hochschule für angewandte wissenschaften
FACHBEREICH ELEKTROTECHNIK UND INFORMATIK hamburg
university of applied sciences

Diplomarbeit

Olaf Christ

Sound Source Localization using an ALN

Studiengang Sofwaretechnik

Betreuender Prüfer: Prof. Dr. Reinhard Baran
Zweitgutachter: Prof. Dr. Wolfgang Renz
Abgegeben am 29. November 2002

Berliner Tor 3 · D-20099 Hamburg

Zusammenfassung

Lokalisation einer Geräuschquelle mittels eines ALN

Stichworte:

ALN, Regression, Bagging, Geräuschquelle, Lokalisation, KEMAR, HRTF.

Zusammenfassung

In dieser Arbeit geht es um die Lokalisation einer einzelnen Geräuschquelle in der sagittalen Ebene unter Verwendung eines ALN. (Adaptive Logic Network). Um den Elevationswinkel später abschätzen zu können, werden ALN's mit dem Ergebnis der Autokorrelation des Ergebnisses der Frequenzanalyse eines Audiosignals mittels eines Cochlea-Modells, trainiert. Der resultierende DTree (Entscheidungsbaum), der den Durchschnitt aller trainierten ALN's repräsentiert, kann dann benutzt werden um den zum Kopf relativen Elevationswinkel einer Geräuschquelle abzuschätzen.

Sound Source Localization using an ALN

Keywords:

ALN, regression, Bagging, sound source, localization, KEMAR, HRTF

Abstract:

In this thesis Sound Source Localization of a single sound source within the sagittal plane is performed using an ALN (Adaptive Logic Network). To be able to estimate the elevation angle, ALN's are trained with the autocorrelation pattern of a cochlea model output, computed from an audio signal. The resulting DTree (Decision Tree) that represents the average of the trained ALN's can then be used to estimate the elevation of a sound sources' spatial position relative to the head.

Contents

Preface

This work has been carried out as a german Diploma thesis over a period of 3 months during the summer / fall 2002 at the University of Applied Sciences Hamburg (HAW). The subject captured my interest, because i have been interested in robotics, biology and medicine for a long time. And, of course, because the problem is challenging and an opportunity to work in this area for the first time.

The work has been edited using MS Word. The code has been written in the Matlab editor. The C/C++ code was compiled using VC++.

Olaf Christ
Hamburg, Germany, November 29th 2002

Acknowledgments

At first, I would like to thank my supervisors Prof. Dr Reinhard Baran and Prof. Dr. Wolfgang Renz for letting me work on this subject.

I would also like to thank Prof. Dr. Richard O. Duda for writing his excellent, very perspicuous tutorial and for the permission to use his graphics, Prof. Dr. Richard F. Lyon for developing the earmodel and Dr. Malcolm Slaney for implementing it for Matlab and putting it on the web for free as a part of the Auditory Toolbox II. Last, but not least, i want to thank Prof. Dr. William W. Armstrong for the Dendronic Learning Engine and many helpful tips on how to use it.

1 Introduction

it has been 55 years since J. Presper Eckert and John Mauchly formed the worlds first computer company, the Eckert Mauchly Computer Company, and it has been over 30 years since Ed David and Oliver Selfridge wrote their paper called "Eyes and Ears for Computers." [David+Selfridge62]. It was a time where it was hard to even get the data into the computer and computing power was expensive and scarce.

They concluded that things like character and spoken word recognition where hard and existing systems were unsatisfactory and possible solutions were expensive.

Fortunately, over the last years, inexpensive yet very powerful technolgies have become availabe to virtually everybody. Thus, tremendous progress has been made in the field of robtics and artificial intelligence over the last few years. Today, projects like e.g. COG and KISMET (both MIT) or Honda's Humanoid Robot are trend-setting with respect to robots that dont much resemble today's industry robots.

These robots mark important steps on the long way to develop humanoid robots.

Although these systems mean a huge step with respect to e.g. very sophisticated vision systems and revolutionary learning concepts, their auditory system still consists of one or two simple microphones for ears.

To achieve the ultimate goal, humanoid robots that see *and* hear, there is still very much to do.

In the past years a lot of work has been done in the area of artificial auditory systems, both binaural and monoaural. The computers are much faster today and recent projects are very promising like e.g. the Kitano project that is mentioned later in this thesis.

Sound source localization, in particular the localization of multiple sources in a non anechoic environment, is essential for simultaneous speech recognition and the subject of heavy research today.

Although, even some of the best ideas are patented and work very well in certain situations, the ultimate solution, a system that works as well as the human auditory system, is still missing.

1.1.Goals

The goal of this thesis is to build a quite simple Sound Source Localizaton System, that is able to locate the spatial position of a single sound source in the median plane and to cover some of the intrinsic problems.
Although, both, azimuth and elevation are covered, this work mainly focuses on elevation estimation.

1.2. Methodology and limitations

The research in this thesis has been conducted mostly in an experimental fashion. After an initial idea is sprung, informal reasoning around the idea gives insights into its soundness. Those ideas that are found to be sound are then implemented and tested to see if they should be further pursued or discarded. If an idea should be further pursued it is refined and further discussed.
Because of lacking proper lab equipment and the unavailability of a KEMAR (Knowles Electronics Mannequin for Acoustical Research) artificial head, the free HRTF measurement testset by [Gardner+Martin94] was used instead of own measurements.
Further processing of these measurements was done in Matlab to produce test data of test signals and training data of training signals.

1.3 Thesis structure

This work is organized as follows:

Apart from chapters 3 and 4, the pupose of chapters 1 to 3 is to not only explain the scientific background, but also to discuss the problems from different views, thus providing a more in depth understanding.

Chapter 1 gives an introduction, presents the research goals and the limitations and includes this structure.

Chapters 2.1 to **2.3** cover the more technical aspects of hearing that are needed to not only build a sound source locator but are also important to interpret the results, the accuracy of sound source localization systems, correctly.

Chapter 2.4 then focuses on the biological and neurophysiological basis. Since we use computional models of actual organs in the human body, it is essential to have some knowledge in this area too.

Chapter 2.5 illustrates a general sound source locator.

Chapter 3 eventually concentrates on the Sound Source Localization using an ALN, defends the usage of the computational components used and their connection and gives an interpretation of the results.

Chapter 4 further discusses the experiments and the performance of the presented approch and how to improve the accuracy in future work.

Appendix A The appendix explains some anatomical terms used in this work.

Appendix B explains what is on the accompanying CD-ROM.

Bibliography includes the bibliography.

2 Background

2.1 Machine perception

Machine perception is still a very hard problem.

Seeing and hearing are our wide-bandwidth ways of perceiving the world. Psychologists make an important distinction between sensing and perceiving. Computers ar very good at sensing, e.g., acquiring digital data like images or sound waveforms.

Processing and enhancing images and sounds with various digital filters and categorizing them with various statistical or neural-network classifiers.is a common practice.

Unfortunately computers have various limitations in perceiving. Computers have difficulies in recognizing what an image bitmap shows as, e.g. two people in a boat on a mountain lake, or a digitital sound recording of a conversation between a woman and a child on a busy street.

An interesting approach to educate a machine, to make the machine perceive, might be the Kismet-Project by Dr. Cynthia Breazeal.
Kismet is an expressive robotic creature with perceptual and motor modalities tailored to natural human communication channels. Kismet was developd to interact and to learn from its infant caretaker.
To facilitate a natural infant-caretaker interaction, the robot is equipped with visual, auditory, and sensory inputs.

Today, no one knows if the problem of machine perception is well stated.
We dont even know if we really want e.g. convert sequences of images into geometrical models of objects. Hence, we dont know if Rodney Brooks and Cynthia Breazeal are right saying that all we want to do is to get a robot to respond reasonably to its environmental stimuli.
Perhaps, somebody will eventually ask the right questions and find the solution.

2.1.1 Sound Perception

Basically, there are two things that can be done with digital data.
We can filter and or classify.
Filtering does not take us far enough
Digital filters, linear and nonlinear, which are important in e.g. audio processing might be used to:

Reduce undesirable noise
Compensate for deficiencies of transducers
Compensate for the room environment.
Restore spatial effects.
Compress or expand the dynamic range.

However, a digitally filtered signal is still a digital signal, thus filtering is not hearing.

Classification, unfortunately, may take us too far.
As an example of a classification system, a speech recognition system takes any input signal and turns it into a code representing english words. It tries to pigeonhole the input no matter what it is.
It tries to do this despite the large amount of variations that might appear in the input without regarding aspects like:

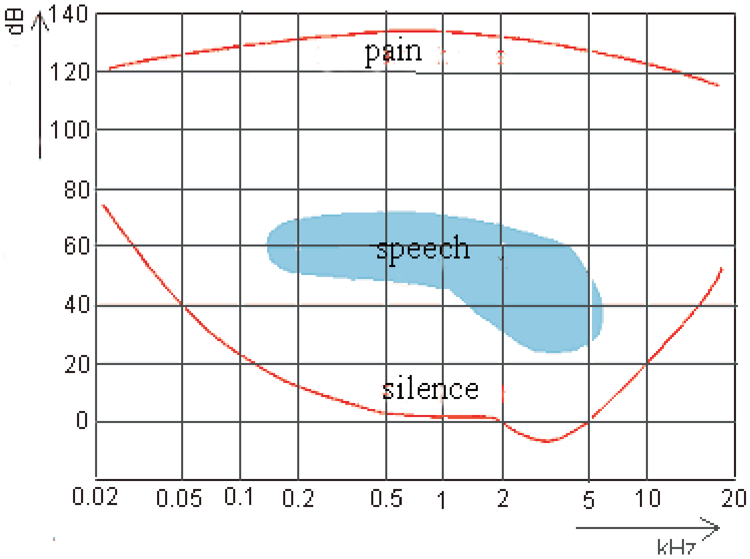
Speaking variations, Loudness, Pitch, Tempo, Isolated-word/continuous, Healthy/sick, Relaxed/stressed, Speaker variations, Male/female, Adult/child, Native/foreign, Environmental variations, Spectral distortion, Echoes and Reverberation, Interference, Random noise and Multiple sound sources
The output of a classifier is just a number, thus: Classification is not perception.
We need something intermediate between filtering and classification.
If computers are ever to hear, they will need methods for representing acoustic inputs in a way that reveals their components, and discloses rather than suppresses their structure [Duda-98].

2.2 Psychoacoustic

Why Psychoacoustic ?

Building a Sound Source Locator requires some psychoacoustical background.
Psychoacoustic is a discipline describing the relation between the physics of sound and the resulting sensations. The transferfunction resulting from this relation is called the HRTF (Head Related Transfer Function), which is used to model the human hearing.
Knoweldge of this function is important to develop or enhance technical appliances for acoustic communication, because in acoustic communication the ear is always the receiver of acoustic information. Psychoacoustical knowledge is also important to become more familiar with the very difficult problems related to sound source localization.

2.2.1 Hearing threshold level

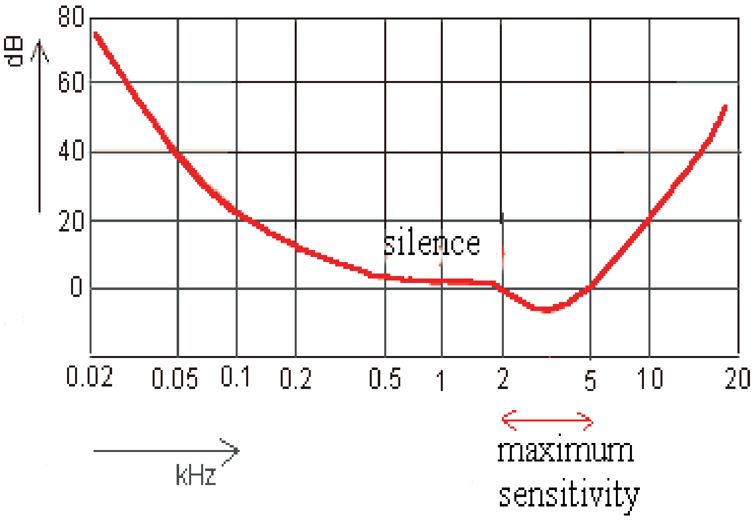


The human hearing is only capable of sensing sound between 20 Hz to 20 kHz.
To be audible, an acoustic pressure of at least 20 µPa is required. This equals a sound level of 0 dB.

The lowest sound level where a human can hear sound is called hearing threshold level. E.g. breathing has a level of approximately 10 dB.

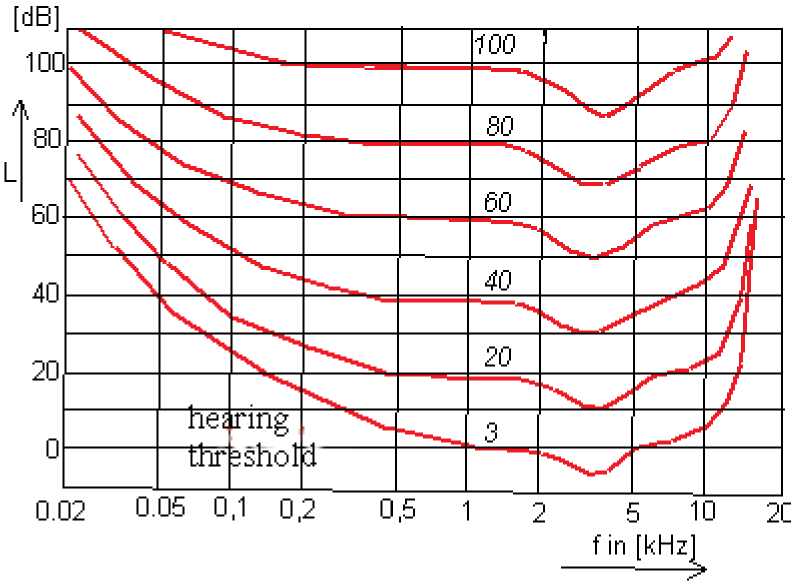
On the other end there is the threshold of pain: A noise level of 120 dB causes pain.

Between both levels are 6 decimal powers. This dynamic range makes it possible to hear a buzzing fly as well as a jack-hammer.



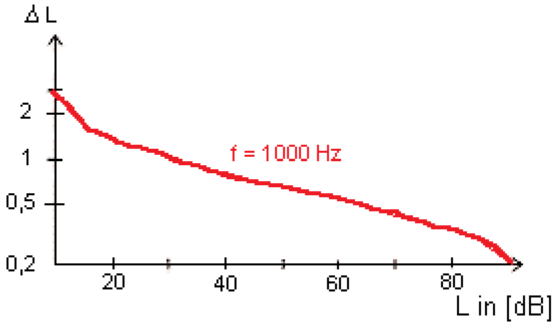
Because the Hearing threshold level differs throughout the frequency range of interest, the sensitivity for different frequencies also differs.The diagram above shows that the range of maximum sensitivity is between 2 khz and 5 kHz.

