs to the set of possible angles of incidence.

4.3.1 Filter bank, rectification, and low-pass filtering

The Slaney filter bank, full-wave rectification and low-pass filtering are the first three processes of the analytical algorithm, as shown in Figure 4.5. These processes, described in Section 3.4.1, approximate the action of the cochlea by converting the two pressure waves of the binaural input into a representation of neural activity in the inner ear. Full-wave rectification is favoured over half-wave rectification because this enables timing and intensity information to be extracted from both positive- and negative-going portions of the input signals.

4.3.2 Interaural intensity differences

Owing to the problems associated with extracting IIDs, binaural localisation algorithms that are not designed to simulate human listening often ignore them altogether, concentrating on ITD-based localisation. However, IIDs are an important cue for human listeners, significantly more salient at high frequencies than ITDs [Macpherson and Middlebrooks 2002]. They must therefore be taken into account in the localisation algorithm.

When a listener hears a natural sound source, the IID cues will be caused by three phenomena:

- At very close distances, an interaural intensity difference occurs because of the different distances that a signal must travel to either ear. For sources that obey the inverse pressure law, the maximum level disparity generated by this mechanism is approximately 1dB at three metres' distance. This mechanism is not frequency-dependent.
- As wavelength decreases, the listener's head becomes more of an obstacle to the sound waves. Reflections from the head cause the sound pressure level to be raised at the nearer ear, and lowered at the farther ear. This head shadowing effect increases monotonically with frequency. It is greatest when the sound is coming from 90° left or right.
- At short wavelengths, the outer ear, and particularly the folds of the concha, creates reflections that effectively comb-filter the incoming signal. This causes a characteristic pattern of peaks and notches in the ear's frequency response. The pattern is heavily dependent on source location, and is termed a *head-related transfer function* (HRTF). Shoulder and torso reflections also contribute weak but perceptible changes in the HRTF.

IIDs may also be complicated by room reflections, as signals arriving from many angles around the head simultaneously can confound the IID cues provided by the unreflected signal path.

IID caused by distance-related attenuation is the easiest and most uniform cue to decode, but the information it conveys about distance and source angle is highly ambiguous, and its influence is confined to small source distances. Furthermore, as distance-related attenuation interacts with head shadowing and HRTF peaks and notches, the pattern of IIDs immediately around the listener's head becomes very complicated. This is shown in Figure 4.4. It is very difficult to localise narrow-band stimuli at close distances reliably using only IIDs.

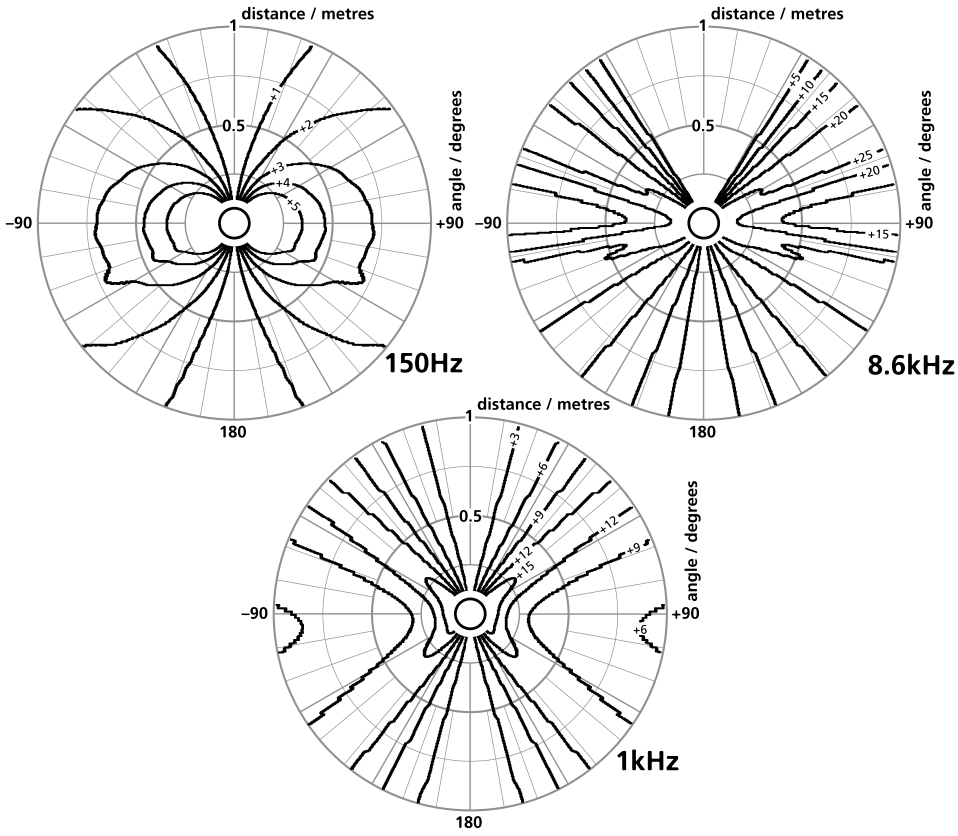


Figure 4.4. Equal-IID contours for distances between 20cm and 1m from the centre of a listener, computed using KEMAR HRTF data and physical measurements [Gardner and Martin 1994; Burkhard and Sachs 1975].

Spectral cues created by pinna and torso reflections convey a wealth of information, since they change quickly with head angle and source elevation. However, for a number of reasons, it is difficult to extract spectral cues reliably. HRTFs vary widely between individuals, and even between dummy heads of standard dimensions from different manufacturers [Møller et al. 1999]. Although certain spectral characteristics are known to concur with certain source directions, the frequency and extent of these characteristics vary between listeners.

Just as humans learn to localise using their own HRTFs, it is possible for artificial listeners to use spectral cues. This is achieved by splitting the signal into frequency bands, analysing the IIDs in each band, and combining the results. Some recent algorithms perform this task using neural networks [Nandy and Ben-Arie 2001; Palomäki et al. 1999]. However, both human and machine listeners exhibit one particular problem when attempting to locate an unfamiliar sound without the freedom of movement: it may be impossible to tell whether the spectral characteristics that are heard are an innate characteristic of the source, or whether they have they been caused by pinna filtering. This is particularly true near the centre line of the head, where the left and right ear cues are similar.

At low frequencies and distances of a few metres, the interaural intensity difference is a simple function from which the source angle can be approximated mathematically. At higher frequencies, where spectral features are manifested, the IID-to-angle mapping ceases to be one-to-one. Thus, the intensity difference information from a small number of frequency bands may indicate a large locus of probable sound locations, rather than a single point in space. The location may have to be derived by obtaining a consensus between several active frequency bands, and by using interaural time difference cues.

Extracting interaural intensity difference information

Figure 4.5 is an expansion of part of Figure 4.1. It presents a general method for extracting interaural intensity differences. The procedure works by integrating signal energy over a short time period in each binaural channel, and comparing the results. This approach is implicit in all existing designs.

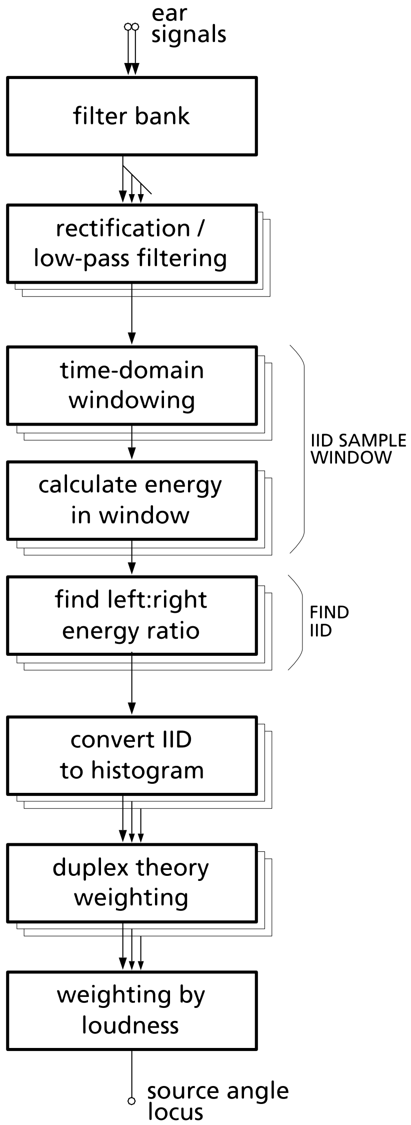


Figure 4.5. A general method for extracting interaural intensity differences.

Unless the signal under analysis is only a few milliseconds long, or the complete signal is available and its interaural cues are fairly static, it must be divided into many smaller parts using a windowing algorithm. Analysing these parts separately will produce a moving representation of IID against time.

In existing designs, the windowing algorithm usually takes one of two forms: either a continuous leaky integration of new information (referred to as a *level meter model* by Hartmann and Constan [2002]), or a slicing algorithm that takes a rectangular window of samples around a certain time. The extracted IID sometimes depends greatly on the windowing or filtering algorithm used to calculate it. This can be seen in the two periodic waveforms analysed in Figure 4.6.

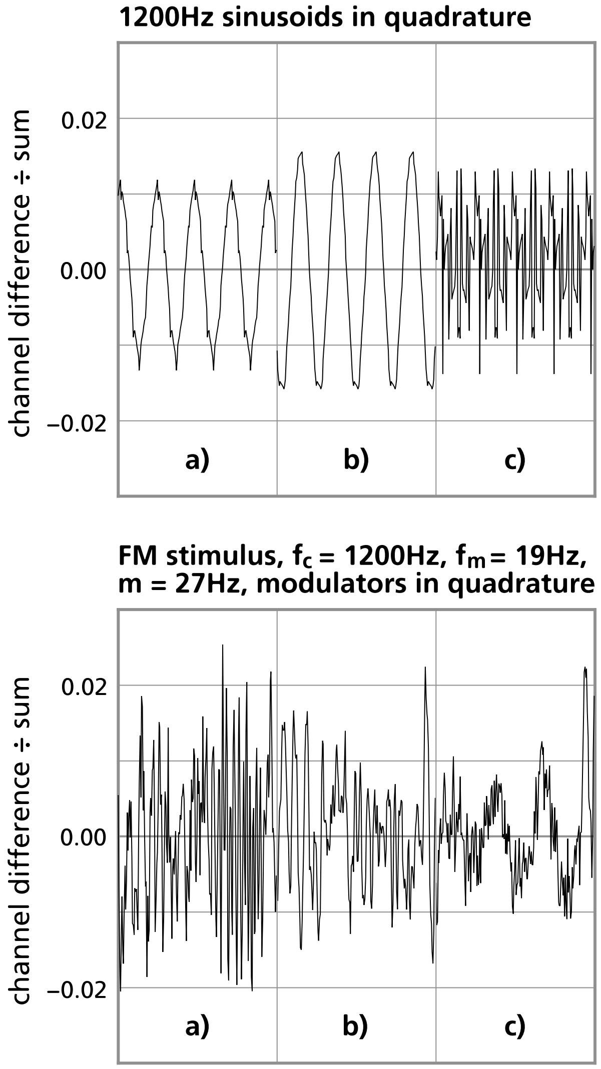


Figure 4.6. A comparison of IID calculations for two stimuli, using three different implementations of the level-meter model. Approximately 85ms of each method is shown. The left and right channels are of equal level. Ideally, the channel difference ÷ sum would be a constant value of zero.
a) Standard level-meter model. Output is second-order Butterworth filtered, f_0 = 800Hz.
b) Square-windowed average of 14 samples (950µs);
c) As above, with time-alignment correction.
Scheme c) is used in this project.

IID-to-angle conversion can be performed in one of two ways. The first strategy centres around a data array that represents physical space. This is pre-loaded with a look-up table of IID data, and might contain many hundreds of datum points. The second approach is to feed all available ITD, IID, and loudness data into a trained neural network, from which will emerge one or two values representing source angle.

When direct-arriving sound energy dominates, and the IIDs and ITDs from all active bands can be considered, the table of probable source angles usually resolves to a few closely-clustered candidates.

A review of existing IID extraction algorithms

There is no standard implementation of a technique for extracting IID by energy integration, so practices differ from researcher to researcher. Macpherson [1991] calculates ‘interaural amplitude difference’ as the ratio of energy (the sum of the squared signal) in the left and right channels of each critical band. The signal is split into 2.5ms windows, and the interaural amplitude difference is expressed in decibels. Frequency bands are combined using a weighted mean, with bands given more weight if they contain more energy. This weighting prevents random results caused by background noise, which can dominate in quiet bands.

Martin [1995b] uses a very similar technique over the same window interval, where signals from incoming bands are squared, then smoothed with a low-pass filter whose frequency is the same as the centre frequency of the incoming band up to 800Hz, and limited to 800Hz thereafter. The rest of the technique is identical to Macpherson’s, and the output value is also expressed in decibels.

The choice of 2.5ms as a window length in both papers is possibly a coincidence, although psychoacoustic and computational criteria dictate a value of this order. According to Hartmann [1997: 198], interaural differences become less salient around 1–1.5ms after onset. To accommodate such small time differences, an IID sampling window of 2.5ms is about the longest allowable — in fact, a sampling resolution twice as fine would be preferable. The demands on a microprocessor of calculating IID are fairly independent of IID sampling frequency in this range, so the practical limit of IID sampling frequency is dictated only by the sampling frequency of the input data, and the way in which IID data is handled subsequently.

Martin suggests that noise should be introduced artificially into the IID results in order that the model can ‘meaningfully be compared with human psychoacoustic data’ [Martin 1995b: 2–3]. The purpose of this noise may be twofold: it would increase localisation blur as the sound pressure level decreases (this phenomenon is observed in human listeners: see Blauert [1997: 155]) and the addition of internal error would also randomise responses and reduce the certainty of results.

To derive ILD, Palomäki et al. [2004] employ a frequency-dependent mapping function, derived from experimental data, to obtain an instantaneous lateral source angle estimate from the ILD of each critical band. This ILD is the energy ratio of the squared envelopes of the left and right ear signals. These envelopes are obtained using a time constant of 15ms. ILD data is then compared with ITD data in order to determine whether the data from the two cues are consistent, and therefore whether a wanted signal or a distracting signal is dominating the sound field at each instant.

When IID is expressed in decibels, it has a theoretically limitless range. This is not important when IIDs are resolved to source angles using look-up tables, as very large outer values of IID can be clipped to fit the tables. However, it would not be possible to introduce data with a large or limitless range into a neural network-based localiser, as the training mechanism of a neural network is based on the back-propagation of error values. Thus neural network models tend to be engineered to accept values in the range [–1, 1] [Gurney 1997].

In order to produce a value in this range with a minimum of computation, the *level meter model* of Hartman and Constan [2002] adapts the ‘judgment function’ exploited by Sayers and Cherry [1957: 982] in their experimental paper. This is produced by dividing the difference of left and right intensity values by their sum. This normalised ratio, which always falls within [–1, 1], is also employed by Backman and Karjalainen [1993], and modified in Karjalainen [1996] to derive IID data directly from the firing rates of two neuron models.

4.3.3 Interaural time differences

There are several current ITD-extracting algorithms. These can be divided broadly into two types: one group is designed principally for efficient computer implementation, and the other aims to model neurophysical activity. A little insight into both types of algorithm will be required to understand the problems of this project. Because each new model tends to incorporate a specific advance, it is easiest to tackle them in historical order.

The Jeffress model

Any sound which contains some amount of low-frequency content, arriving simultaneously at both ears, will be converted into coincident neural impulses. This happens because inner hair cells within both cochleas fire impulses at the same phase in every cycle [Yates 1995]. A coincidence-counting neuron that receives both impulses simultaneously will register a high number of coincidences in this case. Fewer coincidences will be registered by this neuron if the same sound is delayed in one ear. However, if another coincidence-counting neuron is placed in such a way that there is a counteracting propagation delay along the nerve fibres, it will register a high number of coincidences. A suitably large array of coincidence-counting neurons arranged in this way produces a neurophysically viable method of finding interaural time difference: the most active coincidence counter indicates the ITD.

This model, illustrated in Figure 4.7, was first proposed by Lloyd Jeffress [Jeffress 1948]. The neurological scheme that this employs, in which two signals are compared repeatedly as one is successively delayed with respect to the other, is now known as a *labelled line* [Schnupp 2001: 677]. Sayers and Cherry [1957: 980–1] expressed this model mathematically, and it now forms the basis of the running interaural cross-correlation (IACC) algorithm. A flowchart of this diagram is shown in Figure 4.8. The algorithm is simple to implement, produces precise results, and has therefore obtained widespread popularity.

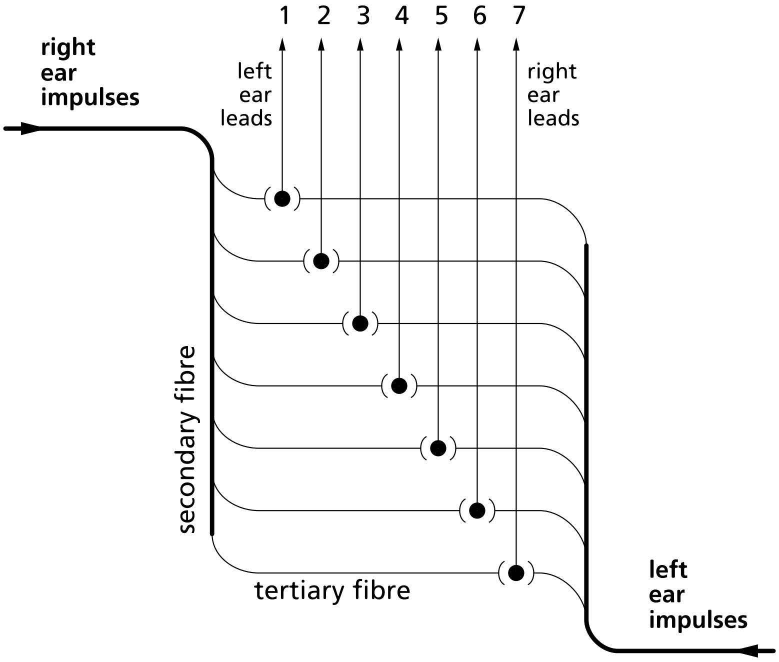


Figure 4.7. The Jeffress model (after [Jeffress 1948]). Coincidence-counting units are arranged along fibres between left- and right-ear neurons. Every fibre possesses a propagation delay.

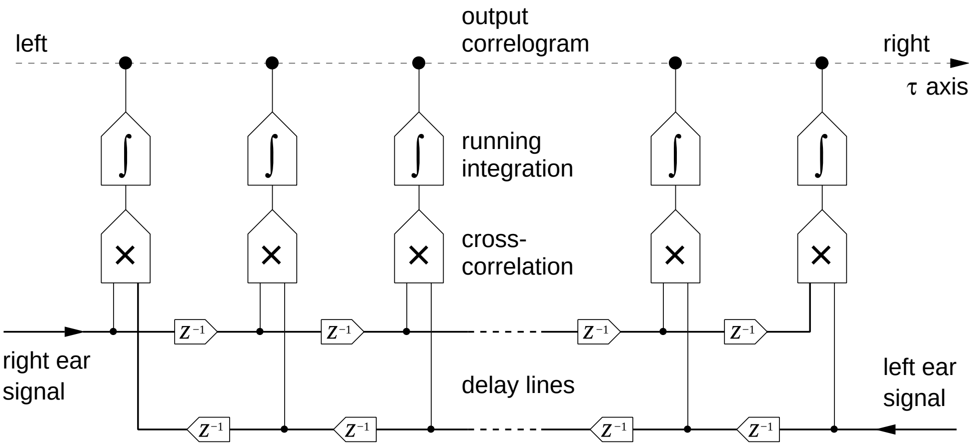


Figure 4.8. Computational version of the Jeffress model, after Lindemann [1986].

The correlogram

The interaural cross-correlation (IACC) of a binaural signal is usually computed separately for every frequency band, so time differences can be weighted differently against intensity differences across the frequency spectrum. IACC data is therefore four-dimensional: cross-correlation versus time-lag is computed for each frequency band, and this happens for every audio sample.

The piecewise product of the left- and right-ear signals of each frequency bands, normalised to compensate for input level, can be plotted against mutual time-lag to form a graph called a *correlogram* [Lyon 1983]. The symbol τ conventionally represents the time-lag axis, and serves this purpose in the standard cross-correlation formula [BS EN ISO 3382:2000].

A correlogram is a useful analytical tool, and is used commonly in auditory localisation research. It provides an indication both of the amount of coincidence between the two ear signals, and the interaural time delay for which this coincidence is maximal.

The purpose of level normalisation in a correlogram is to produce a measurement of signal correlation that is independent of input level. Perfectly correlated signals will produce a peak value of 1 on the correlogram, and this is their interaural cross-correlation. Uncorrelated signals will have a peak IACC close to zero, irrespective of the absolute input level. The standard cross-correlation algorithm [BS EN ISO 3382:2000] takes the root-product-square of the two signals as a denominator to normalise the correlogram:

$$IACF_{T_1 T_2}(\tau) = \frac{\int_{T_1}^{T_2} p_l(t) p_r(t+\tau) dt}{\sqrt{\int_{T_1}^{T_2} p_l^2 dt \int_{T_1}^{T_2} p_r^2 dt}} \tag{4.1}$$

$IACF_{T_1T_2}(\tau)$ represents the interaural cross-correlation function in the time interval T_1 to T_2. $p_l(t)$ and $p_r(t)$ are the sound pressures at the left and right ears against time.

Many algorithms do not apply this normalisation, because the denominator is not quick to calculate. In general, if a normalised correlogram is required, it is far faster to approximate the denominator by taking the sum of running level meters from each ear.

Many cross-correlation ITD extractors, among them the models of Stern and Colburn [1978] and Macpherson [1991], disregard the actual measurement of coincidence: they are concerned only with the interaural time delay for which coincidence is maximal. This is determined either using the peak or the centroid of a correlogram. Under these circumstances, computational efficiency can be improved by discarding normalisation entirely. However, some recent localisation algorithms require the normalised *IACF* as a separate cue. In addition to its use as an architectural acoustics parameter (see Section 2.2), the interaural cross-correlation function has been used to detect auditory onsets by Faller and Merimaa [2004], and as a cue for separating speech sources from distractors by Palomäki et al. [2004].

The running interaural cross-correlation of a binaural signal with a sampling frequency of 44.1kHz, split into 24 bands, and calculated over a time-lag range of ±1ms, generates a set of correlograms containing more than 40 million data points per second, all of which result from separate multiplication operations. Multiplication is generally a slow operation, so generating IACC data is most of the work of a localisation system. Reducing the computational load of ITD calculation is a convenient way of speeding up the algorithm. Unfortunately, the large number of multiplication operations that a running IACC requires, and its high data throughput, makes the IACC inherently slow. This has led to attempts to find other means of extracting interaural time difference from binaural signals, and there are at least two systems in which the IACC process has been modified to perform ITD and IID extraction simultaneously.

EC and Stereausis

For decades after the publication of the Jeffress model, its physical validity was debated among researchers. Evidence for the neural auditory delays that the model requires were found in barn owls, but doubt has always been expressed about its validity in humans [Fitzgerald 2002: 13]. Durlach [Durlach 1963] proposes a theoretical mathematical model that simulates the improvement in masking level differences experienced in binaural listening over monaural listening. In this model, sharper directivity is obtained by cancelling distracting signals in the sound field. Durlach separates this model into two processes. The first is equalisation, in which the distracting signals in the two ears are transformed to be of similar phase and level. The second is cancellation, in which the signals are subtracted, thus cancelling the distracting information and increasing the signal-to-noise ratio of useful signal components. He termed this hypothesis the *EC model* ('equalisation and cancellation').

Schroeder [Schroeder 1977] hypothesised that instead of being generated neurally, the interaural phase shifts necessary for the EC model may be caused acoustically, by exploiting the movement of waves along the basilar membrane. An array of inhibition-type neurons would then perform the necessary level equalisation and cancellation. A system that uses the inherent delays in a simulated cochlea as the basis of a localisation model was introduced and investigated in 1991 [Shamma et al. 1989], and given the name *stereausis.*

In stereausis, each channel of the binaural signal is divided into more than one hundred frequency bands, each of which somewhat overlaps its neighbours. The different filters have different centre frequencies, and bandpass filtering a signal imposes a small group delay that is dependent on frequency. The higher a filter's centre frequency, the shorter its impulse response, and the shorter this delay. By piecewise multiplying the output signals from different bands together, a number of different interaural delays can be probed directly, without the need for the explicit delay lines of Jeffress's model.

The latency hypothesis

Jeffress suggested cautiously that his model could also account for sensitivity to interaural intensity differences, if one supposes that inner hair cells transform a quieter signal into a later neural spike. This hypothesis, later termed the *latency hypothesis*, was eventually refuted. Colburn and Durlach [1978: 469] argue that at high frequencies, the phase-locking sensitivity of the inner ear breaks down so the human auditory system is no longer sensitive to the fine structure of a signal. However, it is very sensitive to interaural intensity difference. Unfortunately, the latency hypothesis accounts only for situations where the human auditory system is sensitive to both time and intensity differences, or to time differences alone. It cannot cope with this exception. The latency hypothesis also cannot explain the inability of a listener to trade time and intensity differences completely, a phenomenon that was first highlighted by Hafter and Carrier [1972]. Problems with the latency hypothesis are covered in more detail by Stern and Trahiotis [1997: 506*ff*].

Lindemann and Gaik

Lindemann's model [1986] extends the computational version of the Jeffress model shown in Figure 4.8. Firstly, it includes two 'monaural processors'. These are simple mechanisms built into the ends of each delay line to improve localisation results when large IIDs are present. The Lindemann model also includes a system of actively-controlled losses along each delay line. Two kinds of loss are imposed:

- Contralateral inhibition: the stronger the signal on one ear's delay line, the more strongly the opposite delay line is attenuated.
- Dynamic inhibition: the greater the output of the multiplier at each tap, the more both delay lines are attenuated. Each of the signals that control this attenuation is passed through a specialised low-pass filter.

These are shown in Figure 4.9. Contralateral inhibition sharpens the peaks of the correlogram, and also adds some sensitivity to IIDs. When a large signal from one ear meets a smaller signal from the other, the small signal will be attenuated severely by the presence of the large signal. The large signal will not be attenuated as heavily, and will propagate further along the delay line. This spreads the correlogram in favour of the greater signal level. Dynamic inhibition, and the low-pass filters that control it, incorporates a simple model of the precedence effect into the delay line.

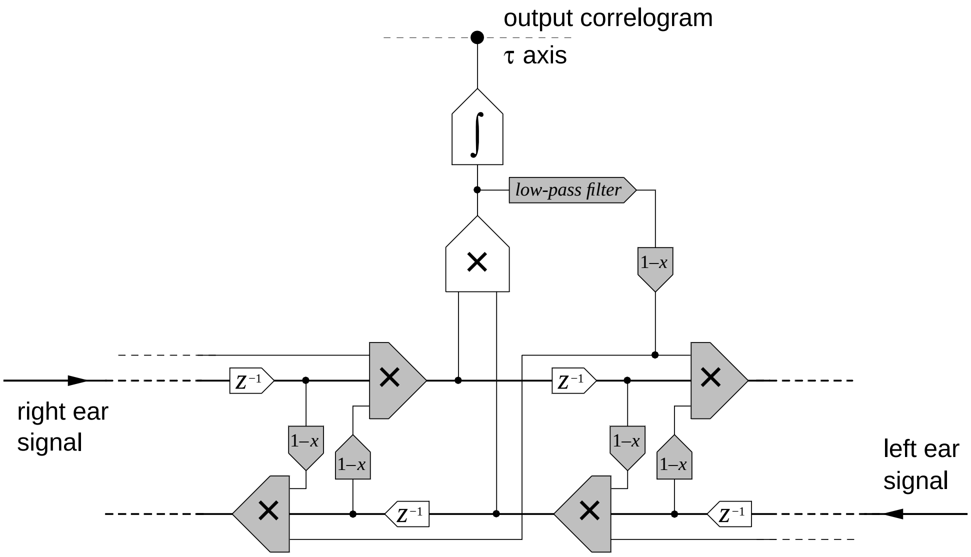


Figure 4.9. One correlogram tap of Lindemann's localisation model [Lindemann 1986]. Contralateral and dynamic inhibition are illustrated here, but the monaural processor is not.

Gaik's implementation of the Jeffress model [Gaik 1993] extends Lindemann's work by distributing a further set of attenuators along each delay line. These attenuators are weighted differently for every critical band, and this optimises results for natural combinations of ITD and IID.

The delay-line attenuators that form part of the Gaik model would require re-calculation whenever the length of the correlogram or the sampling frequency of the incoming data is changed. Also, these revised models produce a result that is not a correlogram in the strictest sense, because the process is not identical to the numerator of the cross-correlation algorithm shown in Equation 4.1. This means that the findings of an experiment conducted with the Lindemann or Gaik models are not directly applicable to those that use interaural cross-correlation.

Just as importantly, there is a growing body of research which indicates that the ITD and IID cues are extracted in different parts of the brainstem ([Schroger 1996], [Pratt et al. 1997]), and that much of the subsequent processing of these cues also takes place separately [Ungan et al. 2001]. Such findings reject the neurophysical validity of an integrated strategy.

EI neurophysical models

Very recent advances in neuroscience have made it possible to isolate individual neurons and study their responses to auditory stimuli. Experiments on anaesthetised cats have revealed coincidence-counting mechanisms that differ substantially from the Jeffress model. The neurons within these mechanisms do not respond symmetrically to stimulation. While impulses from one ear increase their activity, impulses from the opposite ear decrease it. Neurons that behave asymmetrically like this are called EI (excitatory-inhibitory), and form the basis of many newer models of ITD detection, including those by Breebaart et al. [2001], McAlpine et al. [2001], and Hancock and Delgutte [2004]. Older paradigms that are based on Jeffress's model are now referred to as excitatory-excitatory, or EE, models [Breebaart et al. 2001].

In the EI paradigm, cells are tuned to fire maximally at a certain characteristic interaural phase or delay, and this delay is of the order of hundreds of microseconds [Fitzgerald 2002: 14]. It is the rate of neural firing, and not the location of maximum activity along a labelled line, that corresponds to the interaural time difference. The details of workings of recent EI models are beyond the scope of this review.

EI models explain two related phenomena more satisfactorily than the EE paradigm with which they conflict. The first is the ability of human listeners to detect very small interaural differences: typical just-noticeable differences for untrained listeners are of the order of 10–20µs [Blauert 1997: 41; Domnitz 1973]. The second is the sensitivity of a human listener to a range of ITDs that can extend to two milliseconds [Hartmann 1997]. This apparent conflict was difficult to resolve in the Jeffress model without resorting to a model containing many hundreds of coincidence-counting neurons over each frequency band. In EI models, the human auditory system's sensitivity and range can be explained using a smaller number of more specialised cells.

The use of the EI model in localisation has one notable antecedent: an inhibitory mechanism, referred to as 'contralateral inhibition', is built into the delay lines of Lindemann's cross-correlation model [Lindemann 1986]. However, this is used mainly to sharpen the correlogram and to assist the imposition of interaural intensity sensitivity onto the model.

4.3.4 IID and ITD detection algorithms used in this project

A conceptual model of the ITD detection algorithm used in this project, from the rectified and smoothed critical band signals to the ITD output, can be seen in Figure 4.10.

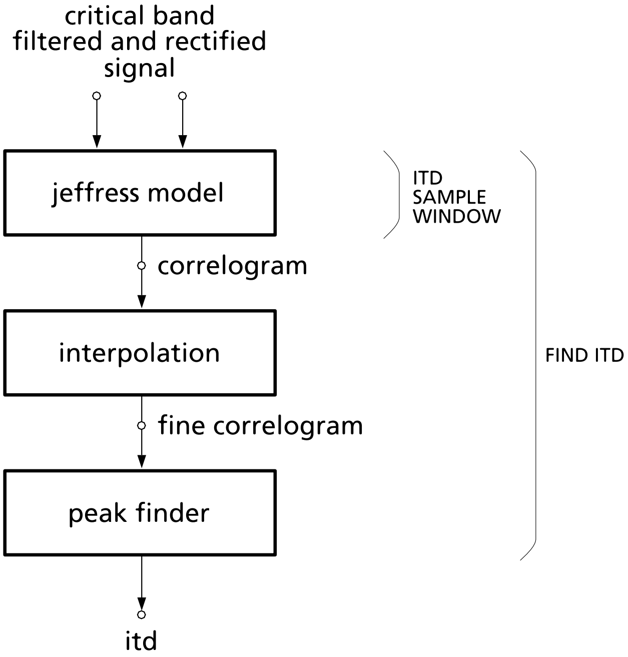


Figure 4.10. Concept for extracting ITD, at the required resolution, from one channel of a critical band filtered binaural signal.

Although interaural cross-correlation is no longer regarded as neurophysically valid, there are many reasons for choosing the Jeffress model over the other methods for this spatial feature extractor. The reasons for this choice over the three valid alternatives — stereausis, the Lindemann-Gaik approach, and EI — are based upon its simplicity, reliability, and speed.

In order for stereausis to work well, a very large amount of data needs to be manipulated. The prototype model demonstrated by Shamma et al. [1989] uses 128 critical band signals. To generate and manipulate this amount of data efficiently would require an FFT-based algorithm. Thus, a frequency-domain approach, rather than a time-domain approach, would have had to be taken in the design of the rest of this project. This would be a considerable departure from the majority of existing approaches. The possibilities of stereausis are still too uncharted for this departure to have been safe. Furthermore, there are no clear advantages, from a computational point of view, of the stereausis technique over the optimised Jeffress cross-correlation model presented here.

The Lindemann-Gaik approach has been rejected for several reasons. By combining ITD and IID sensing, the Jeffress model becomes far more complicated, and one departs further from a neurophysical approach to binaural localisation. In passing the trading of ITD and IID cues to a computer algorithm, not only is flexibility sacrificed, but the extracted IID emerges encoded on a time axis. Thus results from a Lindemann or Gaik model are difficult to compare directly with an IID database, or IID-based psychoacoustic research.

EI lateralisation methods are relatively new. When this research project started, much of the information needed to build an EI style localisation model was still unavailable or strongly contested, and the Jeffress model had not yet been superseded by the new discoveries. To abandon a working IACC model and replace it with a new-style model would have required more advantages than neurophysical validity alone, and would require there to be an established and agreed method of implementation. The different EI localisation models are only now starting to concur to the extent that an efficient computer algorithm, on which the majority of research agrees, can be realised.

The Jeffress interaural cross-correlation technique has been used in auditory processing for more than half a century, and its almost universal adoption during this time has generated a wealth of research on methods for interpreting and processing the IACC and correlogram to account for perceptual phenomena. If the IACC approach is preserved then this literature remains directly applicable.

Optimisation of the Jeffress model

The human auditory system is sensitive to changes in interaural time difference of approximately 10 microseconds around 0° [Blauert 1997: 41; Domnitz 1973]. Because each multiplying tap of the Jeffress model uses signals that are delayed by one sample in one ear and advanced by one sample in the other ear relative to its neighbours, the smallest interaural delay that the model can detect is two samples. At 44.1kHz, this corresponds to $\Delta\tau = 45.4$ microseconds. To give the ITD detector at least the same resolution as the human auditory system, one would need to interpolate either the input or output of the Jeffress model by at least 5:1.

Of these two alternatives, it is more sensible to interpolate the output of the Jeffress model, as this is more than twice as efficient as interpolating both channels of the input and then processing the resulting data.

To convey the full range of ITDs of the human auditory system, each correlogram would need to be computed with τ between ±1ms. This corresponds to ±22 samples when $\Delta\tau = 45.4$ microseconds. Each correlogram will therefore have 45 points.

A simple calculation shows that the data rate generated by this process is:

44100 Hz × 24 bands × 45 correlogram points × 5:1 interpolation factor
≈ 238 million points per second
(47.6 million of these are uninterpolated)

This amount of data can barely be calculated and manipulated in real time, and most of it is redundant. Hence there are many ways in which the computational load of the Jeffress model can be reduced in practice.

Choosing the sampling frequency

It is necessary to find the lowest data rate that the Jeffress model requires to produce an undistorted correlogram, given a generalised input signal.

When two sinusoids with the same frequency and an arbitrary phase relationship are multiplied together, the result will contain two components: a sinusoid of twice that frequency and, as long as the signals are not in quadrature, some DC. This is proved in Equation 4.2:

$$\begin{array}{rlcl}
 & \sin(x+\delta) & \equiv & \sin x \cos\delta + \sin\delta \cos x\,; \\
 & \cos(x+\delta) & \equiv & \cos x \cos\delta - \sin x \sin\delta\,. \\
\Rightarrow & \sin x \cos x & \equiv & \frac{\sin 2x}{2}\,; \\
 & \sin^2 x & \equiv & \frac{1-cos2x}{2}\,. \\
\Rightarrow & \sin(x+\delta)\sin x & \equiv & \sin^2 x \cos\delta + \sin\delta \sin x \cos x \\
 & & \equiv & \frac{\sin\delta}{2}\sin 2x + \frac{\cos\delta}{2}(1-\cos 2x)\,.
\end{array} \tag{4.2}$$

Thus there is a factor of eight that relates the minimum sampling frequency required for the Jeffress model to produce an unaliased output with the maximum audio frequency of interest. This factor is produced by three doublings:

×2 : Doubling of the maximum frequency when two audio signals are multiplied together;

×2 : The halving of time resolution in the Jeffress model, because each successive tap receives an increasingly delayed signal from one ear and a decreasingly delayed signal from the other;

×2 : Nyquist limit. The sampling frequency must be at least twice the highest frequency within the sampled signal.

For example, if the maximum frequency of interest is 2kHz, the minimum sampling frequency of the input data would be 16kHz. In practice, the maximum frequency of interest is a little lower, and the input sampling frequency of 44.1kHz can be reduced by a factor of three to 14.7kHz.

The peak-finding algorithm

The temporal resolution of the correlogram taken at 14.7kHz is approximately 136μs. Since the human sensitivity to interaural time difference is more than an order of magnitude finer than this, the data must be interpolated in order to find a peak. However, there is clearly no need to interpolate the entire data set: only the likeliest data intervals need to be examined more closely. A simple algorithm has been devised to interpolate the data using a cubic regression technique.

Efficient interpolation

The purpose of an interpolation algorithm is to take an arbitrary discrete-time function, $y(n)$, n = 0, 1, ... , $N-1$, and to synthesise data points between the sampling intervals so that n is no longer required to be an integer. There are two classes of algorithm that can be used to perform this task: true interpolators and approximators. Interpolators satisfy the requirement that the output data equals the input data at the sampling intervals, and approximators do not [Lehmann 1999]. The algorithm described here is a true interpolator.

There are a number of methods for interpolating. For discrete-time sampled signals, it is generally the case that the more terms the interpolation algorithm possesses, the more closely the interpolated data will fit the original continuous-time data. A simple class of interpolation algorithms is based on low-order polynomials. The first four polynomials are shown in Figure 4.11.

Interpolation type	Order	General formula between nodes	y	dy/dx	d^2y/dx^2
truncation	zeroth	$y = d$	discont.	zero	zero
linear	first	$y = cx + d$	continuous	discont.	zero
parabolic	second	$y = bx^2 + cx + d$	smooth	continuous	discont.
cubic	third	$y = ax^3 + bx^2 + cx + d$	smooth	smooth	continuous

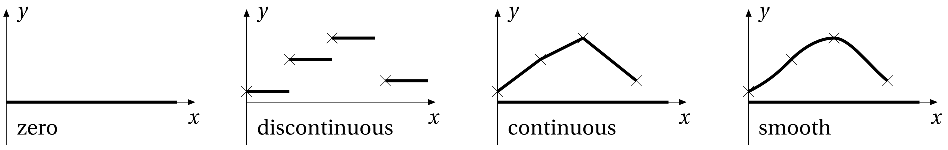


Figure 4.11. A table of polynomial interpolation formulae, with graphs to illustrate the terms 'zero', 'discontinuous', 'continuous' and 'smooth'.

To find a peak that falls between discrete values, a smooth output is required. Otherwise, the discovered peak will always fall on an original data point (in interpolation terminology, these original data points are called 'nodes' or 'knots').

Quadratic interpolation is the simplest kind of polynomial interpolation that fulfils this criterion. A quadratic interpolator was implemented, but found to be unsatisfactory. This is principally because this second-order model does not possess the flexibility to program both start and end data points and start and end gradients, all of which can be derived from the input data, whereas high-order polynomials do.

A cubic method is adopted in this algorithm. Cubic coefficients are relatively easy to derive from input data, and cubic equations may be manipulated and solved rapidly, while higher-order polynomials are harder to solve. Sinusoids with different frequencies and phases were used to test the algorithm. It located the peaks of these signals with a time-domain error of less than one thousandth of a sample.

To derive this equation, we will assume that a peak is to be found in the discrete-time sequence $y(n)$, $n = 0 \ldots N{-}1$. First-order derivatives, $y'(n)$ and $y'(n{+}1)$, are calculated using two more neighbouring samples:

$$y'(n) = \frac{1}{2}\big(y(n+1) - y(n-1)\big) \tag{4.3a}$$

$$y'(n+1) = \frac{1}{2}\big(y(n+2) - y(n)\big) \tag{4.3b}$$

Thus, to ensure that there are no ranging errors, an extra point is calculated at each end of the correlogram. These points are disregarded when the maximum is being located.

To simplify the derivation, it is assumed that the value to be interpolated is in the continuous interval $0 \le n \le 1$, where the integer values correspond to sampling intervals. It is trivial to shift data into this range and back again. The coefficients for the cubic equation are then calculated as follows:

$$\begin{aligned} y(n) &= an^3 + bn^2 + cn + d \\ y'(n) &= 3an^2 + 2bn + c \end{aligned}$$

$$\begin{aligned} \Rightarrow \quad y(0) &= d\ ; \\ y'(0) &= c\ ; \\ y(1) &= a + b + y'(0) + y(0)\ ; \\ y'(1) &= 3a + 2b + y'(0) \end{aligned}$$

$$\begin{aligned} \Rightarrow \quad a &= 2\big(y(0) - y(1)\big) + y'(0) + y'(1)\ ; \\ b &= 3\big(y(1) - y(0)\big) - 2y'(0) - y'(1) \end{aligned} \tag{4.4}$$

This equation is identical to the cubic convolution interpolator derived by Keys [Keys 1981]. The algorithm is reviewed alongside other members of the cubic family, and many other kinds of interpolators, in Lehmann's review [Lehmann 1999] — the $\alpha = -½$ instance of cubic interpolator is the one derived here. Lehmann's analysis focuses on the interpolation kernel. This is the effective impulse response of the interpolator. It can be shown that the kernel function for the equations above is:

$$y(n) = \begin{cases} \frac{3}{2}n^3 - \frac{5}{2}n^2 + 1 & |n| < 1 \\ -\frac{1}{2}n^3 + \frac{5}{2}n^2 - 4n + 2 & 1 \leq |n| < 2 \\ 0 & |n| > 2 \end{cases} \tag{4.5}$$

This function has the time- and frequency- domain characteristics shown in Figure 4.12.

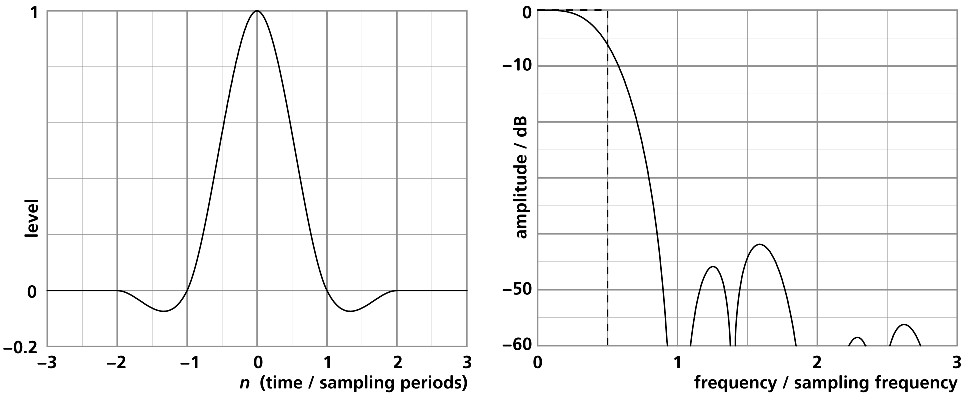


Figure 4.12. Time- and frequency- domain characteristics of the interpolation kernel. The dashed line on the frequency-domain plot shows the magnitude response of the ideal interpolator — a sinc function low-pass filter.

Keys also derives another class of cubic interpolator that requires six data points. In theory, this is more accurate than the four-point technique defined above. However, a six-point cubic interpolator would require the running cross-correlation to be extended by two data points. This would slow the computing of the correlogram by approximately 10%. Maeland [Maeland 1988] demonstrates that the four-point cubic interpolator is also less accurate than the cubic *B*-spline, which has become a very popular interpolation technique. However, Maeland concludes that because the cubic *B*-spline takes so much longer to calculate, and the algorithms are so different, 'no [direct] comparison of the cubic *B*-spline with other interpolation kernels should be done' [Maeland 1988: 216].

The greatest advantage of the cubic interpolator in this instance, however, is that a peak-finding formula is easily derived from the interpolating equation by equating the first-order derivative $y'(n)$ to zero, and solving the resulting quadratic within the interval $0 \le n \le 1$. (In this case, the value of interest is n rather than $y(n)$. Therefore the technique is called reverse interpolation.)

There is one important exception where this formula may not be used. The principal maximum occasionally appears just outside the range of a correlogram. When this happens, $y'(0)$ and $y'(1)$ are of the same polarity, because there is no turning point in the test interval. It is necessary to check for this case explicitly, because it renders the quadratic equation insoluble. Therefore, when $y'(0)$ and $y'(1)$ are of the same polarity, the peak-finding subroutine bypasses the quadratic equation solver, and returns with $n = 0$ or $n = 1$, according to whichever value of $y(n)$ is higher.

When finding the initial maximum, the correlogram is biased in favour of central values. This is achieved by multiplying by a linear ramp function that is 1 at the centre and 0.95 at the edges. A similar practice is employed by Stern and Colburn [1978]: it forces the peak-detecting algorithm to favour central ITDs when faced with highly periodic signals. The periodicity is then prevented from causing peaks at implausible ITDs.

Using this method to process a 19-point correlogram at 14.7kHz, the number of data points that need to be handled becomes far more manageable:

$$\begin{aligned}&14700\text{ Hz} \times 24\text{ bands} \times 19\text{ correlogram points}\\&\quad \approx 6.7\text{ million points per second}\end{aligned}$$

This requires less than 2.5% of the computation demanded by the conceptual model in Figure 4.10.

Windowing method for extracting ITDs

Figure 4.13 shows the flow of binaural data through the spatial feature extractor, and the way in which the input sampling frequency of 44.1kHz is eventually converted to a slower data rate of 2.45kHz.

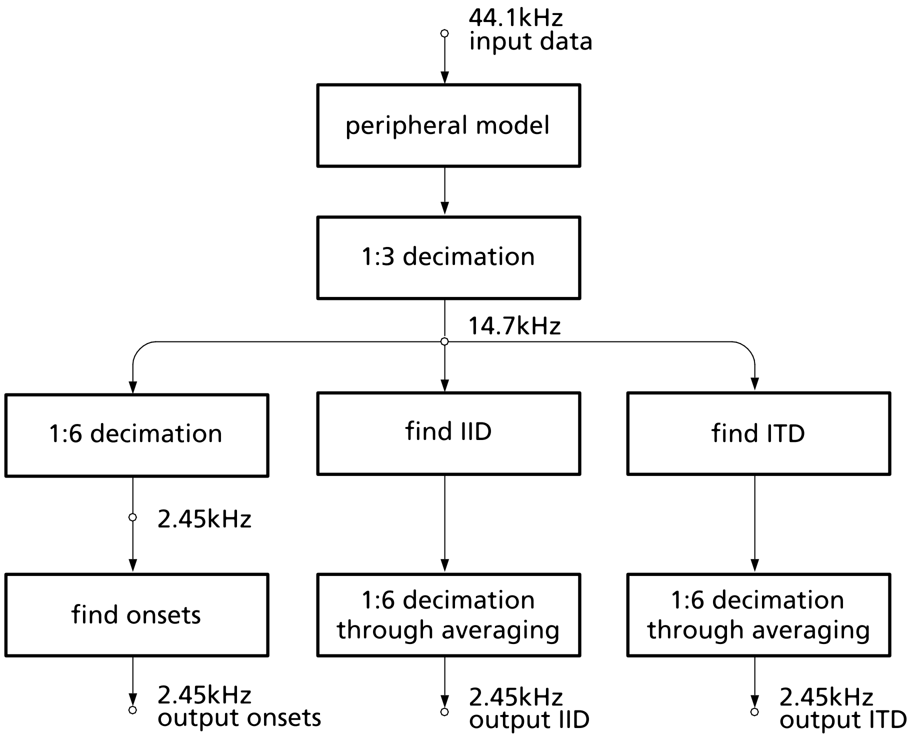


Figure 4.13. Internal sampling frequencies used within the algorithms.

To arrive at the output sampling frequency, the input data must be reduced sixfold. This is achieved using windowed averaging of adjacent correlograms. Since the correlogram's absolute level is of no consequence when finding its peak, the sum is used in place of an average. Summed correlograms emerge at a rate of 2.45kHz, and ITDs are extracted only from these. The windowed averaging algorithm works as follows:

$$\begin{aligned} X(\tau, n) &= \sum_{m=0}^{13} \chi(\tau, 6n+m) \\ &= \sum_{m=0}^{13} \sum_{\tau=-9}^{\tau=9} l(6n+m+\tau)\, r(6n+m-\tau) \end{aligned} \tag{4.6}$$

where $\chi(\tau,n)$ is the component correlogram centred around sample n. The factor of 6 comes from the 1:6 decimation built into this formula: samples are dropped so that X is a sixth of the length of l and r. It can be seen that sum correlograms overlap, sharing most of their data with their neighbours. The minimum window length is 14 samples, which is approximately 950µs. This value is chosen for its closeness to the spatial integration time of the human hearing system, which is approximately one millisecond [Wallach et al. 1949]. However, there is no reason to presume that the square window featured here models the human auditory system optimally.

In lower frequency bands, the number of samples that constitute each sum correlogram is increased to one full period of the centre frequency of the band. For example, in the lowest band, which has a centre frequency of 60Hz, the window length is 14700/60 = 245 samples. This length may seem cumbersome, but if it is not used, the correlogram output will be strongly influenced by the instantaneous phase of the input signal [Ifeachor and Jervis 1993]. As the algorithm is designed not to make use of forthcoming audio data (see Section 3.3), window lengths beyond 14 samples are extended into the past.

Owing to the amount of input data shared by adjacent sum correlograms, a basic implementation of this algorithm would perform the same series of Jeffress computations many times. This would be a particular problem in lower frequency bands. If data is computed afresh for every sum correlogram, about 97% of the correlogram calculations for the lowest frequency band — 239 out of 245 — will be identical in the next iteration. Before the averaging window completes its pass, each component correlogram, $\chi(\tau,n)$, will have been calculated 40 or 41 times.

Without much effort, a far more efficient routine can be implemented, where each component correlogram is calculated exactly twice. The previous sum correlogram is used as a starting point for calculating the next sum correlogram. Its bottom six component correlograms are re-calculated and subtracted, and six new correlograms are added on. This simple optimisation reduces the processing load of this part of the algorithm by 60%, and of the entire localisation algorithm by 45%.

In theory, further optimisation is possible by storing all the component correlograms so that they are calculated only once, but this requires storage of over 13,000 data points. In MATLAB [Mathworks 2004], even when plenty of RAM is available, the overhead caused by memory swapping and administration means that the algorithm takes 40% longer to run, even though there is a theoretical saving in computational demand (see Section 5.5).

Interaural time differences are extracted from sum correlograms using the cubic interpolation method described in the previous section, with the peak ITD rounded to $\tau/16$. This is effectively 16:1 interpolation of the correlogram, yielding a temporal accuracy of 8.5μs.

Method for extracting IIDs

The IID extraction technique uses a square-windowed, time-corrected comparison of sum-squares of the left- and right-ear signals of each band (see Figure 4.6c). The same length of window is used in both time and intensity difference calculations — whichever is the greater of fourteen samples, or a period of the band's centre frequency. However, in IID calculations, left- and right-ear windows are advanced or retarded symmetrically to offset the interaural time difference. This ensures that readings are taken at similar phases in both signals. An example of this method can be seen in Figure 4.6.

The output from this comparison is based on the judgment decision favoured by Sayers and Cherry [1957] (see Section 4.3.2). The difference of the two ear readings divided by their sum is used. This always lies in the range [–1 1]:

$$d = \frac{e_l - e_r}{e_l + e_r} \tag{4.7}$$

where e_l and e_r are the left- and right-ear signal energies. Energy is calculated by squaring and then integrating the signal. To improve contrast when the two signals are similar in level, the square of the decision variable is mapped to a look-up table location, and thus to a lateral angle. It must be remembered that d^2 is unipolar, so the sign of d must be determined before squaring to preserve the distinction between left-heavy and right-heavy signals.

Similar results may have been achieved using a low-pass filter followed by piecewise energy comparison: this would create a level-meter model of the kind implemented by Macpherson [1991] and recommended by Hartmann and Constan [2002]. The working of the model presented here may be seen as a level-meter model where the input windowing performs as the meter's input filter. Running the effective impulse response through a FFT produces a cascaded comb filter and a low-pass filter with a comb frequency of about 1kHz and a 3dB point of approximately 500Hz (Figure 4.14).

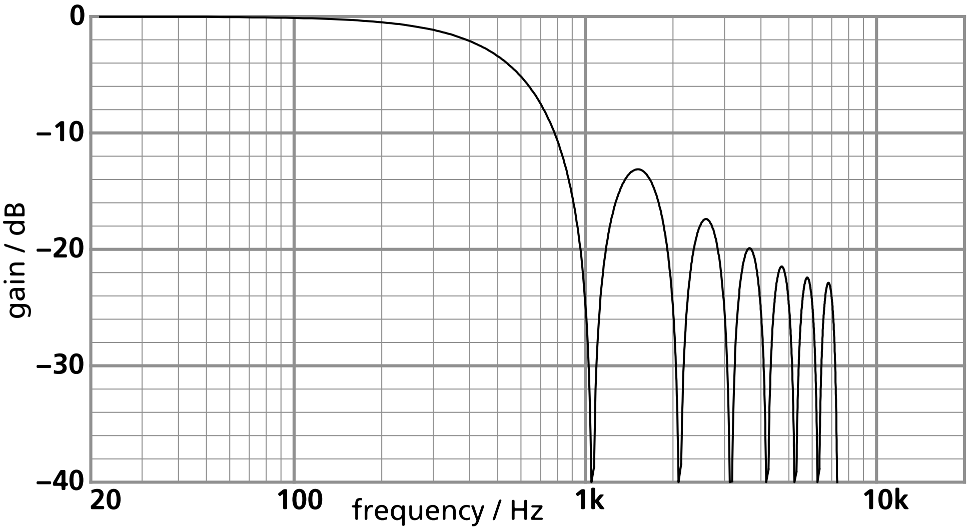


Figure 4.14. Frequency response of a filter that is equivalent to averaging 14 consecutive samples at a sampling frequency of 14.7kHz.

4.3.5 Mapping to histograms

Each interaural time and intensity difference can be mapped via a look-up table to a lateral angle histogram. This is a function of truth value against lateral angle.

This process has an antecedent in Huang’s localisation algorithm [Huang et al. 1997], and a system based on 5°-resolution, one-one look-up tables is used by Palomäki et al. [2004]. Huang’s method considers only ITD, but the argument for the histographic approach is even stronger when IIDs are taken into account because the mapping of IID to lateral angle becomes increasingly complex above about 800Hz, and ceases to be one-one. There are two other motivating reasons for the histographic approach over simple ITD-to-angle or IID-to-angle mapping. Firstly, there is usually some imprecision in interaural difference calculations, and it is better to deal with this imprecision than to ignore it. Secondly, different dummy heads have slightly different acoustical characteristics: even the same head will produce different cues as the source distance is varied. Therefore there will be some uncertainty in mapping from an extracted ITD or IID to an angle of incidence.

To ensure fast execution, it is necessary to store all the possible data tables. However, the price of the convenience of a ready-generated set of tables is a considerable amount of memory. Two different sets of tables must be generated and stored: one set for ITD information, and one for IID information. Each of these sets contains 24 tables: one for each critical band. Every one of these tables is a two-dimensional array of data locations. One dimension is lateral angle. Every angle is represented from –90° to 90° in 1° increments, so there are 181 indices in this dimension. (In practice, owing to the symmetry of the head, only one side of this data needs to be represented, so there are only 91 indices.) The second dimension is either ITD in 8.5µs increments, with 120 indices, or IID in an internal format, with 51 indices. Each data location contains a truth value. In total, each ITD table contains 91×120 = 10920 data entries, and each IID table contains 91×51 = 4641 data entries. Since there is one table per critical band, the two table sets require 373 464 entries. These memory requirements are shown graphically in Figure 4.15.

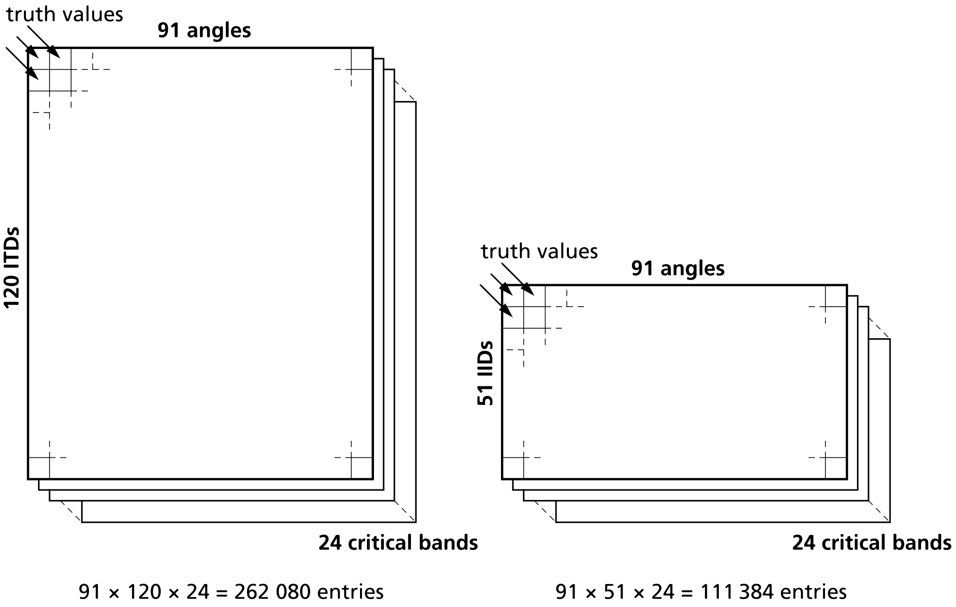


Figure 4.15. Graphical representation of the ITD (left) and IID (right) look-up table sets, showing total memory requirements.

4.3.6 Duplex theory weighting

At low frequencies, the human auditory system weights ITDs more than IIDs. At high frequencies, IIDs dominate over ITDs. This phenomenon has come to be known as the *duplex theory*. Subsequent investigation has refined the duplex theory. The crossover between ITD dominance and IID dominance occurs when the wavelength of sound becomes comparable to the dimensions of a listener's head. When dealing with the duplex theory, though, it is important that it is not over-simplified, because it can be demonstrated that both cues retain some salience across the frequency spectrum.

It is often hypothesised that IIDs are assigned a low priority at low frequencies because human listeners naturally prioritise those cues that furnish information which is clear and unambiguous. This teleological viewpoint is defended, for example, by Hafter and Carrier [1972]. IIDs are very small at low frequencies, owing to diffraction around the head. It can be shown from the KEMAR HRTF data that the maximum diffuse-field IID is approximately 2.3dB at 60Hz. The directional cues provided by low-frequency IIDs will often be poor because small differences will need to be extracted with a high precision for the data to be useful, and this is difficult. The influence of source proximity on IID will also be proportionally higher. This makes IID cues at low frequencies unreliable.

At high frequencies, the signal level fluctuations caused by head shadowing become far greater, and the IID becomes a higher-priority cue. Conversely, ITD-based cues change very little with frequency. However, if the stimulus is highly periodic, high-frequency ITDs are ambiguous because interaural time differences may plausibly extend over several periods, and many head angles may be plausible for any one particular interaural phase difference. This is not a problem at frequencies below about 800Hz, because all plausible ITDs at these frequencies will fall within one half-period of the waveform.

In the past, *trading experiments* were often used to investigate the interplay between ITD and IID at different frequencies. These experiments set the two cues in conflict in order to determine the amount of one that is required to offset an amount of the other. The result is a trading ratio that is usually expressed in µs/dB. For certain reasons, this approach is now generally regarded as obsolete. Firstly, the stimuli used in trading experiments deny naturally-occurring physical relationships between the ITD and IID cues. Very frequently, subjects perceive two distinct sources: one is chiefly time-panned and the other is amplitude-panned. This causes a wide distribution of results in localisation experiments, and listening subjects will, if asked, report the stimuli to sound ‘unnatural’ [Gaik 1993]. With some training, subjects can become accustomed to opposed localisation cues, but distribution of results is still fairly wide when ITD and IID cues are placed in conflict [Hafter and Jeffress 1968]. Thus the ability of trading experiments to provide information about the relative weighting of IID and ITD is uncertain at best.

Hafter and Carrier [1972: 1853] state another issue concerning trading experiments: ‘Implicit in the use of the binaural trading ratio is the assumption that functionally identical values of [ITD and IID] are in some way identical in their neural representation. Otherwise the ratio itself is nothing more than an interesting oddity, bearing no information about physiological processes.’ However, any project that sets out to unite ITD and IID cues in a localisation paradigm is compelled to make this assumption at some point.

A final interesting slant on trading-ratio experiments is provided by Buell et al. [1994], in which narrow-band noise, centred around 500Hz and with an ITD of 1.5ms (which cannot occur naturally), is subjected to manipulation. Using five listening subjects, they discovered that the influence of IID on lateral angle is dependent on the interaural phase difference and bandwidth of the test stimulus. This questions the methodology of any experiment that assumes a simple trade of ITD against IID.

Macpherson and Middlebrooks [2002] investigate the duplex theory using a number of methods that are largely devoid of the problems associated with trading experiments. They conducted a series of listening experiments in which either the ITD or IID was adjusted around a natural combination. The resulting image movement was elicited from thirteen test subjects, and each movement was converted into a dimensionless *bias weight*. Bias weight is the ratio of the amount by which the ITD or IID of a natural sound source would change as it was moved by the reported amount, to the actual amount that was applied. An IID bias weight near zero at a certain frequency would thus indicate that the IID cue is relatively weak, as large amounts of artifically-imposed IID would be needed to move the image by a small amount. Conversely, an IID bias weight of one would mean that it is such a strong cue that the ITD cue is ignored in localisation.

Macpherson and Middlebrooks calculated ITD and IID bias weights at three different frequencies for thirteen subjects, using filtered noise-based stimuli. This is helpful, because when sinusoids or amplitude-modulated tones are used in duplex theory investigations, spatial perception is often controlled overwhelmingly by either the fine detail or signal envelope. This is not a natural listening condition. Noise, containing a natural combination of fine detail and signal envelope cues, would appear to be more directly comparable to natural stimuli.

The findings of Macpherson and Middlebrooks can be summarised as follows:

- In common with trading experiments, considerable inter-subject variation was experienced. Nevertheless, significant trends could be extracted from the data.

- The calculated average bias weights for 2kHz low-pass filtered noise were 0.93 for ITD, and 0.20 for ILD [IID].

- The calculated average bias weights for 4kHz high-pass filtered noise were 0.11 for ITD, and 0.96 for ILD [IID].

These bias weights sum to more than unity, suggesting that the movement experienced when both cues are manipulated naturally is smaller than the sum of the movement of its component ITD and IID manipulations. This may indicate a partial collapse of binaural fusion when unnatural movements are attempted, creating the familiar two-image perception observed by Whitworth and Jeffress [1961], Hafter and Jeffress [1968], Gaik [1993], and many other researchers.

Unfortunately, there is a very limited body of research that deals with the weighting of ITD and IID without assuming simple trading. Furthermore, the majority of trading-ratio experiments focus on a small number of stimuli, and avoid the problematic middle frequency range where dominance is exchanged between IID and ITD.

Using a series of observations and assumptions, a cross-weighting law can be tentatively advanced that is based on the bias weights found by Macpherson and Middlebrooks. The derivation of this law follows.

Calculation of duplex theory weighting coefficients

The cross-weighting law employs two cut-off points, specified by two critical band numbers. Below and including the first band, b_l, the ITD weighting coefficient, $w_{ITD}(b)$, is greater than the IID weighting coefficient, $w_{IID}(b)$. Above and including the second band, b_h, $w_{IID}(b)$ is greater than $w_{ITD}(b)$. In the range between b_l and b_h, called the *crossover range*, the weighting coefficients change monotonically between their low- and high-frequency values. For reasons that will become clear, the sum of $w_{IID}(b)$ and $w_{ITD}(b)$ will not necessarily be unity. b_l is set to 8 (centre frequency 845Hz), and b_h to 14 (centre frequency 2160Hz) — this agrees with widely-observed behaviour.

Wightman and Kistler [1992] advance the suggestion that when a stimulus is presented with little low-frequency content, its localisation is dominated by IID. When low-frequency content is present (specifically below 2kHz), localisation is controlled by ITD. Since the weighting coefficients in this model are not adjusted when the stimulus changes, this behaviour must be encoded by ensuring that the sum of values of $w_{ITD}(b)$ below band b_h is sufficiently greater than the sum of values of $w_{IID}(b)$ above and including it. This will be termed the *ITD dominance condition.*

The position of the centroid of the weighted sum of ITD and IID histograms is equal to the weighted average of the centroids of the two histograms. Thus, if the centroid is allowed to act as an indicator for source location in these experiments, the bias weights may be used directly. Because these bias weights are dimensionless, they will be approximately the same whether they are calculated from ITD and IID data or from the lateral angle.

A first approximation would be to use the unprocessed bias weights below b_l and above b_h, and to ramp linearly between the values in the crossover range. However, this violates the ITD dominance condition, because the sum of $w_{ITD}(b)$ below b_h would be 9.52, and the sum of $w_{IID}(b)$ above and including band 14 would be 0.96 × 11 = 10.56. This is remedied by multiplying $w_{ITD}(b)$ by a linear ramp function below b_h. An empirical law is used to generate this function: it is calculated so that the sum of values of $w_{ITD}(b)$ below b_h is 3dB greater than the sum of values of $w_{IID}(b)$ above it. The final formulae are shown in Equations 4.8 and 4.9:

$$w_{ITD}(b) = \begin{cases} 1.597 - 0.047b & : \quad b \leq 8 \\ 0.0102b^2 - 0.437b + 4.06 & : \quad 8 < b < 14 \\ 0.11 & : \quad b \geq 14 \end{cases} \tag{4.8}$$

$$w_{IID}(b) = \begin{cases} 0.20 & : \quad b \leq 8 \\ 0.152b - 1.016 & : \quad 8 < b < 14 \\ 0.96 & : \quad b \geq 14 \end{cases} \tag{4.9}$$

The two functions are depicted in Figure 4.16.

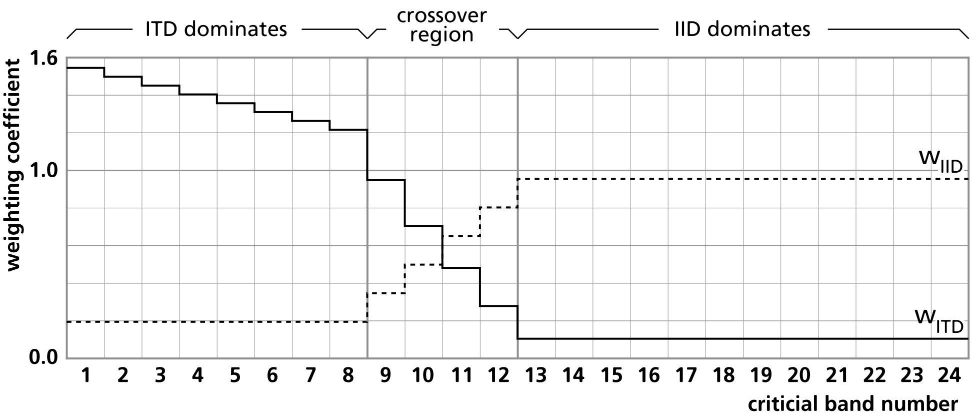


Figure 4.16. Cross-weighting coefficients for ITD and IID histograms.

4.3.7 Loudness weighting

Loudness weighting is applied to the histograms before they are combined arithmetically. Critical bands in which the signal level is high are weighted over those in which there is little active content. The localisation data from relatively inactive bands is often dominated by low-level noise, and this noise is thereby prevented from degrading the final histogram. The algorithm thus serves as a simple, static masking model.

Some simplifications of loudness perception theory have been necessary to enable simple, fast calculation of weighing coefficients. These simplifications are as follows:

- It is assumed that the louder a frequency band, the greater its relative influence on the perception of auditory space. It is then assumed that the relationship between loudness and spatial weighting is linear. Although these assumptions make sense intuitively, they have not been tested formally.

- Dynamic changes that affect loudness perception, such as pre-masking, are not included in the algorithm.