Because the Hearing threshold level is not equally spaced, sinus sounds of equal sound levels are sensed differently with respect to the sound level.

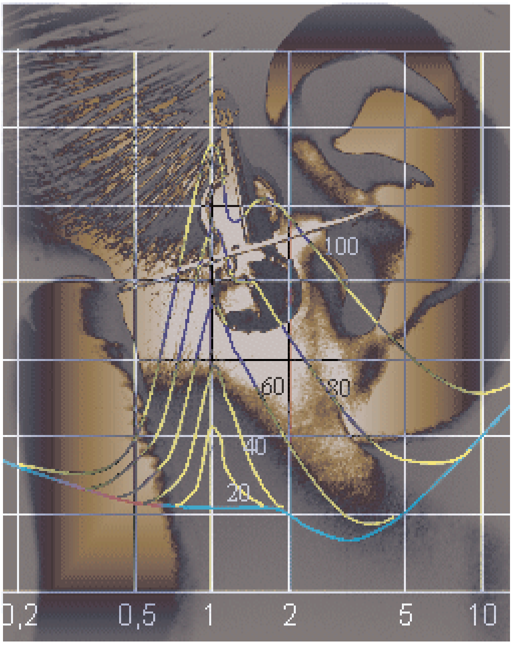


2.2.2 Minumum distinguishable changes in sound levels

The maximum level change between two sounds depends on the frequencies of the sounds and their sound level. Human hearing is more or less responding to ratios rather than differences. 1 dB is about the smallest perceptible change in loudness, no matter what the starting intensity level is, and a one-octave frequency change sounds like the same musical interval, no matter what the starting frequency is.

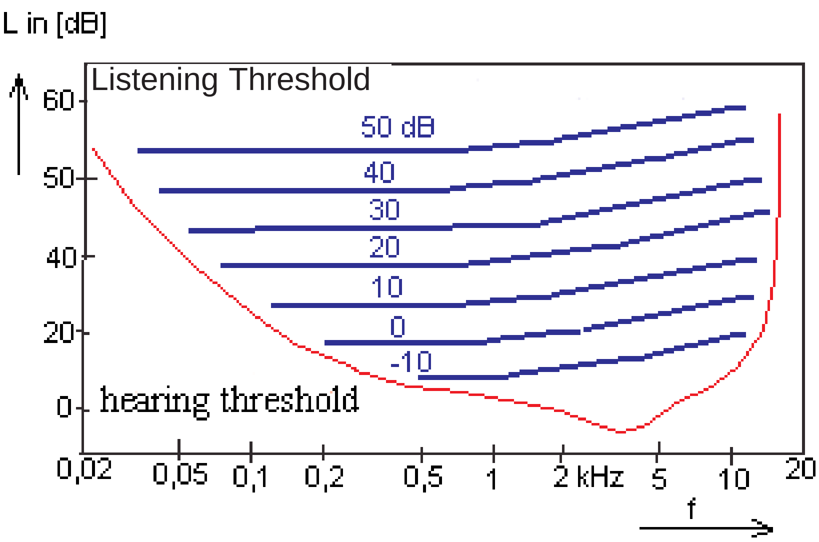


2.2.3 Masking Effects



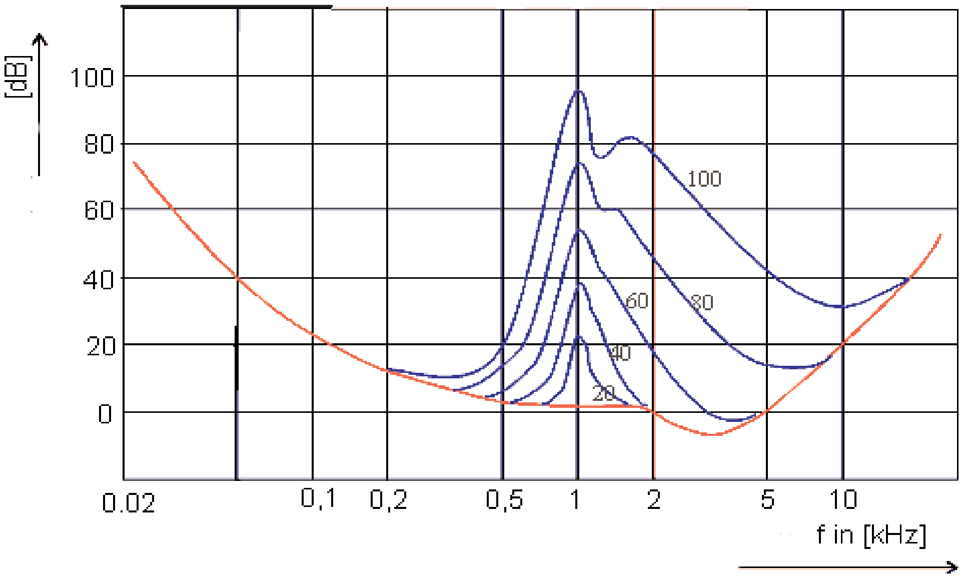
Masking effects are caused by the influence of the audibility of sound by the superposition of interfering signals.

This occurs in everyday conversations quite often and background noise is usually the reason, that makes understanding difficult. The interference of the voice signals by background noise is the reason why it is hard to follow a conversation e.g. next to a jack-hammer or a running air conditioning. Besides the sound level, the frequency spectrum is also important with respect to the ability to interfere a signal. In general a sound with higher frequency is only able to interfere with another sound with lower frequency if the frequency separation is small. A sound with lower frequency is only capable of interfering with another sound of higher frequency if the sound level is significantly higher.



The diagram above shows the ability to hear a sinus sound with respect to superposed white noise.

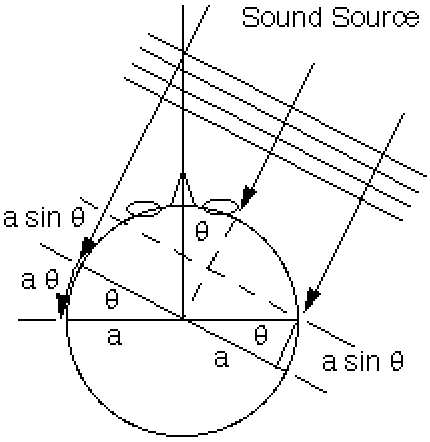
The diagram below shows the ability to hear a sinus sound superposed by bandlimited white noise.



2.2.4 Azimuth Cues

The Duplex Theory

About 100 years ago one of the pioneers in spatial hearing research John Strutt, better known as Lord Rayleigh developed his so-called Duplex-Theory.
The Duplex-Theory names two primary cues for azimuth, the Interaural Time Difference (ITD) and the Interaural Intensity Difference (IID).



Lord Rayleigh’s explanation for the ITD is rather simple. Sound travels at a speed c of about 343 m/s. Let the direction of a sound wave striking a spherical head of radius a from a distant be defined as the azimuth specified by θ.
The Sound will arrive at the right ear before the left ear, since it has to travel the extra distance a θ + sin θ to reach the left ear.
Simply deviding the distance by the speed of sound gives us the ITD, the interaural Time Difference. $\mathrm{ITD} = \frac{a}{c}(\theta + \sin\theta)$, $-90^\circ \leq \theta \leq +90^\circ$

The ITD is zero when the source is directly ahead, and is a maximum of (a/c) (π/2+1) when the source is off to one side.
This represents a difference of arrival time of about 0.7 ms for a typical size human head, and is easily perceived.

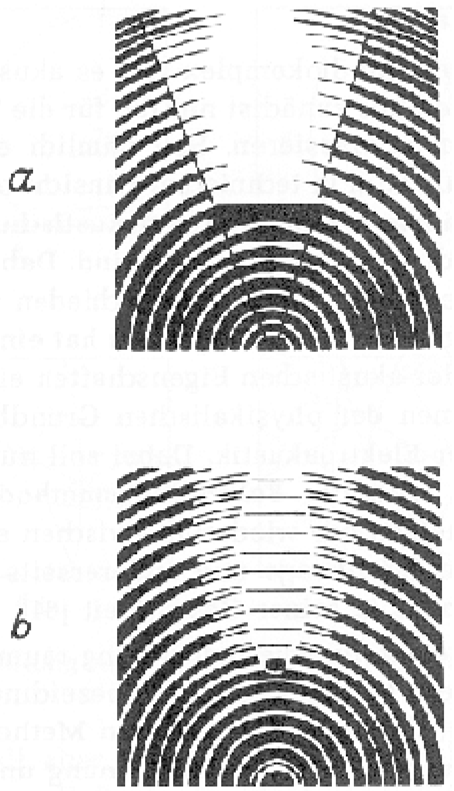
Of course, the accuracy with which this can be done depends on the circumstances.
For speech in normally reverberant rooms, typical human accuracies are on the order of 10° to 20°. Under optimum conditions, a accuracy (on the order of 1°) is possible if the problem is to decide merely whether or not a sound source moves.
This is rather remarkable, since it means that a change in arrival time of as little as 10 microseconds is perceptible.
(For comparison, the sampling rate for audio CD's is 44.1 kHz, which corresponds to a sampling interval of 22.7 microseconds. Thus, in some circumstances, less than a one-sample delay is perceptible.)

Publication	Type of Signal	Precision
Klemm (1920)	Impulses (clicks)	0.75°-2°
King and Laird (1930)	Impulse (click) train	1.6°
Stevens and Newman (1936)	Sinusoids	4.4°
Schmidt et al. (1953)	Sinusoids	>1°
Sandel et al. (1955)	Sinusoids	1.1°-4.0°
Mills (1958)	Sinusoids	1.0°-3.1°
Stiller (1960)	Narrow-band noise, cos2 tone bursts	1.4°-2.8°
Boerger (1965a)	Gaussian tone bursts	0.8°-3.3°
Gardner (1968a)	Speech	0.9°
Perrott (1969)	Tone bursts	1.8°-11.8°
Blauert (1970b)	Speech	1.5°
Haustein and Schirmer (1970)	Broadband noise	3.2°

The table shows some measurements for different signals from a single sound source in the tranverse plane. (Blauert [Bla97])

Lord Rayleigh also observed that the incident sound waves are diffracted by the head.

He actually solved the wave equation to show how a plane wave is diffracted by a rigid sphere. His solution showed that in addition to the time difference there was also a significant difference between the signal levels at the two ears -- the Interaural Intensity Difference (IID).



The IID is highly frequency dependent. At low frequencies, where the wavelength of the sound is long relative to the head diameter, there is hardly any difference in sound pressure at the two ears.

At high frequencies, where the wavelength is short, there may well be a 20-dB or greater difference. This is called the head-shadow effect, where the far ear is in the sound shadow of the head.(a) However if the wavelength is long, the soundwaves are affected very little by the obstacle and easily travel around the head. This is called diffraction.(b)

[Web94]

The Duplex Theory asserts that the IID and the ITD are complementary. At low frequencies (below about 1.5 kHz), there is little IID information, but the ITD shifts the waveform a fraction of a cycle, which is easily detected. At high frequencies (above about 1.5 kHz), there is ambiguity in the ITD, since there are several cycles of shift, but the IID resolves this directional ambiguity.

Rayleigh's Duplex Theory claims that the IID and ITD taken together provide localization information throughout the audible frequency range.

The Clifton Effect and Room Modeling

The common understanding of the precedence effect is, that when there are multiple reflections of a sound, the auditory system uses the first signal for localization and suppresses the reflections that follow. However, in 1987, Clifton showed that the suppression process is a dynamic one, taking time to become established, and that it does not merely suppress anything and everything that follows a first onset. The task of creating a model of the listening space is an important job for the auditory system, and is a part of sound localization that is not addressed by the Duplex Theory.

At first, Rayleigh's duplex theory seems to answer the question of how people localize sounds quite well. Unfortunately, there are at least eight serious problems with the duplex theory.
(For a detailed explanation of all these problems see [duda98])

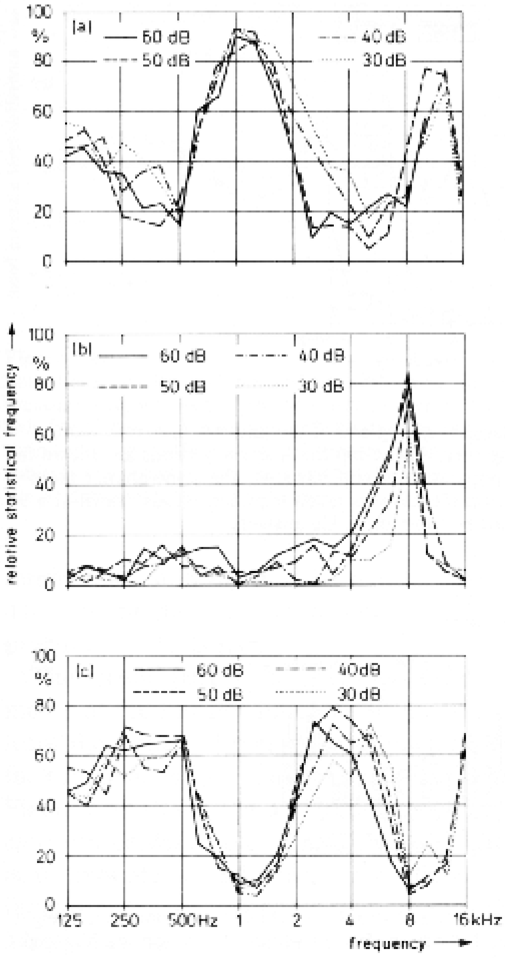
1. It is based on a model that fails at high frequencies
2. It neglects the effect of the acoustic environment
3. It neglects the importance of onsets and the precedence effect
4. It does not address the Clifton effect and room modeling [Clifton87]
5. It fails to account for vertical location
6. It fails to account for front/back discrimination
7. It fails to explain how we experience sounds as coming from outside of our head.
8. It does not address the question of multiple sound sources

Thus, to sucessfully localize sound sources, we need to know more about both the physical source of the sound and biological mechanisms people and animals use to locate sounds.

2.2.5 Elevation Cues

Although, in the sagittal plane (median plane) there almost are no changes between the signals in terms of sound level and time differences, a sound source can be located by changes in the timbre of the sound.
These changes are mostly caused by the shape of the head and the pinnae.
This causes the level of certain frequency bands to rise, depending on the angle to the sound source.
The precision in the sagittal plane is significantly lower than in the transverse (horizontal) plane.
Localization works best if the signal is a broadband signal and the structure of the signal is familiar to the brain.
This transfer characteristic of the outer ear (pinnae) enables us to to discriminate between front, back, top and down (sagittal plane)
Although the impact of the frequencies reaching the pinnae from a certain angle slightly differ from person to person, there are some frequencies that have been shown to cause a sensation of direction perception in the sagittal plane.