- To approximate the loudness of a single critical band, it is assumed that the content of this band can be represented by an amplitude-modulated sinusoid at this band's centre frequency. The instantaneous amplitude of this sinusoid equals the instantaneous peak signal level.

The loudness weighting model is based on the findings of Marks [1978], which extend Stevens's model [Stevens 1957] to binaural listening conditions. Stevens found that a power law connects perceived loudness with sound pressure, such that 10dB increase in SPL corresponds roughly to a doubling of loudness. Marks finds that perceived binaural loudness can be modelled simply by adding the sound pressures at the left and right ears. The extension into this application is tentative, because Marks tests the hypothesis only with pure tones.

As a basis for the loudness algorithm, Table B.2 of the International Standard equal-loudness contours is referenced [BS ISO 226:2003]. This table maps sound pressure level in dB to loudness in phons. Some alterations have been made to the standard table:

- A 0 phon floor and a 100 phon ceiling have been imposed on the data;
- Critical band 24, which has a centre frequency of 13 750Hz, falls outside the domain defined by the standard. The standard values for 12 500Hz have therefore been substituted;
- The table has been interpolated in the frequency domain using cubic splines, and re-sampled at the critical band centre frequencies.

The re-sampled look-up table is shown in Figure 4.17.

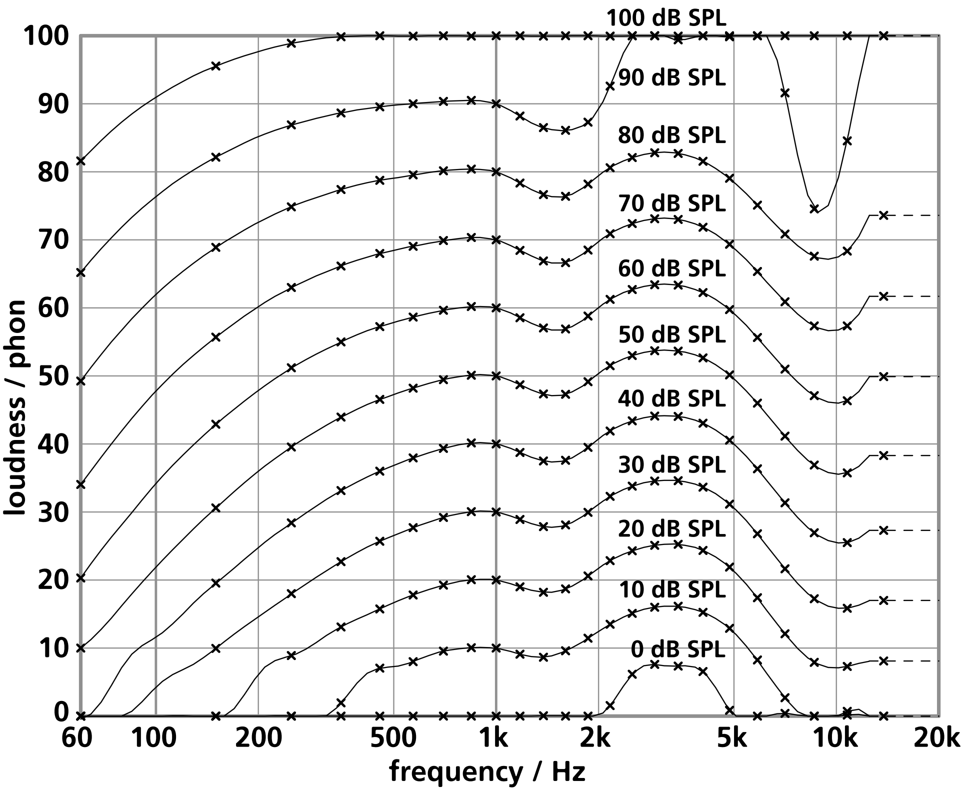


Figure 4.17. Interpolated BS ISO 226:2003 equal-loudness contours, used in loudness weighting. Crosses show the centre-frequency data points used within the algorithm.

The interpolated look-up table has eleven entries for each critical band. These map input sound pressure levels (SPL) to loudness levels in phons. Data points range from 0dB SPL to 100dB SPL in 10dB steps. An arbitrary SPL is converted to phons using inverse-distance weighing for the two nearest neighbours in the table:

$$L(s) = \frac{(s_+ - s)L(s_-) + (s - s_-)L(s_+)}{10} \tag{4.10}$$

In this equation, s_- and s_+ are the lower and higher nearest multiples of 10 to the input SPL, s. $L(s)$ is the corresponding phon level. $L(s_-)$ and $L(s_+)$ can be read directly from the look-up table.

The SPL is calculated by finding the maximum absolute signal level obtained from the sum of the offset windowed left and right signals. The resulting level is converted to dBFS, and 90dB is added to produce a reasonably-ranged value for the input SPL.

Based on an approximation of the Stevens model, the level in phons is converted to a weighting coefficient so that a 10 phon reduction in level reduces the weighting coefficient by a factor of two. An equation that converts the maximum signal level in each ear into such a weighting coefficient is:

$$w_L = 2^{L/10} \tag{4.11}$$

One further rule is applied: if L is zero, the frequency band will effectively be too quiet to be audible, so the weighting coefficient, w_L, will be forced to zero.

Two further effects have not been incorporated into the loudness weighting algorithm, but they may make worthwhile extensions in the future. Firstly, when the stimulus level is low, localisation blur is seen to increase [Blauert 1997: 155]. Martin [1995b] therefore recommends adding low-level noise to the input signals before lateral angle calculations are performed, so that localisation blur is increased when the are no critical bands with loud content. Spatial impression has also been shown to increase linearly at higher listening levels [Barron and Marshall 1981: 230]. This, however, may be an epiphenomenon of a more familiar process: the time for which the reverberant decay tail of a sound stays above the threshold of perception will be greater for a loud source than a quiet source.

4.4 The generative algorithms

Both table sets, ITD and IID, are generated using the KEMAR HRTF data of Gardner and Martin [Gardner and Martin 1994] as raw data. This data set was chosen because it is well-documented and freely available.

To convert the KEMAR data set into a usable set of histogram look-up tables requires several stages of processing. The ITD and IID data preparation processes are fairly similar. The KEMAR data contains complete sets of monophonic head-related impulse responses (HRIRs) for two different sizes of pinna. The characteristics of each data set are nearly identical at low frequencies, but differences are more pronounced at higher frequencies where the audio wavelength is comparable to, or smaller than, the size of the outer ear.

The ITD and IID look-up tables are saved on disk, so that each generative algorithm needs to be run only once. Therefore, unlike the analytical algorithm, a generative algorithm does not need to perform its task efficiently.

In order for data from both sets to be considered, the look-up table generation task is divided between a hierarchy of two routines. The subordinate routine converts the input HRIRs of the small- and large-pinna sets into tables of interaural time differences against angle of incidence. The main routine uses the resulting data to populate the set of look-up tables. Flowcharts for the ITD look-up table generating routines are shown in Figures 4.18 and 4.19.

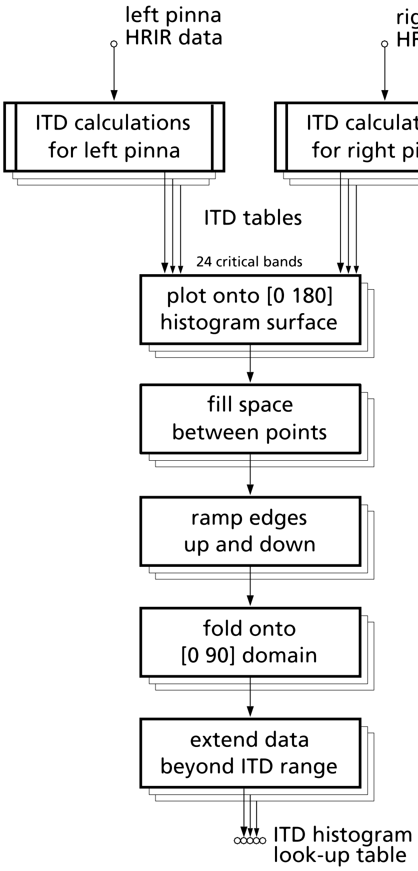

left pinna
HRIR data
right pinna
HRIR data
ITD calculations for left pinna
ITD calculations for right pinna
ITD tables
24 critical bands
plot onto [0 180] histogram surface
fill space between points
ramp edges up and down
fold onto [0 90] domain
extend data beyond ITD range
ITD histogram
look-up table

Figure 4.18. Generation of ITD histogram look-up table data. Main routine.

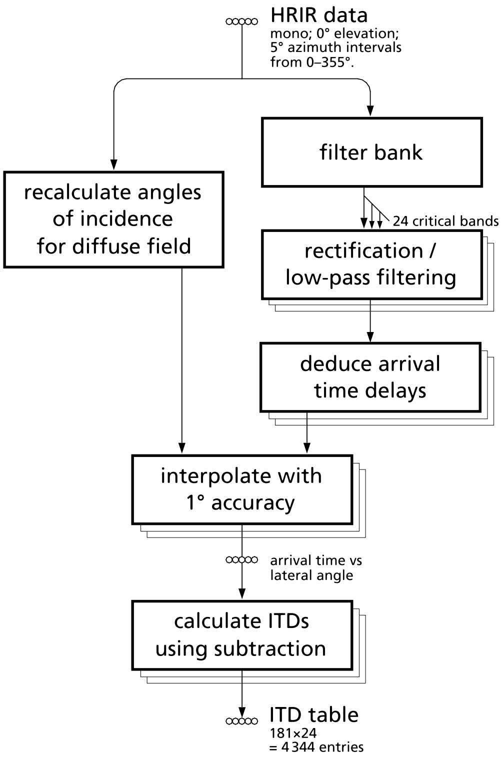

HRIR data
mono; 0° elevation;
5° azimuth intervals
from 0–355°.
filter bank
recalculate angles of incidence for diffuse field
24 critical bands
rectification / low-pass filtering
deduce arrival time delays
interpolate with 1° accuracy
arrival time vs lateral angle
calculate ITDs using subtraction
ITD table
181×24
= 4 344 entries

Figure 4.19. Generation of ITD histogram look-up table data. Sub-routine.

Each head-related impulse response is filtered through the same auditory peripheral model that is used throughout the system. The HRIR is first divided into 24 frequency bands, and to each of these bands full-wave rectification is applied, followed by a low-pass filter, 0 and decimation by 1:3 to 14.7kHz (Section 3.4.1 covers the filter bank and cochlear model; Section 4.3.4 justifies the decimation). Using the standard processing path ensures that each channel of the impulse response is subjected to the same internal delays that would be imposed on a test stimulus during signal analysis.

4.4.1 Recalculation of angles of incidence for the diffuse field

The Gardner and Martin HRTF data set was recorded 1.4 metres from the recording head in an anechoic chamber. This is too close to the head to make the data applicable to the diffuse field, or sources at a distance, without further processing. The easiest way to correct the data is to work out the equivalent diffuse-field angle for each near-field point, and then interpolate the data using the new set of angles. The concept behind this correction can be seen graphically in Figure 4.20, and the mathematics for this is as follows:

$$\theta' = \tan^{-1}\frac{\sin\theta - (r/d)}{\cos\theta} \tag{4.12}$$

where θ' is the equivalent diffuse-field angle of the near-field angle θ, r is the radius of the recording head (0.076 metres), and d is the recording distance (1.4 metres).

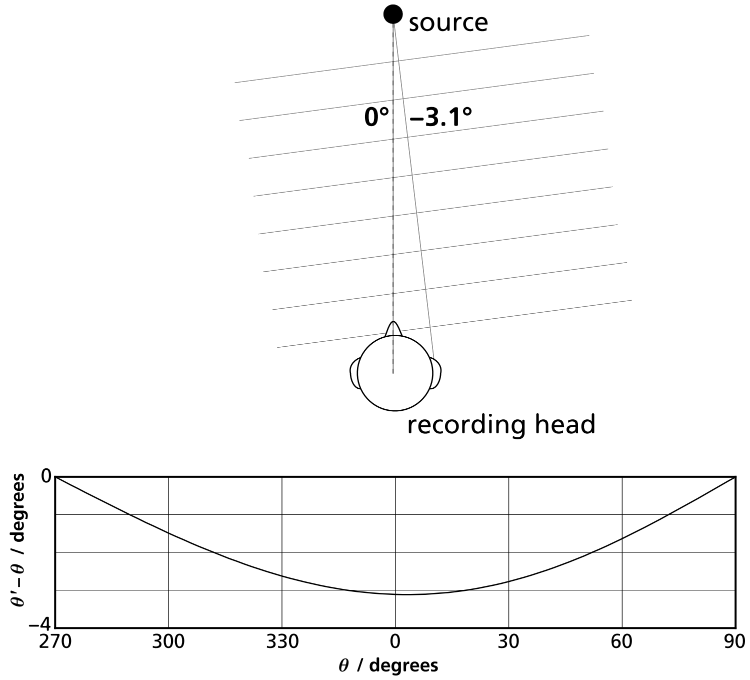


Figure 4.20. Diffuse-field correction.
Top: At 1.4 metres, a near-field source impinging on the head at 0° is equivalent to a plane wave (diffuse-field source) arriving at the right ear from 3.1° left.
Bottom: The extent by which diffuse-field corrected angle θ' differs from near-field angle θ across the horizontal plane. θ' is always biased towards the left when considering the right ear, and vice-versa.

4.4.2 Delay estimation

The original Gardner and Martin HRTF data is arranged in complementary pairs. However, once diffuse-field compensation is made to the data set, this is no longer the case. For example, the original data values 10° and 350° are symmetrically arranged, and could be used directly to form a stereo head-related impulse response. Using the diffuse-field compensation formula, however, these points become 6.9° and 347.0°, which are no longer symmetrical. In fact, only one pair of symmetrical readings exists after diffuse-field compensation (90°, 270°), so the interaural time difference cannot be computed directly using an interaural cross-correlation algorithm.

In this system, absolute arrival time is estimated by finding the point in each head-related impulse response where it reaches 5% of its maximum value. The look-up table requirements demand that this arrival time be found to an accuracy of one eighth of a sample, so 8:1 interpolation is applied to every impulse response using the cubic spline interpolator from MATLAB's signal processing toolbox [Mathworks 2004]. Reducing execution time is not important, because the look-up tables are not generated in run-time.

Once a set of arrival times has been calculated, the arrival-time data is interpolated using cubic splines, after wrapping the first and last data points to improve interpolation. (The advantage of wrapping the data is that problems regarding the treatment of end-points, for which the terminal gradient is usually unknown, can be avoided.) Output data is mapped to all integers in the domain [0° 359°]. Interaural time differences are then found by subtracting each arrival time from its symmetrical partner. This halves the data set, as ITDs from 1…180° are the negative of ITDs from 359…180°, and need not be calculated. Output data is in the domain [0° 180°].

4.4.3 Derivation of ITD histograms from input data

This section describes the remaining operations, shown in Figure 4.17, that are required to create the look-up tables. In each critical band, the ITD-versus-lateral angle data from both pinnae are transferred to a two-dimensional table. Initially, a truth value of one is placed wherever an ITD maps to a lateral angle, and all other locations are set to zero.

The data is now processed in individual ITD slices. A slice is extracted by fixing the critical band and the lateral angle, so that in each processing pass, the algorithm concentrates only on the variation of truth value against ITD. Every possible critical band and lateral angle is processed in such slices.

Initially, each slice will contain two truth values of one, and the other values will be zero, as shown in Figure 4.21a. One of the initial ones will correspond to the small-pinna data, and the other to large-pinna data. The first step is to fill the interval between these values with ones, as Figure 4.21b. This is because the look-up table sets are intended to be as generic as possible. Intervening values represent plausible ITDs for intervening pinna sizes.

The next stage extends the data beyond the upper and lower boundaries, as it is also plausible that some situations — either those where the sources are close, or those that use head and pinna metrics that differ from KEMAR — will create localisation cues in this range. The further that data falls outside the boundaries established by the KEMAR set, the less plausible it is. Therefore, a ramp is applied to these truth values so that they fall away with distance from the KEMAR boundaries. The width of the ramp is equal to the number of filled spaces (this includes the KEMAR boundaries). The ramp function equation is:

$$x = \begin{cases} \dfrac{n+1-s}{n+1} & : \quad s \leq n \\ 0 & : \quad s > n \end{cases} \tag{4.13}$$

where x is the truth value, s is the distance from the nearest KEMAR datum point, and n is the total number of values between and including these points. When the small-ear and large-ear ITDs are the same for a given lateral angle and critical band, the spaces above and below this boundary are thus assigned a truth value of 1/2. The effect of this function is shown in Figure 4.21c.

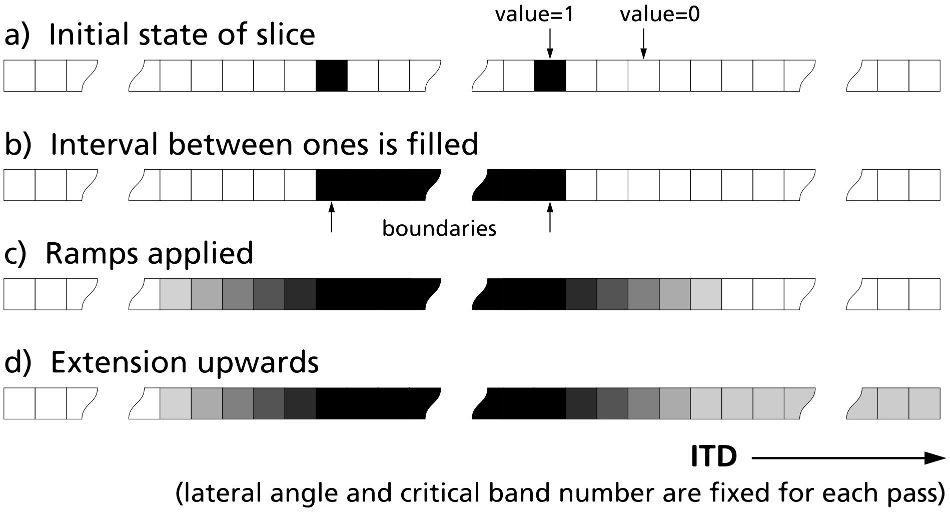

a) Initial state of slice
value=1
value=0
b) Interval between ones is filled
boundaries
c) Ramps applied
d) Extension upwards
ITD
(lateral angle and critical band number are fixed for each pass)

Figure 4.21. The effect of filling and ramping each slice of the ITD look-up table.

After all slices have been processed, the remaining operation is to convert the data from the [0° 180°] domain to the [0° 90°] domain. This is achieved by folding the data over, so that the rear hemisphere is reflected onto the front. Data over 90° is added to the corresponding data at 180°–θ, and all values are clipped to 1.

Finally, in imitation of the plausibility hypothesis of Hartmann [1997], some slices are filled in the way shown in Figure 4.21d, where the truth value above the upper boundary does not decay to less than 0.2. This happens only to the slices with the highest non-zero truth values in each table. The reason for this process is that if an abnormally large ITD is detected during analysis, the look-up table will be able to give a sensible result: it will localise the source over a range of lateral angles that have the highest ITDs, but the relatively low truth values will reflect the uncertainty of the prediction.

These operations can be seen in detail on band 16 (centre frequency 2925Hz) in Figure 4.22, and the complete ITD look-up table data set is reproduced over two pages in Figure 4.23.

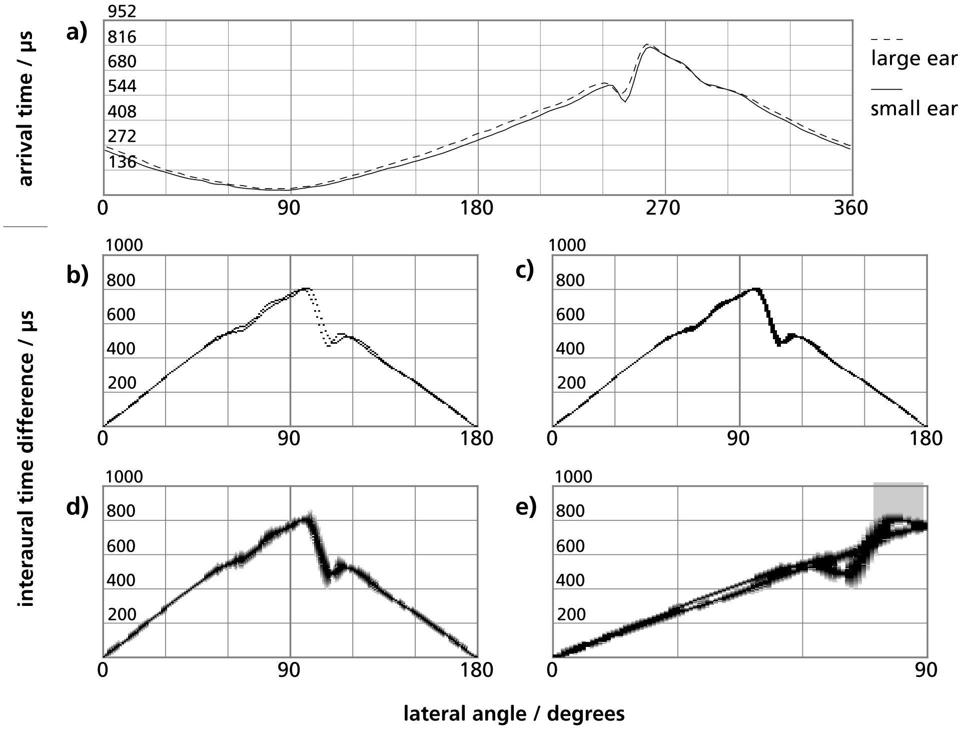


Figure 4.22. The generation of the ITD look-up table for critical band 16, centre frequency 2925Hz. The slices referred to in the text run vertically in these examples. a) Calculated and interpolated monaural arrival time versus diffuse-field lateral angle. b) ITD histogram plotted from a), unfilled. c) Filled. d) Filled with ramp function imposed. e) Filled, ramped, folded to [0° 90°], with the highest ITDs extended upwards.

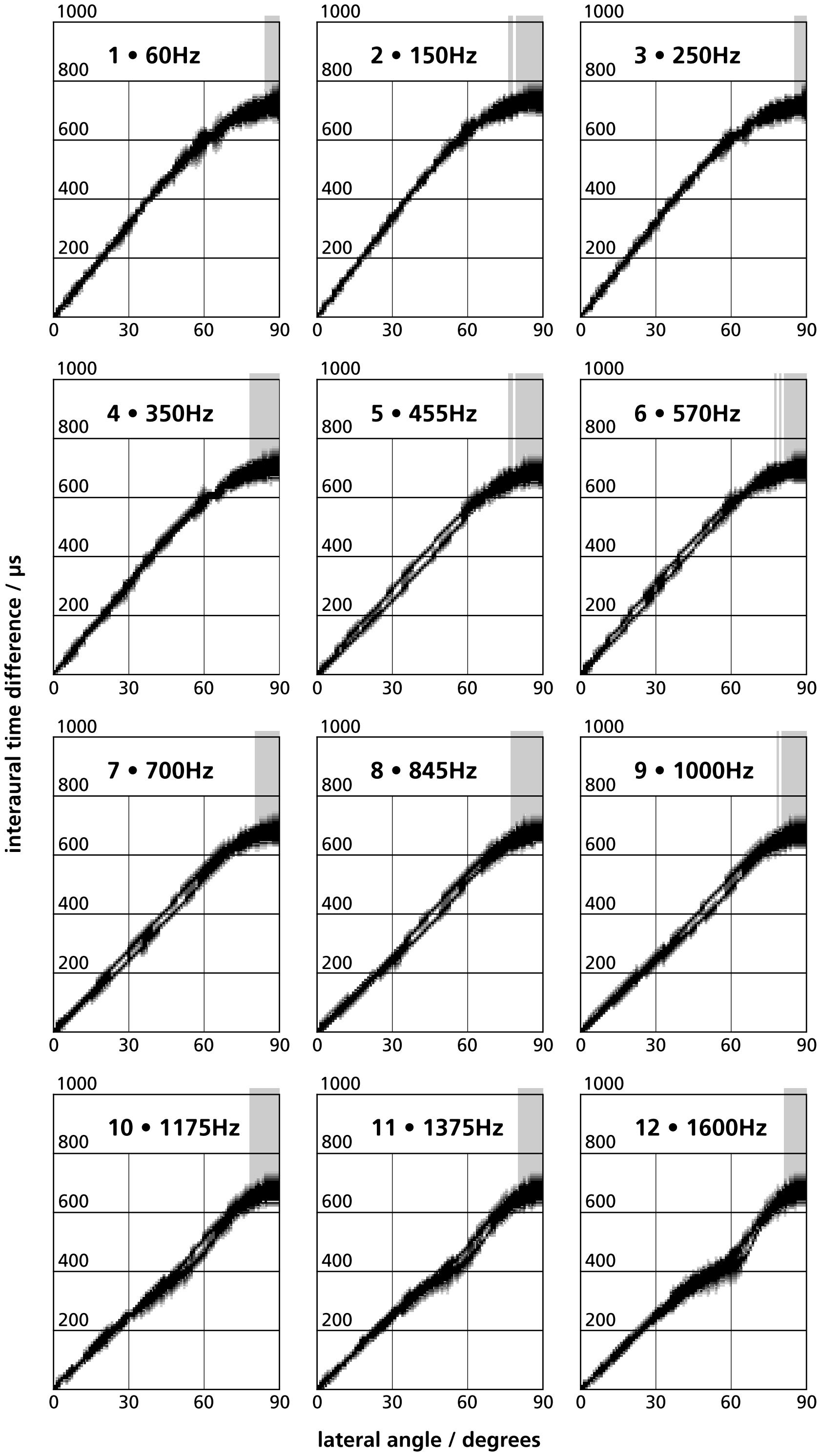


Figure 4.23a. ITD histogram look-up table data set. Critical bands 1–12.

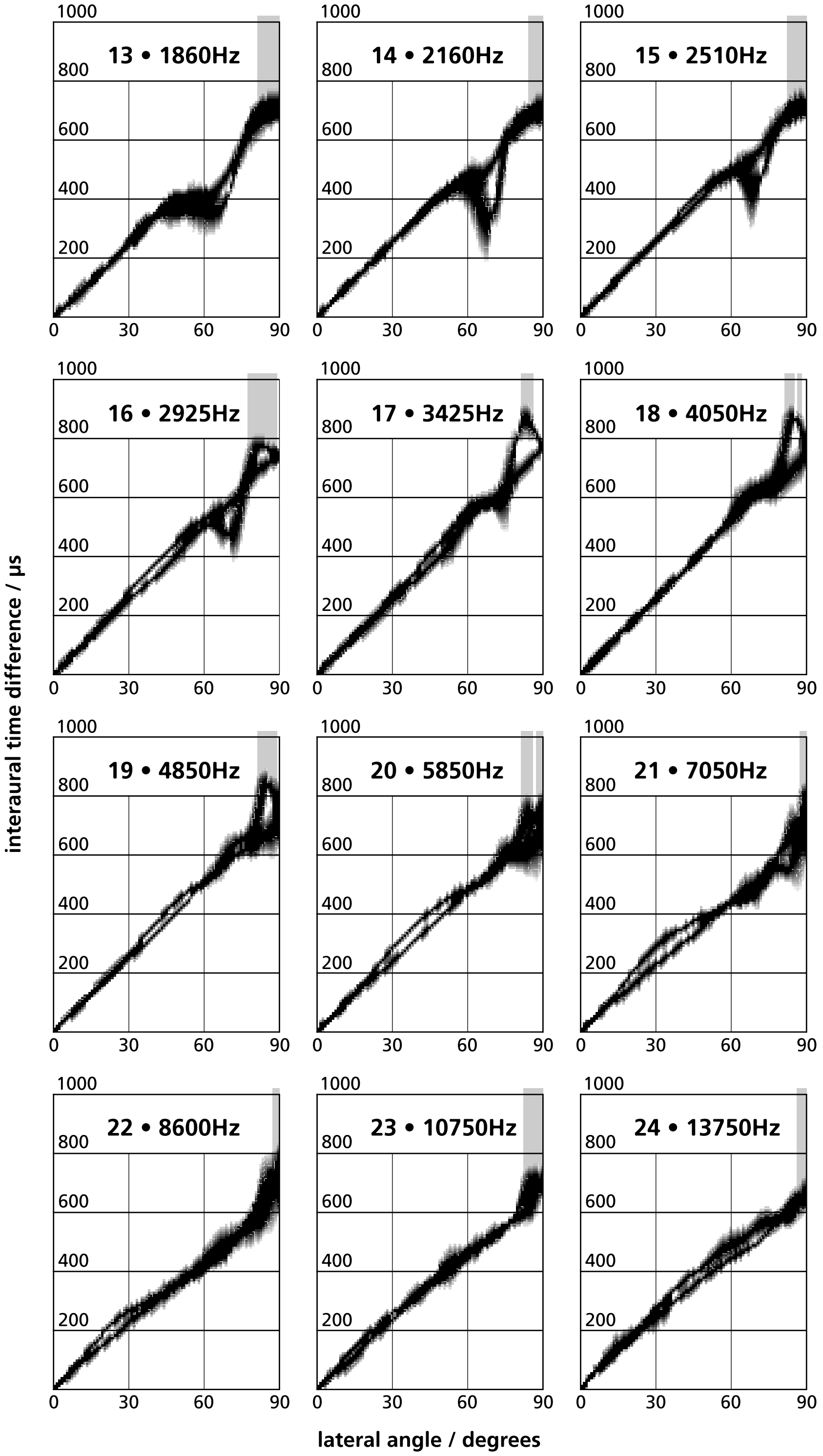


Figure 4.23b. ITD histogram look-up table data set. Critical bands 13–24.

4.4.4 Generating the IID look-up table

The routines for generating the IID look-up table are very similar to those behind the ITD. However, there is one important additional process that can be seen in the subordinate routine flowchart in Figure 4.24, in which IIDs are derived from the KEMAR HRTF data.

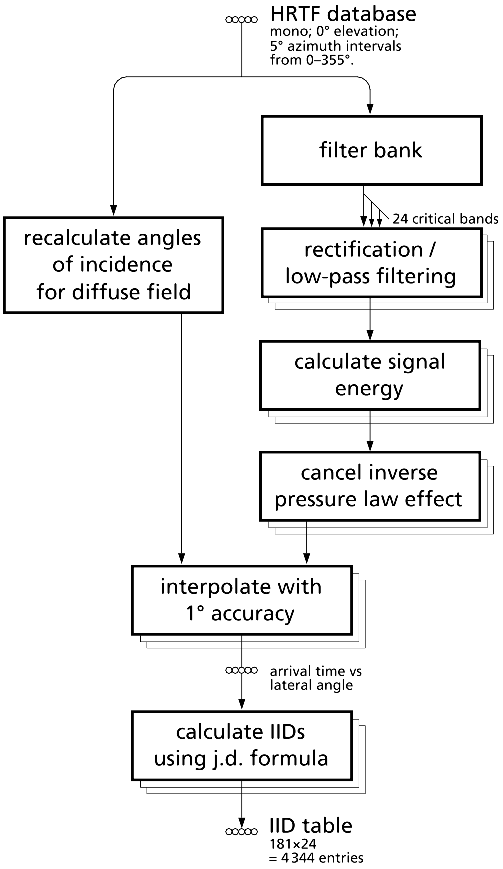

HRTF database
mono; 0° elevation;
5° azimuth intervals
from 0–355°.
filter bank
24 critical bands
recalculate angles of incidence for diffuse field
rectification / low-pass filtering
calculate signal energy
cancel inverse pressure law effect
interpolate with 1° accuracy
arrival time vs lateral angle
calculate IIDs using j.d. formula
IID table
181×24
= 4 344 entries

Figure 4.24. Generation of ITD histogram look-up table data. Sub-routine.

Owing to the proximity of the source to the dummy head, the inverse pressure law exerts some influence that would be absent in the true diffuse field case. This effect will be maximal when the source impinges from the direction of one ear, where the IID will be biased by approximately 1dB in favour of the nearer ear. Although the influence is relatively small at high frequencies, at low frequencies it dominates other phenomena and is therefore worth compensating. This is achieved for every energy calculation by dividing the sum-square of each HRTF by the square of the instantaneous distance of that ear from the source. IID calculations using pairs of values are then normalised for the diffuse field.

IIDs are calculated using the judgment decision formula of Equation 4.7. This function, in common with the ITD, is bipolar and symmetrical. If the left-ear and right-ear inputs are swapped, the result will be identical in magnitude to unswapped data, and of opposite sign. Therefore only the positive data needs to be stored. The square of the judgment decision is used as an index in the look-up table, rather than the judgment decision itself. This uses the tables more evenly, whilst preserving the [0 1] range of data. Of course, the sign of the judgment decision must be ascertained before squaring.

The routine that converts the IID data into the histogram data set is functionally identical to the ITD's main routine. It comprises the same filling, ramping, folding and extending processes. The IID histogram look-up table data set is shown, over two pages, in Figure 4.25. The differences between low- and high- frequency IIDs are particularly clear.

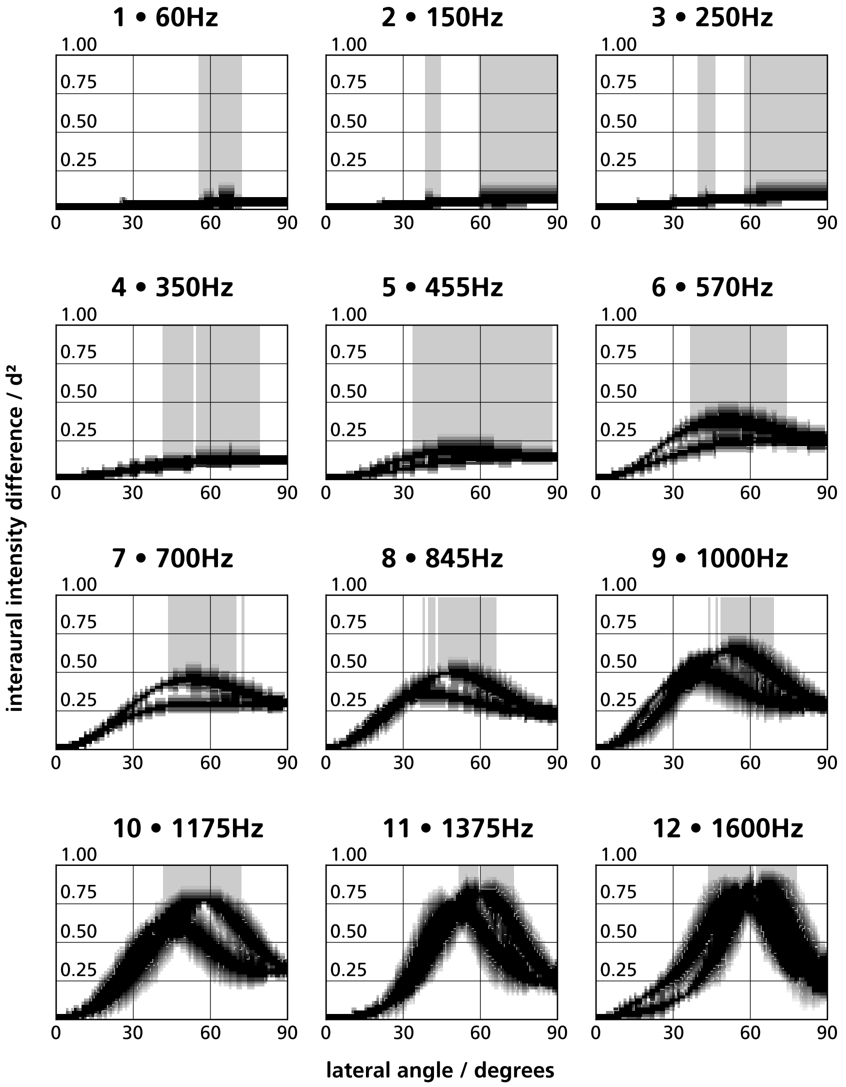


Figure 4.25a. IID histogram look-up table data set. Critical bands 1–12.

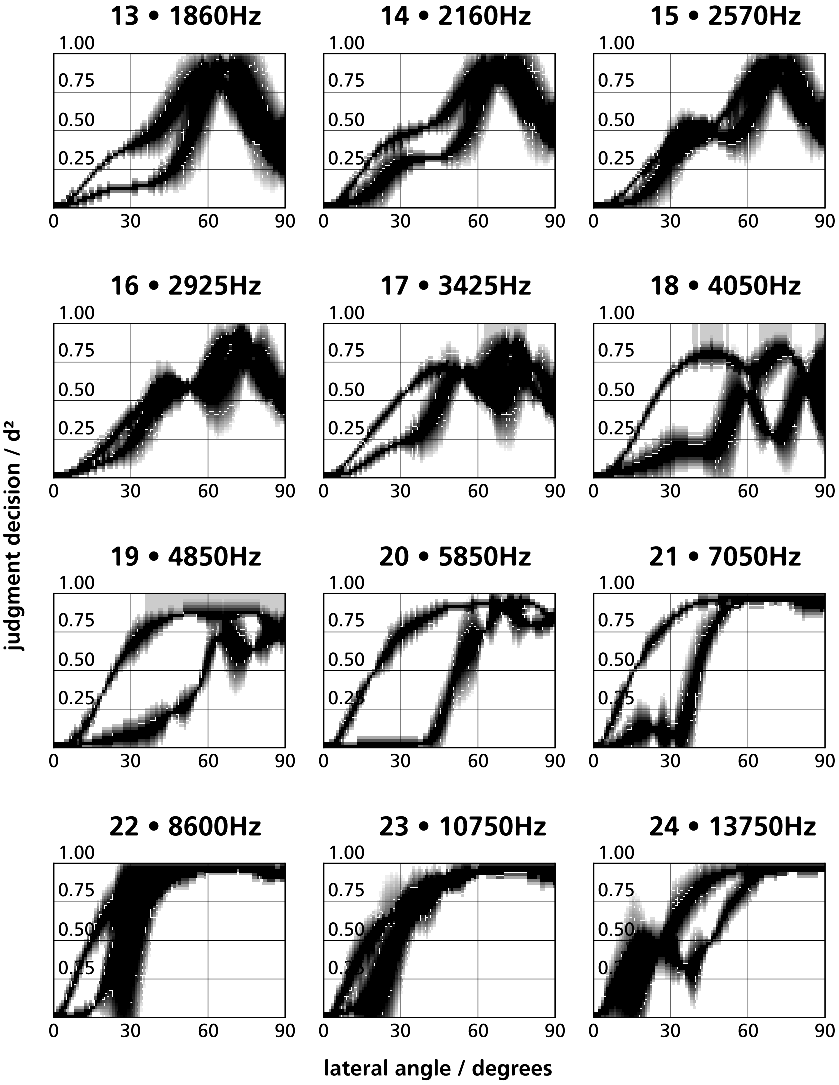


Figure 4.25b. IID histogram look-up table data set. Critical bands 13–24.

4.5 Summary

This chapter presented the localisation algorithm. This converts the left- and right-ear input signals of a binaural stream into a three-dimensional map comprising truth values against lateral angle and time. The inherent limitations of the binaural format mean that only one-dimensional localisation is attempted at 1° resolution, on an axis from −90° to +90°. This avoids ambiguities and front-back confusion errors, which would otherwise occur frequently. All sources are therefore assumed to be unelevated, frontal, and in the diffuse field. For most real or recorded sources, these approximations may be made without much detriment to the results.

The 181-point function relating truth value to lateral angle is referred to in this thesis as an output histogram. The localisation algorithm calculates histograms at 2.45kHz (approximately every 410µs), using sliding windows to extract interaural time and intensity differences (ITDs and IIDs) from the filtered and rectified critical band signals. Both cues play important roles in sound localisation. However, ITDs convey less ambiguous information than IIDs, may more easily be used to simulate cues than IIDs, and are also easier to extract precisely. Furthermore, the unconscious decisions made by a human subject in combining of ITD and IID data into a single percept appears to be very complicated. Hence, in spite of their importance, IIDs do not feature at all in many existing localisation models.

ITD and IID are calculated separately for each of the 24 critical bands used in this algorithm. ITD is found using interaural cross-correlation. As it is a computationally demanding algorithm, the cross-correlation process is optimised here by reducing the input sampling frequency as much as possible, and then using a special interpolation formula to locate the correlogram's peak to a very high temporal resolution. IID is extracted by integrating energy in ITD-corrected windows. Using look-up tables, the two types of interaural differences can be converted into component histograms that are then combined.

The formulation of look-up tables and weighting coefficients used to generate and combine the histograms is also described in this chapter. In total, there are three sets of look-up tables: one converts ITD values to lateral angle histograms, another converts IID values, and a third is used to find loudness weighting coefficients for each critical band.

The ITD and IID histogram look-up table data sets contain unique data for each critical band, which are derived from the KEMAR HRTF database of Gardner and Martin [1994]. Loudness weighting data, used to combine histograms across all critical bands, is based on BS ISO 226:2003. The loudness of each critical band is approximated as a amplitude-modulated sinusoid at the band's centre frequency. A reduction in level of 10 phons within any band halves its weighing coefficient. A further set of weighting coefficients are generated to unite ITD and IID data. Owing to the complexity of the relationship between ITD and IID, the method behind the generation of these coefficients is highly speculative. Nevertheless, it is based heavily on other authors' listening experiments where this is possible. The loudness and duplex theory weighing coefficients are used to combine the 48 histograms generated at each sampling point into an overall histogram. Spatial attributes are extracted only from this output.

5 INVESTIGATION

More than one hundred different excerpts from binaural recordings are analysed in this chapter. These form four main groups of stimuli, which will be studied in separate experiments to investigate specific properties of the spatial analysis algorithms. The assertions that are made in the previous two chapters will be tested.

All experiments in this chapter are designed to investigate the localisation of single instruments in a reverberant room. The first experiment of this chapter investigates the precision of the localisation algorithm using two sets of stimuli: tamborim hits [see glossary] and replayed square waves, at a number of known source angles.

The second experiment of this investigation tests the onset detector. The methodology for this section is based on the experiment of Supper et al. [2005]. Instead of assessing the onset detector in the frequently-encountered manner, by regarding its performance as a note-counter, it is tested in conjunction with the localisation algorithm. The ability of the system to locate a variety of complex instrumental and speech sources in a reverberant environment is assessed, first in the absence of any other sound sources, and then with an artificially-generated distracting noise source.

Two further experiments test the potential for expanding the spatial analyser to extract secondary spatial attributes. In the first of these experiments, four piano stimuli with different physical widths are analysed, and a formula is advanced tentatively to relate the extracted data to the actual angular source width. The second experiment focuses on the extraction of source distance information. A function that reliably correlates the input data to source distance is found.

Finally, the computational demand of the current MATLAB implementation will be determined. This will demonstrate that the possibility exists for creating a real-time version of these algorithms on existing equipment.

An implicit purpose of this chapter is to effect a general validation of the spatial auditory algorithms. It will be shown that the localisation algorithm and the onset detector combine to produce meaningful results for a wide range of programme material in a challenging acoustic environment.

All recordings in these experiments were made in Studio 1 at the University of Surrey, using a Cortex Instruments MK2 dummy head microphone using

44.1kHz sampling frequency and 24 bits resolution. Studio 1 is a classical recording studio and performance hall, with a floor area of 250m². The recordings were made with the audience seating present: this reduced the hall's reverberation time. As Figure 5.1 shows, the 60dB reverberation time of the hall is approximately one second.

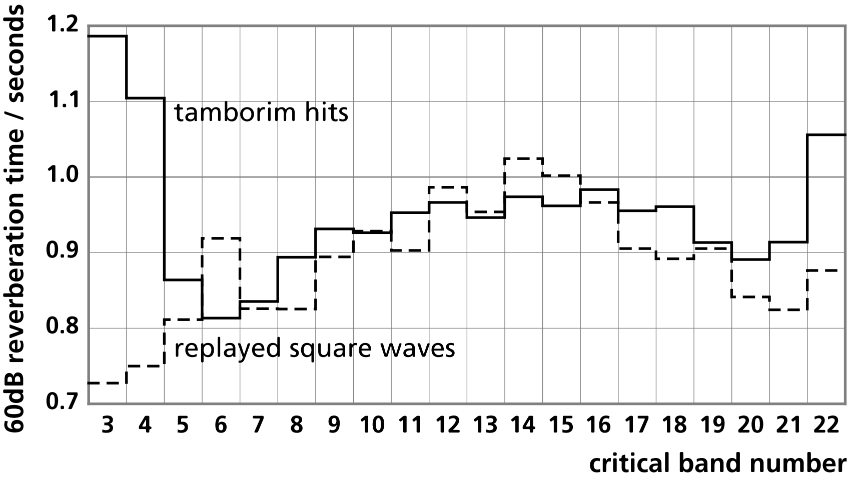


Figure 5.1. Reverberation time of Studio 1. This data is averaged from 36 tamborim hits and 27 replayed square waves at four metres from the dummy head, with various source angles. Decay curves were extrapolated to find the reverberation time wherever 60dB of dynamic range was not available. Reverberation time data is unreliable in the very lowest and highest bands, owing to limited energy content in the stimuli at these frequencies.

5.1 Visualising the data

It is worth illustrating some of the features of the output histograms that will be shown in this chapter. The localisation algorithm generates an output at 2450Hz: approximately every 408µs. This time interval, which will henceforth be referred to as the spatial sampling period (SSP), will be used frequently in this chapter. According to the precedence effect, two or three SSPs contain most of the pertinent data that concerns source location [Wallach et al. 1949]. For comparison, a video frame spans approximately 80 SSPs, and late reflected energy can affect spatial perception even 100ms — 250 SSPs — after onset [Barron and Marshall 1981]. Such a large ratio between the shortest and longest periods of interest can present challenges regarding the visual representation of the data.

Figure 5.2 shows raw output data from the localisation algorithm when presented with a 250ms square wave burst, positioned 20° left in front of the dummy head. The output data depicted here spans 500ms, which is 1225 SSPs.

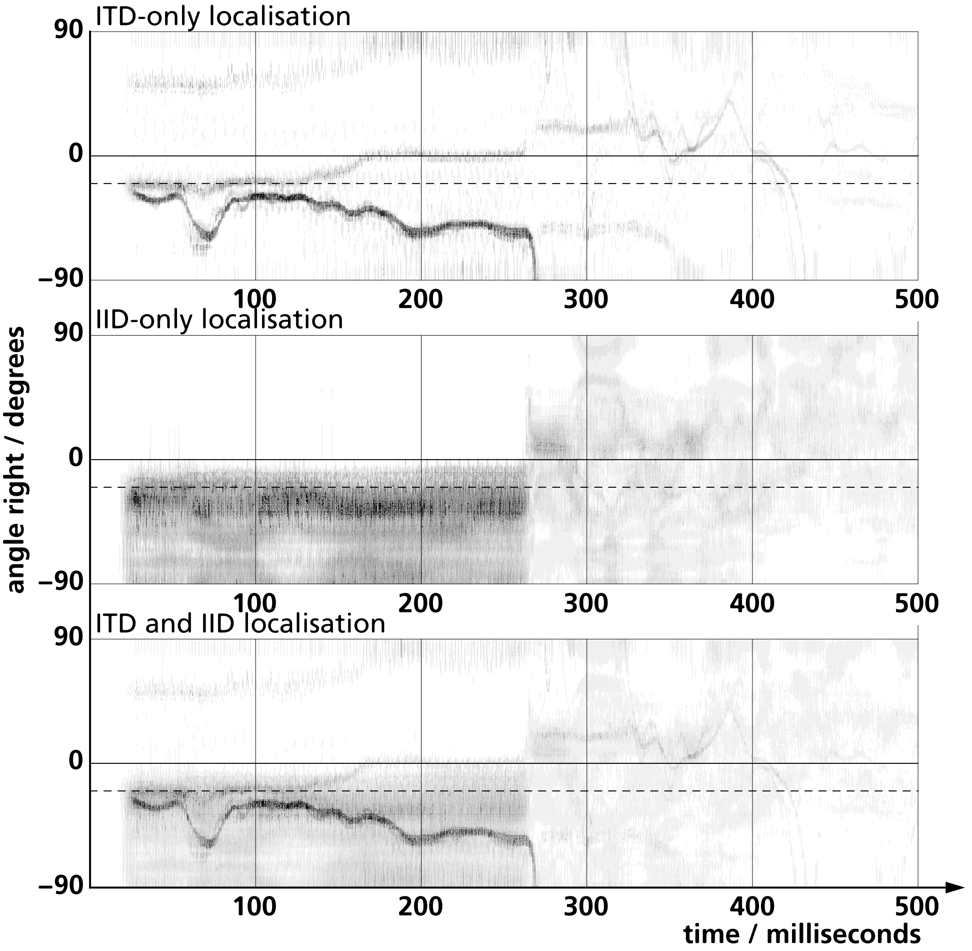


Figure 5.2. Raw output from the localisation algorithm: truth value versus lateral angle over time. The stimulus is a square wave at middle C lasting 250ms, with 10ms fades in and out, replayed through a large Lentek loudspeaker in Studio 1. The dotted horizontal line is at 20° left: the actual source location.

Although the data is difficult to manage in detail at this scale, for sustained stimuli the source position generally remains apparent while the stimulus is active. However, for impulsive stimuli, it is usually necessary to use a finer scale to study the detail around an onset. This is where the onset detector is useful. Figure 5.3 shows a magnified version of Figure 5.2 around the detected onset at 63 SSPs (approximately 26ms). Each segment in this plot is loudness-normalised so that the signal is still clear when its level is low.

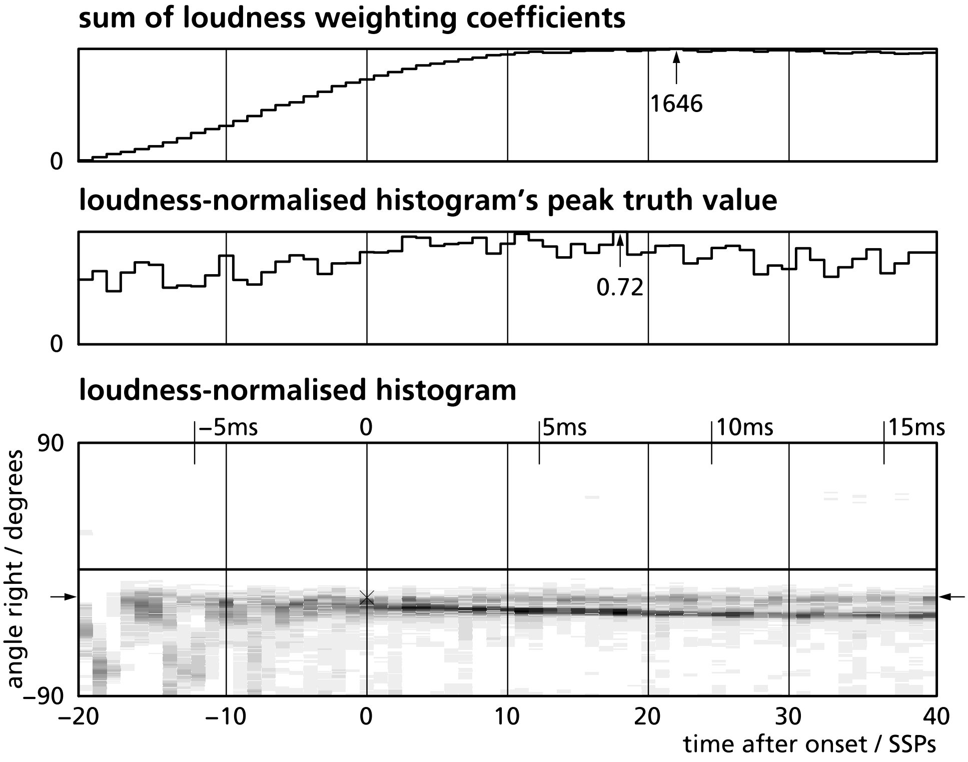


Figure 5.3. Magnified localisation algorithm data: combined ITD and IID cues for the square wave stimulus of Figure 5.2 (bottom graph). This output histogram has been normalised by dividing each histogram by the sum of loudness weighting coefficients (top graph). Zero time in this figure refers to the detected onset, and corresponds to about 26ms in Figure 5.2. The cross and arrows indicate the actual source location: 20° left.

By averaging the three output histograms from 0–2 SSPs after the detected onset and finding the peak of the result, a source location of 25° left can be calculated. An error of 5° exists between the calculated and actual source locations. These visualisation and analysis techniques will be used in the next section to examine the characteristics of localisation error in the algorithm.

5.2 Localisation precision

There are many factors within the algorithm that could affect the precision with which a source is localised. Many untested assertions were made in

Chapter 4 during the formulation of the localisation algorithm. The least safe of these assertions — those that have not been thoroughly tested by other researchers and are not mathematically provable — are listed below:

- The physical parameters of the KEMAR and Cortex MK2 heads are very similar, so databases recorded using KEMAR should be compatible with stimuli gathered using the Cortex MK2 dummy head.
- The creation method and the resolution of the HRTF database are sufficient to render and store ITD and IID cues accurately.
- The diffuse-field correction of head angles will entirely compensate for the most important near-field effects within the KEMAR data (see Section 4.4.1).
- Histogram representation provides an effective way of combining ITD and IID data across frequency bands to produce a unified localisation decision. (Faller and Merimaa employ a more elaborate method for combining cues from multiple onsets [Faller and Merimaa 2004]. In their paradigm, truth values are plotted against ITD and IID for a number of onsets. This effectively produces a three-dimensional histogram. A centroid can then be extracted comprising an ITD and an IID. However, this method is useful only when data points from several onsets are available. See Section 2.4 for more details of this algorithm).

Using the method presented in Section 5.1, the localisation error has been calculated for a recorded set of 63 unelevated tamborim and square-wave stimuli. These were played at a distance of 4m from the head, and at 1° resolution. The angular error can be seen in Figure 5.4.

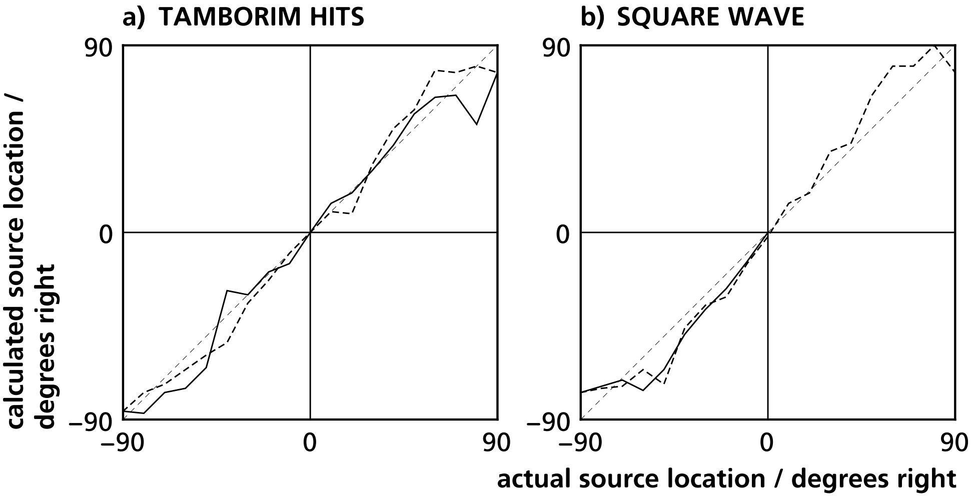


Figure 5.4. Localisation precision for two sets of stimuli. Dotted lines represent rear sources; solid lines represent frontal sources. Results from 63 test stimuli are shown in these graphs. The maximum deviation from normal is 18°. The average deviation from the normal for all data sets is 9°.

Localisation accuracy data for the two stimuli show that the discrepancy between the actual and calculated source angle is small for small angles, and larger for larger angles. However, these errors are not unreasonably large when compared with the performance of human listeners. The localisation discrepancies bear strong similarities in pattern, but are slightly lower in magnitude, than localisation variability data collected from six trained human listeners by Makous and Middlebrooks [1990], although the localisation error of around 10° at the aural axis matches data from a number of experiments summarised by Blauert [1987: 41]. They also match the pattern of localisation blur observed by Boone and Helleman [2004] for an artificial reflection, delayed by 30ms, of a central speech stimulus.

The localisation error encountered in this algorithm can be attributed to a number of causes. Firstly, the shape of IID data, particularly at high frequencies, is heavily dependent on fine detail of the recording head and ears [Møller et al. 1999]. Thus, it may not be safe to assume that localisation cues will be similar for different manufacturers' dummy heads with similar proportions. Secondly, owing to the geometry of the head, a small change in lateral angle for a central sound source has a much greater effect on interaural cues than the same angular change for a source positioned near the aural axis. The interaural differences between stimuli positioned around the aural axis may be too subtle to detect reliably and repeatedly. Thirdly, the correlogram peak-finding localisation algorithm, used to derive the ITD, favours central sources. To compound this bias, the ITD-to-angle look-up tables do not contain phase-wrapped data, that would counteract the ambiguities that occur when one ear leads the other by a whole wavelength. Very periodic signals in the mid-frequency range, positioned near the aural axis, may easily be mislocalised more centrally.

The action of IID bias is shown in Figure 5.5, in which the results of Figure 5.4 are split into time and intensity information. The ITD-based localisation accuracy plot follows a very linear characteristic with a slightly larger gradient than the normal. This would be symptomatic of the recording [Cortex] head having a larger effective diameter than the analysed [KEMAR] head, but this is not demonstrated by the head metric data provided in Chapter 4, Figure 4.3.

(It can also be noted from Figure 5.5 that the ITD and IID data are complementary: localisation ability is impaired in the absence of either of these cues).

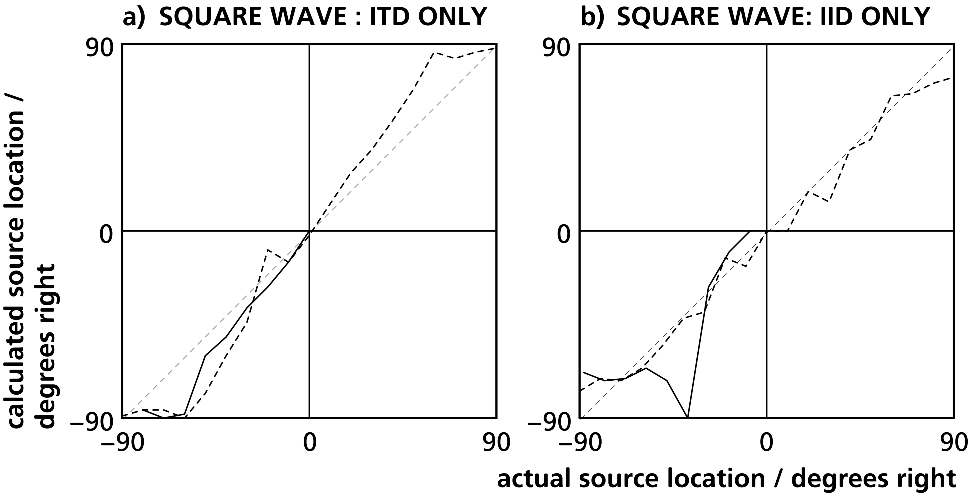


Figure 5.5. Angular error of square wave stimuli in Figure 5.4, separated into ITD and IID components. The dotted lines represent rear sources; solid lines represent frontal sources.

5.2.1 Investigation of a large localisation error

A small number of anomalous results are evident in Figure 5.4. One of the largest discrepancies is generated by the 80° front right tamborim hit. This is localised at 52° right: a 28° discrepancy between actual and localised angles. The output histogram from this tamborim hit is shown in detail in Figure 5.6, so that the anomaly can be investigated further. Figure 5.6 also contains a graph of the histogram's peak truth value against time, and the sum of loudness weighting coefficients used within the localisation algorithm.

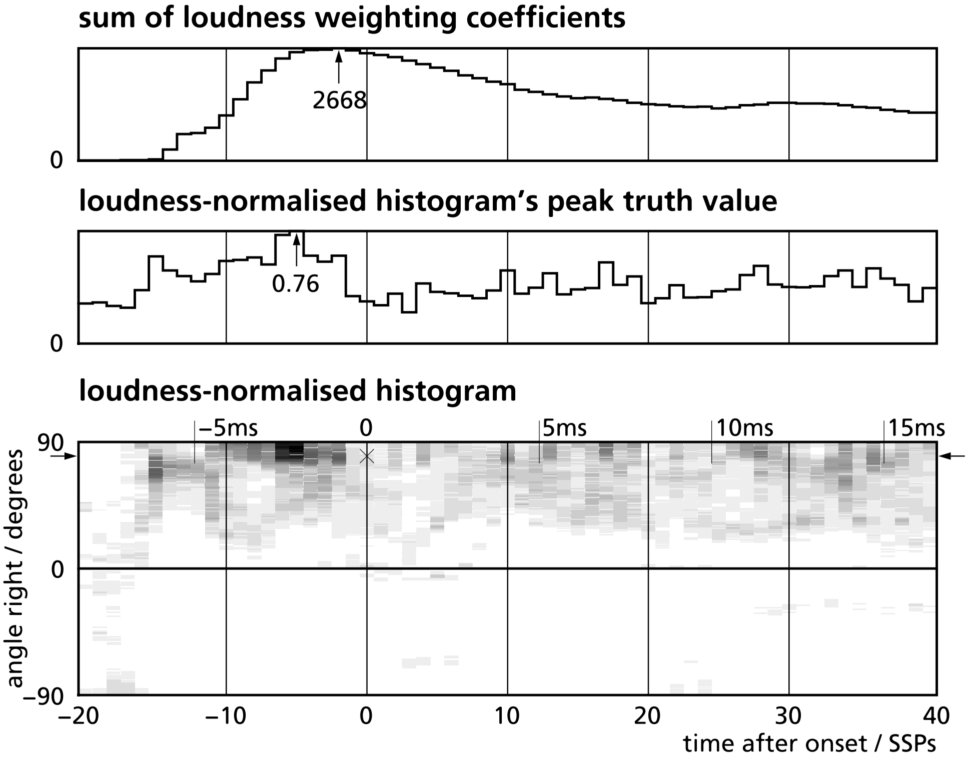


Figure 5.6. Output histogram for the front 80° right tamborim hit that localises anomalously at 52°. Comparing the loudness weighting and peak truth value graphs reveals that the onset was detected between 2ms and 7ms too late.

The loudness-sum function and the peak truth value function (which relates to the confidence of the localisation decision) both reach their maxima 1–2ms before the detected onset, although the first local peak occurs 7ms before. This proves two important things: firstly, that the onset detector is triggered at least 2ms too late; secondly, that the tamborim hit was so impulsive in this example that its auditory onset, during which it generates useful localisation data, was only 6ms long. Both phenomena are required simultaneously to generate such an anomaly.

To investigate the phenomenon of late onset detection more closely, another set of localisation data has been produced for the tamborim and square wave stimuli. This data, shown in Figure 5.7, was produced by finding the peak truth value of the histogram within 25ms of the detected onset, and treating this peak as the true onset. Figure 5.8 summarises the adjustment in onset times in order to produce this figure. Surprisingly, Figure 5.7 contains more anomalies than the original data in Figure 5.4, and these anomalies are more serious. However, the rest of this data follows the normal line a little more closely than the original onset-guided data. Close inspection has revealed that the more serious anomalies are all caused by atypical localisation data just before onset, causing focussed localisation in the ’wrong’ position for between three and five SSPs.

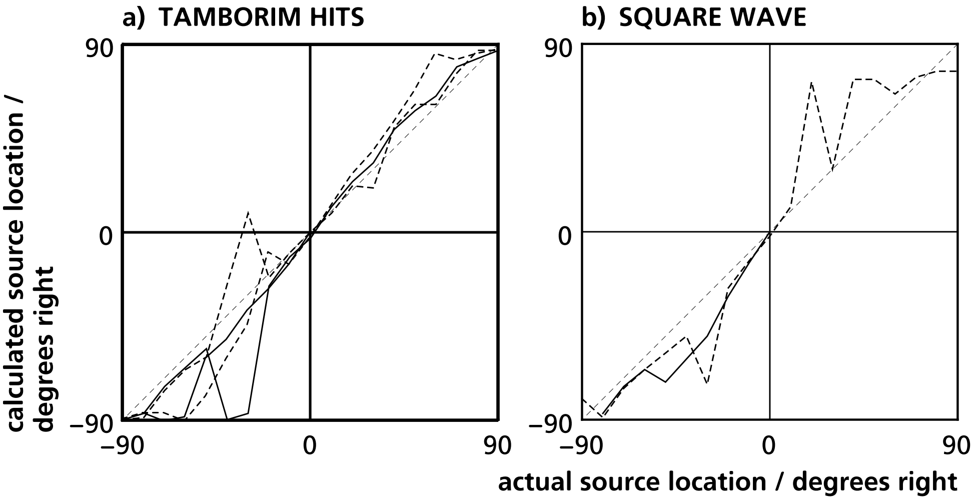


Figure 5.7. Actual versus calculated localisation angle for the tamborim and square wave stimuli. In this figure, analysis points were determined by the peak truth value 25ms around the detected onset.

From these results, it can be concluded that the onset detector determines lateral source angle more consistently than the use of peak truth value alone, although in a few cases the latter method will produce an output that is closer to the actual source position. It can be hypothesised that because most of the onsets were adjusted by shifting them backwards in time, there is a general drift of lateral angle away from the centre of the sound field over time. This drift can be seen explicitly in Figure 5.3, and can also be noted in other examples of the square wave stimulus. The cause of this drift cannot satisfactorily be explained, except that it may be a caused by a specific environmental reflection. Its characteristics cannot, however, be attributed to a first-order reflection from the studio floor, as this would tend to pull the image towards the centre. Is it necessary to justify the attribution of this anomaly to a room acoustics phenomenon with a further example.

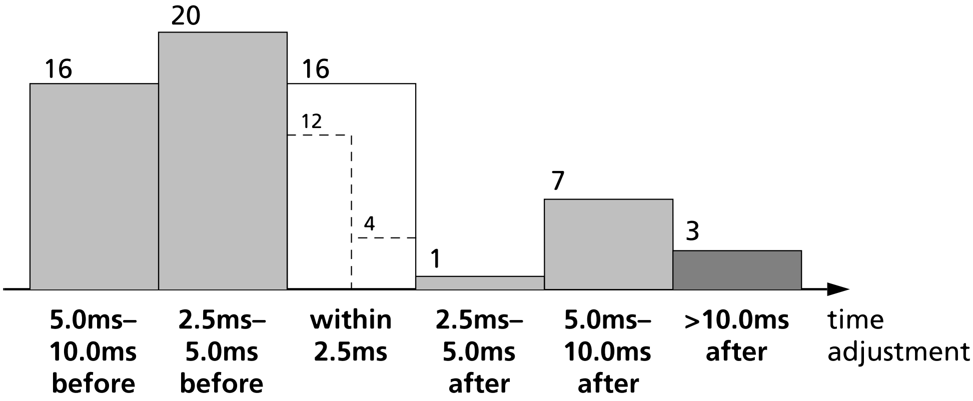


Figure 5.8. Bar chart of the amount of time adjustment applied to the detected onsets of the 63 tamborim and square wave stimuli in order to plot Figure 5.7. The permitted extent of this adjustment was arbitrarily set as [–20SSP +40SSP], which is approximately [–8ms +16ms].

5.2.2 Localisation of an anechoic source

Figure 5.9 is the output histogram for a monophonic square wave that has been convolved with two of Gardner and Martin's anechoic KEMAR impulse response recordings [Gardner and Martin 1994] to position it 20° left in a binaural sound field.

The short-term drift of the calculated lateral angle away from the centre line, which is clear in Figures 5.2 and 5.3, is absent in this example. Hence it can be inferred that this drift is not inherent within the analytical algorithm. Although this single example does not prove that the drift is caused by an acoustic reflection, it does indicate that it is produced by an environmental characteristic. The calculated lateral angle in Figure 5.9 is approximately 25° left. This 5° departure from the actual position is caused mostly by the proximity of the impulse response generator to the dummy recording head when the KEMAR signals were made. The analytical algorithm has been diffuse-field compensated, so this small lateral shift is expected (see Section 4.4.1).