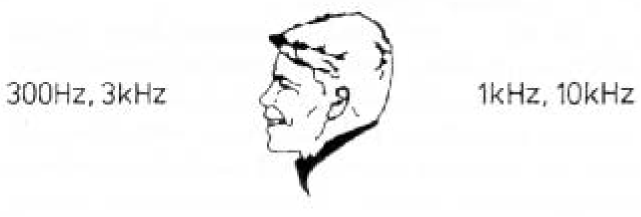
However, while the primary cues for azimuth are binaural, the primary cues for elevation are often said to be monaural. Due to its unique shape, the outer ear or pinna acts like an acoustic antenna. Its resonant cavities amplify some frequencies, whereas its geometry leads to interference effects that attenuate other frequencies. Moreover, its frequency response is directionally dependent.

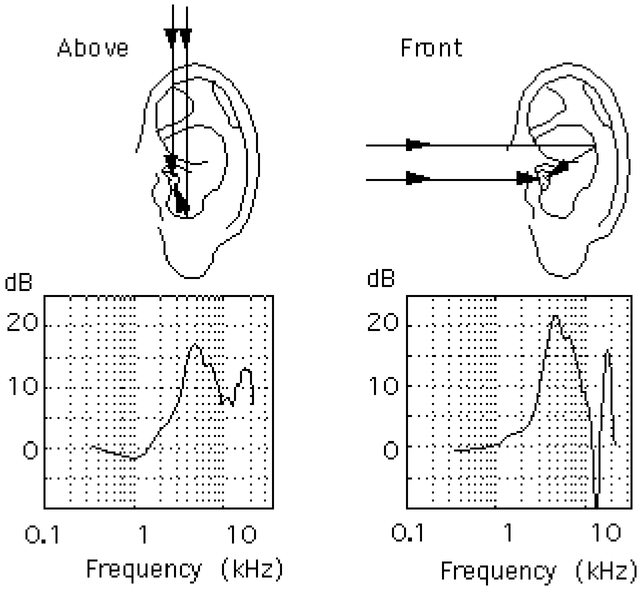


The diagram to the left shows the relative frequency of the answers back, above and front with respect to the center frequency of a noise signal coming from the front or the back. (Blauert [Bla97])

Approx. 90 % of the probands have located a signal with a center frequency of 1 kHz behind them whereby a signal with a center frequency of 8 kHz has been located to be coming from above. 4kHz causes a "front-sensation".

The knowledge about these frequencies is important not only for sound source localization, but also for systems like Microsofts Direct3D Sound (TM) which incorporates this knowledge to position a sound sources in a virtual space, causing sensations of direction e.g. to the player in computer games.





[Duda98]

The figure above shows measured frequency responses for two different directions of arrival. In each case we see that there are two paths from the source to the ear canal -- a direct path and a longer path following a reflection from the pinna. For more on the anatomy of the ear please refer to the medical / biological chapters.
At low frequencies, the pinna collects additional sound energy, and the signals from the two paths arrive in phase. At high frequencies, the delayed signal is out of phase with the direct signal, and destructive interference occurs. The greatest interference occurs when the difference in path length d is a half wavelength, i.e., when f = c / 2d. In the example shown, this produces a "pinna notch" around 10 kHz. For more on this refer to the physics chapter. With typical values for d, the notch frequency is usually in the 6-kHz to 16-kHz range.

Because of its orientation the pinna reflects sound coming from the front much more effectively than sound coming from above.
The resulting notch is much more distinct for sources in front than for sources above.
While the path length difference changes with elevation angle, the frequency of the notch moves with elevation.
Although there is still a debate about what features are perceptually most important (for example, see Han [Han94]), it is well established that the pinna provides the primary cues for elevation.

Shoulder echo

Frequencies in the range of 1 to 3 KHz get reflected from the upper torso. Its not a very significant cue, but contributes to some extent in elevation determination.

2.2.6 Range Cues

When localizing a source, azimuth can be estimated with highest precision, elevation with second highest precision and range with worst. The cues for azimuth are quite well understood, the cues for elevation are less well understood, and the cues for range are least well understood.
Frequently mentioned cues are:

Loudness
Motion parallax
Excess interaural intensity difference (IID)
Ratio of direct to reverberant sound

The loudness falls of inversely with the source of range. The problem is, that we need to have some experience about the characteristics of the sound source to use its loudness as a cue of range. Humans usually have knowledge gathered throughout their life, especially in the case of human speech.Thus, a simple one to one relation between energy and range cannot be determined.
The Motion parallax refers to the fact that, if a listener translates his or her head, the change in azimuth will be range dependent.

For sources that are very close, a small shift causes a large change in azimuth, while for sources that are distant there is esentially no azimuth change. An effect also occuring with objects moving very close in front of a camera or eyes (e.g. if we take a look out of a fast train) Moreover, as a sound source gets very close to the head, the IID will increase.
This increase becomes noticeable at ranges under about one meter.

Although in ordinary rooms, sound reaching the ear is reflected and scattered many times from environmental surfaces, the reverberant energy does not change much with the distance between source and listener, thus the ratio of direct to reverberant sound is an important cue for range. Its obvious that for close ranges, the ratio is very large, while at long ranges it is quite small.

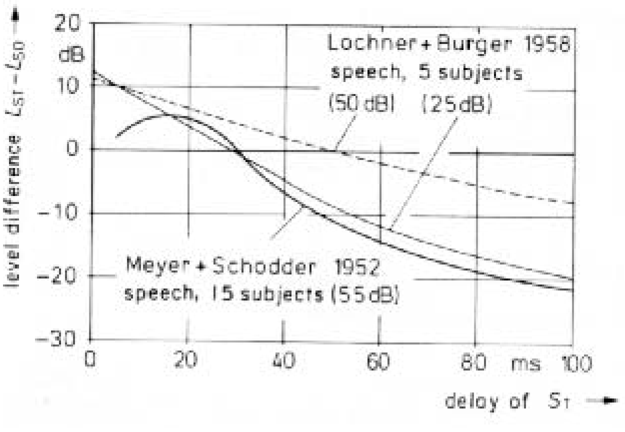
2.2.7 Reverberation and Echoes

Ususally we are unaware of how much the sound we hear is affected by the envrionment.
A significant amount of sound energy is reflected by walls, the ground or vegetation.

These reflections only become noticeable if the time delay is greater than 30-50 ms, the echo threshold, in which case we become consciously aware of them and call them echoes.
To eliminate echoes, special rooms, called anechoic chambers are build to absorb sound energy, so that only the directly radiated energy reaches the ears

In everyday life the auditory system quickly adapts to new acoustic environments using only partially understood mechanisms to suppress the effects of reflections and reverberation.
The fact that we are able to localize on the basis of the signals that reach our ears first is known as the precedence effect or the Law of the First Wavefront.(see also the Clifton Effect) [Clifton87]
See [Angel+Algazi+Duda2002] to find out more about how a model for late reverberation, the spatially-diffuse late portion of a room response, can be used to increase localization accuracy.

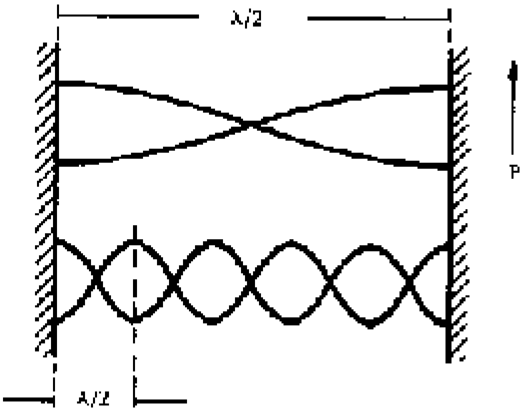
The figure below shows Measurements of the echo threshold regarding speech ([Bla97])



Reflections are subconsciously used to estimate range and have very little effect on the ability to localize sounds unless reverberation becomes severe,

The precedence effect forces us to revise Rayleigh's Duplex Theory. In a reverberant room, low-frequency information is essentially useless for localization, because for a low-frequency sound whose period is longer than the time for reflections to arrive, the reflections begin to arrive before the completion of one cycle.
By the time several cycles have arrived and the auditory system can begin to estimate pitch, the sound pattern in the room is a jumble of standing waves, and it is now impossible for the auditory system to estimate interaural time differences.
Thus, in a reverberant room, low-frequency information is useless for localization.

2.2.8 Standing waves



Pressure distribution in standing waves between
two parallel walls for two different frequencies ([Dic87])

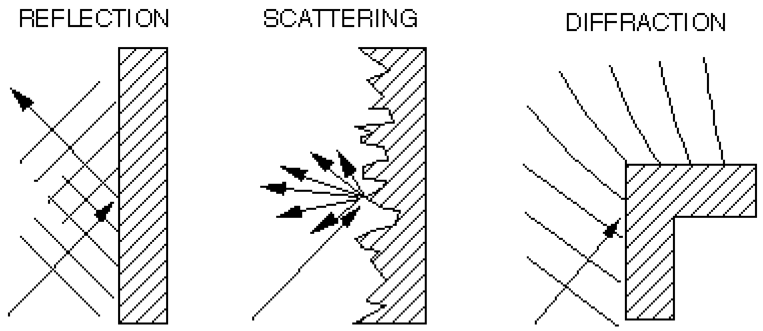
If a sound wave is hitting a wall at a perpendicular angle with the wall, the incident and the reflected sound wave superpose. If we take two parallel walls, like in the figure shown above, with a distance of a multiple of a half wave length, the direct sound wave and the reflected sound wave will superpose in phase.
What we get is a stationary sound field, with an acoustic pressure of 0 at the nodes, and a maximum acoustic pressure at the antinodes. Because these nodes and antinodes are stationary, they are called standing waves.
Standing waves caused by multiple reflection are called standing waves of higher order if the distance is a multiple of the half wave length.

Interaural timing differences are important.
The important timing information comes from the Interaural Envelope Difference (IED), e.g., from the transients at the onset of a new sound. This is demonstrated by the Franssen Effect:
If a sine wave is abruptly turned on and a high-pass-filtered version is sent to loudspeaker A while a low-pass filtered version is sent to loudspeaker B, most listeners will localize the sound at loudspeaker A.

The starting transient provides unambiguous localization information, while the steady-state signal is very difficult to localize, and in this circumstance the auditory system simply ignores the ambiguous information. Maybe oversimplified, we can generalize and say that in reverberant environments it is the high-frequency energy, not the low-frequency energy, that is important for localization.

2.3 The Physics of Sound

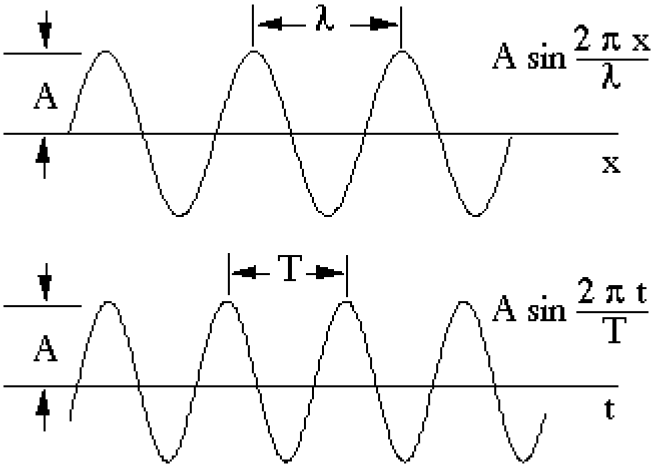
Sound is an acoustic wave produced by a vibrating source propagating through a medium from one location to another. When a sound wave reaches a listener's ear drum, the vibrations are transmitted to the inner ear, where the impulses are converted into nerve signals that are sent to the brain and result in the sensation of sound.



([Duda98])
In a homogeneous medium, sound waves travel at a constant speed c, 343 m/s for air. The amplitude falls off inversely with distance.
Usually these waves are reflected or by smooth and scattered by rough surfaces. A “smooth” surface has irregularities with a size that is small relative to the wavelength and considered “rough” otherwise.

Sound waves are diffracted by objects. If the object is small relative to the wavelength, the wave passes around the object undisturbed, whereas if the object is large, a "sound shadow" appears behind the object and a significant amount of energy is reflected back.
If the wavelength and the object are about the same size, complicated diffraction patterns occur. (see also page 20)

2.3.1 Sine Waves



In music and acoustics, sine waves are often called pure tones.
In Physics, the source is a steady-state sine wave with frequency f, then the response at any other point in space is also a sine wave of frequency f; only the amplitude and phase changes as one moves around, which is only true for sine waves.
For spatial sine waves, we specify the wavelength λ, the distance for one cycle. For temporal sine waves, it is common to specify the frequency f (in Hertz or cycles per second).
The angular frequency ω (in radians per second), or the period T (in seconds). These quantities are linked to the speed of sound c through the basic equations:

$$c = f\lambda \qquad f = \frac{1}{T} = \frac{\omega}{2\pi}$$

2.3.2 Fourier Analysis

Most natural sounds are not sine waves.
Sine-waves have a very narrow-band frequency spectrum and because they set up standing-wave patterns in rooms, they are very difficult to localize. Every waveform can be represented as a superposition of sine waves. In particular, a periodic signal x(t) with a fundamental frequency can be represented as a complex Fourier series.

$$x(t) = \sum_{n=-\infty}^{\infty} X_n \, e^{jn2\pi f_0 t} \quad \text{where} \quad X_n = \int_{\text{period}} x(t) \; e^{-jn2\pi f_0 t} \; f_0 \, dt$$

X(f) is called the Fourier transform of x(t).
In general, the Fourier transform is complex, having both a magnitude |X| and a phase /_X.

The squared magnitude of X gives the power for a periodic signal and the energy density for a finite-energy signal.For the power or energy of a signal in different frequency bands, X is commonly refered as the spectrum of x.Physically this makes sense for periodic signals, but for aperiodic signals such as speech, the usual practice is to only consider a short time spectrum by multiplying the x(t) by an appropiate window function w(t). W(t) x(t) is called the short-term spectrum.
The right window function is very important when using FFT for spectral analysis. A poor choice for a window function can produce results that are quite different from what is desired.

(See also [Bracewell 86], [Oppenheim89] and [Schafer89], [Steiglitz96])

2.3.3 Correlation and coherence

Two signals, are coherent, if one signal was shifted against itself and and scaled, e.g. if both signals originate from the same source. Thus, time and level differences do not matter.
Two signals are incoherent if they show no similarities at no time.
Correlation denotes the relation between two signals at one point of time. Signals are correlated if they match, except a factor of scale.
The grade of correlation denotes the amount of similarity on a scale of [- 1;1]. 1 is eqal, -1 out of phase and 0 is uncorrelated.

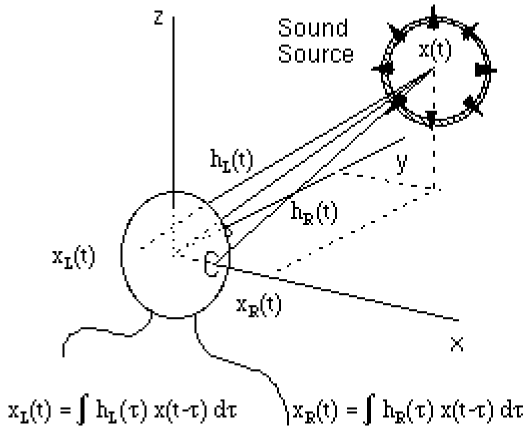
Two incoherent signals are always uncorrelated whereas correlated signals are always coherent. The inversion ist not generally true !

E.g. noise is incoherent to all signals (except to itself) and therefore uncorrelated. A sinus and a cosinus wave are uncorrelated to each other, but coherent.

One might think, that a sound field with no standing waves, in particular a field where at any point the sound level is the same, might help to localize a sound source. Unfortunately, those sound fileds, called diffuse sound fields are ones that encompasses the listener, filling the listening space without being very directional (having low directivity). Thus, and because diffuse sound fields are created by room characteristics (reverberations and echoes) they play havoc to localization.

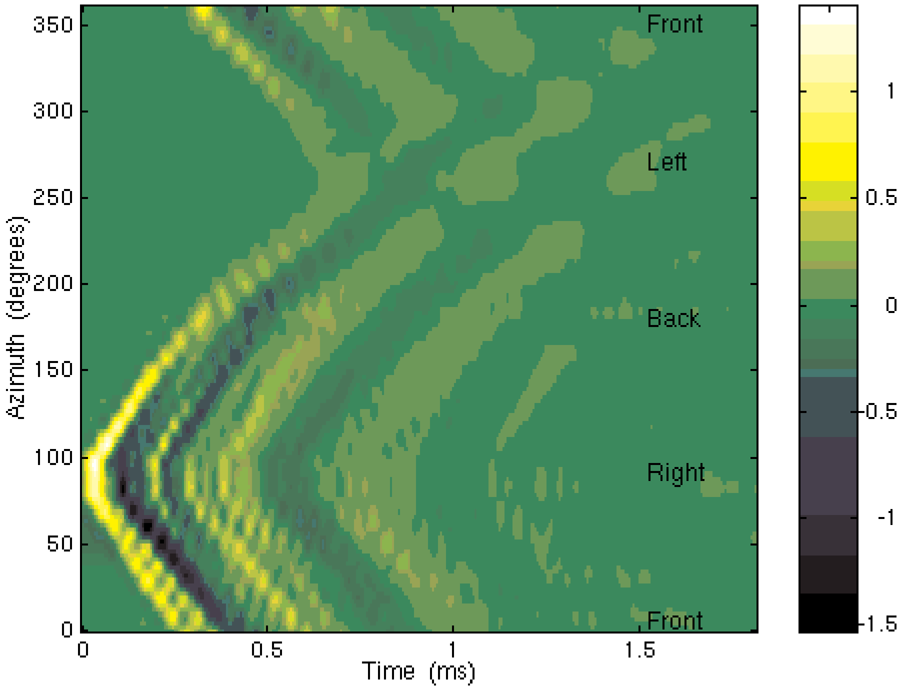
A diffuse sound is often used in surround channels to create an enveloping effect and create a believable sonic environment without drawing attention from the directive sound of the front channels (left, center and right). These or similar techiques are used in surround formats like THX or DTS.

2.3.4 Head-Related Transfer Functions


z
Sound Source
x(t)
$h_L(t)$
y
$h_R(t)$
$x_L(t)$
$x_R(t)$
x
$x_L(t) = \int h_L(\tau)\, x(t-\tau)\, d\tau$
$x_R(t) = \int h_R(\tau)\, x(t-\tau)\, d\tau$

H(t), the impulse response from the source to the ear drum denotes the sound pressure that a source x(t) produces at the eardrum and is called the Head-Related Impulse Response (HRIR). Its Fouirier transform h(t) is called the Head Related Transfer Function (HRTF). The HRTF provides all the physical cues to source localization. (It is assumed that the HRTF is measured in an anechoic chamber, such as the one used in this work, and thus does not include effects of environmental sound reflections, which also provide localization cues)
[Duda98] made a series of HRIR measurements on an acoustic manikin known as KEMAR, which stands for Knowles Electronics Manikin for Auditory Research. To get an idea of how KEMAR's response varies with azimuth and elevation see figures below.

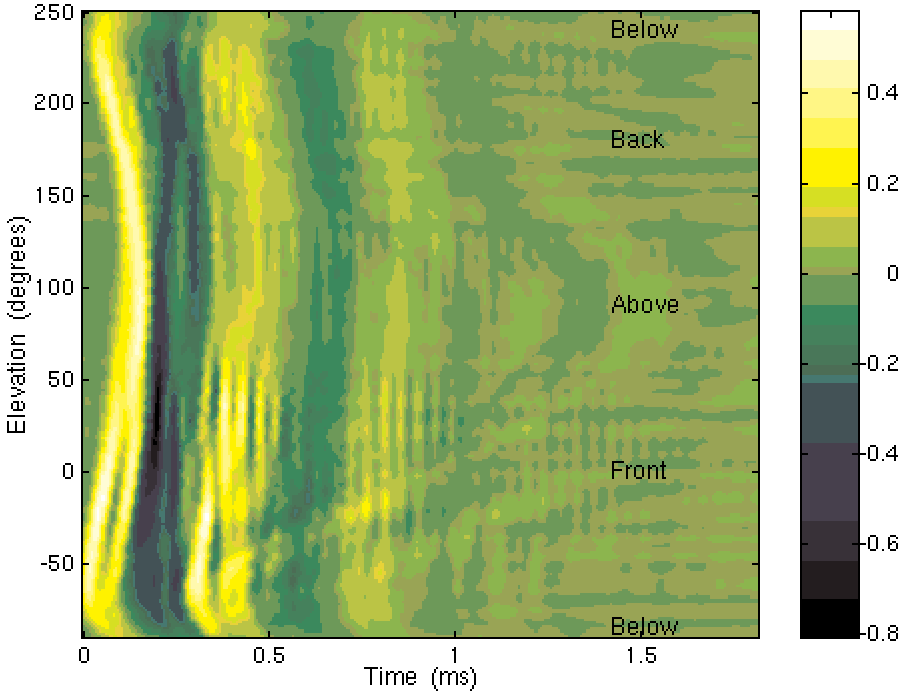
HRIR: Horizontal Plane



The figure above is an image representation of KEMAR's experimentally measured head-related impulse response of the right ear to an impulsive source in the horizontal plane. [Duda98]

The strength of the response is represented by brightness.
The sound is strongest and arrives soonest when it is coming from the right side (azimuth = 90°).
The difference between the shortest and the longest arrival times is about 0.7 ms, the theoretical value.
The response when the source is in front is similar to the response when the source is in back. Exisiting differences show up as a lack of perfect symmetry about a horizontal line at 90°.
People unconsciously use additional small head motions to discriminate between front and back in difficult unclear situations.

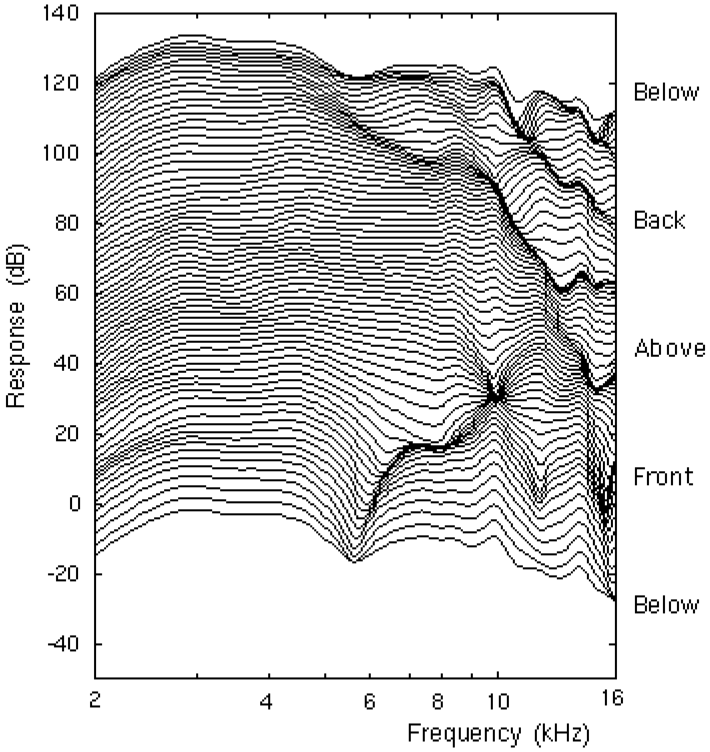
HRIR: Median Plane



source moving around the head in the median plane [Duda98]

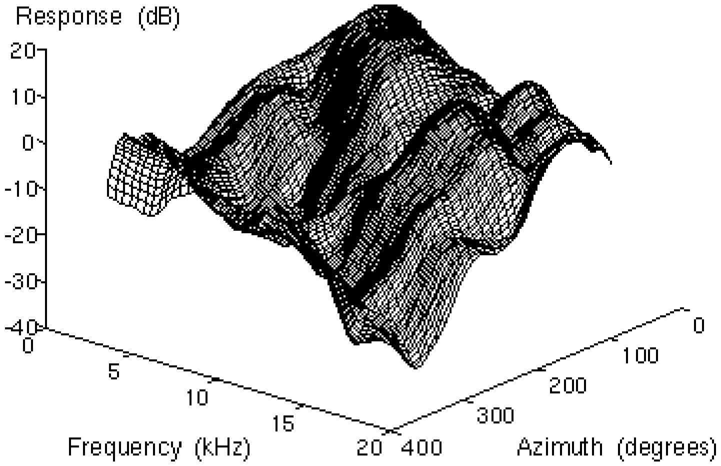
It is much harder to estimate the position in the median plane with high precision, therefore it is hardly suprising that the changes are much more subtle. As expected the arrival time is almost the same and the main changes occur in the strenghts of the pinna reflections. This can be seen as a notch whose frequency changes with elevation. The difference between front and back shows up in the moderate but clear lack of symmetry about a horizontal line at 90 degrees elevation.
(the artefacts seen at about 30° have been traced back to floor echo [Duda98])

HRTF: Median Plane



The frequency response looks as expected.
The broad ear-canal resonance around 4 kHz doesn't change.
The frequency is very flat and if the source is directly above the head and the notch has almost disappeared.
The behavior of the pinna features is quite sensitive to pinna shape, and varies considerably from person to person.

HRTF: Horizontal Plane

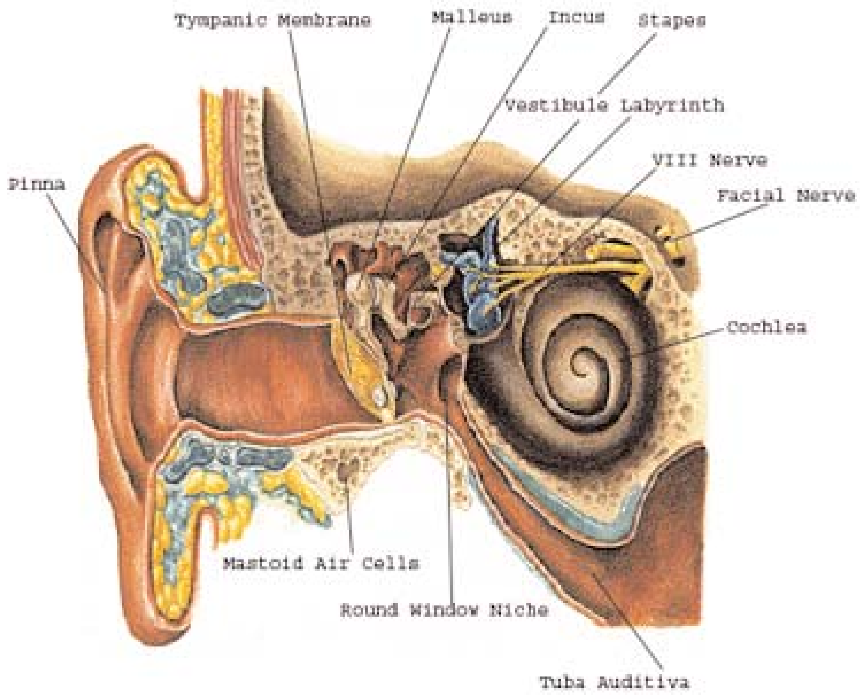


The complexity of the geometry makes solving the HRTF analytically an impossible task. Moreover, the wide range of wavelengths involved (17m at 20 Hz to 17mm at 20 kHz) discourages numerical solutions. Thus, most of what we know about the HRTF has come from direct experimental measurements

Despite of sound source localization, the HRTF can also be used to build 3D-audio systems like e.g. Microsofts Direct3D Sound to position a sound source anywhere in the virtual space.
(The DirectX SDK has some quite impressing examples).
For more on the HRTF please refer to R. O Duda’s tutorial on spatial hearing [Duda98] or [Begault94] and [Wright96].

2.4 The biological / neurophysiological basis

2.4.1 An overview of the anatomy of the ear.



How Do We Hear?

2.4.2 The Outer and Middle Ear.

Overview:
The outer part of the ear is involved with the collection of sound and its transmission into the inner ear. In a normal ear the process is quite efficient and enables the relatively small amount of energy in airborne sound to be transmitted into the fluids of the inner ear. Here it is converted into electrical energy and passed up to the brain via the auditory nerve.

The energy needed to hear is surprisingly small. [Zemlin97] notes that our ears are sensitive enough to hear the random movements of air molecules as they strike the eardrum (Brownian motion).

The hearing mechanism is made up of three parts:

1. Outer Ear
2. Middle Ear
3. Inner Ear

Vibrations in the molecules of the air are converted, first to vibration in solids (ossicles or ear bones), then to vibrations in liquids (cochlea), and finally to tiny electrical impulses in the nervous system which go up to the brain.

The outer ear has 2 parts:

The pinna,
which is composed of cartilage. Its presence on both sides of the head allows us to localize the source of sound from the front vs. the back. Our ability to localize from side to side depends on the relative intensity and relative phase of sound reaching each ear and the analysis of the phase / intensity differences within the brainstem.

The External Auditory Meatus or ear canal.
Vibrations in the air go via the pinna into this canal which leads to the eardrum. The first part goes through cartilage and the final part through bone. (see also the psychoacoustics chapter for more on the pinna and the ear canal). Its function is to conduct sound down to the eardrum. It is about 25 mm long (in adults). The wall of the outer 1/3 is cartilaginous and the inner 2/3 is bony.
Its resonance frequency of approximately 3400 Hz is an important frequency region for understanding speech.