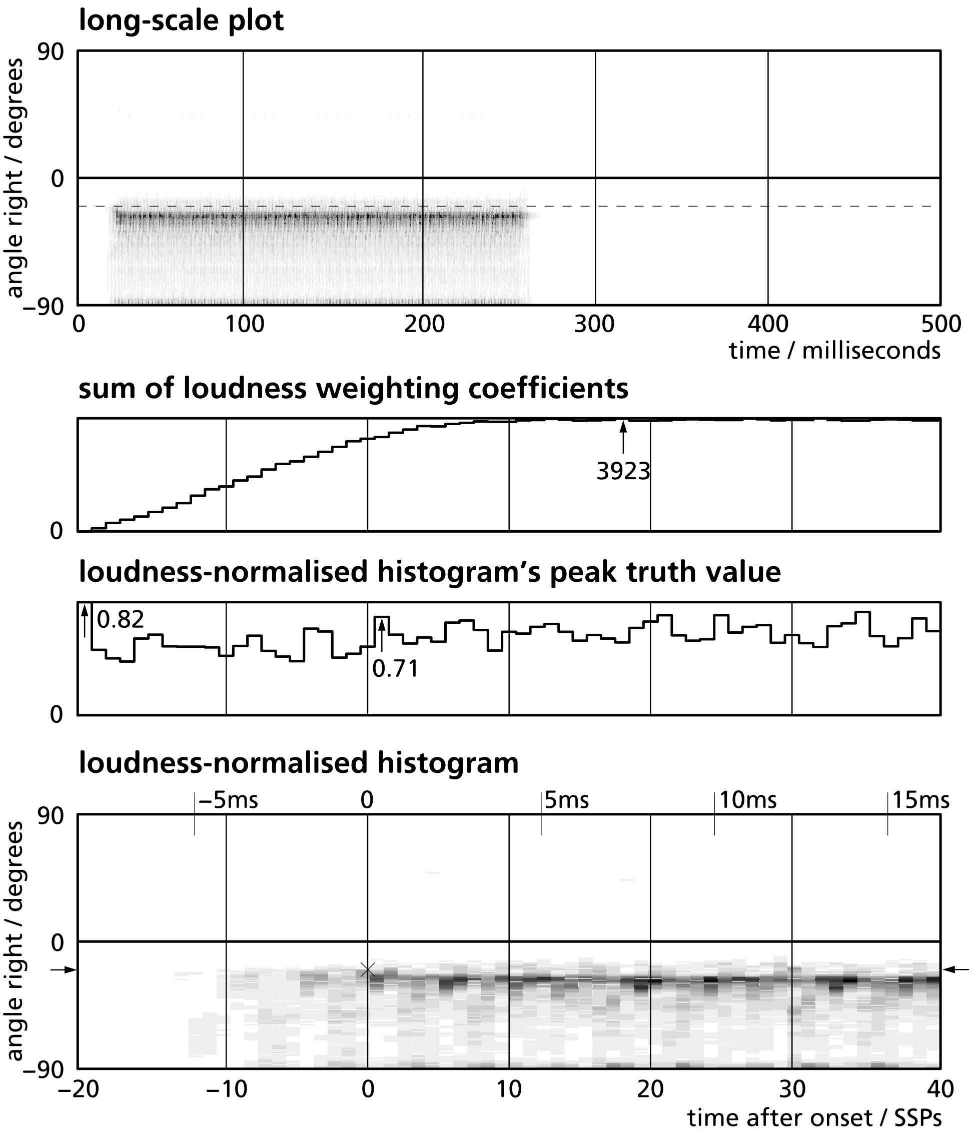


Figure 5.9. Long-range and zoomed in views of the raw output from the localisation algorithm. The input is a synthesised square wave at middle C, positioned 20° left by convolving it with a KEMAR head-related impulse response. The stability of the calculated lateral angle in this data may be compared with that of the recording in Figures 5.2 and 5.3. The initial spike in the loudness-normalised peak truth value is caused by the very low loudness quotients at the beginning of this example.

Only a small number of cues have been tested in this section. Anomalous results have not been re-measured using different recordings, so generalisations regarding anomalies must be treated carefully. The onset detector and localisation algorithms have now been shown to determine source location effectively for two types of sound source, both of which have fast rise times. A tamborim hit also has a fast decay time. Tamborim onsets are therefore only a few milliseconds long, so precise onset detection is important for localisation of the instrument. Localisation anomalies, which are caused by late onset detection, become more problematic the closer the source is positioned to the aural axis.

Attempting to correct anomalies from late onset detection by using the histogram as an 'onset corrector', to find the peak truth value around a detected onset, produces anomalies of a different kind by detecting false truth value peaks. These are produced before onset and could not be detected by a human listener. For the rest of this chapter, the original onset-guided solution, from which Figure 5.4 was plotted, will be maintained.

An effect can be seen in the square wave results where the waveform is pulled away from the centre line the longer it continues, and this can affect localisation by several degrees if an onset is detected slightly too late. This phenomenon cannot be explained, but can cautiously be attributed to the acoustics of the recording studio.

5.3 Onset detector performance

So far, the onset detector's performance has been assessed only as far as its ability to extract information from single musical notes. There is a limit to the number of general conclusions that may be inferred from the analysis of one onset. This section will test the onset detector using extended musical and oratory examples containing many onsets.

Many onset detectors are assessed on their performance as note-counters. A performance benchmark can then be formulated by manually counting the number of musical events in a passage, and then comparing this count with the number of events that the algorithm misses, and the number of spurious detections.

The note-counting approach assumes that musical events are equivalent to auditory onsets. However, when the definition of an auditory onset is extended to suit the spatial analysis task (see Section 3.1), it is no longer satisfactory to assume this relationship. Auditory onsets are now defined as intervals of time during which correct localisation information may be extracted from the binaural stream. For reference, however, a simple comparison of the output of the onset detector against sections of four binaural waveforms can be seen in Figure 5.10.

A more relevant method for testing the onset detector is used in this section. This extends the methodology of Supper et al. [2005], which investigated the detector as a component of the spatial analyser. In the first part of this experiment, the onset detector is tested by assessing the ability of the spatial analyser to localise a complex source over an extended period of time. The effectiveness of such localisation relates to the precision and sensitivity of the onset detector. The second experiment investigates the robustness of the onset detector further, by adding coherent white noise to the binaural stimuli.

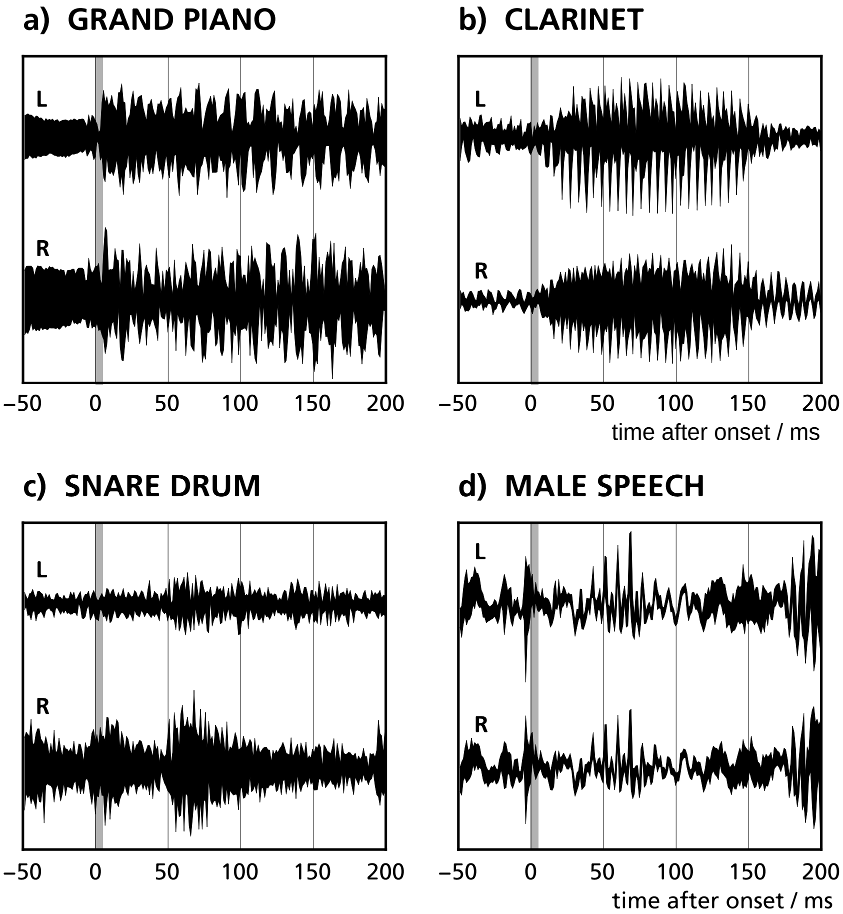


Figure 5.10. Binaural input waveforms around detected onsets, for four different sound sources. None of these examples are initial or final onsets from the excerpts. The grey box encloses 2ms of audio. This is the amount by which the onset detector is permitted to look ahead.

5.3.1 Onset-guided localisation of an individual source

Figure 5.11 presents localisation data from a 14-second passage of fast solo clarinet music. The clarinet was positioned in the frontal hemisphere, 60° left of the recording head, and four metres away. Analysing this signal, 53 onsets were detected. To produce the results shown, the histograms have been loudness-compensated, and data from every detected onset added together. To improve the realism of the results and to suppress secondary triggering of the onset detector when strong level fluctuations are encountered, an additional rule has been applied in plotting the diagrams in this section: any onset detected less than 20ms after another onset is excluded from calculations.

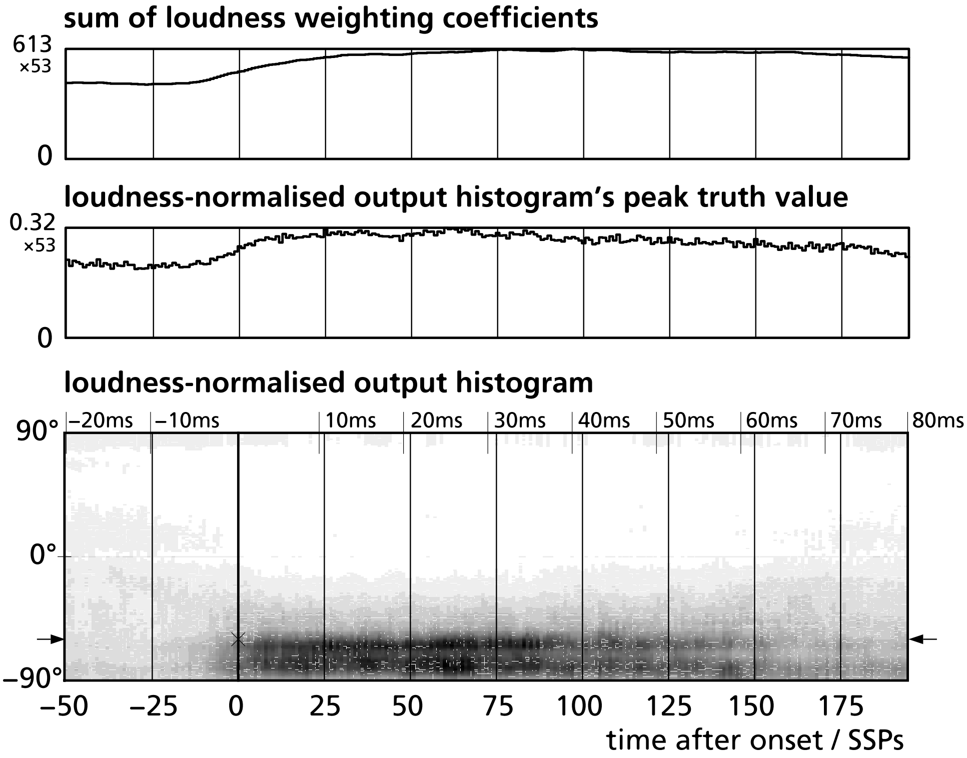


Figure 5.11. Sum of loudness-normalised output histograms around the 53 detected onsets in a 14-second clarinet excerpt.

A number of features of Figure 5.11 indicate that the onset detector is performing well. The most noticeable of these is the strong convergence in localisation data around the 0ms point, giving rise to the horizontal bars between 60° and 80° left in the output histogram. This is accompanied by significant increases in the peak truth value and the sum of loudness weighting coefficients. Investigation of the individual onsets shows that the large apparent width of the histogram peak is a phenomenon of the localisation algorithm: it is not caused by the onset detector picking up strong lateral reflections.

One other interesting aspect of Figure 5.11 is that the histogram becomes far less focussed, and its peak lowers, well within the scope of the time range shown. The transition from the state when localisation is possible to the state when it is not — the *auditory offset* — occurs between 50ms and 80ms after the beginning of the auditory onset. The majority of clarinet notes in the extract are longer than this, so this transition is most likely governed by the presence of late acoustic reflections. The 50–80ms range of auditory onset is consistent with the time interval that has been chosen for existing measures of spatial impression (for example, Bradley and Soulodre [1995]), to separate source-related secondary attributes such as apparent source width from environment-related attributes such as listener envelopment. This time constant grows from properties of early reflections within the recording environment.

Although Figure 5.11 indicates that the onset detector works satisfactorily for the clarinet tone, the visualisation method it employs is such that detection of genuine auditory onsets cannot be separated easily from the detection of spurious onsets. Figure 5.12 is a more critical representation of the localisation data. It shows the location of the maxima of every output histogram. When all 53 onsets are included, the distribution of localisation data in the histograms can be inspected more precisely. The different methods of representing onset-guided localisation data will be distinguished by referring to the method used in Figure 5.11 as the *histogram sum technique*, and the method Figure 5.12 as the *hit count technique*.

As expected, at 0ms (the detected onset time) there is a pronounced change in the distribution of localisation data. Before onset, a vaguely left-heavy distribution can be seen, with an approximately 3:1 distribution of histogram peaks between the left and right hemispheres of audition. This ratio shifts heavily from approximately 2ms before the detected onset, so that a 3:1 ratio now separates histogram peaks that appear 30° or more to the left from those that do not. It can also be seen that between 60% and 70% of the histogram peaks are distributed between 60° and 90° left.

The obverse of this statement is that 30–40% of the detected onsets in the clarinet excerpt do not occur near the actual source location. However, this observation indicates only where each histogram peak occurs — it does not take into account the magnitude of each truth value maximum, and therefore the ‘confidence’ of each localisation decision. Neither does it convey the importance of other information, such as the signal loudness at the time of onset. This data is incorporated into Figure 5.11, in which the focussing of histogram data around the onset is more striking.

Owing to these differences between Figures 5.11 and 5.12, it can be inferred that much of the mislocated data shown in Figure 5.12 is based either on short-lived, low-loudness anomalies of the kind that are manifested in Figure 5.7, or on data from room reflections. These would be masked by a human listener, but the spatial analyser that is tested here lacks a masking model. The decision not to include a masking model was taken during the development of the onset detection algorithm, since there is no reason to believe that spatial auditory processing is subject to the same masking mechanisms that affect conscious sound perception. In fact, there is more reason to believe that the opposite is true. Basic spatial processing, such as ITD and IID extraction, takes place at a very low level of the human auditory system. However, a simple long-term suppression mechanism, intended to mask strong early reflections, may improve the localisation data.

To suppress the detection of spurious sources in a future implementation of the spatial analyser, parameters within the onset detector could be adjusted automatically and continuously, to adjust the sensitivity of the algorithm to suit each stimulus and environment. This has been investigated informally by adjusting the onset detector, specifically the parameters $\alpha_0(t)$ and $\alpha_1(t)$, to control the rate of onset within maximum and minimum limits. (The results of this experimentation have been promising, but a formal evaluation has not yet been conducted).

A more refined implementation of the onset detection algorithm will also incorporate data besides the angle at which the histogram peak occurs. For example, the peak value of the output histogram could be used as an indication of confidence. This would imitate the now standard use of the interaural cross-correlation function as an indicator of interaural coherence, and hence as a measure of source width and localisability (the inverse relationship between coherence and localisability is discussed in Hartmann [1983]). This idea is investigated in the forthcoming sections.

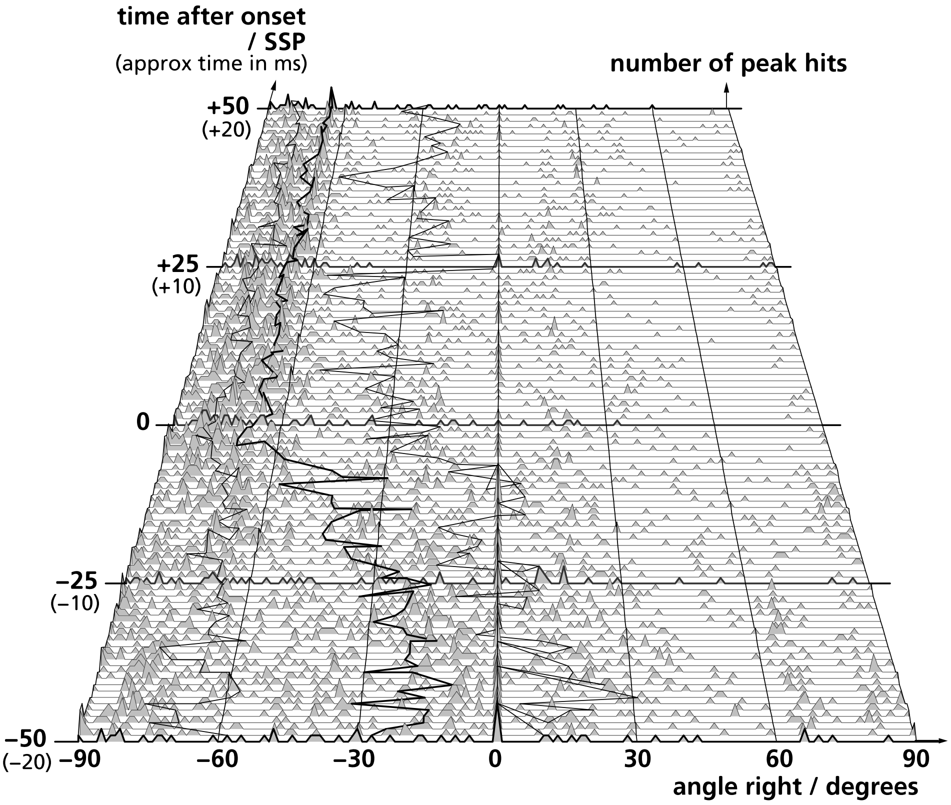


Figure 5.12. Distribution of truth value maxima around the 53 detected onsets of the clarinet excerpt. The time axis has been rotated by 90° from its familiar orientation for clarity. The three transverse lines trace the quartiles and median, to make the data distribution clearer to see. Thus the data is divided into four sections, each of which contains approximately one quarter of the 53 total maxima in each row.

5.3.2 Onset-guided localisation of different sources

It is now clear that different instruments, with different envelope characteristics and hence different lengths of onset, present different challenges for a localisation algorithm, and for spatial analysis in general. In Figures 5.13–5.16, the localisation algorithm's performance is presented for four more musical and oratory sources. All are positioned 60° left, four metres from the dummy head: a solo played on a melodica [see glossary], a passage from a piano menuet, male speech, and a latin rhythm played on the tamborim.

Many general observations can be made from this data. The long sustain characteristics of the clarinet and melodica are clearly visible in the histogram sums, and offset is seen to begin about 50ms after onset. Relatively few onsets are detected in the clarinet and melodica examples compared with the remaining excerpts. In fact, the number of counted onsets is a little fewer than the number of notes in the two excerpts. The histogram sum and hit count methods both work well for estimating source position.

In contrast, the stimuli that have shorter notes and sharper decay characteristics cause the spatial analyser to respond very differently. All are localisable using the histogram sum technique, and the short onset times can clearly be seen. In the speech and tamborim examples, onset is very short: all useful localisation data from the tamborim sum histogram has ceased 30ms after the detected onset; in the speech extract, this time is approximately 50ms. Owing to the chaotically fluctuating nature of the piano, speech and tamborim signals, and also to the shortness of their auditory events and the high speed of their decay, many onsets are detected. The average rate of onset in each example is between 5 and 10Hz. It is likely that many of these detected onsets are caused by strong reflections. This is particularly true for the tamborim stimulus, which is the loudest, fastest-decaying, most noise-based instrument, and therefore the most likely to create strong, discrete reflections that are included erroneously as onsets.

These characteristics explain the results from the hit count technique. In the case of the tamborim hit, the only indication of the presence of the instrument in the hit count graph (Figure 5.16 bottom) is that the quartile and median lines change by 20° from their positions for random data. This shift would require one third of onsets to identify source position as 60°, or 22% of onsets to identify the source position as 90°, amongst randomly-distributed data from early reflections. In the other quickly-decaying instruments, about 30–40% of counted hits occur around the actual source location, and about 50% in the correct lateral quadrant (from 45° in front of the listener to 45° behind).

Mellinger [Mellinger 1991: 58] uses a simple method to address the problem of dealing with different instruments with different attack characteristics. A specific kind of onset detection algorithm — essentially a specialist type of band-pass filter — is run four times with four different integration time constants, approximately spanning the interval between 2ms and 20ms. When all four iterations are completed, an output is chosen which represents the best compromise between strength of detection and sharpness of contrast.

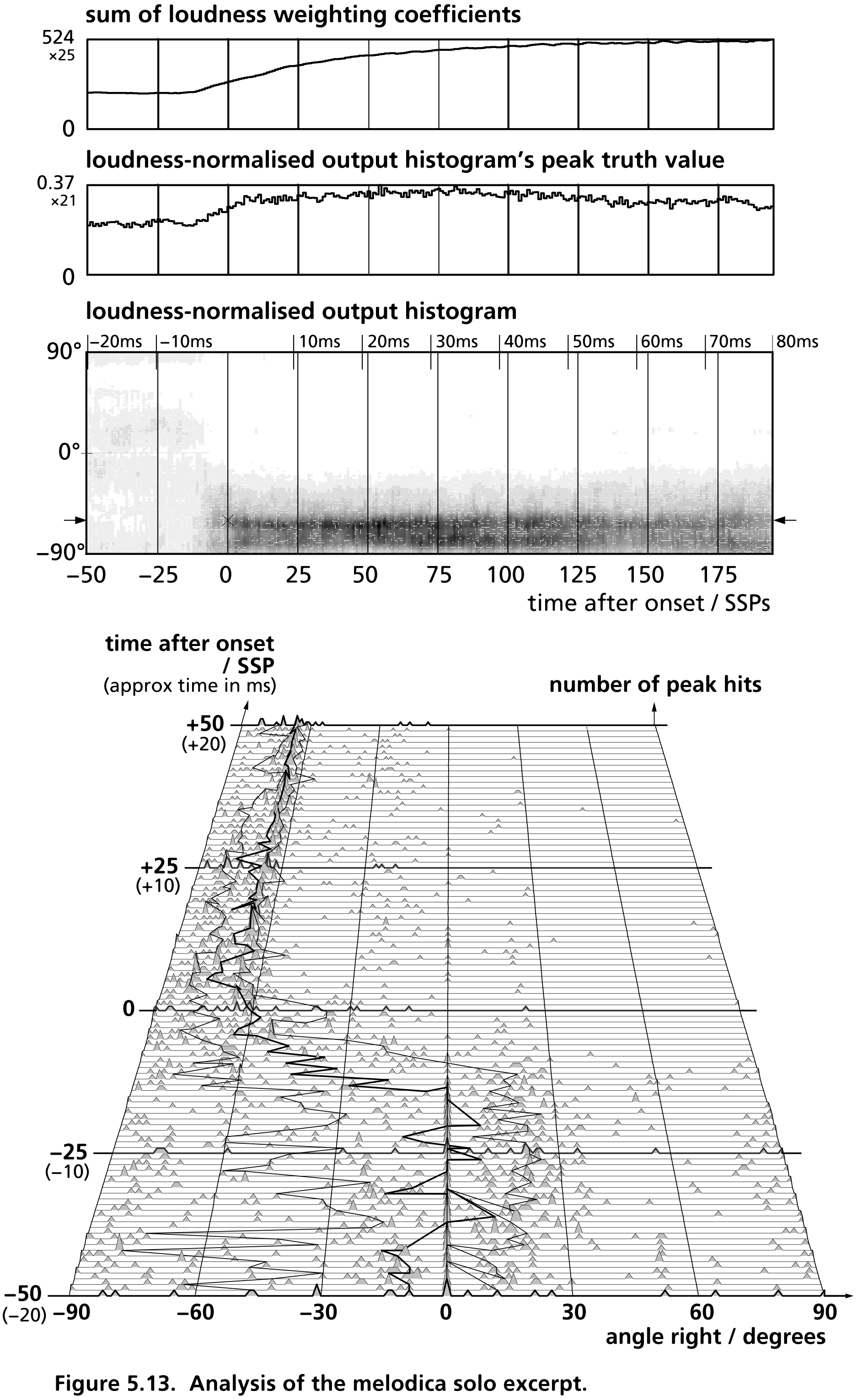


Figure 5.13. Analysis of the melodica solo excerpt.

This excerpt is 28.5 seconds long. 21 onsets are detected and included.

Actual source position 60° left, frontal hemisphere, distance of 4 metres.

Top: histogram sum technique. Bottom: hit count technique.

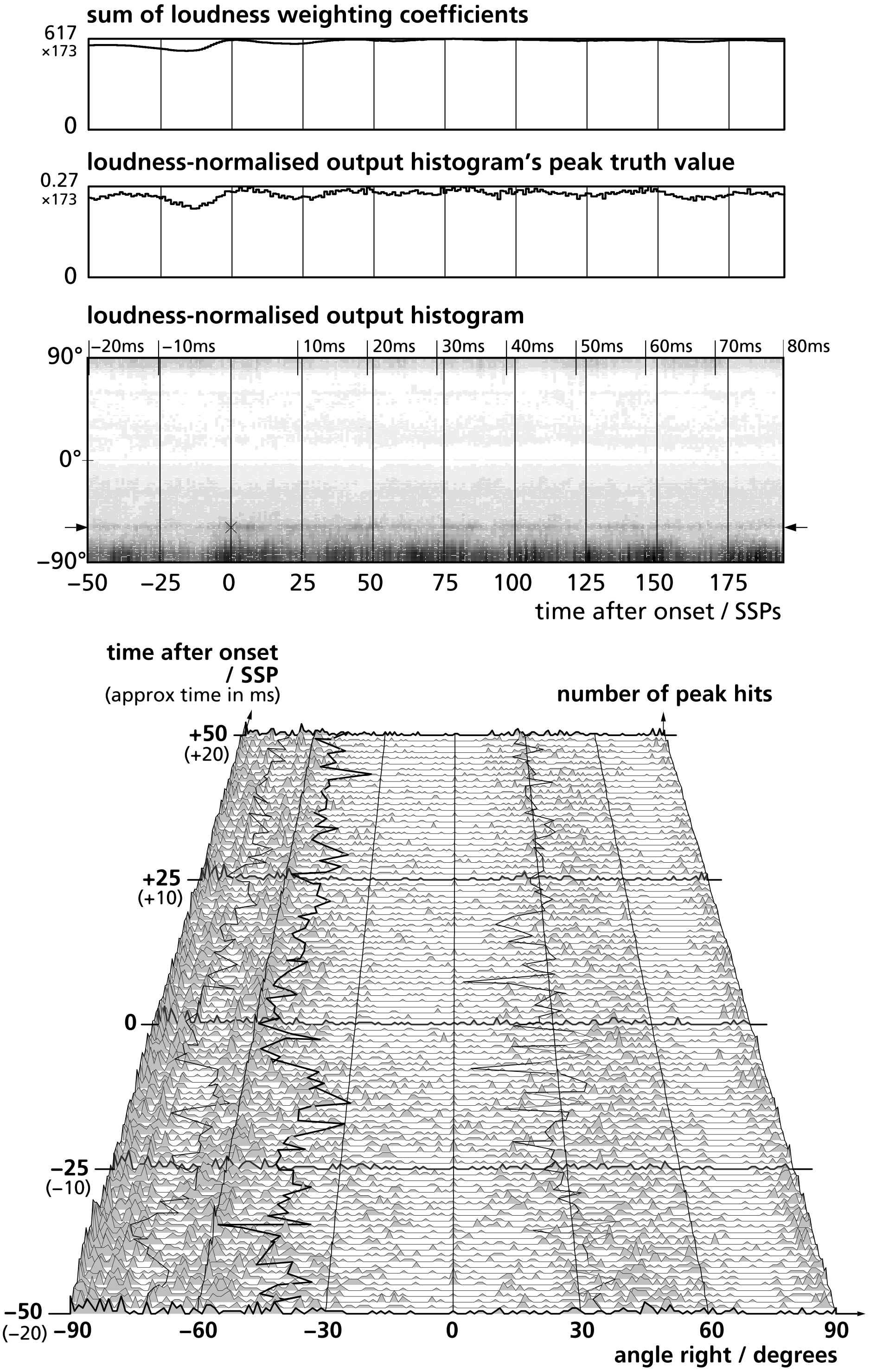


Figure 5.14. Analysis of the piano menuet excerpt.
The excerpt is 24 seconds long. 173 onsets are detected and included.
Actual source position 60° left, frontal hemisphere, distance of 4 metres.
Top: histogram sum technique. Bottom: hit count technique.

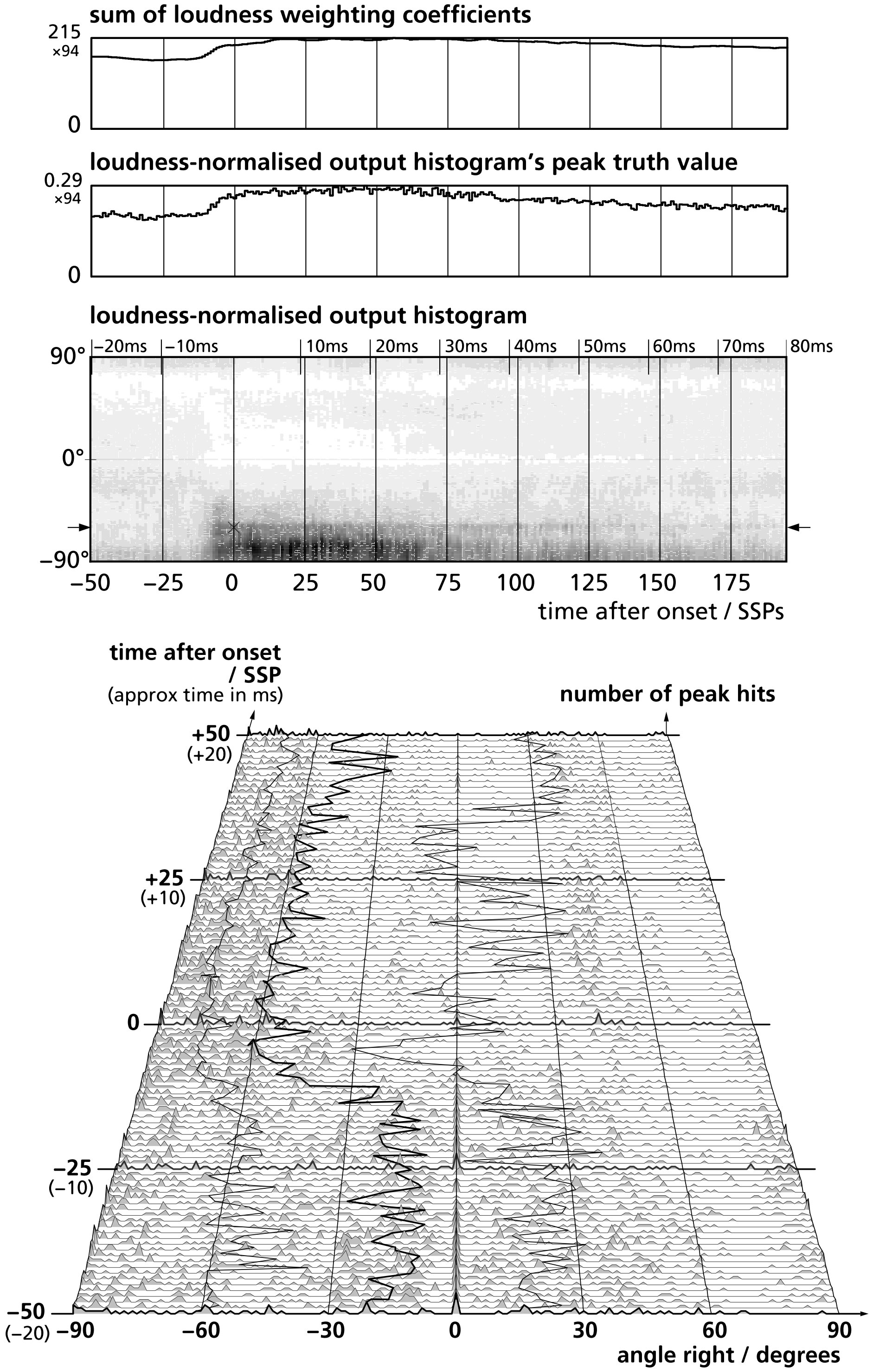


Figure 5.15. Analysis of the male speech excerpt.

The excerpt is 15.3 seconds long. 94 onsets are detected and included.

Actual source position 60° left, frontal hemisphere, distance of 4 metres.

Top: Histogram sum technique. Bottom: hit count technique.

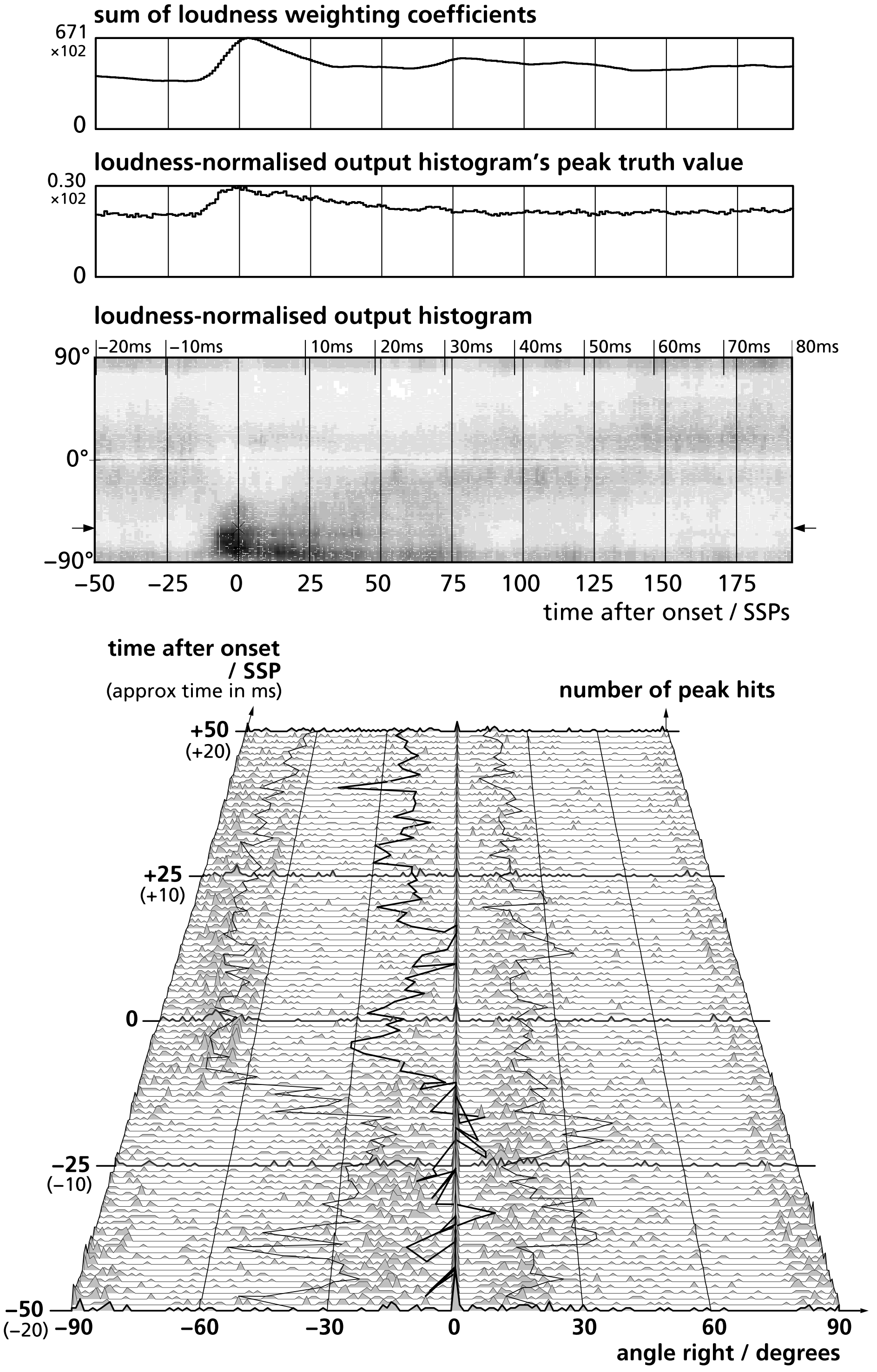


Figure 5.16. Analysis of the tamborim latin rhythm excerpt.
The excerpt is 11.3 seconds long. 102 onsets are detected and included.
Actual source position 60° left, frontal hemisphere, distance of 4 metres.
Top: histogram sum technique. Bottom: hit count technique.