The Middle Ear has 2 main parts:

1. The Eardrum or Tympanic Membrane is a tough but very flexible membrane which stretches across the ear canal.

The tympanic membrane consists of three layers, with the outer layer continuous with the skin of the outer ear canal.
The upper portion of the TM is called the pars flaccida, while the lower portion is called the pars tensa. The central portion of the pars tensa provides the active vibrating area in response to sound.

2. In general, the ossicles or ear bones are tiny bones. Their purpose is to carry sound across the ear. Thes bones are called after the things they look like: The MALLEUS (hammer), the INCUS (anvil) and the STAPES (stirrup).

The functionality is straightforward. Air molecules strike the eardrum causing it it to move it to and fro. Although, very small, this is enough to cause the chain of ossicles to vibrate. This ossicular chain vibrates and carries the vibration right across the middle ear to the inner ear.

The malleus is the most lateral (toward the side of the head) of the ossicles in the middle ear. The long process of the malleus is attached to the inner layer of tympanic membrane. When the TM vibrates in response to sound, the malleus vibrates in concert.

The incus is attached to the malleus, and so vibrates as the malleus vibrates. The long process of the incus is also attached to the head of the stapes. Because the long process of the incus is slightly shorter than the long process (manubrium) of the malleus, incoming sound is given a slight (2.5 dB) boost in energy. This is referred to as the lever advantage.

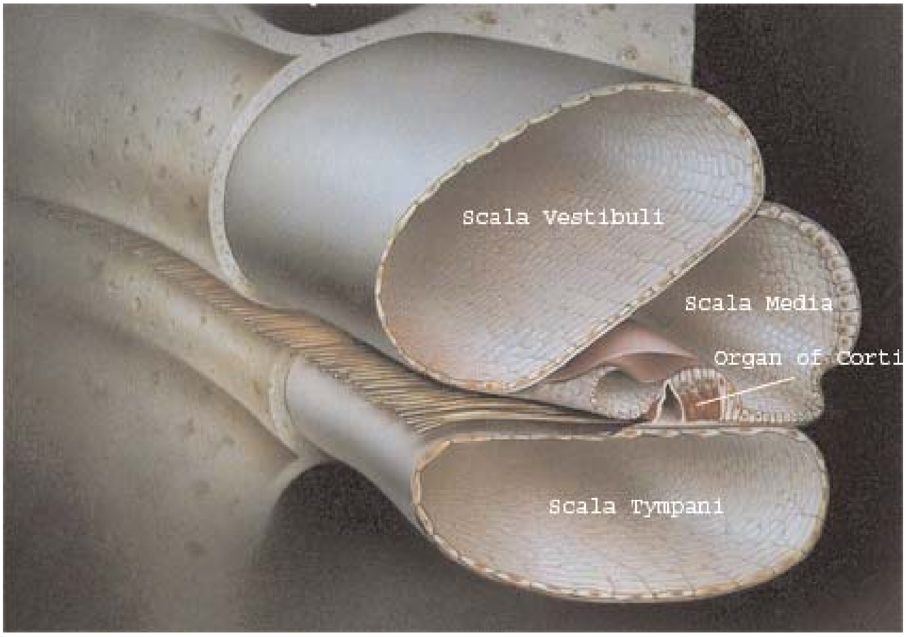
The stapes has a footplate and a superstructure. Its footplate is seated in the oval window. The oval separates the middle ear from perilymph of the inner ear. As the long process of the incus vibrates, so does the footplate of the stapes. Because the vibrating area of the tympanic membrane is larger than the area of the stapes, incoming sound is given a significant boost in energy of over 20 dB. This is referred to as the hydraulic advantage.

The round window is located within the round window niche. The round window is the most basal end of the scala tympani, and allows release of hydraulic pressure of perilymph that is caused by vibration of the stapes within the oval window. All ossicles are held in place by tiny muscles.These muscles have an important protective function.

The tensor tympani muscle is attached to the malleus and its contraction increases the tension of the eardrum. The stapedius muscle is attached to the head of stapes. Contraction of the stapedius muscle occurs as a response to loud sounds (about 80-90dB) and thus protecting the inner ear from damage. Unfortunately, protection is limited as there is a time lag of 10-30 msecs until contraction takes place. and only lower frequencies are affected and because only lower frequencies are affected. The last energy transformation occurs in the inner ear. Mechanical energy is transformed into the traveling wave pattern of the Membrana basilaris (basilar membrane), which is located in the Cochlea.

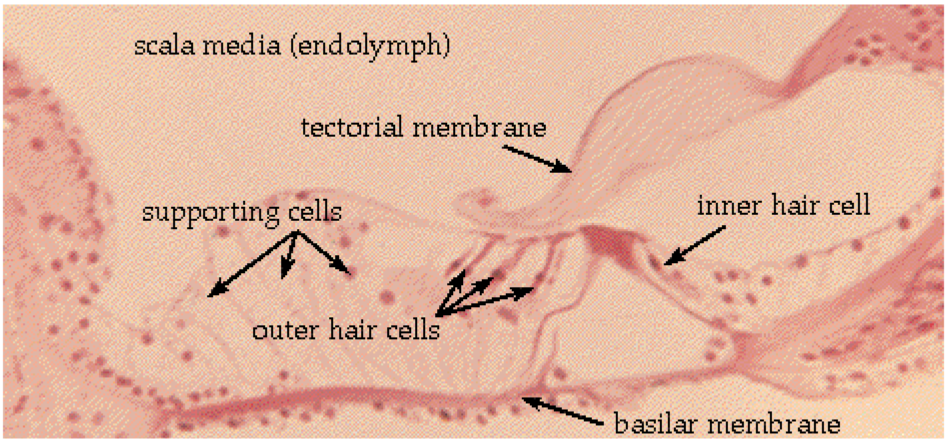
2.4.3 Cochlea

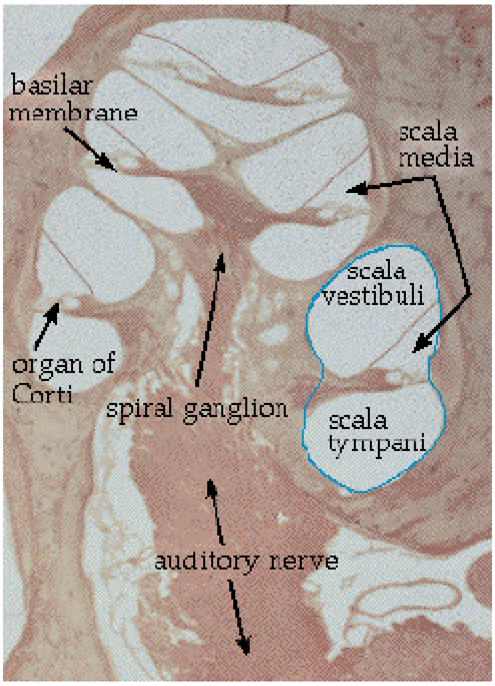
The Cochlea is a snail shaped structure and the sensory organ of hearing. The vibrational patterns that are initiated by vibration of the stapes footplate cause a traveling wave pattern within the cochlea. This pattern causes a shearing of the cilia of the outer and inner hair cells. This shearing causes hair cell depolarization resulting in all or none neural impulses that the brain interprets as sound.



the cochlea

2.4.4 The Organ of Corti





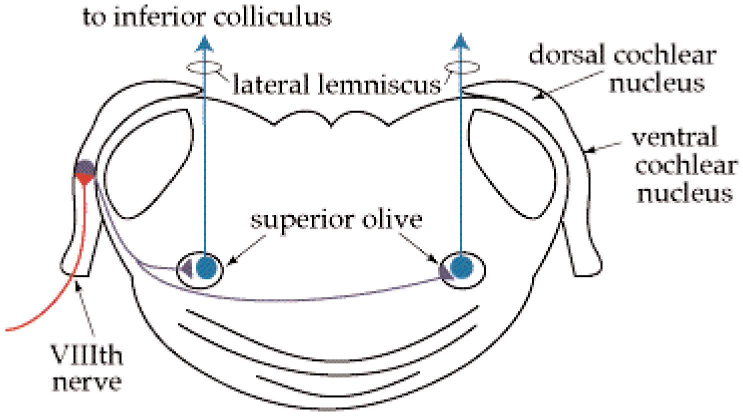
The Organ Of Corti, consisting of the Membrana Tectorialis, outer hair cells, inner hair cells, membrana basilaris, and supporting cells, is the sense organ of hearing. The outer and inner hair cells of the Organ of Corti change vibrational energy into neural energy, that is transmitted via the VIII nerve to the brain. The Membrana Tectorialis (Tectorial Membrane) is a delicate flexible gelatinous membrane overlying the sensory receptive inner and outer hair cells. The cilia (hair like tufts that extend from both outer and inner hair cells) of the outer hair cells are embedded in the Membrana Tectorialis. For inner hair cells, the cilia may or may not be embedded in the tectorial membrane. When the cochlear partition changes position in response to the traveling wave, the shearing of the cilia is believed to be the stimulus that causes the hair cells to produce an action potential.

The Membrana Basilaris itself is about five times wider at the apex than at the base. An important feature of the Membrana Basilaris important for the conduction of sound is that the base is about a hundred times stiffer than the apex. The Membrana Basilaris change as its shape changes. This Membrana Tectorialis vibrates to high frequencies at the base of the cochlea and to low frequencies at the apex. A hair cell at the base of the cochlea will respond best to high frequencies, since at those frequencies the Membrana Basilaris underneath it will vibrate the most.There are three rows of approximately 12000 outer hair cells. These cells contain muscle-like filaments that contract upon stimulation and fine tune the response of the Membrana Basilaris to the movement of the traveling wave.
There is one row of approximately 3500 inner hair cells .While the outer cells receive only about 5% of the innervations of the nerve fibers from the acoustic portion of the VIII nerve, the inner cells receive about 95%. These cells have primary responsibility for producing our sensation of hearing.

The signal carried by the auditory nerve into the brainstem and synapses in the cochlea nucleus splits into at least two streams.The ventral cochlea nucleus cells then project to a collection of nuclei in the medulla called the superior olive.
Ranjan Batra, Shigeyuki Kuwada, and Douglas C. Fitzpatrick have shown that estimation of the ITD, the interaural Temporal Disparities, a cue for localization of sounds along the azimuth takes place in the human brain by a an area called the Low- and High-Frequency Neurons in the Superior Olivary Complex. Sensitivity to ITDs originates in the main nuclei of the superior olivary complex (SOC), the medial and lateral superior olives (MSO and LSO, respectively).

[The Journal of Neurophysiology Vol. 78 No. 3 September 1997]

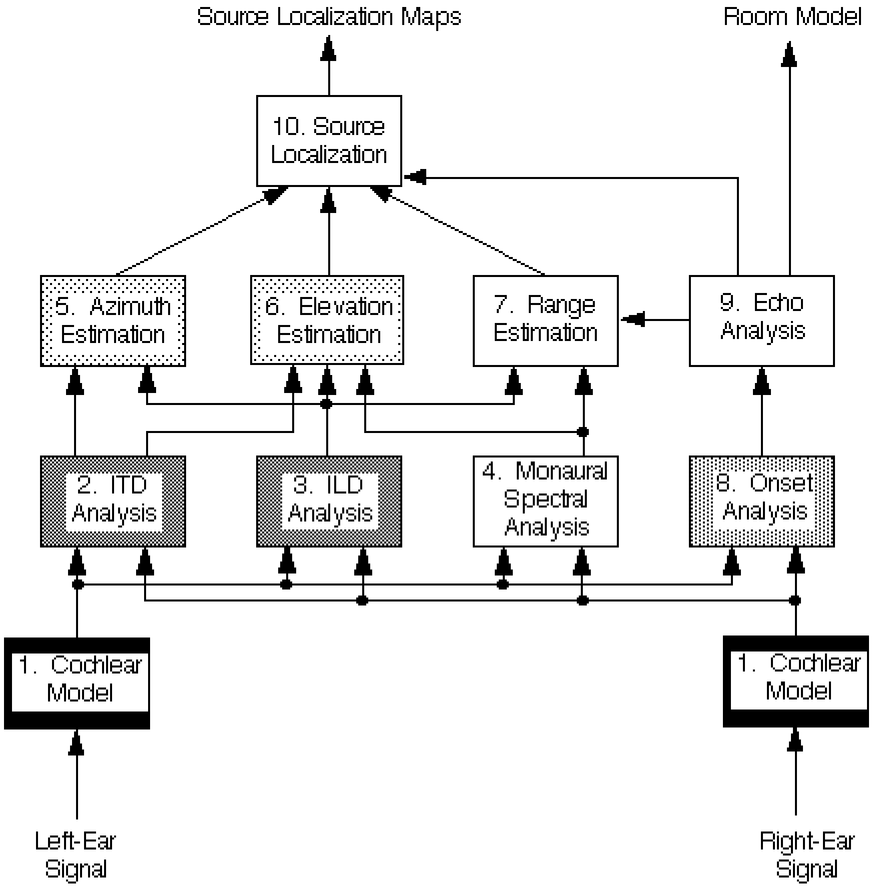
This supports Jeffress [Jeffress48)] classic cross-correlation model of horizontal localization, proposed long before neural correlation circuits were identified in the Medial Superior Olive (MSO)
Although primarily of historical interest, it is stiil in use, because the inclusion of automatic gain control makes it a robust method that is insensitive to amplitude scale. [Duda98]



location of the SOC

(for more on cochlea mechanics and neuroscience please visit the Cochlear Fluids Research Laboratory [CFRL02] and the Neuroscience Tutorial [NT97] of the Washington University School of Medicine), repectively.

2.5 Sound localization in general



[Duda98] suggests these major modules in a sound localization system. Shading indicates current degree of completeness.

Much work has already gone into developing functional models of the cochlea [Allen85], and this part of the model is essentially complete. The computational model used in this thesis is the model developed by Richard F. Lyon [Lyon82; Lyon88] and implemented by Malcolm Slaney.

The usage of his implementation, which is part of the Auditory Toolbox II, offers some convenient advantages:

1: It is available as part of the Auditory Toolbox II from the web for free.

2: It has also been implemented by Lyon in an analog VLSI chip to allow real-time execution
It properly accounts for such major phenomena as frequency selectivity and automatic gain control (AGC).
The dynamic range of human hearing exceeds 100 dB. While standard, commercial 16-bit recordings provide a dynamic range up to 96-dB. Even if background noise limits this to a useful 60-dB or so, that is still a 1000-to-1 ratio of amplitudes that must be accommodated, something that shoul not be ignored. E.g. in determining interaural time differences, the results should be insensitive to the amplitude scales in either of the channels. For more on AGC refer to [Lyon90] and to [Duda98] There are several serious problems related to a general sound source localization model.

2.5.1 ITD Analysis:

ITD, interaural time difference, can be determined e.g. using a combination of Jeffress crosscorrelation approach and AGC [Duda98].
See also Fourcin [Fourcin70], [Fourcin58] or [Akeroyd98] and [Summerfield98] (A fully-temporal account of the perception of dichotic pitches).

Lindemann [Lindemann86] and Gaik [Gaik93] propose introducing echo suppression at this point.
But, because of the Clifton-Effect [Clifton87] suppression should occur higher in the processing sequence [Duda98].

2.5.2 ILD-Analysis:

Reed, M. C, and J. J. Blum [Reed+Blum90] have developed a computer model to show that a model similar to Jeffress approach could generate a spatial map of ILDs analogous to the spatial map of ITDs in the MSO. Besides that, there are two general approaches: Lyon's cochlear model [Lim+Duda94] or an FFT e.g. Duda and Chau [Duda95+Chau95]. Lyon's cochlear model offers an advantage because it does not depend on spectral "fine structures" that people cannot hear.
(For more on auditory filters see ILD Analysis [Duda98])

2.5.3 Monaural Spectral Analysis:

ILD analysis provides the primary information for estimating elevation, According to Middlebrooks and Green [Middlebrooks+Green91] monaural cues dominate for sources in or very near the sagittal plane. (Carlile and Pralong 1994; Middlebrooks et al. 1989; Searl et al. 1975) reported small asymmetries in humans. Both monoaural and binaural spectral cues may be available for localizing in the midsagittal plane, thus the binaural approach is used in this work. Experiments with ferrets (Carlile and King 1994; Schnupp et al. 1998) [Carlile+King94], [Schnupp98] have shown that pinnae play a prominent role in resolving front-back ambiguities.

These experiments have shown that animals with pinnae removed make significantly more front-back errors compared to unaltered ferrets.
However, localization in the midsagittal plane, when a 500-ms stimulus was used, the pinnae-removed ferrets performed nearly as well as the normal ferrets in the midsagittal plane task, but performed comparatively poorly at shorter stimulus durations (200 ms).
Assuming slight headmotions of the ferrets, 500 ms signals might provide more information thus helping the brain of the modified ferrets to compensate the lack of pinnae.
But, it could also mean that for longer stimuli both the unaltered group performed as bad as the altered group.
This might be an explanation the for large errors reported by [Duda96+Chau96].

Because headmotions are believed to improve accuracy, more and more attention is paid to this subject [Angel+Algazi+Duda2002]

Despite of headmotions, elevation estimation, in partiicular, requires knowledge of the (monaural) spectral shape that is needed to achieve higher accuracy and robustness. Consequently, a general model should also include such a module.

2.5.4 Azimuth Estimation:

It is an important, but unsolved problem how best to combine ITD and ILD information. [Stern88] For a single wideband source in an anechoic environment, either the ITD alone or the ILD alone can provide very accurate azimuth estimates.

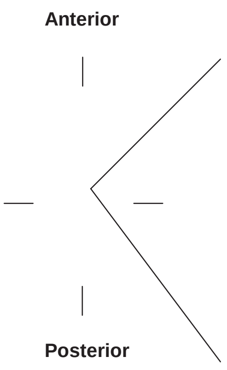
The problems are:
1. Room reflections severely limit the usefulness of ITD information at low frequencies,
2. Bandwidth limitations can severely limit the usefulness of ILD information at high frequencies.
3: Multiple simultaneously active sources confuse both methods.

ITD is relatively range-independent, whereas ILD changes significantly at close ranges.
Humans hev very few problems to cop with inconsistent cues, because they seem to ignore inconsitent values. [Wightman+Kistler93], [Hartmann95]

2.5.5 Elevation Estimation

Combining ILD and monaural elevation estimates, that are needed in the median plane is similar to raises similar problems to the ones occuring when combining ILD and ITD information for azimuth estimation.
Knowledge about how far the source is away from the midsagittal plane combined with elevation estimation might provide an increase in localization accuracy.

ITD and ILD alone, the primary interaural cues, are supposed to not provide enough information for localization. The positions where sources produce the same interaural cues occur in the so called cone of confusion which is shown to the left. To further improve localization accuracy, additional, dynamic cues, are needed. Cues like headmotions have recently beeing investigated. For results refer to [Angel+Algazi+Duda2002]



2.5.6 Range estimation:

Estimating range is the least well understood of the sound localization problems. The following cues to range are
commonly mentioned. (see Psychoacoustic) For distant sources, results of echo analysis to determine the ratio of direct to reverberant energy is needed. To reliably estimate range values, knowledge of the signal characteristics is needed.

2.5.7 Onset Analysis

Because of experiences of Tran and Tran [Tran+Tran93] with Lindemann's model [Lindemann86] and the Clifton Effect Duda [Duda98] believes that echo suppression cannot be implemented during the cross-correlation process and requires a more complex mechanism.

2.5.8 Echo Analysis

Clifton's experiments suggest that the first occurrence of an echo is not suppressed, but used to create a model of the acoustic environment to establish expectations that enable the suppression of subsequent echoes (see The Clifton Effect and [Duda98]).

2.5.9 Source Localization

Considering all the problems, locating a sound source in a normal non-anechoic environment is inherently difficult.

2.5.10 Characteristics of a sound source locator

There are several ways to build a good sound source locator that are more or less appealing to engineers depending on their scientific background.

1: An optimal locator.

Engineers with an analytical background might want to develop a mathematical model of the processes of sound propagation and diffraction by the head and pinnae to design an optimal locator. Unfortunately, this ignores that the results are critically dependent on knowlegde of geometrical and acoustic characteristics of the sources and the environment. They also ignore the bandwith of over nine octaves and that some wavelenghths are small whereas others are large relative to the size of the head and pinnae. All of this makes mathematical analysis and optimization extremely difficult if not impossible.

2: A clever sound locator

Ignoring the head and pinnae and inventing some kind of scanning or beam-forming system that can lock onto the main source and reject reflections and other sources and use an neural network to tune the system to particular environments may sound appealing to practically inclined engineers. But, If it is the goal to develop a system to resemble the human auditory system, ignoring the impact of the head and pinnae is clearly a very bad idea.

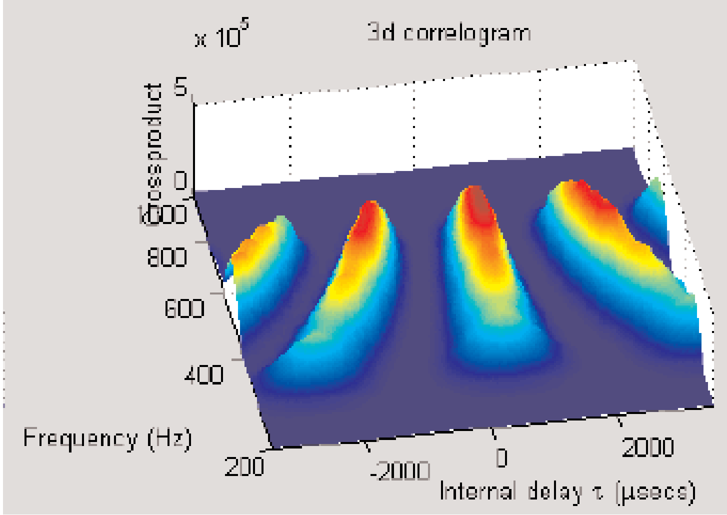
3: A biological sound locator

The problem with this is clearly the lack of fully understanding the neurophysiological background of how people localize sounds. However, much is known about how acoustic events are transduced and transformed by biological auditory systems [Pickles88].
Unfortunately, neurophysiological studies have not yet provided us with all the facts and reveiled secrets of hearing.
However. available information is sufficient for designing a general architecture for a localization system. Biological studies of barn owles, suggests a processing architecture that progresses through a series of tonotopic maps to a final spatial map.

I believe if we want robust performance that is comparable to the human auditory system we have to find ways to employ the same general mechanisms that are used by the human auditory system and the usage of ALN's might provide a solution to address those problems, thus we use a more biological approach here.

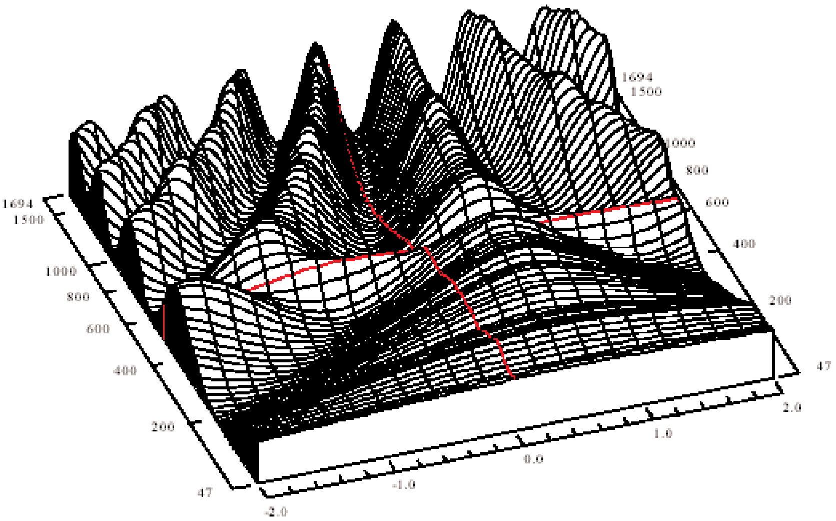
2.5.11 The importance of time (see also [Slaney+Lyon93])

The position in the transverse plane primarily depends on two cues, the ITD and the ILD To compute the azimuth we need to establish an adequate temporal representation of the incoming signal(s). This temporal representation is achieved by computing the so called correlogram (autocorrelation) from the output of the cochleamodel.
The correlogram is a three dimensional function of time, frequency and periodicity.



[Akeroyd01]

The azimuth between two signals (left ear and right ear) is then computed from a crosscorrelogram (shown below).



[Akeroyd01]

2.5.12 Lyons Cochlea Model

All Cochlea models are used to resemble the nonlinear charcteristic of human hearing.The model used in this work is called a “passive long-wave model”, as opposed to the “active short-wave model.” [Lyon+CarverMead88]. Differing in their underlying assumptions, approximations, and implementation structures, they share three primary characteristics.

Both models perform filtering by modeling the propagation of energy as waves on the Membrana basilaris (MB) utilizing a broadly tuned cascade of lowpass filters.

A detection nonlinearity converts MB velocity into a representation of inner haircell (IHC) receptor potential or auditory nerve (AN) firing rate.(Detection)

Compression is performed by Automatic Gain Control (AGC) continuosly adapting the operating point of the system in response to its level of activity, thus compressing varying sound levels into a limited range of MB motion, IHC recptor potential and AN firing rate.

The passive long-wave model is based on a one-dimensional (long-wave) hydrodynamic approximation with a lightly-damped resonant membrane [Zweig+Lipes+Pierce76].

The active short-wave model is based on a two-dimensional hydrodynamic approximation (emphasizing the short-wave region) with active undamping and negligible membrane mass [Lyon+CarverMead89].

The passive long-wave model uses time-invariant linear filters followed by a variable gain to functionally model the AGC. The active short-wave model varies the filter pole Q over time to effect a gain variation and to model the mechanical AGC in terms of active adaptive hydrodynamics.[Slaney+Lyon93]
See also [CFRL02] for more on cochlea fluids.

(For a detailed description of the passive long wave and the active shortwave model see [Slaney+Lyon93] and [AppleTechnicalReport #13 1988].

2.5.13 Auditory Scene Analysis

At a crowded party there are many competing voices and other interfering sounds such as music. Nonetheless, we are able to concentrate on a particular voice or sound in these situations. How does the ear achieve this apparently effortless separation of concurrent sounds ? E. C. Cherry in 1953, noticed this phenomenon and called it the “cocktail party problem”. Since then, it has been the subject of extensive psychological research. In 1990 a psychologist named Albert Bregman of McGill University, Montreal published his book about the characteristics of sounds that cause people either to fuse them into a single sound "object " or to separate them into different sound "objects " called sound streams. Bregman argues that the mixture of sounds reaching the ears is subjected to a two-stage auditory scene analysis (ASA). In the first stage, the acoustic signal is decomposed into a number of “sensory elements”. Elements that are likely to have originated from the same environmental source are grouped into perceptual structures that can be interpreted by higher-level processes (such as those involved in speech understanding). Bregman makes a distinction between two types of auditory grouping; primitive grouping and schema-driven grouping. Primitive grouping is driven by the incoming acoustic data, and is probably innate. In contrast, schema-driven grouping employs the knowledge of familiar patterns and concepts that have been acquired through experience of acoustic environments. The ability to adapt to the auditory environment and to concentrate on a particular sound source is an important aspect of robust speech recognition. An ultimate speech recognition systems should be capable of robust, speaker-independent or speaker-adaptive, continuous speech recognition. Thus we need to have methods that are robust against voice variation due to individuality, physical and psychological condition of the speaker, room- microhone and telephone characteristics as well as background noise and speaking styles. To solve these problems we need to develop automatic adaptation techniques. For robust speaker recognition or verification there are still some questions unanswered. By now we do not know how humans correctly recognize speakers. We do not know either what feature parameters are appropriate for speaker recognition as well as we do not know how can we fully exploit suprasegmental features of speech and how to cope with long-term variability in people's voices

2.5.14 Multipath Effects

Normally, we are surrounded by a mixture of various signals. Let h(t) be the transfer function. Remarkably, we are able to recover individual components from the convolution of the sum p1(t) + p2(t) + p3(t) + p4(t) with h(t), quite easily, which is even more remarkable since h(t) due to reverberations can be unbelievable complicated. These secondary signals caused by reflections become only audible if the delay exceeds about 30 to 40 ms compared to the direct sound.

2.5.15 Localization of multiple moving sound sources: The KItano project.

HRTF's are usually measured in an anechoic chamber. Consequently, room characteristics should be measured in advance to use HRTFs. In addition HRTFs are available only at discrete points due to discrete measurement. Therefore, such sound source localization methods are running into severe difficulties when used with a robot that changes its position by moving or rotation. Sound source separation techniques, such as beamformers with a microphone array [Asano+Goto+Itou+Asoh01], blind source separation and computational auditory scene analysis (CASA) techniques based on human auditory system [Mizumachi+Akagi98] are still not providing satisfying perormance and therefore not applicable to the above problem.