5.3.3 Localisation in the presence of noise

The results presented so far are from individual sources recorded in a reverberant room. A more critical indication of the robustness of the onset detector can be obtained by analysing the same stimuli in the presence of a distractor signal. Coherent [centrally-localising] white noise will be summed with the binaural signal, to behave as a distractor. It is largely free from the prolonged increases in signal energy across frequency bands that signify an onset, so onset-guided localisation should not be perturbed by its presence. The level of the distractor for each stimulus has been set so that its rms level is approximately 15dB lower than the rms level of the source.

To test the operation of the onset detector, two sets of data have been derived from the noisy stimuli. The first takes localisation data from the times identified by the onset detector, and the second extracts data from randomly-generated times. Provided the onset detector is serving its purpose, it can be expected that marked differences will be seen between the two sets of data, with the distractor signal strongly present in the random-time data, and the onset-guided data demonstrating robustness against the distractor. Figure 5.17 presents the sum histograms from all the results, and Figure 5.18 summarises the hit count data.

In Figure 5.17, the effect of the coherent distractor can clearly be seen as a centrally-positioned stripe and cloud, in both the random-time examples and the onset-guided examples before 0ms. For all examples but the piano, however, the effect on onset-guided localisation after 0ms is minimal. The control histogram data, taken at random times, confirms that the distractor noise is substantial enough to disrupt localisation in the absence of onset detection. This is especially noticeable within the speech and tamborim stimuli, as these contain long periods of inactivity which are entirely taken over by the distractor signal. In every example, there are a few extra onsets introduced by the distractor. In all cases but the piano and the melodica, however, these onsets constitute fewer than 8% of the total number of onsets.

In addition to the creation of these spurious onsets, two further indications of performance impairment can be seen in Figure 5.18. There is the amount by which the median localised angle in the distractor-present data deviates from the distractor-absent case, and there is also the widening of the statistical distribution of these results when the distractor is introduced. The more robust the onset detector is against the distractor signal, the smaller this widening will be, and the less visible the effect of the distractor will be.

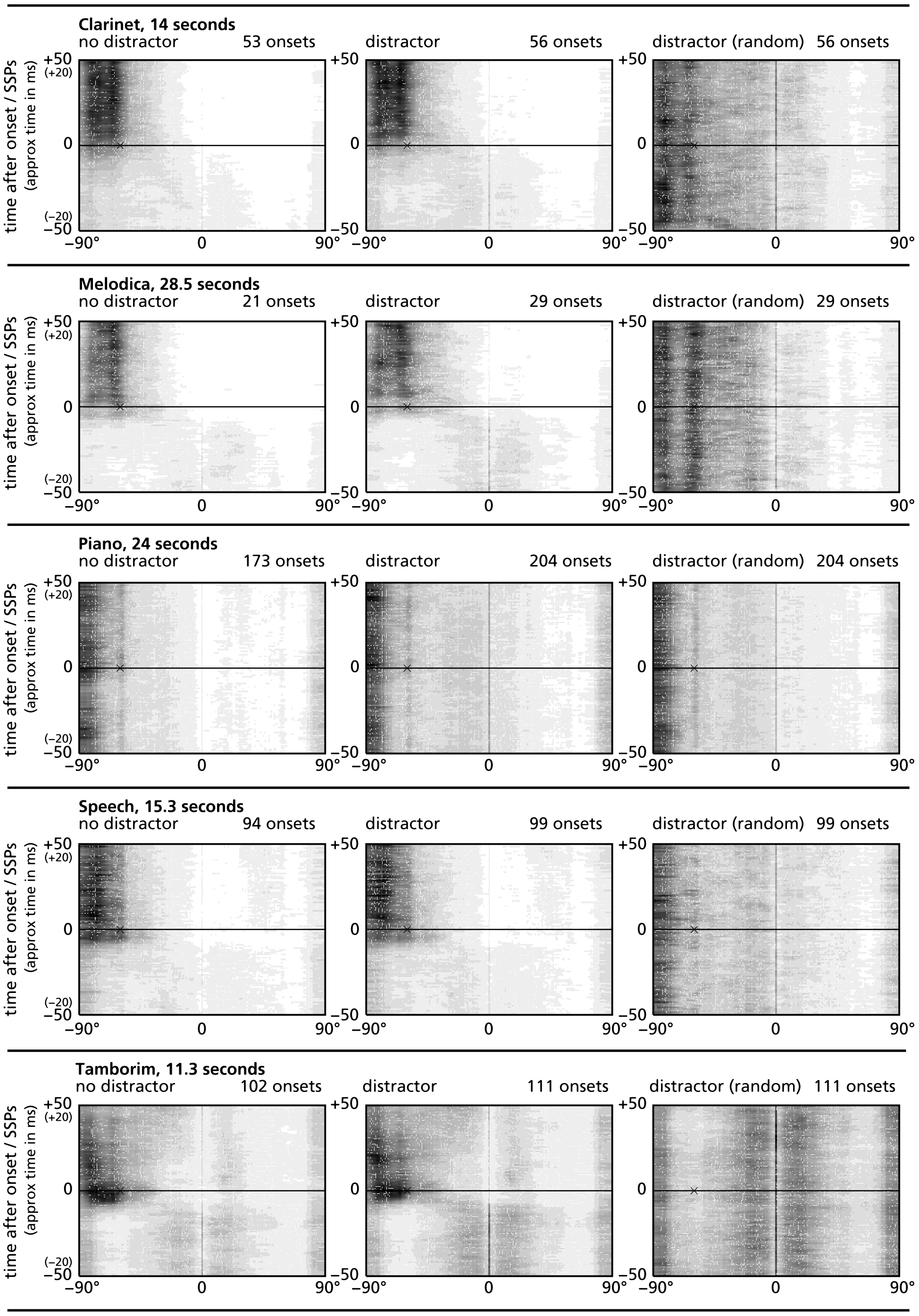


Figure 5.17. Sum histogram representation of the five stimuli. Left: without central distractor. Centre: with central distractor. Right: with central distractor, onset times randomised. For clarity, the axes have been rotated and the time axis now increases upwards, in the same direction as Figures 5.12–5.16. The 60° left position is marked by a cross on each graph.

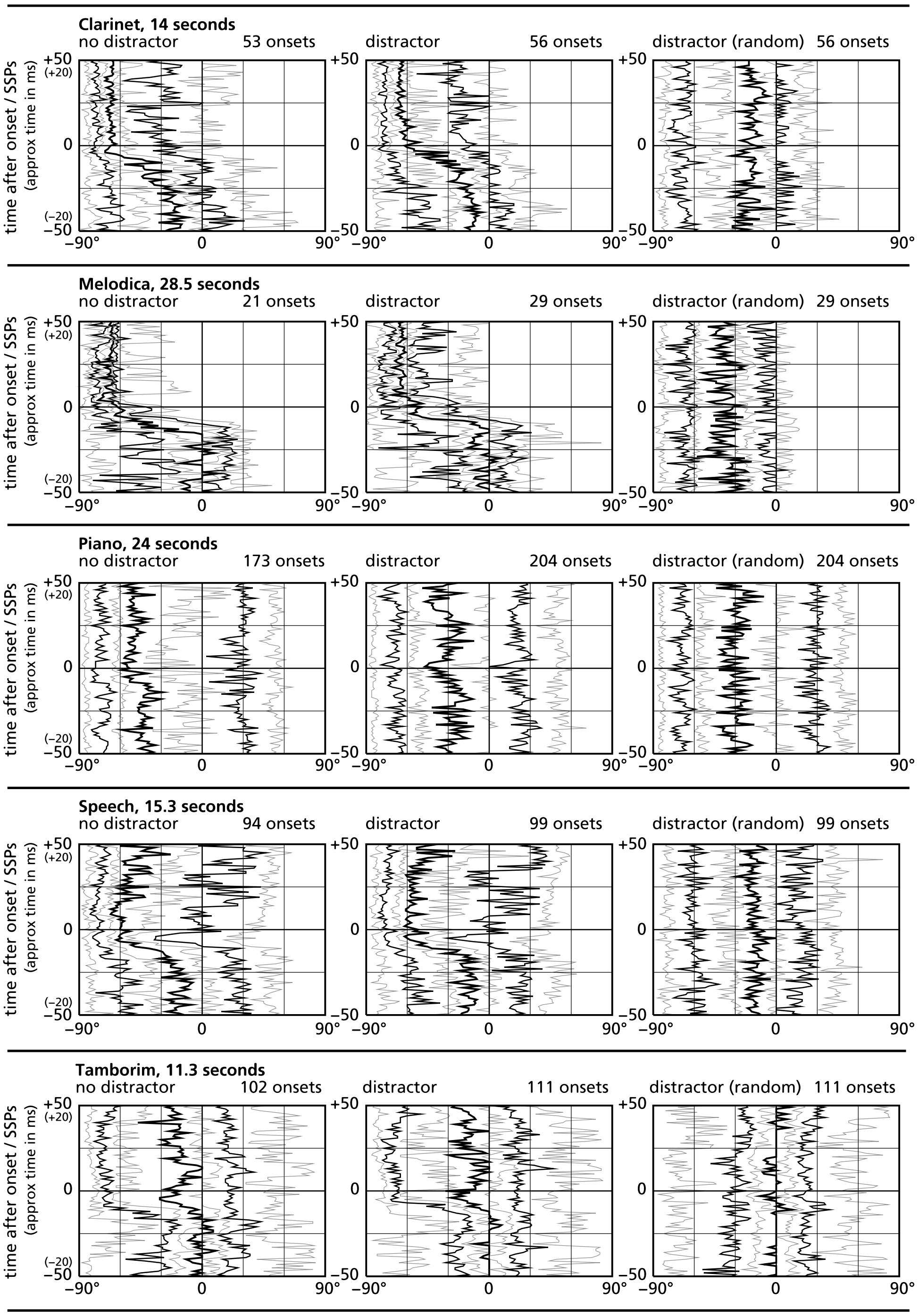


Figure 5.18. Hit count representation of the five stimuli. Left: without central distractor. Centre: with central distractor. Right: with central distractor, onset times randomised. The time axis increases upwards, in the same direction as Figures 5.12–5.16. The raw data cannot be displayed at this scale, so the hit count quartile and median lines [black] are supplemented with octile lines [grey].

For all sources except the piano, the degree of localisation impairment that the distractor causes is minimal, as the sources continue to be clearly localisable. Of these sources, the melodica is most affected: this sensitivity is caused by the sparseness of detected onsets within the stimulus. In the distractor-present case, 28% of the detected onsets are caused by the presence of the noise. However, the false detections contribute only a 5° detraction away from the distractor-absent source location. It is likely that many of the additional onsets in the distractor-present case occur at times when the source dominates the distractor, and can be attributed to the statistical influence of the noise on the audio signal.

Of these examples, only the piano signal highlights a problem with the onset detector. This excerpt is a good illustration of a case where the onset detector would require variable sensitivity in order to perform its intended task. The grand piano is a challenge to onset detection and localisation algorithms, as it is a physically wide instrument, has a complex radiation pattern, and possesses sudden and chaotic fluctuations in level across different frequency bands owing to sympathetic resonances in strings and the soundboard. A high rate of onsets is detected: an average of one every 140ms in the distractor-absent case, and one every 85ms in the distractor-present cases. Many of these onsets will occur during times when room reflections, and distracting noise, dominate. The location of the grand piano can be seen in all examples in Figure 5.18 as a statistical shadow on the left-hand side of the image. While the onset detector's sensitivity to different stimuli remains fixed, its ability to localise a piano effectively cannot be assured.

It is acknowledged that any future development of this algorithm must include a method for varying the onset detector's sensitivity automatically. The fixed-threshold onset detector shows considerable versatility for a wide range of stimuli, even when the problems with the piano excerpt are considered. Excepting these, onset-based localisation with the distractor present compares very well with the distractor-free case.

5.4 Secondary spatial attributes

A principal aim of this project, which distinguishes this method of spatial analysis from a large majority of others, is that it should be compatible with the extraction of secondary spatial attributes from the binaural stream. In this section, the response of the localisation algorithm to two controllable secondary spatial attributes will be examined. The two chosen attributes, source width and source distance, are manipulated by changing the position of orientation of sound sources within the recording environment. In other words, *physically-based* changes are applied to the actual source width and distance in these experiments, as opposed to *perceptually-based* changes, which could be invoked by processing a stimulus in certain ways (for example, [Neher 2004]). It is assumed that unidimensional physical changes in secondary spatial attributes translate to unidimensional perceptual changes, so that these investigations may be considered an investigation into the analyser's ability to extract spatial information in a way that is peceptually relevant.

5.4.1 Manipulation of actual source width

To investigate the effects of source width changes, four separate recordings have been analysed. Two of these feature a grand piano, centrally positioned, at two metres' distance. One performance was recorded with the piano oriented sideways-on (its usual orientation for a recital), and the other with the piano in keys-on orientation. The two remaining recordings are of the same musical passage replayed through a floor-standing loudspeaker, oriented horizontally and then vertically. (This pre-recorded passage was performed on an upright piano in a small practice room, using a cardioid microphone at a distance of approximately two feet.) The loudspeaker output was aligned to be of similar loudness in the room to the grand piano. Mechanical drawings of the four orientations, referred to respectively as 'piano wide', 'piano narrow', 'loudspeaker wide', and 'loudspeaker narrow', are shown in Figure 5.19.

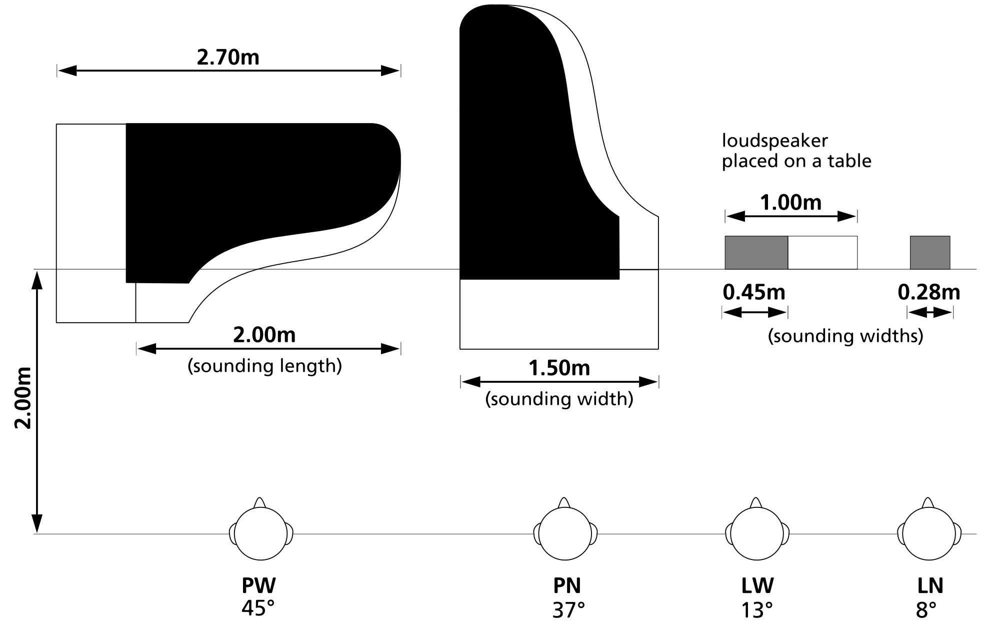


Figure 5.19. Mechanical drawing, with approximate dimensions, of the four recording set-ups used to obtain the different width stimuli. Approximate angular widths for the stimuli are also given. PW = piano wide; PN = piano narrow; LW = loudspeaker wide; LN = loudspeaker narrow.

The following differences might be observed in the output histogram data as the source width is varied:

- The narrower the source, the easier it is to localise. This is because localisation accuracy increases with interaural coherence [Hartmann 1983], while auditory source width varies inversely with interaural coherence [Hidaka et al. 1995]. Thus, an inverse correlation should be observed between source width and maximum peak truth value, and also between source width and average peak truth value during onset.
- A piano is a physically wide instrument, and its width should be apparent from the output histogram. This width will be manifested during, and slightly after, the onset region. It will be visible in two ways: as a blurring of the histogram, and also as angular fluctuations with a period of perhaps 10–125Hz: approximately 8–100ms (see Mason's IACCFF, Section 2.2.2). These phenomena will be almost entirely lacking from the loudspeaker stimuli.

In Section 5.2.2, it was demonstrated that the onset detector does not work reliably on the piano stimulus. Therefore, an isolated middle C has been taken from the beginning of each performance as an example. Processed data can be seen in Figures 5.20–5.23. These figures introduce a slightly different method of representing the histogram data, which allows the width of its distribution to be seen more clearly. The middle graph of each figure shows the median [centroid] and lower and upper quartiles of the histogram data. Thus the sum of histogram truth values between two adjacent lines on this graph equals one-quarter of the sum across the whole segment. In these examples, the peak truth values are shown on a numbered scale: this eases the comparison of the four sets of results.

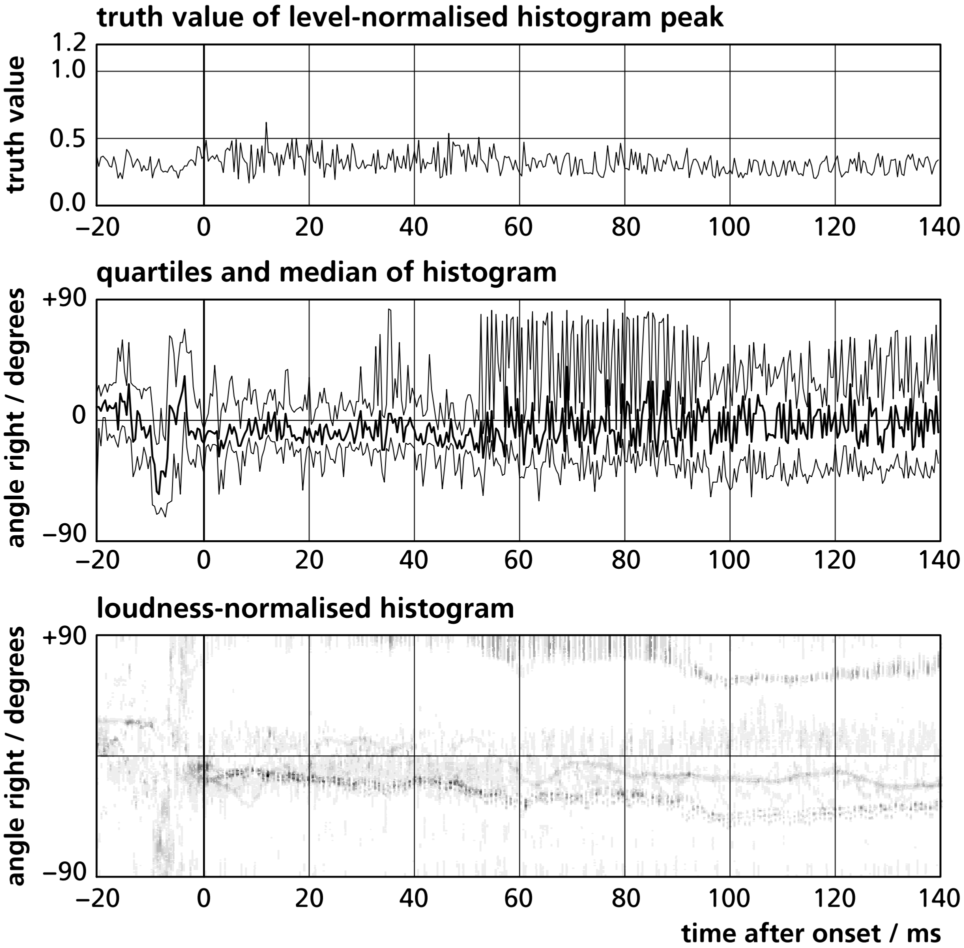


Figure 5.20. Localisation data for middle C, 'piano wide' orientation. In this example, the sound source subtended 45° of the field of audition.

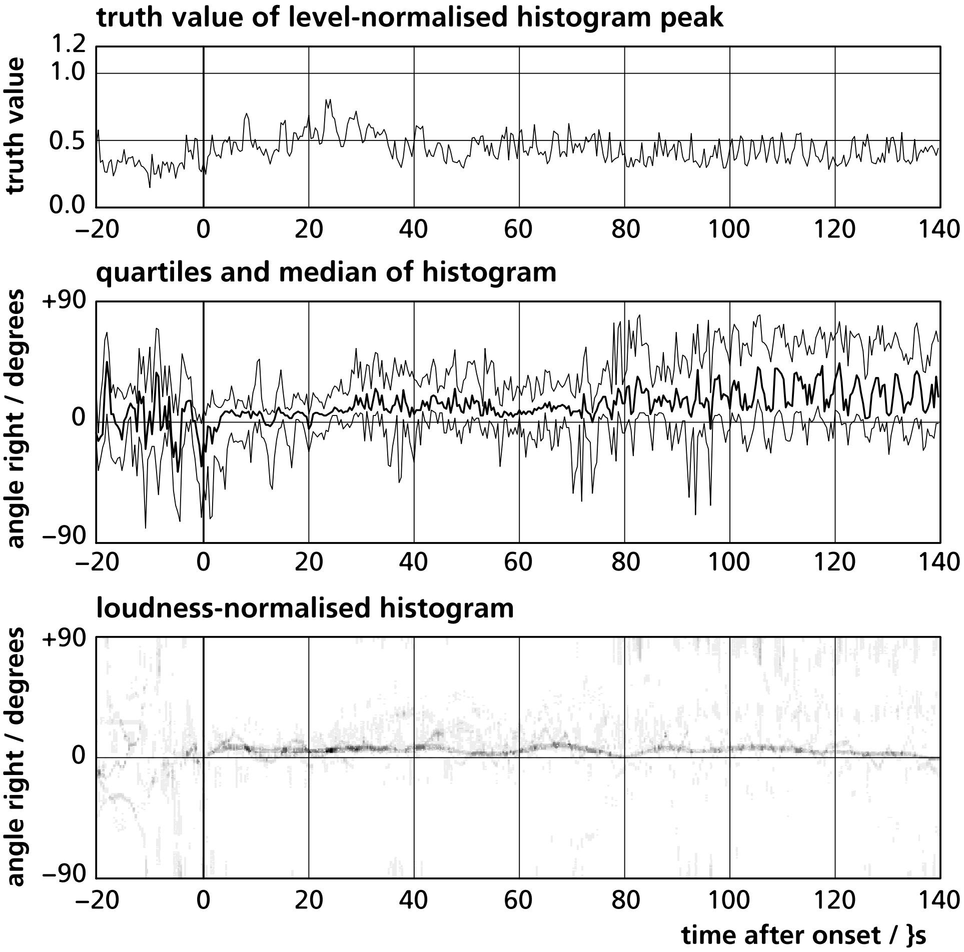


Figure 5.21. Localisation data for middle C, 'piano narrow' orientation. In this example, the sound source subtended 37° of the field of audition.

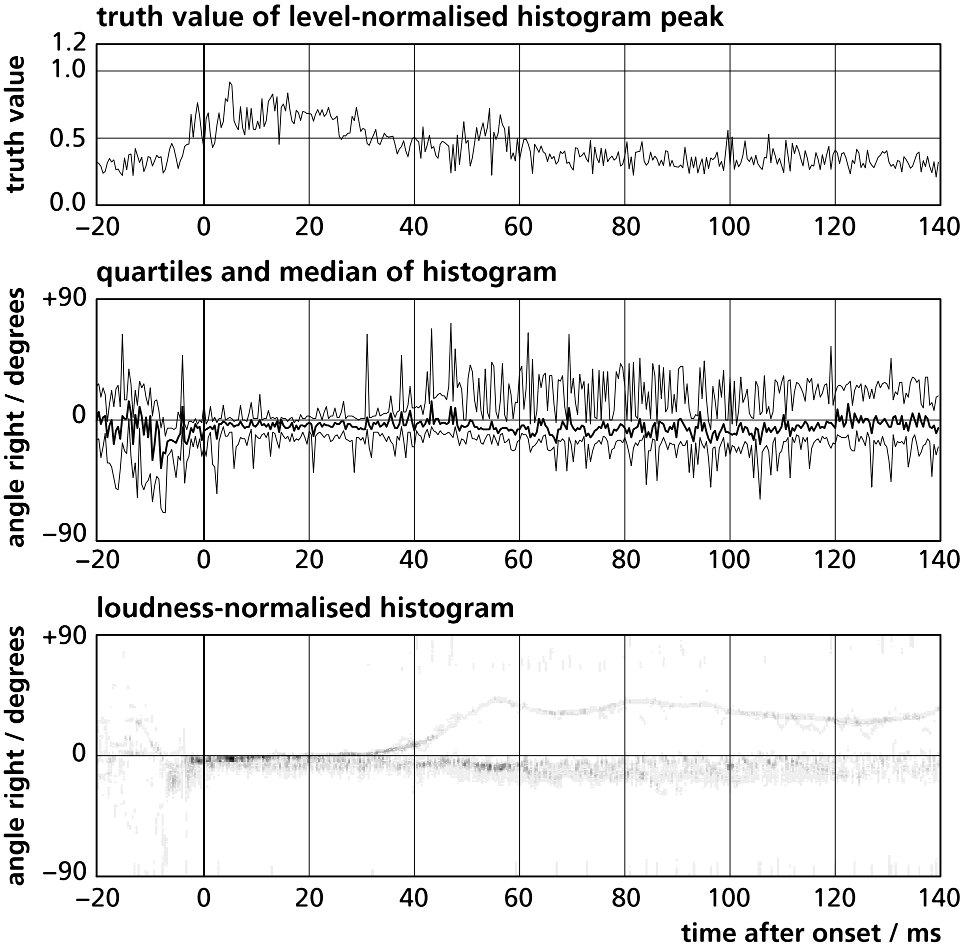


Figure 5.22. Localisation data for middle C, 'loudspeaker wide' orientation. In this example, the sound source subtended 13° of the field of audition.

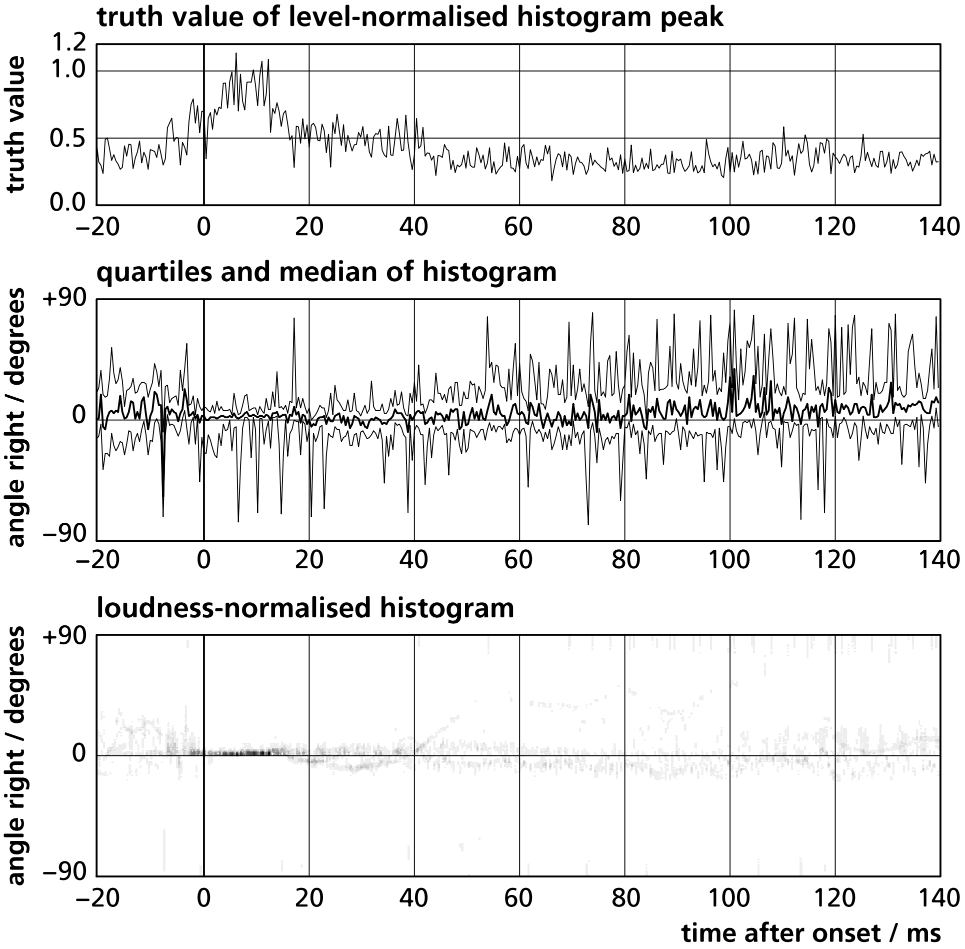


Figure 5.23. Localisation data for middle C, 'loudspeaker narrow' orientation. In this example, the sound source subtended 8° of the field of audition.

This data set is clearly limited in scope: only one note on one class of musical instrument has been tested. However, there are some striking features of the data which demonstrate that the localisation algorithm may be highly suited to extracting source width information. What is most apparent about these results is the strong negative correlation between the source width and the maximum truth value shown within 20ms of onset. The maximum peak value itself is very vulnerable to noise and can be very short-lived, so a more robust statistical measure, the 90th percentile, is employed to relate to the peak value. This is given as the truth value that 10% of peak values within 20ms of the detected onset must equal or exceed. The 90th percentile values are given in Table 5.1.

Stimulus	Actual source width / degrees	Maximum peak within 20ms	90th percentile peak within 20ms
Piano wide	45	0.62	0.49
Piano narrow	37	0.70	0.63
Loudspeaker wide	13	0.92	0.81
Loudspeaker narrow	8	1.13	0.99

Table 5.1. Maximum and 90th percentile peaks of the four width stimuli.

A formula that roughly relates angular width to 90th percentile truth value for the four piano notes is:

$$\phi = \frac{7.5}{A^{2.5}} \tag{5.1}$$

where ϕ is the source width in degrees, and A is the 90th percentile peak truth value. However, other sources, at other angles, will exhibit different characteristics. The peak values observed in Figures 5.3 and 5.6, for example, are not consistent with this rule. Unfortunately, obtaining controllable stimuli to extend this investigation may prove difficult. It is not trivial to contrive controlled stimuli of different widths for any instrument except one that is physically wide in one dimension and narrow in another (such an instrument may have its angular width continuously adjusted by rotating it). This problem could be overcome by processing the audio to create perceptually unidimensional stimuli using the methods formulated by Neher [2004], but the risk of this approach is that any width-detecting algorithm that is formulated may end up simply reverse-engineering the processing, and might not extend validly to other stimuli.

The second suggested approach for extracting source width, by inferring it from the distribution of output histogram data, seems less promising from these results. A difference can be distinguished in the spread of the histograms from the piano and those from the loudspeaker, but this difference is subtle compared with the peak truth value cue, and it is not possible to distinguish the sounds in any more detail. Moreover, it has been observed throughout this chapter that histograms generated from sources placed near the aural axis appear spread between the angle of localisation and the aural axis. It is unlikely that a small increase in the distribution of data that is already widely scattered could be detected reliably and repeatedly.

5.4.2 Manipulation of actual source distance

There are a number of possible ways of estimating source distance in a reverberant room. A method based on the inverse pressure law, that makes informed judgements regarding the loudness of each source, will not be particularly reliable unless the listener or analyser is already familiar with the source.

Conveniently, the source distance attribute may be extracted more reliably from using spatial features of the sound sources. All of the following cues are source distance indicators that could be extracted by the spatial analyser:

- The direct-to-reverberant sound ratio. This may be determined by comparing the loudness of the direct sound (during onset) to the loudness when the peak truth value function has fallen and stabilised, usually after 50–100ms.
- The angle subtended by a source decreases as its distance increases, so its apparent width is an additional distance cue. This change in width is augmented by the deviation of ITD and IID cues from their diffuse-field values when the source becomes closer than about three metres (see Section 4.3.1 for examples of the manifestation of this effect, and Section 4.4.1 for the method used to correct it for diffuse-field conditions). This proximity should cause a stronger reduction in the peak truth value function. Therefore, the method for determining source width is also applicable to estimating distance.
- The onset time of a stimulus will be longer at small distances. This is a manifestation of the increased direct-to-reverberant sound ratio, and also occurs because, in certain circumstances, the act of bringing source and listener closer together removes them away from the

walls and into the centre of the room.

Owing to the simplicity of producing stimuli at various source distances, a large number have been recorded for analysis. Two instruments are included in the data set: a clarinet note, and a single tamborim hit recorded at six different distances (50cm, 1m, 1.5m, 2m, 4m, 6m) and at three different lateral angles (0°, 30° left and 90° left). This makes 36 stimuli: rather too many to illustrate separately, so the data has been processed according to three metrics, each based on one of the hypotheses above, and presented as metric-versus-distance graphs.

The first metric is the 90th percentile peak truth value function. Because this value was discovered to correlate inversely with source width for the piano example in the previous section, it should also correlate directly with source distance. Thus a distant source should have a larger 90th percentile peak truth value than the same source recorded at proximity.

The second metric, called 'direct/reverberant', is the maximum value of the sum of loudness coefficients within 20ms after the detected onset, divided by the average sum of loudness coefficients between 80ms and 100ms after onset. This time constant gives the clarinet note a chance to decay. The metric provides a rough indication of direct-to-reverberant sound ratio.

The final metric, called 'onset sustain', is the last time in the region 50ms after onset for which the peak histogram truth value function is sustained at or above its 0–20ms 90th percentile truth value. This metric is another symptom of the direct-to-reverberant ratio, and it should vary inversely with distance.