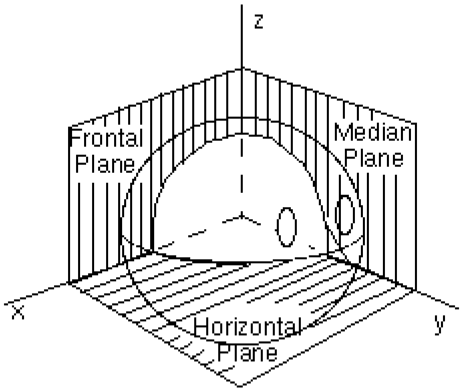
To cope with these technical problems the Kitano project proposes a method called the active direction-pass filter (ADPF). The ADPF does sound source localization by auditory epipolar geometry without using HRTFs or the measurement of acoustic environments in advance [Nakadai+Lourens+Okuno+Kitano2000]. The ADPF is implemented as a part of the real-time multiple speaker tracking system installed on an upper-torso humanoid called SIG. Motor noises and other noises caused by its movement are canceled by using cover acoustics. The tracking system attains accurate localization of multiple face recognition, and active motion control of SIG [Nakadai+Okuno+Kitano2001].

The approach to localize sound sources used by the Kitano project is called Sound Source localization by Integration. At first a STFT is applied to the input sound. A peak on the spectrum is extracted by a subband selection, a subband with a frequency between 90 Hz and 3 KHz. If its power is a local maximum and more than a threshold, which is automatically determined by stable auditory conditions of the room. Then, extracted peaks are clustered according to harmonicity. A frequency of Fn is grouped as an overtone (integer multiple) of F0 if <= 0.06.The constant 0.06 was determined by trial and error.

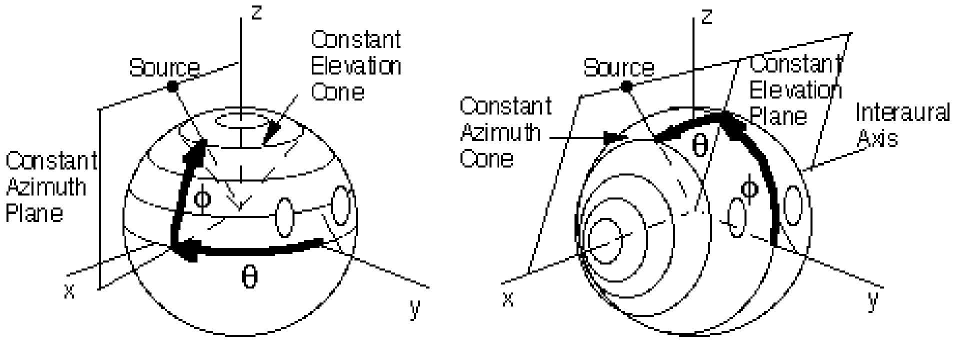
Applying an inverse FFT to a set of peaks in harmonicity, a harmonic sound is reconstructed, and thus seprated from a mixture of sounds.
Sound Source Localization is done in four steps.
1. Localization by interaural phase differnce IPD and auditory epipolar geometry.
2. Localization by interaural intensity difference (IID).
3. Integration of overtones.
4. Integration of 2. and 3. by Dempster-Shafer theory. (for a detailed description refer to [Hiroshi+Okuno+Nakadai+Kitano2002]

Visual input provides a very useful cue to localization of a source out of multiple sound sources.
However, without visual input, the correct source has to be separated and located in a different way from the mixture of sounds. This “way” is called Blind Source Separation. Blind Source Separation is particularly important for simulatenous speech recognition and to cope with the cocktailparty effect or cocktailparty problem. The Kitano project is using cameras to identify and localize the person.
This raises the question how visual cues affect actual human localization abilities. [Zwiers+Opstal+Cruysberg2001] have investigated the two-dimensional sound-localization behavior of early-blind humans. And they have not found support for the hypothesis that blind people can hypercompensate for the loss of vision in the frontal hemifield by developing superior sound-localization abilities, therfore it can be concluded that visual feedback is not required to calibrate the available localization cues. Moreover, this also includes the complex spectral shape cues for elevation. The localization abilities of blind people show that the putative supervising role of vision may be supported, or replaced by other non-visual feedback systems
However, utilizing visual input seems to provide an excellent cue.

2.5.16 Coordinatesystems



To specify a spatial position we need some kind of coordinate system. The coordinate shown to the top is the head-centered rectangular-coordinate system. The x-axis goes (approximately) through the right ear, the y axis points straight ahead, and the z axis is vertical to the xy-plane. The transverse-plane or horzontal-plane defines up/down, the frontal-plane front / back and the mid saggital-plane or median-plane left/right separation. The head is roughly spherical, so we can use a spherical coordinate-system. The three standard coordinates are azimuth, elevation and range. There is more than one coordinate-system, so we have to chose that best fits our needs.



Vertical-Polar Coordinates

Interaural-Polar Coordinates

The system to the left, the Vertical-Polar-System, is the most popular one. It is also the one and the one used in this thesis. The azimuth is measured as the angle from the median plane to a vertical plane containing the source and the z azis, and then measures the elevation as the angle up from the horizontal plane. Surfaces of constant azimuth are planes through the z axis, and surfaces of constant elevation are cones concentric about the z axis. An important alternative is the interaural-polar coordinate system, shown on the right. Elevation is measured as the angle from the horizontal plane to a plane through the source and the x axis, which is the interaural axis; the azimuth is then measured as the angle over the saggital plane or median plane.

3 Sound Source Localization using an ALN

The approach in this thesis uses a rather simple model to estimate the elevation. What follows is a description of the components, their relationships and why it makes sense to use them.

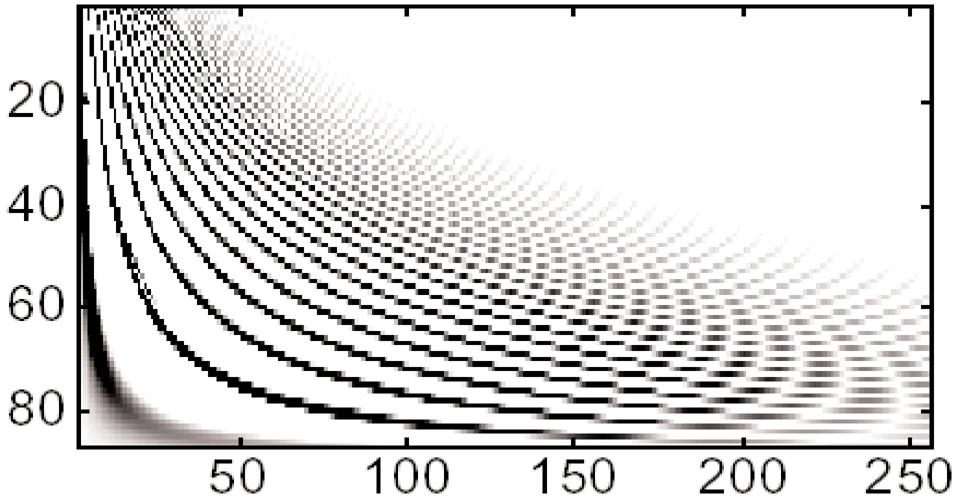
3.1 Cochlea and Correlogram

The aproach utilizes Malcolm Slaneys implementation of Lyon's Cochlear Model from his Auditory Toolbox Version 2.

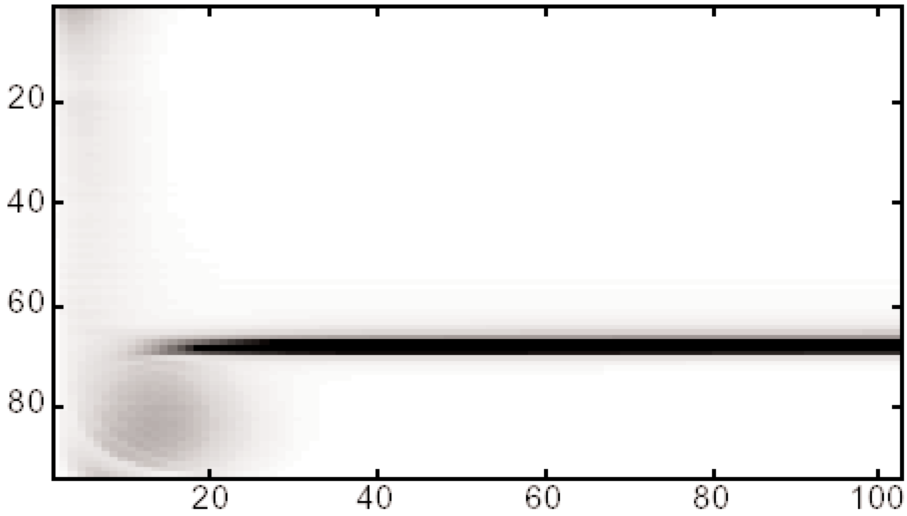
This auditory model is based on a transmission line model of the Membrana Basilaris followed by several stages of adaptation. The model is configured to resemble the functionality of the human ear as close as possible.Usage of a Cochlea Model instead of a short time is advantageous, because normal short-time-Fourier-Transform (STFT or Spec-trogram) reveals high frequency spectral details that are beyond what people can resolve. Also, these narrow-band high-frequency components provide little energy for everyday sounds. This does not automatically mean that using STFT is generally a bad idea. STFT has been used in sound source localization systems several times with remarkable results e.g. [Duda95+Chau95] [Kitano02] However, modeling a system that resembles human performance should implement computational models resembling the human auditory system, a cochlea that is. After the Time-Frequency Analysis the output of the Cochlea Model, the Cochleagram, is further processed using autocorrelation to summarize the periodicities of the cochlea output. The correlogram can be used as a way for understanding multiple sounds and to perform auditory scene analysis. Inversion of these stages can be effectively used for sound separation, in particular for separating speech from background noise. [Slaney94]

As an input for a neural network or Adaptive Logic Network correlograms are better for estimation than cochleagrams because they represent the time periodicities in a spatial way. The correlogram changes slowly, giving a pattern that shows the periodicities that are happening in the cochleagram on a faster scale. The correlogram "demodulates" the cochleagram and gives the correlation patterns in each frequency band. Each column number denotes an index within the temporal analysis window. Precisely, the column number indicates the time delay in the correlation analysis. The first column is zero delay, the autocorrelation. The second column is one sample delay and so on.

The correlogram is an example of an auditory map. Throughout the higher auditory system, it appears that important acoustic parameters are represented in two- or three-dimensional arrays of cells. The value of the parameter at a particular frequency is indicated by the firing rate of the cell at the appropriate position in the neural array.
The approach presented in this thesis utilzes methods successfully used in the past. The important distinction is, that instead of a simple neural network system utilizing a maximum-likelihood [Chuck Lim94], an ALN, Adaptive Logic Network, is used.



A cochleagram



The correponding correlogram of the above cochleagram

3.2 ALN’s

The ALN is a simple concept applied to fitting empirical data: It basically uses a form of linear regression to fit the data. Linear regression, more generally, allows it to form linear combinations of any functions to approximate given empirical data. The limitation is that with a finite set of basis functions and hence a finite set of weights, the set of functions that can be produced is limited, and may not be adequate for a given problem. However, a large number of basis functions is able to fit given data well. Unfortunately, the result may not be good for fitting new data generated from the same source, because the result is likely to not "generalize" well. An ALN, as in linear regression, fits data by first starting off with one linear function, which is called a (lnear) piece. If the fit is not good enough, the piece splits into two, initially the same as the one that split, and each piece fits its own set of data points. The combination of the pieces is the maximum or the minimum of two linear functions, resulting in a V-shape or upside-down V. To make the junction of the two pieces smoother, a quadratic "fillet" is inserted. The process of splitting continues unless the fit is “good”. Eventually, the result is a tree of maximum and minimum functions combining linear functions. This, with fillets inserted at every maximum or minimum node is the final form of the ALN. The ALN grows dynamically as necessary to fit the data. The technique of bagging, developed by L. Breimann [Breiman99] is used to effectively improve generalization.
([Armstrong2001])

3.2.1 Bagging

Bagging (Breiman 1997) works as follows: First, training pairs, consisting of the data at the initial time and the forecast target of certain months (lead time) later, are formed. The available training pairs are separated into a training set and a test set. The test set is reserved for testing only and not used for training. The training set is used to generate a set of neural network models; each member of the set is trained by only a subset of the training set. The subset is drawn at random with replacement from the training set. The subset has the same number of training pairs as the training set; some pairs in the training set appear more than once in the subset, and about 37% of the training pairs in the training set are absent in the subset. The final model output is the average of the outputs from all members of the set. The advantage of bagging is a reduction of the variance, or instability, of the neural network. The error surface of neural network training is full of local minima; trainings with different initial weights and training data are usually trapped in different local minima. These local minima reflect partly the fitting to the regularities of the data and partly the fitting to the noise in the data. Bagging tends to cancel the noise part as it varies among the set members, and tends to retain the fitting to the regularities of the data. Using adaptive bagging to debias regressions [Breiman96] showed that bagging could effectively reduce the variance of regression predictors, while leaving the bias unchanged.

3.2.2 Adaptive Bagging

A new form of bagging, called adaptive bagging, is effective in reducing both bias and variance. The procedure works in stages--the first stage is bagging. Based on the outcomes of the first stage, the output values are altered and a second stage of bagging is carried out using the altered output values. This is repeated until a specified noise level is reached. [Breiman99].