This data is presented in Figure 5.24. In these examples, only the first detected onset has been used for calculation. The direct/reverberant metric is particularly promising, as it behaves expectedly for the tamborim stimulus: there is a clear inverse relationship between the metric and distance. However, this relationship does not hold for the clarinet stimulus. This is because the direct sound from the clarinet is still increasing in level 80ms after onset. This explains why the clarinet's direct-to-reverberant metric is maintained around unity, implying that the reverberant sound is louder than the direct sound. What is being measured is the ratio of initial direct sound to late direct sound plus reverberant sound. To provide good direct-to-reverberant data, the auditory offset should be analysed rather than the onset.

The two remaining metrics exhibit random characteristics, irrespective of the input stimulus. This may be expected for the 90th percentile truth value data. Neither the tamborim nor the clarinet possess significant physical width, so a source width metric might not be expected to change with distance. It may still prove effective for wide stimuli.

Disappointingly, the onset sustain metric does not appear to be suited to source distance analysis. For the clarinet stimulus, the 50ms ceiling of the calculating algorithm is reached too frequently for this metric to be useful: the direct sound is still gaining strength after this time, so the peak truth value function continues to increase. A small negative correlation can be seen for the tamborim onset sustain data, but the data has neither the dynamic range nor the strength of correlation of the direct-to-reverberant data.

From these results, it appears that the most effective strategy for estimating the source distance from input data would be to complement the auditory onset detector with an offset detector. and to apply the direct/reverberant ratio metric to auditory offsets.

Estimating the direct/reverberant ratio is a practicable method for determining source distance. This ratio is known to be an important distance cue in human audition [von Békésy 1960; Bronkhurst and Houtgast 1999]. As the direct-to-reverberant ratio data in Figure 5.24 falls with increasing distance and approaches a practical asymptote at around 1.0, it demonstrates consistency with the *auditory horizon* phenomenon [Mershon and Bowers 1979; Bronkhurst and Houtgast 1999]. This states that beyond a certain distance, the value of which is a function of the room radius, changes of actual source distance have a greatly diminished influence on the perceived source distance.

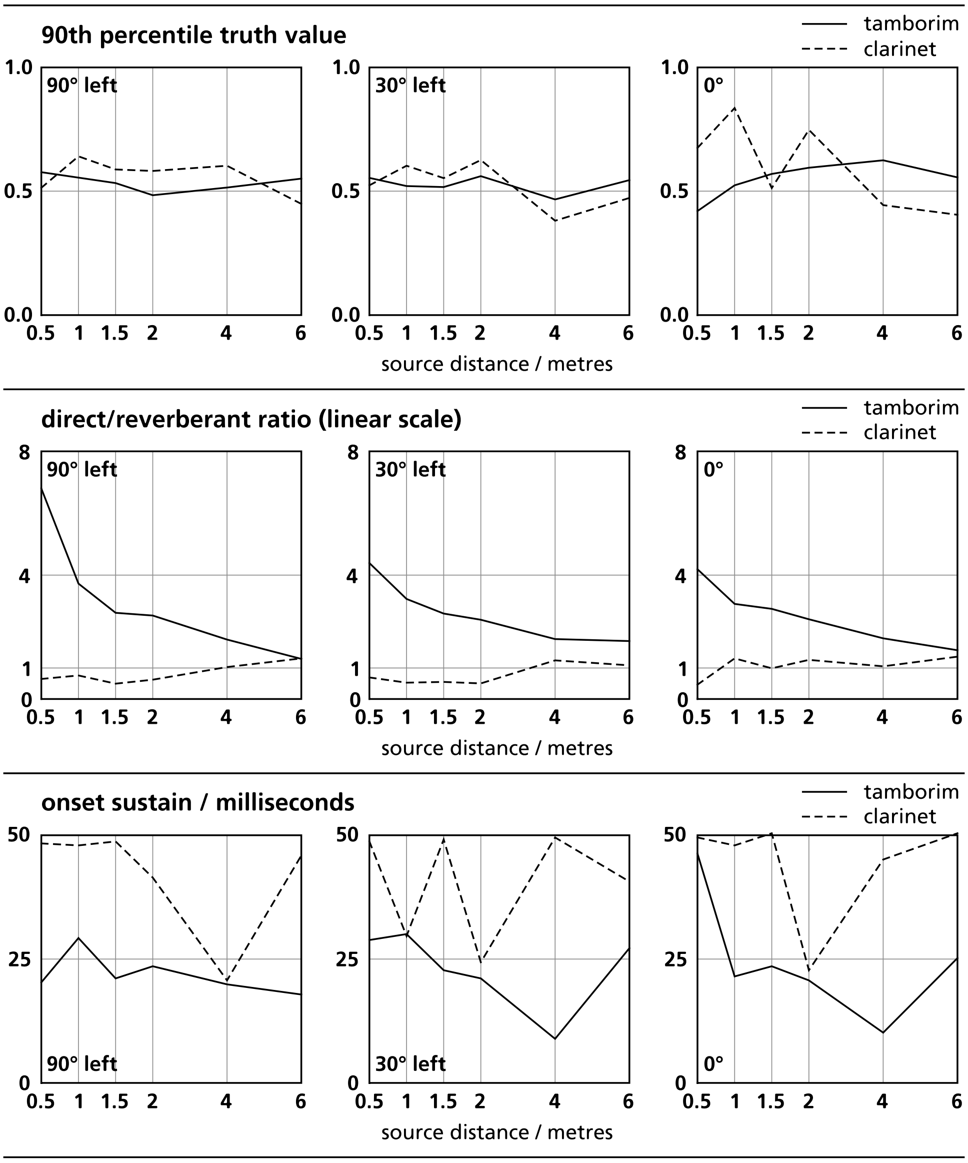


Figure 5.24. Responses of three different proposed source distance metrics to varying source distance. Tamborim and clarinet stimuli.

5.5 Computational demand

An aim of this project is that a real-time implementation of the spatial analyser should be within reach. In order to examine the possibility of a real-time implementation of these algorithms, a table is presented in Table 5.2 of their computational demands in MFLOPS (millions of floating-point operations per second).

Task	Computational demand / MFLOPS	
Filter banks	85.2	
Rectification, filtering, decimation	21.2	
Total for input stage	**106.4**	
Onset detector	**14.8**	
ITD detection	69.8	(50.2)
IID detection	8.1	(8.1)
Loudness calculation	3.2	(3.2)
Lookup, normalisation, summing	42.9	(43.9)
Total for localisation algorithm	**124.1**	**(104.5)**
Grand total	**245.3**	**(225.7)**

Table 5.2. Computational demand of the prototype MATLAB algorithms. MFLOPS data in brackets is for an alternative version of the ITD-extracting algorithm that stores all component correlograms (see the end of Section 4.3.4 for details). However, the memory-swapping overheads of this implementation causes it to run more slowly under MATLAB than the original one.

These data were obtained by applying MATLAB's *flops* command [Mathworks 2004] to each stage of processing of a five-second excerpt of piano music, and dividing all the results by five. For reasons that will be explained, this generates only an approximate indication of the computational requirements of these algorithms, so the results should be interpreted carefully.

The overall computational demand of all prototype algorithms is seen to be approximately 250 MFLOPS. This falls within the published specification of many of the more recent designs of digital signal processor, including the SHARC and C55x processors [Analog Devices 2005; Texas Instruments 2005], both of which have published performance data in excess of 500 MFLOPS. However, an algorithm's load in MFLOPS is not a complete indication of its compatibility with a particular processor. Some operations, such as the repeated multiply-accumulate cycles executed during filtering, can usually be performed on a processor more rapidly than more complicated conditional processing employed, for example, in the onset detector. Although they also take time to perform, memory storage and retrieval are not usually counted in floating-point operation calculations. Furthermore, some arithmetical instructions are not accounted for at all in Table 5.2. (The MATLAB documentation for the *flops* instruction states that 'It is not feasible to count all floating point operations, but most of the important ones are counted'. Note, however, that documentation for *flops* is no longer published as the instruction is obsolete in more recent versions of MATLAB.)

It is important to observe that the algorithms in Table 5.2 generate onset and localisation data, but do not process these in order to make sense of them. However, any such interpretation is expected to be far less demanding than any of those that are currently listed.

A real-time version of the spatial analyser, working on a stream of data, would be a little simpler than this prototype. For example, it would not have to compensate for the limited availability of audio data at the start and the end of the data arrays, as this algorithm does. There are other opportunities for the algorithms to be improved. For example, the filter bank is the most demanding single processing stage, but even the low-frequency filters operate at 44.1kHz. It is probable that a multi-rate approach would run considerably faster. Because this project does not extend to the formulation of a real-time algorithm, this level of optimisation has not been considered.

In order to comment further on the possibility of a real-time implementation, it would be necessary to build a more precise profile of the run-time code, of its demands on memory, and of the kind of instructions that it would employ. Again, an instruction-level profile is beyond the scope of this thesis, but the results in Table 5.2 show that the requirements of the spatial analyser are within the capabilities of current technology.

5.6 Summary

The four experiments in this chapter have demonstrated that the spatial analyser satisfies a number of the objectives of this project. The first two experiments demonstrate that the spatial analyser can localise a variety of sources, and can usually determine the source angle to within 10°. This precision is comparable to human localisation performance. Larger inaccuracies in localisation are attributed to late onset detection or false detection of strong early reflections as separate onsets.

Currently, sources with very short auditory onsets are most vulnerable to mislocalisation. However, localisation ability is improved by the presence of the onset detector. The assumptions made in formulating the localisation algorithm and the ITD and IID look-up table databases were not over-generous, and have had little effect on localisation accuracy.

Two persistent phenomena that affect the output histograms are the drift of the localisation histogram away from the centre of the field of audition the longer onset persists, and the widening and blurring of the edges of the histogram for source angles near the aural axis. These have been attributed, respectively, to the effect of early lateral reflections, and to properties inherent in the trigonometry of spatial hearing.

Problems noted in the localisation of extended stimuli are a result of a high-rate of onset detection: the observed ‘high rate’ is between 5 and 10 onsets per second. Stimuli with fast attack and decay times are affected, and the phenomenon causes spurious onsets to be analysed. This has a detrimental effect on localisation performance. This may be ameliorated by making a number of small improvements to the onset detector. These include the implementation of a masking model that would influence longer-term behaviour of the onset detector, and the implementation of a method for automatically controlling the sensitivity of the onset detection algorithm according to the amount of activity detected within the input signal. These modifications will make the onset detector respond with greater consistency and intelligence to more impulsive sources.

The adaptability of the spatial analyser for secondary spatial attributes has been investigated briefly, with positive results. For the four piano examples tested, the relationship between auditory source width and 90th percentile peak truth value was strikingly inversely proportional, and a simple formula could be used fit the 90th percentile results to the actual width of the source.

Source distance is a more difficult attribute to extract, and of the three metrics that were tested, the direct/reverberant ratio works most satisfactorily. This metric compares integrated loudness measurements at times that are dictated by the spatial analyser. The direct/reverberant ratio is also known to be an important distance cue in human audition. To apply this metric to all sound sources, an offset detector would have to be built to complement the onset detector.

The current implementation of the spatial analyser demands slightly fewer than 250 million floating-point operations per second. It is not possible to comment on the compatability of the prototype routines with any particular digital signal processing hardware without analysing both of these in far more detail, but the spatial analyser's requirements appear to fall firmly within the capabilities of existing technology.

6 CONCLUSION

This conclusion presents a chapter-by-chapter summary of the thesis, drawing out the main conclusions. It ends with suggestions for the continued development of the spatial analyser, and a consideration of the contributions that this thesis has made to the field of spatial auditory analysis.

This research question, proposed in Chapter 1, asked how a spatial attribute extractor may best be realised that suits the requirements of broadcast monitoring. This has been answered by developing the spatial analyser. Its component algorithms meet the requirements of the broadcast monitoring application, and it can localise sound sources and extract secondary spatial attributes.

The closing sections of this chapter describe ways in which the versatility of this spatial analyser can be improved, and propose strategies for developing the secondary spatial attribute extraction mechanisms that were created in this thesis.

6.1 Summary of Chapter 1: Introduction

The aim of this project was to design and present a system, called the *spatial analyser*, that can extract fluctuating spatial attributes from binaural data. These spatial attributes include the most basic auditory scene information — the lateral position of the sound source — and secondary attributes such as apparent source width and distance.

Principal requirements for this system were derived from its intended application in broadcast monitoring technology, and from the need for perceptually accurate spatial attribute extraction. The first of these two applications imposed criteria that are concerned largely with implementation. Processing algorithms must be compatible with streamed data, produce an output with the minimum of delay (less than 33ms) and, to make a real-time implementation practical, be computationally efficient. The second of these applications requires that wherever it is possible, only psychoacoustically-valid algorithms should be used.

In addition to these restrictions, the output from the system must be precise and reliable enough to be useful to an engineer when the source material cannot be heard.

Input audio data is presented in binaural format to achieve the aim of psychoacoustic validity. This requires the algorithm to work by finding and interpreting fluctuations of interaural time and intensity differences — the same cues that are available to the human auditory system. Representation in any other audio format would necessitate a departure from these fundamental psychoacoustic cues, and thus from a perceptually valid representation.

Unfortunately, the binaural format has inherent limitations. When only two ear signals are available for analysis, it is not possible to discriminate reliably between sounds arriving from the front and rear hemispheres, or between elevated sources and those on the horizontal plane. Humans can accomplish this task for two reasons. Firstly, a listener learns the transfer functions of his or her head in response to sounds from different directions. A computer can mimic this task, but algorithms that attempt this are limited to those individual binaural recording heads whose characteristics they have 'learned'. Secondly, even the most sophisticated algorithms cannot perform this task reliably, because human listeners can interact with environments in a way that a computer cannot. Even without visual cues, a human listener can use head movements and other motion-related cues to localise natural sound sources precisely. The drawback of the binaural format is that only one-dimensional localisation (lateralisation) is possible. Every source, including elevated and rear-hemisphere sources, are essentially folded onto the frontal semicircle of the horizontal plane (see Chapter 4.2).

The Zurek model (see Chapter 1.4) is applied as a framework, both for the implementation of the spatial analyser and the chapter divisions of this thesis. Zurek's model requires a dedicated onset detector to make sense of spatial information. Its layout is thus dictated by the precedence effect. This phenomenon was examined in Chapter 2.

6.1.1 Conclusions from Chapter 1

In order to meet the requirements of broadcast monitoring, the spatial analyser must be compatible with streaming audio, and the delay between input and output should be less than the period of one video frame (33ms). The analyser should also work as efficiently as possible, so that a real-time implementation is practicable. The Zurek model provides the most suitable starting point for the design of such an analyser.

Designing an analyser to take binaural data as its input will limit its capabilities to one spatial dimension, but will allow a psychophysically motivated approach to be taken.

6.2 Summary of Chapter 2: Early reflections and spatial impression

The precedence effect is a psychoacoustic inhibition mechanism that affects the perception of sound immediately after the onset of an auditory event. This enables a listener to locate a sound source in a room without being distracted by early reflections, which arrive from different angles. Within 2ms after a fast onset, the precedence effect diminishes both the perceived loudness of further sound energy and the weighting applied to spatial information. Sound arriving at the ears during this period of sensory inhibition is perceived as being spatially fused with the earlier onset.

The period of strongest inhibition lasts approximately 1ms, and the recovery time depends on the amplitude envelope of the input stimulus. The human auditory system regains its acuity within about 10ms when clicks and other short transients are presented. If the sound source is sustained or slow to decay, recovery from the precedence effect can take more than 50ms.

Early reflections cause decorrelation of the two ear signals. Lateral reflections — for example, those from the side walls of a room — cause greater decorrelation than the centrally-positioned first-order reflections from the front wall, back wall, floor and ceiling. Because many of these arrive within the regime of the precedence effect, they are perceptually fused with the direct sound and generate a sense of source breadth.

A number of objective measures attempt to quantify this broadening effect. Most of these rely on the analysis of room impulse responses. In three early measures, the L_f, LF_E, and IACCF, the energy of early reflections up to 80ms is weighted according to angle of incidence, integrated, and then compared with the total unweighted energy of the impulse and its reflections. A cosine or cosine-squared weighting law favours lateral reflections over frontal reflections.

Later-arriving sound energy is not generally perceived as fusing with the direct sound, but is heard as an environmental effect that is termed ‘listener envelopment’ (LEV), so a number of similar measures exist for integrating the later-arriving reflections of an impulse response with directional weighting.

Impulse responses differ from more natural sounds because they possess very quick attack and decay envelopes, and are not sustained. Perceived spatial impression is dependent on factors such as reverberation time, the speed of a music or speech stimulus, and amplitude envelope characteristics of sound sources. Hence, little can be learned from impulse response analysis that can be applied to the human auditory system without heavy interpretation. Similarly, computational impulse response analysis techniques cannot be applied to natural sounds without extensive modification.

At least two attempts have been made to create more sophisticated measures of auditory source width, based on instrumental sounds and continuous white noise instead of impulse responses. Currently, these suffer from a subset of the deficiencies of impulse response testing: the values obtained from these metrics depend heavily on the characteristics of the source material used.

Griesinger has extended the Zurek model. The extra requirements of the Griesinger model include an offset detector, and a method of extracting background spatial impression (BSI). BSI is Griesinger's terminology, but is possibly equivalent to listener envelopment.

A localisation algorithm has been proposed by Faller and Merimaa that obtains its cues entirely from interaural cross-correlation data. The computational simplicity of this approach is attractive, because onset information is obtained from localisation data with very little extra computation. However, the realisation of an onset detector that uses this paradigm, and can adapt without human intervention to different source stimuli, is still as formidable a target as a level-based adaptive design. Furthermore, this mechanism cannot work under monaural conditions without a separate level-based auditory onset detector. However, Faller and Merimaa's approach is worthy of future investigation as an additional onset cue.

6.2.1 Conclusions from Chapter 2

The precedence effect exerts a strong influence on spatial perception. A computer simulation of the precedence effect is therefore a required component of the spatial analyser. This necessitates an auditory onset detector, with sensitivity to different types of auditory event.

Griesinger's theoretical model of spatial perception presents a number of ways of extending Zurek's model to extract secondary spatial attributes. Faller and Merimaa's model presents an interesting approach that could serve as an additional — but not alternative — paradigm.

6.3 Summary of Chapter 3: Onset detection algorithm

An auditory onset is usually defined as beginning of an auditory event. This definition is somewhat simplistic, and a clearer definition was required so that an onset detection algorithm could be formed. Therefore the term *auditory onset* was redefined, for the purposes of this thesis, as the period of time for which directly-arriving sound dominates reflected energy. During an auditory onset, the sound arriving at a listener contains sufficient cues for correct localisation.

The revised definition of auditory onset is compatible with sound localisation and the precedence effect, and means that an onset now signifies a region of time rather than an instant. The detection of the auditory onset at any moment during this region will be sufficient for successful source localisation. The new onset detector that is formulated for this thesis can therefore be assessed according to the precision and consistency of its localisation performance.

This onset detection algorithm combines two different onset metrics. These are calculated individually for the 24 critical bands of each ear signal. The first metric employs a linear regression model. This determines the logarithmic rate of ascent or descent of the input signal. The second effectively uses a non-linear band-pass filter to remove low-frequency changes and short-term fluctuations. The resulting streams of data are thinned and combined into one function of time using fuzzy logic methods and a cascaded 'hold-and-decay' implementation of the precedence effect. This output function's value equals unity at the beginning of a detected onset, and zero at all other times.

6.3.1 Conclusions from Chapter 3

Auditory onsets need to be detected at intervals of 100ms or more to allow the extraction of secondary spatial attributes. An onset occurs whenever direct sound energy dominates over reflected energy. This usually coincides with a rise in input signal level.

No existing algorithm could be found that is both highly sensitive to auditory onsets and robust against noise and spurious input level fluctuations, while maintaining the relatively low detection rate required by spatial attribute extraction. The approach formulated in this chapter fulfils these requirements.

6.4 Summary of Chapter 4: Localisation algorithm

The localisation algorithm finds the instantaneous ITD and IID of each critical band of the binaural input signal at a rate of 2.45kHz. This data is mapped onto an output histogram: a simple map of truth value against lateral angle, with a resolution of one degree. The higher the truth value for a particular lateral angle, the higher the likelihood of localisation to that angle. The ITD-to-angle and IID-to-angle conversion routines use look-up tables derived from a library of KEMAR head-related impulse responses.

This localisation algorithm is divided into two components. The *generative algorithms* build the look-up tables for interaural cue conversion. The *analytical algorithm* is the running spatial analyser. Effort has been invested in simplifying and optimising the component routines of the analytical algorithm, as this must run in real time.

Particular attention has been focused on speeding up the ITD extraction routine, which is based on the interaural cross-correlation function. This is usually computationally intensive, but by reducing the data rate at the input to this function and interpolating the output carefully, high-accuracy localisation is possible with a minimum of data processing.

Each set of ITD and IID data from the 24 critical bands are integrated carefully to reduce the output rate to 2.45kHz. Loudness weighting is applied to each histogram. A reduction in level of 10 phons within a critical band halves its weighting coefficient. This weighting is combined with a novel cross-weighting process that accounts for the relative sensitivity of interaural time and level differences across the frequency spectrum.

6.4.1 Conclusions from Chapter 4

ITD and IID both play important roles in sound localisation. Very few localisation algorithms are sensitive to IIDs, because they cannot easily be converted to localisation data or combined with ITDs.

The two cues can be unified by windowing and integrating ITD and IID data, converting them both to a histogram representation, and then cross-weighting these according to frequency. The localisation algorithm can be made more efficient by carefully reducing the sampling frequency of the input data.

6.5 Summary of Chapter 5: Investigation

Four experiments were conducted, analysing more than one hundred different binaural recordings in order to test the onset detector and localisation algorithms thoroughly.

6.5.1 Localisation precision

The average deviation between actual and calculated source directions is 9°, for a test set of 63 tamborim and square wave stimuli from source directions spread evenly about the horizontal plane. Generally, the localisation error is largest when the source is close to the aural axis. Whether the error pulls the source towards the mid-line or the aural axis depends on the analysis method used.

Some localisation anomalies as large as 18° were found in the data set. All of these occurred near the aural axis. They are attributed to a combination of late onset detection and very short instrumental onset times. An attempt to correct these anomalies by making small automatic changes to the analysis point, depending on the time at which the data was maximally correlated, generated more anomalies than it removed. These anomalies were caused by short-lived chaotic characteristics of the localisation data.

6.5.2 Onset detector performance

The onset detector was tested using a method that is based on the revised definition of *auditory onset* presented in Chapter 3. This method was formulated to test the onset detector's ability to work with the localisation algorithm to locate complex sound sources. Five solo music and speech excerpts were used for analysis, between 11 and 29 seconds long. Musical stimuli were played on different instruments. The onset detector performed well. For all sources except the piano, the difference between the average data before and after the detected onset was striking.

To make the localisation task more difficult, coherent white noise was added to each stimulus. This noise was amplitude-balanced according to the rms level of the source signal. Onset detection and spatial analysis were then performed again. As a control, the spatial analyser was run a third time: this time, the onset detector was disabled, and onset times were randomised. All sources except the piano showed considerable robustness to the noise, and the control signal verified the effectiveness of onset detection. The excerpts with the fewest detected onsets showed the most effective rejection of the distracting noise.

Unfortunately, onset detection did not assist in the localisation of the piano. This instrument has one of the highest average rates of detected onsets in the experiment: 204 in 24 seconds, 31 of which were introduced when the noise signal was added. For the piano stimulus, no significant difference could be detected between the onset-guided localisation data and the randomised-onset data.

6.5.3 Secondary spatial attributes: source width

Existing source width metrics are based on the value of the interaural cross-correlation function (IACCF) or on fluctuations in interaural time difference. By extending this rule, an attempt is made to extract source width information from four different recordings of a piano note. The piano was positioned directly in front of the recording head. Its width was controlled by rotating it by 90° from a sideways-on to a keyboard-on orientation, and then by replacing the piano with a loudspeaker, replaying a prior performance with the loudspeaker in two orientations.

By using the peak truth value function of the histogram, an accurate ranking of the four stimuli was achieved. The peak truth value appears to be a good indicator of auditory source width. It is representative of the interaural coherence of the input signal, and therefore is analogous to the interaural cross-correlation function.

An inverse power formula models the relationship between peak truth value and width in degrees fairly closely. Without alteration, however, this formula is not expected to hold for other instrumental sources, and would not model the same source moved to another listening position.

6.5.4 Secondary spatial attributes: source distance

Three methods were proposed for extracting source distance, each of which corresponded to a theoretically-valid distance cue. These metrics were calculated for a library of 36 stimuli: a clarinet and a tamborim recorded at six distances and three lateral angles.

The direct-to-reverberant ratio was found to correlate best with source distance, particularly for the tamborim stimulus. However, the clarinet stimulus did not produce useful results. This is because its sound level was still increasing several milliseconds after onset. It is hypothesised that source distance can be extracted reliably only from the offset of an auditory event.

Of the remaining distance metrics, peak truth value performed least well. In the previous experiment, this measure was found to vary predictably against source width. It is likely that peak truth value performed poorly as a distance indicator because the tamborim and clarinet are both physically narrow sound sources.

The third proposed metric, 'onset sustain', was based on the length of time for which the truth value stays above a certain threshold after onset. A correlation was found between source distance and onset sustain for the tamborim, but this metric correlated less strongly with actual source distance than the direct-to-reverberant ratio.

Although the direct-to-reverberant ratio metric works as a distance indicator in this implementation, nothing can be inferred about the manner in which the human auditory system determines apparent source distance. This would have to be ascertained using a formal listening experiment.

6.5.5 Computational demand

A real-time version of the MATLAB prototype of the spatial analyser would demand 250 MFLOPS in order to work in real time. Approximately one third of this demand is required by the band-pass filter, and about a quarter is used by the interaural cross-correlation algorithm. It would be possible to code more efficient implementations of both of these algorithms to reduce the required processing power, but the demands of the spatial analyser already fall comfortably within the published performance characteristics of commercially available digital signal processors.

These findings are reassuring, but they are not a complete indication of the spatial analyser's real-time compatibility with existing processors. The demand of these algorithms cannot be completely expressed as a value in MFLOPS, as their speed also depends on their complexity. Furthermore, MATLAB does not count many processes as floating-point operations, although they occupy processor time.

6.5.6 Conclusions from Chapter 5

The localisation algorithm performs well, and is particularly accurate for source positions in the front quadrant of the field of audition. This validates the histogram approach for source localisation. No clear compatibility problems are encountered when analysing stimuli recorded with a Cortex MK2 head using a database based on data recorded with a KEMAR head.

The onset detector is compatible with a variety of source signals, and is largely resistant to the effects of spurious onsets. However, while successful localisation of the piano was possible, the rate of onset detection was too high, and problems ensued during analysis. This could be solved by altering the design of the detector, so that it changes its onset thresholds dynamically.

Peak truth value performs as an excellent source width metric for the four piano stimuli. The direct-to-reverberant ratio, applied to the offset of an auditory event, is the most successful source distance metric of the three that were tested. An offset detector will need to be added to the spatial analyser before source distance can be estimated for non-impulsive stimuli.

The prototype implementation of the spatial analyser would demand 250 MFLOPS to run in real-time. This is quite feasible on modern processors, but owing to the limitations of the MFLOPS measurement, more work would need to be done either to prove this claim entirely satisfactorily, or to achieve a real-time implementation of the spatial analyser on a single processor.

6.6 Continuing and extending the research project

The spatial analyser has been demonstrated to support onset-guided auditory event localisation, and evidence has been presented to prove its compatibility with secondary spatial attribute extraction. However, some system modifications have already been proposed. These would improve the performance of the spatial analyser in its current form. There are also a number of extensions that could be used to develop the model further.

6.6.1 Improving existing functionality

The most important proposed improvement is to the onset detection algorithm, to allow it to adapt the decision maker's parameters continuously to maintain the rate of detected onsets within predefined minimum and maximum limits. This has already been attempted informally (see Section 5.3.1), but will need to be tested and developed further.

Another useful enhancement to the onset detection could cause it to mark auditory onsets as regions in time rather than as singular points. The localisation system could then integrate spatial information over the entire duration of an auditory onset using a precedence effect model, and produce a smoother output. This is a complicated refinement, as it would entail a more advanced precedence effect model.

It is clear that the existing implementation of the precedence effect is rather simplistic, and does not imitate the human echo suppression mechanisms accurately. This could be improved by using a more elaborate algorithm. Depending on the design of the variable-threshold modifications to the onset detector, this refinement may require considerable changes to be made so that the detector continues to detect onsets relatively infrequently.

There is a clear need for an offset detector, which will complement the onset detector. It is hypothesised that this will enable precise source distance and listener envelopment (LEV) extraction. This hypothesis can be tested experimentally, but only when a functional offset detector has been built.

6.6.2 Adding new functionality

A fundamental limitation of the spatial analyser is that it is confined to two spatial dimensions: lateral angle and source distance. This is caused by the use of the binaural recording format (see Chapter 4.2). However, if the model is extended by two audio channels, to include another horizontal head position, the analyser's scope will improve considerably. It will then be possible to differentiate between frontal and rear sources, and to ascertain their angles of elevation. (This system could still not differentiate between dipped sources and elevated sources. This would require a third head position.) Changing the analyser in this way would necessitate a more sophisticated internal representation of auditory space within the model, and a more elaborate technique would be needed to display the output data. The lateral angle histogram can represent only one dimension of space, and would no longer be sufficient.

Unfortunately, adding two audio channels presents several disadvantages. It approximately doubles the signal processing demand, and is a departure from the psychophysically-motivated simplicity of the two-channel spatial analyser. However, these problems could be eliminated by enabling the spatial analyser to select only one head position at any time, in order to correct front-back or elevation ambiguities.

When these problems are overcome, there is still one more: the task of acquiring test stimuli becomes problematic. Four-channel dummy heads have been proposed before (for example, [Kahana et al. 1997]) but none is commercially available. There are no recording techniques from which binaural data can be convolved that can preserve the spatial separation of the ears in the recording environment. Data for analysis could be convolved fairly easily from a multichannel loudspeaker format, but analytical data would then need to be synthesised instead of recorded, and this would affect the validity of any conclusions made about the new analyser.

6.7 Specific contributions to the field of spatial auditory analysis

A number of individual innovations have been necessary to create onset detection and localisation algorithms that would be compatible with secondary spatial attribute extraction.

The onset detector was formulated to solve the problem of secondary spatial attribute analysis. This is a new problem, which consequently required a special definition of *auditory onset* for the purposes of the spatial analyser. The onset detector is a novel algorithm that combines two methods, each of which has been adapted from a number of existing onset detectors. An original model of the precedence effect prevents re-triggering at a rate that would upset secondary spatial attribute extraction.

The localisation algorithm also contains several innovative aspects. Firstly, the histogram conversion using specially-generated databases is unique. Other existing models employ a set of modelling formulae or pre-programmed neural networks to achieve a similar aim. The former technique oversimplifies IID data, and the latter requires a library of training material more extensive than Gardner and Martin’s HRTF database. Furthermore, the individual decision-making mechanisms of a neural network are often impenetrable, and are uninformative about the characteristics of human audition.

The optimisation of the ITD computing algorithm for efficiency is a simple combination of existing techniques — interaural cross-correlation and cubic interpolation — in a novel way. This generates high-quality data using a small fraction of the processing power that would otherwise be required.

To combine ITD and IID histograms across frequency bands, a unique set of weighting coefficients have been formulated, based on inferences from a survey of auditory literature. This *duplex theory weighting* is combined with loudness weighting for each critical band. The algorithm that calculates loudness coefficients has been created specifically for this system. The combination of the two cross-weighting mechanisms works well for a variety of stimuli.

6.7.1 Revisions to Zurek's and Griesinger's models

In the summary of Chapter 5, an extension was suggested to Zurek's model to include auditory offset detection, to assist in the extraction of source distance. The proposed extension necessitates a more elaborate treatment of auditory offsets than Griesinger's model includes (see Section 2.3). Griesinger's model considers only the late-arriving energy that is perceived as background spatial impression (BSI).

An approach that has been taken implicitly in this project also conflicts with the workings of Griesinger's model. While Griesinger assumes that two separate mechanisms are responsible for determining source location and secondary spatial attributes, they are integrated in this thesis: the same mechanism has been used to determine both source location and secondary spatial attributes. A proposed redrafting of Griesinger's model, that includes these findings and alternative approaches, is shown in Figure 6.1. This paradigm would form the basis of future revisions to the spatial analyser.

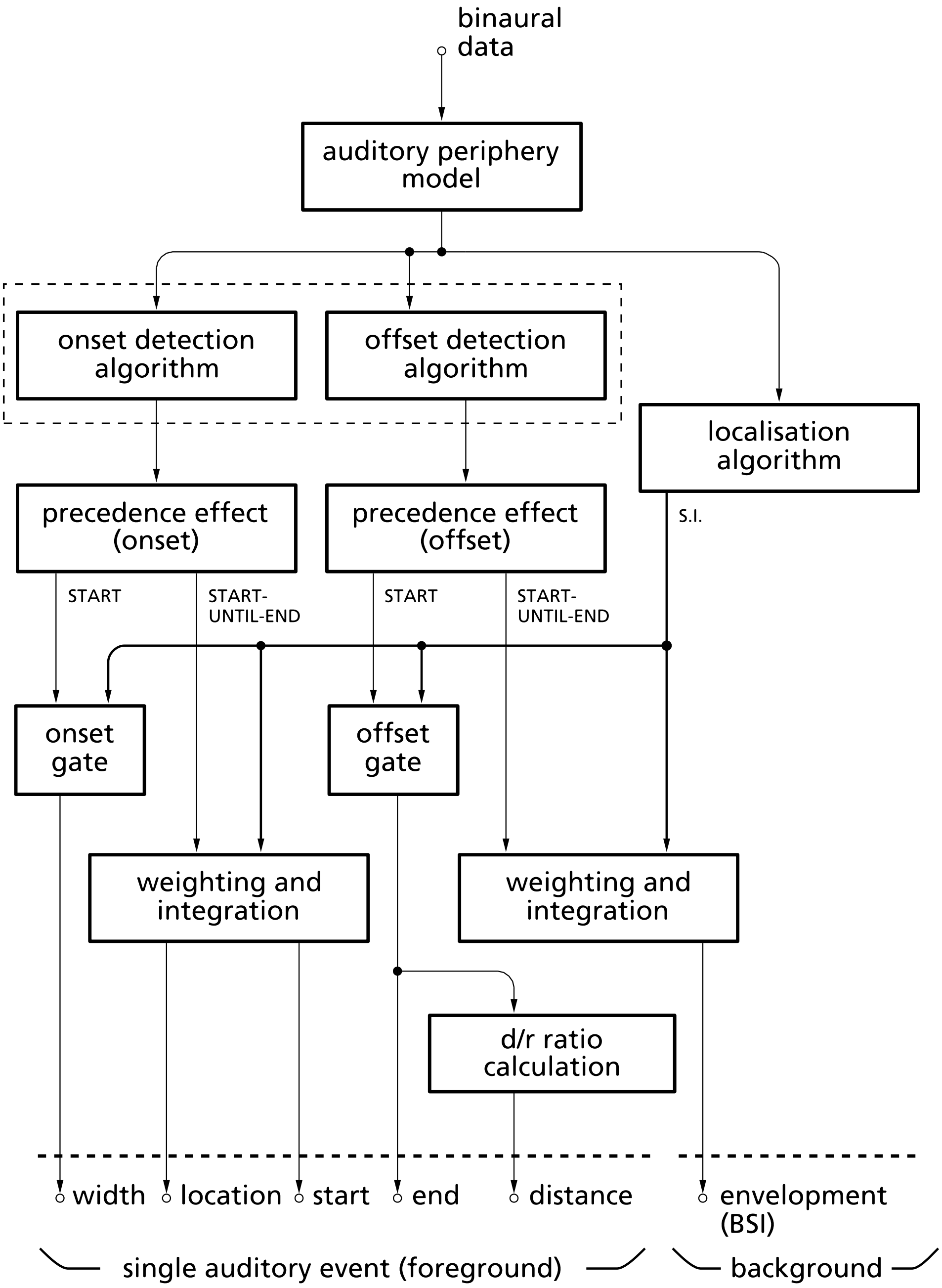


Figure 6.1. A redrafting of Griesinger's model, based on the conclusions of this thesis.

6.8 General contributions to the field of spatial auditory analysis

The spatial analyser (SA) designed in this project is original in several ways. In the field of secondary spatial attribute extraction, the SA is one of only two existing algorithms that have demonstrated an ability to extract secondary spatial attributes from arbitrary binaural signals. (The other is Mason's IACCFF: see Section 2.2.2.) Therefore, it is the only computer algorithm to employ an onset detector and a precedence effect algorithm to assist in the estimation of source width.

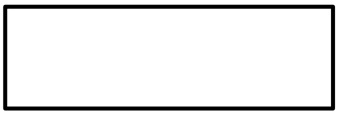
There are a number of localisation algorithms that set out to imitate the human auditory system more accurately than the SA (these were considered in Chapter 4). Currently, none of these are explicitly sensitive to auditory onsets. Hence, unlike the SA, they cannot be expanded to consider secondary spatial attributes.

With its ability to unite sensitivity to IIDs and ITDs, its awareness of auditory onsets, and its implementation in a real-time compatible algorithm, there are very few projects that are comparable to this spatial analyser. Its workings, its design objectives, and its combination of techniques, are therefore unique.

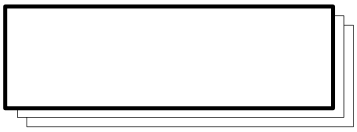
7 APPENDICES

A Key to flowcharts

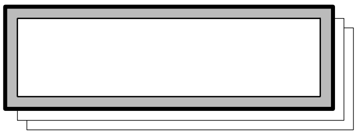
A.1 Boxes



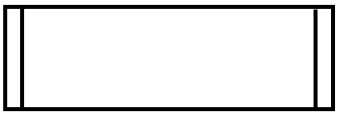
Standard (serial) process



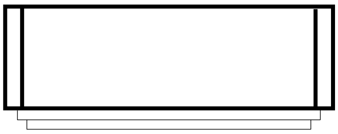
Parallel process
(a serial process that is executed on each element of parallel input data)



Data stored in a look-up table

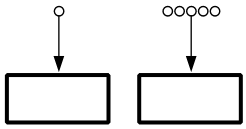


A sub-routine: these are documented in their own flowcharts

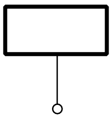


A sub-routine that generates parallel data

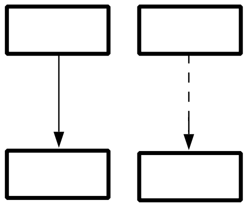
A.2 Flow symbols



Input data / Multi-dimensional input data

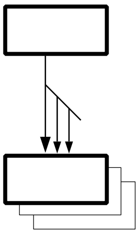


Output data

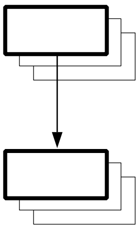


Standard flow between processes

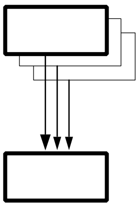
(dotted lines are sometimes used for clarity, to denote less important processes in complicated flowcharts)



Parallel data emerging from a serial process and entering a parallel process



Data flowing between parallel processes



Parallel data channelled into a serial process

B Publications resulting from this research

B.1 Journal paper

Supper, B., Brookes, T., and Rumsey, F. 'An auditory onset detection algorithm for improved automatic source localization'. *IEEE Trans. Speech and Audio Processing.* Accepted for publication.

B.2 Conference papers

Supper, B., Brookes, T., and Rumsey, F. 'A lateral angle tool for spatial auditory analysis.' Presented at the *AES 116th Convention*, Berlin, Germany, May 2004, preprint 6068.

Supper, B., Brookes, T., and Rumsey, F. 'A new approach to detecting auditory onsets within a binaural stream'. Presented at the *AES 114th Convention*, Amsterdam, The Netherlands, March 2003, preprint 5767.

8 GLOSSARY

ASW

Auditory (or apparent) source width. A secondary spatial attribute (q.v.).

aural axis

An imaginary line that passes through both ears. When an external sound source is positioned on the aural axis, it is placed either 90° left or 90° right.

binaural

A recording system designed for headphone reproduction. Binaural recordings are made using a model of a human head and torso with microphones inside its ears.

BSI

Background spatial impression: another term for auditory source width.

concha

The bowl-like portion of the outer ear that lies below the ear canal.

correlogram

A function of interaural piecewise product against interaural time difference (this axis is usually symbolised by τ) and running time (symbolised by t). The peak value of this function yields interaural time difference on the τ axis.

cross-correlation

A measure of similarity of two signals, taken by measuring their piecewise product at different time offsets, and normalising this data for signal level. Two identical signals will have a cross-correlation of 1, whereas two unrelated signals will have a cross-correlation close to zero.

dBFS

Decibel with respect to full-scale deflection. Hence, the largest signal that can be represented by a digital system has an amplitude of 0dB.

echo suppression

The neurophysical inhibition of energy that arrives immediately after the onset of a new auditory event. This complicated mechanism prevents early reflections from disrupting localisation when listeners are placed in small rooms.

head related transfer function

A characteristic pattern of frequency-domain and phase distortion that the outer ear and ear canal apply to a sound coming from a particular direction.

head shadowing

The attenuation of sound waves at one ear, and their reinforcement at the other, when a listener's head behaves as an acoustic baffle. This only happens at higher frequencies (greater than approximately 800Hz).

HRIR

Head-related impulse response. A binaural recording of an impulse response coming from a certain direction.

HRTF

Head-related transfer function (q.v.).

IACCF

Interaural cross-correlation function.

IID

Interaural intensity difference.

ITD

Interaural time difference.

j.d.

Judgement decision: a dimensionless quotient in which the energy difference of the two ear signals is divided by their energy sum.

JND

Just-noticeable difference. An experimental method of determining auditory sensitivity.

KEMAR

Knowles Electronics Manikin for Acoustic Research. A dummy recording head and torso.

lateralisation

The localisation of a sound expressed as an angle relative to the aural axis.

LEV

Listener envelopment. A secondary spatial attribute.

melodica

A blown instrument with vibrating metal reeds and piano-style keys. Its sound is similar to that of a mouth-organ.

MFLOPS

Millions of floating-point operations per second. An approximate indication of processing speed or computational demand.

pinna

The external, visible part of the outer ear.

precedence effect

A low-level, fast-acting, short-lived form of echo suppression (q.v.). As a result of this phenomenon, the localisation cues of early reflections do not have the perceived strength of those from the direct sound.

pre-masking

The phenomenon by which an otherwise audible auditory event is rendered inaudible by the presence a louder event that slightly precedes it.

reversal

A specific problem in binaural localisation. A reversal has occurred when the sound source is localised in front of a listener when it has actually been placed behind, or vice versa.

room radius

The distance between a source and observer in a given room, at which the direct sound energy equals the incident reverberant sound energy.

secondary spatial attribute

A spatial attribute of sound other than source direction (the primary spatial attribute).

SSP

Spatial sampling period, equal in this algorithm to one period of a 2.45kHz cycle (approximately 410µs).

tamborim

A small frame drum used in samba percussion. It has no snares or jingles, so it produces an impulsive, slightly tuned sound.

9 REFERENCES

[Analog Devices 2005]

Analog Devices. *SHARC Processor Home.* Web site, 2005, http://www.analog.com/processors/processors/sharc/

[Ando 1977]

Ando, Y. ‘Subjective preference in relation to objective parameters of music sound fields with a single echo’. *J. Acoust. Soc. Am.*, Vol. 62, No. 6, December 1977, pp. 1436–1441.

[Ando and Gottlob 1979]

Ando, Y., and Gottlob, D. ‘Effects of early multiple reflections on subjective preference judgements of music sound fields’. *J. Acoust. Soc. Am.*, Vol. 65, No. 2, February 1979, pp. 524–527.

[Backman and Karjalainen 1993]

Backman, K., and Karjalainen, M. ‘Modelling of human and spatial hearing using neural networks’. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP’93)*, 1993, pp. I.125–I.128

[Barron 1971]

Barron, M. ‘The subjective effects of first reflections in concert halls — the need for lateral reflections’. *J. Sound and Vibration*, Vol. 15, No. 4, April 1971, pp. 475–494.

[Barron and Marshall 1981]

Barron, M., and Marshall, A. H. ‘Spatial impression due to early lateral reflections in concert halls: the derivation of a physical measure’. *J. Sound and Vibration*, Vol. 77, No. 2, 1981, pp. 211–232.

[**Begault et al. 2001]**

Begault, D. R., Wenzel, E. M., and Anderson, M. R. ‘Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source’. *J. Aud. Eng. Soc.*, Vol. 49, No. 10, October 2001, pp. 904–916.

[von Békésy 1960]

von Békésy, G. *Experiments in Hearing.* New York: McGraw-Hill, 1960.

[Bello and Sandler 2003]

Bello, J. P., and Sandler, M. 'Phase-based note onset detection for music signals'. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP-03),* April 2003.

[Blauert 1997]

Blauert, J. *Spatial Hearing — The Psychophysics of Human Sound Localization.* Cambridge, Massachusetts: MIT Press, 1987.

[Blauert and Cobben 1978]

Blauert, J., and Cobben, W. 'Some consideration of binaural cross-correlation analysis'. *Acustica,* Vol. 39, 1978, pp. 96–104.

[Blauert and Lindemann 1986]

Blauert, J., and Lindemann, W. 'Auditory spaciousness: some further psychoacoustic analyses'. *J. Acoust. Soc. Am.*, Vol. 80, No. 2, August 1986, pp. 533–542.

[Blauert et al. 1986]

Blauert, J., Möbius, U., and Lindemann, W. 'Supplementary Psychoacoustical Results on Auditory Spaciousness'. *Acustica,* Vol. 59, 1986, pp. 292–293.

[Bodden 1998]

Bodden, M. 'Auditory Models for Spatial Impression, Envelopment, and Localization'. *Proc. AES 15th International Conference,* October–November 1998, pp. 150–156.

[Boone and Helleman 2004]

Boone, M. M., and Helleman, H. W. 'Audibility thresholds of spatial variations in a single acoustic reflection'. Presented at the *AES 116th Convention,* Berlin, Germany, May 2004, preprint 5999.

[Bradley 1994]

Bradley, J. S. 'Comparison of concert hall measurements of spatial impression'. *J. Acoust. Soc. Am.*, Vol. 96, No. 6, December 1994, pp. 3525–3535.

[Bradley and Soulodre 1995]

Bradley, J. S., and Soulodre, G. A. 'The influence of late arriving energy on spatial impression'. *J. Acoust. Soc. Am.*, Vol. 97, No. 4, April 1995, pp. 2263–2271.

[Breebaart et al. 2001]

Breebaart, J., van de Par, S., and Kohlrausch, A. 'Binaural processing model based on contralateral inhibition. I. Model Structure'. *J. Acoust. Soc. Am.*, Vol. 110, No. 2, August 2001, pp. 1074–1086.

[Bronkhurst and Houtgast 1999]

Bronkhurst, A. W., and Houtgast, T. 'Auditory distance perception in rooms'. *Nature*, Vol. 397, 11 February 1999, pp. 517–520.

[BS ISO 226:2003]

BS ISO 226:2003. 'Acoustics. Normal equal-loudness-level contours'. British Standards Publishing Limited.

[BS EN ISO 3392:2000]

BS EN ISO 3382:2000. 'Acoustics. Measurement of the reverberation time of rooms with reference to other acoustical parameters'. British Standards Publishing Limited.

[Buell et al. 1994]

Buell, T. N., Trahiotis, C., and Bernstein, L.R. 'Lateralization of bands of noise as a function of combinations of interaural intensive differences, interaural temporal differences, and bandwidth'. *J. Acoust. Soc. Am.*, Vol. 95, No. 3, March 1994, pp. 1482–1489.

[Burkhard and Sachs 1975]

Burkhard, M. D., and Sachs, R. M. 'Anthropometric manikin for acoustic research'. *J. Acoust. Soc. Am.*, Vol. 58, No. 1, July 1975, pp. 214–222.

[Clifton and Freyman 1997]

Clifton, R. K., and Freyman, R. L. 'The Precedence Effect — Beyond Echo Suppression'. In *Binaural and Spatial Hearing in Real and Virtual Environments* , R. Gilkey, T. Anderson, Eds. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc., 1997, pp. 233–255.

[Colburn and Durlach 1978]

Colburn, H., and Durlach, N. 'Models of Binaural Interaction'. In *Handbook of perception — Volume 4: Hearing.* R. Carterette, and M. Friedman, Eds. New York: Academic Press, 1978, pp. 467–518.

[Dixon 2001]

Dixon, S. 'Learning to Detect Onsets of Acoustic Piano Tones'. Presented at *MOSART Workshop on Current Directions in Computer Music*, Barecelona, 2001.

[Domnitz 1973]

Domnitz, R. ‘The interaural time jnd as a simulataneous function of interaural time and interaural amplitude’. *J. Acoust. Soc. Am.*, Vol. 53, No. 6, June 1973, pp. 1549–1552.

[Durlach 1963]

Durlach, N. I. ‘Equalization and Cancellation Theory of Binaural Masking-Level Differences’. *J. Acoust. Soc. Am.*, Vol. 35, No. 8, August 1963, pp. 1206–1218.

[Faller and Merimaa 2004]

Faller, C., and Merimaa, J. ‘Source localization in complex listening situations: Selection of binaural cues based on interaural coherence’. *J. Acoust. Soc. Am.*, Vol. 116, No. 5, November 2004, pp. 3075–3089.

[Fitzgerald 2002]

Fitzgerald, R. ‘Inihibition in the Brain Plays a Key Role in Sound Localization’. *Physics Today,* October 2002, pp. 13–14.

[Franssen 1960]

Franssen, N. V. *Some considerations on the mechanism of directional hearing.* Ph.D. thesis, Technische Hogeschool, Delft, The Netherlands, 1960.

[Gaik 1993]

Gaik, W. ‘Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling’. *J. Acoust. Soc. Am.*, Vol. 94, No. 1, July 1993, pp. 98–110.

[Gardner and Martin 1994]

Gardner, B., and Martin, K. ‘HRTF Measurements of a KEMAR Dummy Head Microphone’. Web site, 1994,
http://sound.media.mit.edu/KEMAR.html

[Griesinger 1997]

Griesinger, D. ‘The psychoacoustics of apparent source width, spaciousness and envelopment in performance spaces’. *Acta Acustica,* Vol. 83, No. 4, 1997, pp. 721–731.

[Griesinger 1998]

Griesinger, D. ‘General Overview of Spatial Impression, Envelopment, Localization and Externalization’. *Proc. AES 15th International Conference,* October–November 1998, pp. 136–149.

[Griesinger 1999]

Griesinger, D. ‘Objective Measures of Spaciousness and Envelopment’. *Proc. AES 16th International Conference*, April 1999, pp. 27–41.

[Gurney 1997]

Gurney, K. *An Introduction to Neural Networks.* London: UCL Press, 1997.

[Haas 1951]

Haas, H. ‘The influence of a single echo on the audibility of speech’. *J. Aud. Eng. Soc.*, Vol. 20, No. 2, March 1972, pp. 146–159. (First published as ‘Über den Einfluss des Einfachechos auf die Hörsamkeit von Sprache’. *Acustica*, Vol. 1, 1951, pp. 49–58).

[Hafter and Carrier 1972]

Hafter, E. R., and Carrier, S. C. ‘Binaural interaction in low frequency stimuli: the inability to trade time and intensity completely’. *J. Acoust. Soc. Am.*, Vol. 51, No. 6, 1972, pp. 1852–1862.

[Hafter and Jeffress 1968]

Hafter, E. R., and Jeffress, L. A. ‘Two-Image Lateralization of Tones and Clicks’. *J. Acoust. Soc. Am.*, Vol. 44, No. 2, August 1968, pp. 563–569.

[Hancock and Delgutte 2004]

Hancock, K. E., and Delgutte, B. ‘A Physiologically Based Model of Interaural Time Difference Discrimination’. *J. Neuroscience*, Vol. 24, No. 32, 11 August 2004, pp. 7110–7117.

[Hartmann 1983]

Hartmann, W. M. ‘Localization of sound in rooms’. *J. Acoust. Soc. Am.*, Vol. 74, No. 5, November 1983, pp. 1380–1391.

[Hartmann 1997]

Hartmann, W. M. ‘Listening in a Room and the Precedence Effect’. In *Binaural and Spatial Hearing in Real and Virtal Environments* , R. Gilkey, T. Anderson, Eds. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc., 1997, pp. 191–210.

[Hartmann and Constan 2002]

Hartmann, W. M., and Constan, Z. A. ‘Interaural level differences and the level-meter model’. *J. Acoust. Soc. Am.*, Vol. 112, No. 2, September 2002, pp. 1037–1045.

[Hidaka et al. 1995]

Hidaka, T., Beranek, L. L., and Okano, T. ‘Interaural cross-correlation, lateral fraction, and low- and high-frequency sound levels as measures of acoustical quality in concert halls’. *J. Acoust. Soc. Am.*, Vol. 98, No. 2, Pt. 1, August 1995, pp. 988–1007.

[Horbach et al. 1999]

Horbach, U., Karamustafaoglu, A., Pellegrini, R., Mackensen, P., and Theile, G. ‘Design and Applications of a Data-based Auralisation System for Surround Sound’. Presented at the *AES 106th Convention,* Munich, Germany, April 1999, preprint 4976.

[Huang et al. 1997]

Huang, J., Ohnishi, N., and Sugie, N. ‘Sound Localization in Reverberant Environment based on the Model of the Precedence Effect’. *IEEE Trans. Instrumentation and Measurement,* Vol. 46, No. 4, August 1997, pp. 842–846.

[Ifeachor and Jervis 1993]

Ifeachor, E., and Jervis, B. W. *Digital Signal Processing: A Practical Approach,* Harlow: Addison-Wesley, 1993.

[Jeffress 1948]

Jeffress, L. A. ‘A place theory of sound localization’. *J. Comparative Physiology and Psychology,* Vol. 41, 1948, pp. 35–39.

[Kahana et al. 1997]

Kahana, Y., Nelson, P. A., Kirkeby, O., Hamada, H. ‘Multichannel Sound Reproduction Using a Four-Ear Dummy Head’. Presented at the *AES 102nd Convention,* Munich, Germany, March 1997, preprint 4465.

[Karjalainen 1996]

Karjalainen, M. ‘A binaural auditory model for sound quality measurements and spatial hearing studies’. *Proc. IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP’96)*, 1996, pp. 985–988.

[Keys 1981]

Keys, R. G. ‘Cubic Convolution Interpolation for Digital Image Processing’. *IEEE Trans. Acoustics, Speech, and Signal Processing,* Vol. ASSP-29, No. 6, December 1981, pp. 1153–1160.

[Klapuri 1999]

Klapuri, A. ‘Sound Onset Detection by Applying Psychoacoustic Knowledge’. In *Proc. ICASSP 1999*, Vol. VI, pp. 3089–3092, 1999.

[Kleiner 1989]

Kleiner, M. ‘A new way of measuring the lateral energy fraction’. *Applied Acoustics*, Vol. 27, 1989, pp. 321–327.

[Kunz and Bodden 1996]

Kunz, O., and Bodden, M. ‘Ein rechenzeiteffizientes Modell zur Lokalisation von Schallquellen in Realzeit’. *Fortschritte der Akustik — DAGA’96*, DPG-GmbH, Bad Honnef, pp. 364–365.

[Lehmann 1999]

Lehmann, T. M. ‘Survey: Interpolation Methods in Medical Image Processing’. *IEEE Trans. Medical Imaging*, Vol. 18, No. 11, November 1999, pp. 1049–1075.

[Lindemann 1986]

Lindemann, W. ‘Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals’. *J. Acoust. Soc. Am.*, Vol. 80, No. 6, December 1986, pp. 1608–1622.

[Lyon 1983]

Lyon, R. F. ‘A Computational Model of Binaural Localization and Separation’. *Proc. IEEE Conf. on Acoustics, Speech and Signal Processing (ICASSP’83)*, April 1983, pp. 1148–1151.

[McAlpine et al. 2001]

McAlpine, D., Jiang, D., and Palmer, A. R. ‘A neural code for low-frequency sound localization in mammals’. *Nature Neuroscience*, Vol. 4, No. 4, April 2001, pp. 396–401.

[Macpherson 1991]

Macpherson, E. A. ‘A Computer Model of Binaural Localization for Stereo Imaging Measurement’. *J. Aud. Eng. Soc.*, Vol. 39, No. 9, September 1991, pp. 604–622.

[**Macpherson and Middlebrooks 2002]**

Macpherson, E. A., and Middlebrooks, J. C. ‘Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited’. *J. Acoust. Soc. Am.*, Vol. 111, No. 5, May 2002, pp. 2219–2236.

[Maeland 1988]

Maeland, E. ‘On the Comparison of Interpolation Methods’. *IEEE Trans. Medical Imaging*, Vol. 7, No. 3, September 1988, pp. 213–217.

[Makous and Middlebrooks 1990]

Makous, J. C., and Middlebrooks, J. C. ‘Two-dimensional sound localization by human listeners’. *J. Acoust. Soc. Am.*, Vol. 87, No. 5, May 1990, pp. 2188–2200.

[Marks 1978]

Marks, L. E. ‘Binaural summation of the loudness of pure tones’. *J. Acoust. Soc. Am.*, Vol. 64, No. 1, July 1978, pp. 107–113.

[Marolt et al. 2002]

Marolt, M., Kavcic, A., and Privosnik, M. ‘Neural Networks for Note Onset Detection in Piano Music’. Web site, 2002,
http://lgm.fri.uni-lj.si/~matic/research.html

[Martin 1995a]

Martin, K. D. *A Computational Model of Spatial Hearing.* Masters Thesis, Massachusetts Institute of Technology, 1995.

[Martin 1995b]

Martin, K. D. ‘Estimating Azimuth and Elevation from Interaural Differences’. Presented at *IEEE Mohonk workshop on Applications of Signal Processing to Acoustics and Audio*, October 1995.

[Mason 2002]

Mason, R. *Elicitation and measurement of auditory spatial attributes in reproduced sound.* Ph.D. thesis, Institute of Sound Recording, University of Surrey, 2002.

[Mason and Rumsey 2001]

Mason, R., and Rumsey, F. ‘Interaural time difference fluctuations: their measurement, subjective perceptual effect, and application in sound reproduction’. *Proc. AES 19th International Conference*, June 2001, pp. 252–271.

[Mathworks 2004]

The Mathworks. Signal Processing Toolbox. Web site, 2004,
http://www.mathworks.com/products/signal/

[Meddis et al. 1990]

Meddis, R., Hewitt, M. J., and Shackleton, T. M. ‘Implementation details of a computation model of the inner hair-cell/auditory-nerve synapse’. *J. Acoust. Soc. Am.,* Vol. 87, No. 4, April 1990, pp. 1813–1816.

[Mellinger 1991]

Mellinger, D. K. *Event formation and separation in musical sound.* Ph.D. thesis, Center for Computer Research in Music and Acoustics (CCRMA), Department of Music, Stanford University, December 1991.

[Mershon and Bowers 1979]

Mershon, D. H., and Bowers, J. N. 'Absolute and relative cues for the auditory perception of egocentric distance'. *Perception,* Vol. 8, 1979, pp. 311–322.

[Minnaar et al. 2001]

Minnaar, P., Olesen, S. K., Christensen, F., and Møller, H. 'Localization with Binaural Recordings from Artificial and Human Heads'. *J. Aud. Eng. Soc.*, Vol. 49, No. 5, May 2001, pp. 323–336.

[Møller et al. 1995]

Møller, H., Sørensen, M. F., Hammershøi, D., and Jensen, C. B. 'Head-Related Transfer Functions of Human Subjects'. *J. Aud. Eng. Soc.*, Vol. 43, No. 5, May 1995, pp. 300–321.

[Møller et al. 1996]

Møller, H., Sørensen, M. F., Jensen, C, B., and Hammershøi, D. 'Binaural Technique: Do We Need Individual Recordings?' *J. Aud. Eng. Soc.*, Vol. 44, No. 6, June 1996, pp. 451–469.

[Møller et al. 1999]

Møller, H., Hammershoi, D., Jensen, C. B., and Sørensen, M. F. 'Evaluation of Artificial Heads in Listening Tests'. *J. Aud. Eng. Soc.*, Vol. 47, No. 3, March 1999, pp. 83–100.

[Moore 2000]

Moore, B. C. J. *Introduction to the Psychology of Hearing.* London: Academic Press, 2000.

[Morimoto 1997]

Morimoto, M. 'The role of Rear Loudspeakers in Spatial Impression'. Presented at the *AES 103th Convention,* New York, September 1997, preprint 4554.

[Morimoto 2002]

Morimoto, M. 'The relation between spatial impression and the precedence effect'. Presented at *International Conference on Auditory Displays (ICAD 2002),* Kyoto, Japan, July 2002.

[Mountain and Hubbard 1996]

Mountain, D. C., and Hubbard, A. E. ‘Computational Analysis of Hair Cell and Auditory Nerve Processes’. In *Auditory Computation,* H. L. Hawkins, T. A. McMullen, A. N. Popper, and R. R. Fay, Eds. New York: Springer-Verlag, 1996, pp. 121–156.

[Nandy and Ben-Arie 2001]

Nandy, D., and Ben-Arie, J. ‘Neural Models for Auditory Localization Based on Spectral Cues’. *Neurological Research,* Vol. 23, No. 5, July 2001, pp. 489–500.

[Neher 2004]

Neher, T. *Towards a Spatial Ear Trainer.* Ph.D. thesis, Institute of Sound Recording, University of Surrey, 2004.

[Palomäki et al. 1999]

Palomäki, K., Pulkki, V., and Karjalainen, M. ‘Neural network approach to analyze spatial sound’. *Proc. 16th AES International Conference: Spatial Sound Reproduction,* March 1999, pp. 233–245.

[Palomäki et al. 2004]

Palomäki, K. J., Brown, G. J., and Wang, D. ‘A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation’. *Speech Communication,* Vol. 43, 2004, pp. 361–378.

[Pratt et al. 1997]

Pratt, H., Polyakov, A., and Kontrovich, L. ‘Evidence for separate processing in the human brainstem of interaural intensity and temporal disparities for sound lateralization’. *Hearing Research,* Vol. 108, No. 1, June 1997, pp. 1–8.

[Rakerd and Hartmann 1985]

Rakerd, B., and Hartmann, W. M. ‘Localization of sound in rooms, II: The effects of a single reflecting surface’. *J. Acoust. Soc. Am.*, Vol. 78, No. 2, August 1985, pp. 524–533.

[Rakerd and Hartmann 1986]

Rakerd, B., and Hartmann, W. M. ‘Localization of sound in rooms, III: Onset and duration effects’. *J. Acoust. Soc. Am.*, Vol. 80, No. 6, December 1986, pp. 1695–1706.

[Ren 2002]

Ren, T. ‘Longitudinal pattern of basilar membrane vibration in the sensitive cochlea’. *Proc. National Academy of Sciences USA,* Vol. 99, No. 26, December 2002, pp. 17101–17106.

[Rumsey 2002]

Rumsey, F. ‘Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm’. *J. Aud. Eng. Soc.*, Vol. 50, No, 9, September 2002, pp. 651–666.

[Saberi and Perrott 1990]

Saberi, K., and Perrott, D. R. ‘Lateralization threshold obtained under conditions in which the precedence effect is assumed to operate’. *J. Acoust. Soc. Am.*, Vol. 76, No. 4, April 1990, pp. 1732–1737.

[Sayers and Cherry 1957]

Sayers, B. McA., and Cherry, E. C. ‘Mechanism of binaural fusion in the hearing of speech.’ *J. Acoust. Soc. Am.*, Vol. 29, No. 9, September 1957, pp. 975–987.

[Scharf 1978]

Scharf, B. ‘Loudness’. In *Handbook of Perception — Volume 4: Hearing,* R. Carterette and M. Friedman, Eds. New York: Academic Press, 1978, pp. 187–242.

[Schnupp 2001]

Schnupp, J. ‘Of delays, coincidences and efficient coding for space in the auditory pathway’. *Trends in Neurosciences,* Vol. 24, No. 12, December 2001, pp. 677–678.

[Schroeder et al. 1974]

Schroeder, M. R., Gottlob, D., Siebrasse, K. F. ‘Comparative study of European concert halls: correlation of subjective preference with geometric and acoustic parameters’. *J. Acoust. Soc. Am.*, Vol. 56, No. 4, October 1974, pp. 1195–1201.

[Schroeder 1977]

Schroeder, M. R. ‘New viewpoints in binaural interactions’. In *Psychophysics and Physiology of Hearing,* E. F. Evans, J. P. Wilson, Eds. London: Academic Press, 1977, pp. 455–467.

[Schroger 1996]

Schroger, E. ‘Interaural time and level differences: integrated or separated processing?’ *Hearing Research,* Vol. 96, No. 1, July 1996, pp. 191–198.

[Schwartz et al. 1999]

Schwartz. O., Harris, J., and Principe, J. ‘Modeling the precedence effect for speech using the gamma filter’. *Neural Networks,* Vol. 12, No. 3, 1999, pp. 409–417.

[Shamma et al. 1989]

Shamma, S. A., Shen, N., and Gopalaswamy, P. 'Stereausis: Binaural processing without neural delays'. *J. Acoust. Soc. Am.*, Vol. 86, No. 3, September 1989, pp. 989–1006.

[Slaney 1993]

Slaney, M. 'An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank'. *Apple Computer Technical Report*, #35, 1993.

[Smith 1994]

Smith, L. S. 'Sound Segmentation Using Onsets and Offsets'. *J. New Music Research*, Vol. 23, No. 1, 1994, pp. 11–23.

[Smith 2001]

Smith, L. S. 'Using depressing synapses for phase locked auditory onset detection'. In *Artificial Neural Networks — ICANN 2001,* G. Dorffner, H. Bischof, K. Hornik, Eds. *Lecture Notes in Computer Science* 2130. Berlin: Springer, 2001.

[Soulodre et al. 2003]

Soulodre, G. A., Lavoie, M. C., and Norcross, S. G. 'Temporal Aspects of Listener Envelopment in Multichannel Surround Systems'. Presented at the *AES 114th Convention,* Amsterdam, The Netherlands, March 2003, preprint 5803.

[Stern and Colburn 1978]

Stern, R. M., and Colburn, H. S. 'Theory of binaural interaction based on auditory-nerve data. IV. A model for subjective lateral position'. *J. Acoust. Soc. Am.*, Vol. 64, No. 1, July 1978, pp. 127–140**.**

[Stern and Trahiotis 1997]

Stern, E. M., and Trahiotis, C. 'Models of Binaural Perception'. In *Binaural and Spatial Hearing in Real and Virtual Environments.* R. Gilkey, T. Anderson, Eds. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc., 1997, pp. 499–531.

[Stevens 1957]

Stevens, S. S. 'On the psychophysical law'. *Psychological Review,* Vol. 64, 1957, pp. 153–181.

[**Stinson 1990]**

Stinson, M. R. 'Revision of estimates of acoustic energy reflectance at the human eardrum'. *J. Acoust. Soc. Am.*, Vol. 88, No. 4, October 1990, pp. 1773–1778.

[Supper et al. 2005]

Supper, B., Brookes, T., and Rumsey, F. ‘An auditory onset detection algorithm for improved automatic source localization’. *IEEE Trans. Speech and Audio Processing.* Accepted for publication.

[Texas Instruments 2005]

Texas Instruments. *DSP Platforms : C5000™ DSPs.* Web site, 2005.
http://dspvillage.ti.com/ (Requires manual navigation.)

[Ungan et al. 2001]

Ungan, P., Yagcioglu, S., and Goksoy, C. ‘Differences between the N1 waves of the responses to interaural time and intensity disparities: scalp topography and dipole sources’. *Clinical Neurophysiology*, Vol. 112, No. 3, March 2001, pp. 485–498.

[Wallach et al. 1949]

Wallach, H., Newman, E. B., and Rosenzweig, M. R. ‘The Precedence Effect in Sound Localization’. *Am. J. Psychol.*, Vol. 42, 1949, pp. 315–326. Reprinted *J. Aud. Eng. Soc.*, Vol. 21, No. 10, December 1973, pp. 817–826.

[Whitworth and Jeffress 1961]

Whitworth, R. H., and Jeffress, L. A. ‘Time vs Intensity in the Localization of Tones’. *J. Acoust. Soc. Am.*, Vol. 33, No. 7, July 1961, pp. 925–929.

[Wightman and Kistler 1992]

Wightman, F. L., and Kistler, D.J. ‘The dominant role of low-frequency interaural time differences in sound localization’. *J. Acoust. Soc. Am.*, Vol. 91, No. 3, March 1992, pp. 1648–1661.

[Wojcik and Cardinal 1999]

Wojcik, J. J., and Cardinal, P, G. ‘New Advanced Methodology for Near Field Measurements for SAR and Antenna Development’. PDF file, 1999.
http://www.spectrum-sciences.org/research/Measure SAR IEEE 1999.PDF

[Yates 1995]

Yates, G. K. ‘Cochlear Structure and Function’. In *Hearing*, B. C. J. Moore, Ed., London: Academic Press, 1995, pp. 41–73.

[Zurek 1980]

Zurek, P. M. ‘The precedence effect and its possible role in the avoidance of interaural ambiguities’. *J. Acoust. Soc. Am.*, Vol. 67, No. 3, March 1980, pp. 952–964.

[Zurek 1987]

Zurek, P. M. ‘The Precedence Effect’. In *Directional Hearing*, W. Yost, G. Gourevitch, Eds., New York: Springer-Verlag, 1987, pp. 85–105.