3.3 ALN versus ML methods

It is important to say that in general ALN and ML methods are not competitors. It depends more on whether or not the assumptions of the probability distributions are known, which is why ML-methods are likely to produce relatively poor results. However, if the the assumptions are indeed reasonable, an empirical non-linear method like ALNs with bagging is likely to not perform better than the ML-method. A clear and important advantage of ALNs with bagging over “regular” neural networks is that those “regular” neural networks can be overtrained relatively easy. Those overtrained networks tend to produce unbelievably good results unless the conditions remain unchanged. Unfortunately, slight changes, that usually are no problem to the network result in unbelievably bad results. Of course, overtraining an ALN does also produce an ALN that does not generalize very well, but using the technique of bagging, developed by L. Breimann mitigates the effect of overtraining. Another mitigating factor is, that overtraining is never done to completion, just to a very small error,

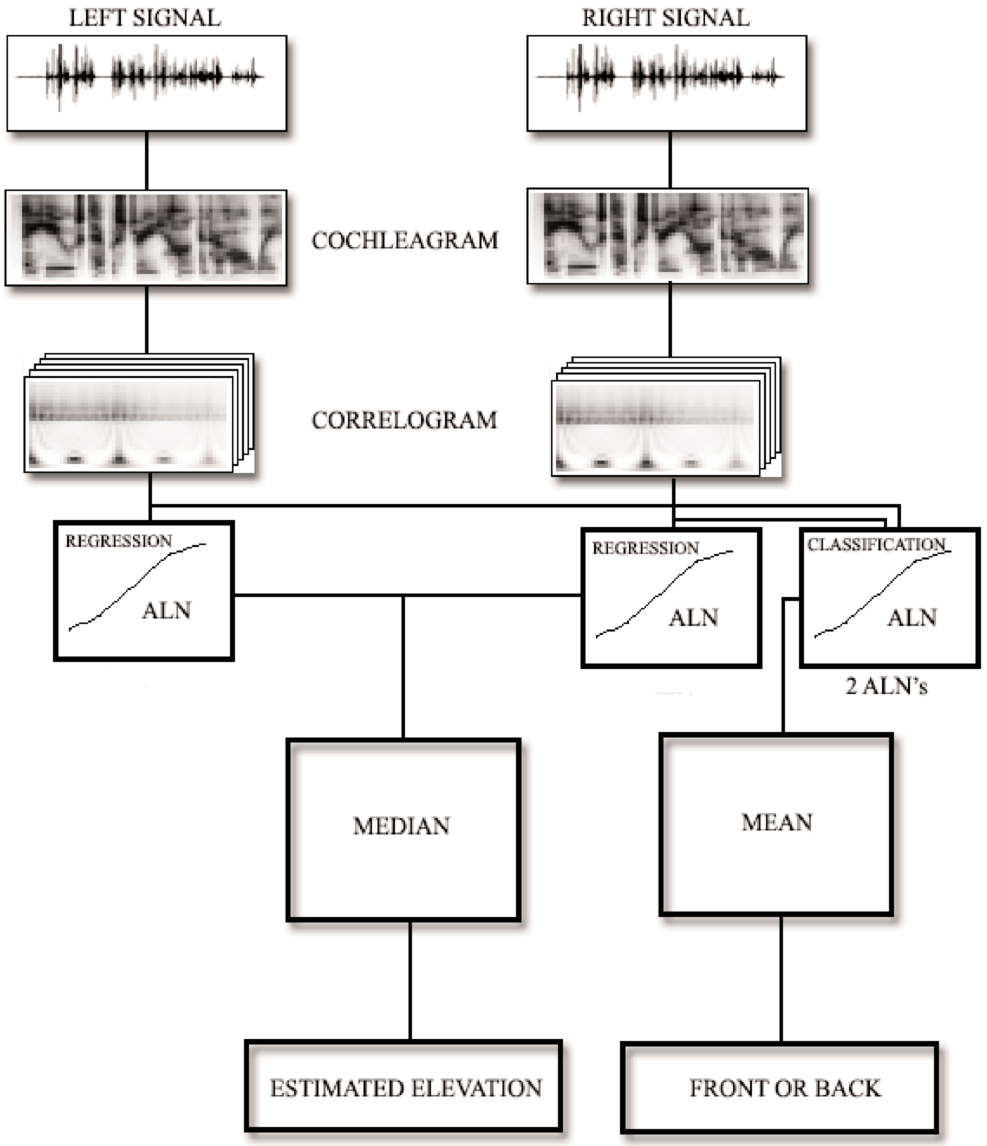
3.4 ALN’s for elevation estimation

To utilize an ALN and the Dendronic Learning Engine, the SDK used in this work, a datafile / matrix with several rows of input data is built.
Each row in this matrix represents a normalized corellogram pattern built from cochleagram data of input data taken at several azimuthal coordinates and elevation angles from -40 to 90 degrees around the head. The rightmost column of each row is the ouputvalue, the elevation angle. Each column is the autocorrelation result of one cochleagram channel. After training several ALN’s with this data, an average ALN is generated from these ALN’s (see Bagging). Finally, a data structure called DTree is computed from the average ALN and used to estimate the elevation angle for known and unknown signals. The signals used in this work for training are the HRTF measurement itself, white noise, and a humming “bee” sound. To use the DTree we need to generate a row of a normalized correlogram pattern and then use this as an input for the DTree and take the result as the estimated elevation angle. These Dtrees can estimate the elevation very fast (less than 1 second on a 1 Ghz Athlon), For even higher speed they can be tranformed into hardware (e.g. FPGA’s).

3.5 ALN’s for front / back classification

To distinguish whether or not the sound source is located in the front or in the back, two additional ALN’s (one for each “ear”) are used for his purpose. These ALN’s were specifically trained to act as classifiers, thus a 0 and a 1 were used in the rightmost column of each row. A 0 denotes that the sound source is located in the front, a 1 that the sound source is located in the back. Because of the subtle changes in the sagittal plane it is expected that the amount of errors increases, the closer the source is located to the cone of confusion. (also see the psychoacoustic chapter for more)

3.6 Experiments and results



The diagram shows how the elevation is estimated and how front / back distinction is performed.

All signals used in this Thesis were filtered using the complete HRTF by [Gardner+Martin94] taken at various degrees of azimuth around the head . The HRTF measurements were taken at different steps of azimuth. Elevation ranges from -40 to 90 degree in 10 degree steps, (above the head) The number of azimuthtal measurements depends on the degree of elevation they correspond to. At elevation of 0 degree there are 72 measurements, whereas at 90 degrees there are only 2. (for each ear). The total amount of measurements is 1422 (711 for each ear).
The training and test sets, the autocorrelation outputs (see diagram), for each "ear" were then precomputed and saved to a text-file, one for each ear. Each row in the file corresonds to one position of elevation and a degree of azimuth. The rightmost column denotes the desired degree of elevation. When used as for testing, the desired ouput value is, of course, ignored. However, the textfile was then used to train and compute the ALN's / Dtrees to perform classification and regression at very high speed using a Matlab program that loads the Dtres, loads the testsets and estimates the elevation and direction (front/back) for each degree of azimuth used in the testset. Before we come to the actual testing results, we have to take a closer look at some data gathered during training the Dtrees. These errors indicate what we can expect from the resulting Dtrees. All Dtrees were computed from the average ALN of 20 ALNs using bagging for good generalization.

1: Dtrees trained with only the HRTF
Dtrees to classify between front and back

Left ear:

```
RMS error between DTREE and the desired output on test set = 0.310080
RMS error between DTREE and the average of ALNs on the test set = 0.116984
The number of misclassified members of the test set is 29 or  8.285714 percent
The number of misclassifications on the user's test set is 46  or 6.469761percent
```

Right ear:

```
RMS error between DTREE and the desired output on test set = 0.283277
RMS error between DTREE and the average of ALNs on the test set = 0.094309
The number of misclassified members of the test set is 28 or 8.000000 percent
The number of misclassifications on the user's test set is 38 or 5.344585 percent
```

Dtrees to estimate the elevation:

Left ear:

```
RMS error between DTREE and the desired output on test set = 10.389974
RMS error between DTREE and the average of ALNs on the test set = 3.165660
```

Right ear:

```
RMS error between DTREE and the desired output on test set = 9.757914
RMS error between DTREE and the average of ALNs on the test set = 3.458636
```

2: Dtrees trained with the HRTF and white noise
Dtrees to classify between front and back

Left ear:
RMS error between DTREE and the desired output on test set = 0.304940
RMS error between DTREE and the average of ALNs on the test set = 0.117513
The number of misclassified members of the test set is 58 or 8.285714 percent
The number of misclassifications on the user's test set is 105 or 7.383966 percent

Right ear:
RMS error between DTREE and the desired output on test set = 0.281503
RMS error between DTREE and the average of ALNs on the test set = 0.115388
The number of misclassified members of the test set is 53 or 7.571429 percent
The number of misclassifications on the user's test set is 86 or 6.047820 percent

Dtrees to estimate the elevation:

Left ear:
RMS error between DTREE and the desired output on test set = 11.194749
RMS error between DTREE and the average of ALNs on the test set = 5.069472

Right ear:
RMS error between DTREE and the desired output on test set = 11.494710
RMS error between DTREE and the average of ALNs on the test set = 5.663293

3: Dtrees trained with the HRTF, white noise and the bee sound
Dtrees to classify between front and back

Left ear:
RMS error between DTREE and the desired output on test set = 0.321162
RMS error between DTREE and the average of ALNs on the test set = 0.100274
The number of misclassified members of the test set is 79 or 11.285714 percent
The number of misclassifications on the user's test set is 177 or 8.298172 percent

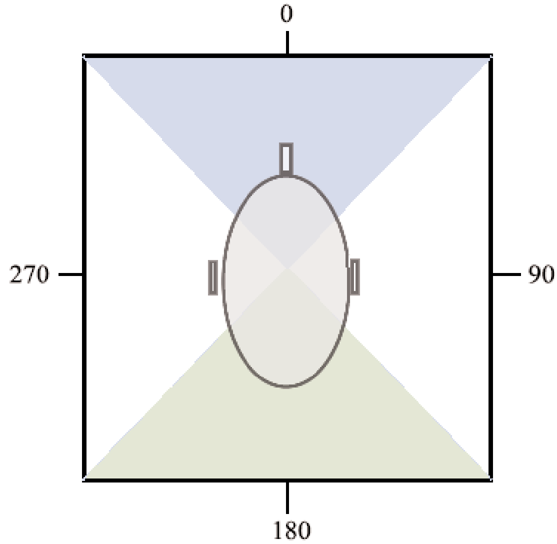
Right ear:
RMS error between DTREE and the desired output on test set = 0.327860
RMS error between DTREE and the average of ALNs on the test set = 0.100061
The number of misclassified members of the test set is 84 or 12.000000 percent
The number of misclassifications on the user's test set is 188 or 8.813877 percent

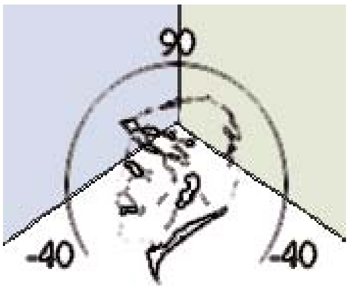
Dtrees to estimate the elevation:

Left ear:
RMS error between DTREE and the desired output on test set = 15.261611
RMS error between DTREE and the average of ALNs on the test set = 4.344428

Right ear:
RMS error between DTREE and the desired output on test set = 15.303272
RMS error between DTREE and the average of ALNs on the test set = 4.688738

Please note:
Testing data was never used for training. The test signals used in the experiments were not used at all during training to avoid cheating.
E.g. the bee signal is not identical with the one used for training.
Also, please note that because of lacking a KEMAR, the accuracy is expected to be higher than what one would expect within a real anechoic room, since no real anechoic chamber in the world is perfect (E.g. there might be some floor echo) The bee sample is not an actual sound of a real bee, it is more a silly bee impersonation of a student.





Two sets of experiments were conducted with one set using test signals only near the mid sagittal plane (blue and green, when looking from the top) and the second set that included the white lateral area of the head for testing.

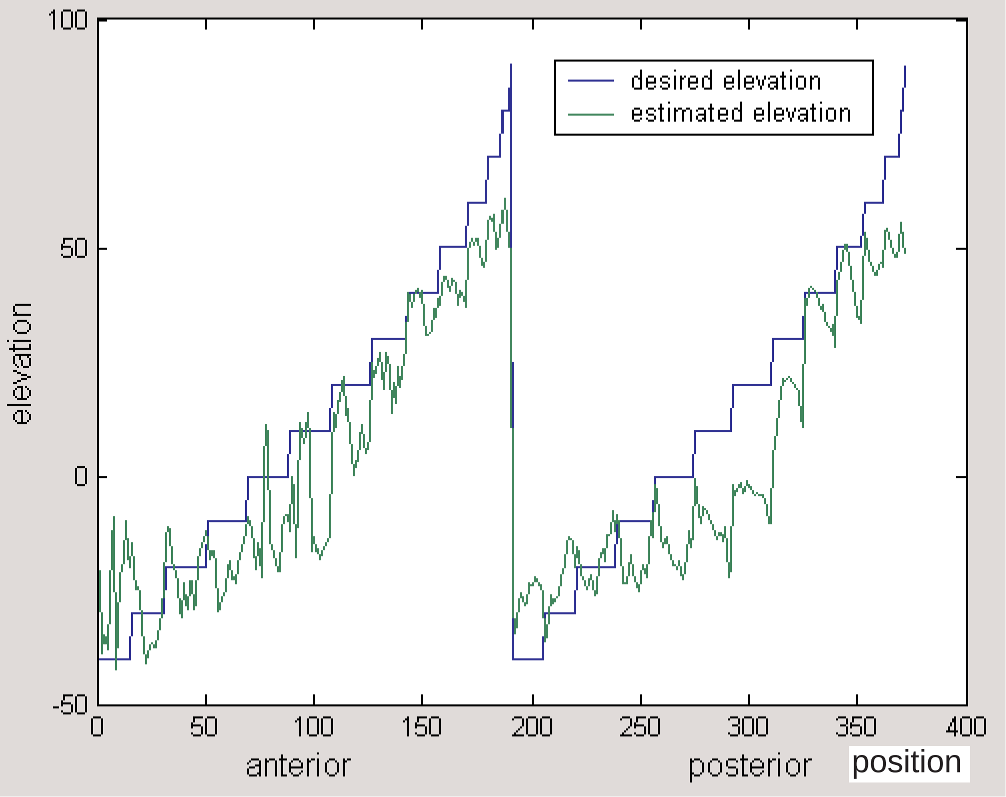
Except for these differences both tests were conducted the same way

First, the accuracy of Dtrees trained solely with pure HRTF measurements was tested on the bee sound and white noise.
Second, the accuracy of Dtrees trained with HRTF measurements and white noise was tested on the bee sound and white noise.
And, third, the accuracy of Dtrees trained with HRTF measurements, white noise and the bee sound was tested on a similar bee sound and white noise. The table on the next page shows the results of these experiments. The results above the red line are the ones gathered by omitting the lateral areas at testing. The Dtrees were of course the same for both sets of experiments.

DTrees trained with:	min. elev. error	max. elev. error	avg. elev. error	min. RMSE	max. RMSE	avg. RMSE	front / back errors in %
pure HRTF							
bee sound	0,21	102,58	23,80	2,26	33,75	22,46	48,92%
white noise	0,37	1102,40	32,84	1,54	91,46	58,18	19,36%
Number of positions	372						
DTrees trained with:							
pure HRTF and white noise							
Tested on:							
bee sound	0,00	93,23	24,86	10,35	31,13	23,59	48,92%
white noise	0,22	75,16	33,44	0,75	36,28	30,41	22,04%
Number of positions	372						
DTrees trained with:							
pure HRTF white noise and bee sound tested on:							
bee sound	0,05	81,78	16,31	2,49	24,75	16,96	48,92%
white noise	0,11	41,34	11,82	9,12	19,11	13,09	2,69%
Number of positions	372						
DTrees trained with:							
pure HRTF							
Tested on:							
bee sound	0,15	232,63	24,52	2,26	52,30	32,80	47,96%
white noise	0,03	120,02	38,76	16,11	48,23	36,59	46,13%
Number of positions	711						
DTrees trained with:							
pure HRTF and white noise							
Tested on:							
bee sound	0,00	93,23	23,61	10,35	30,51	23,41	47,96%
white noise	0,65	110,00	40,65	10,70	47,60	36,50	38,26%
Number of positions	711						
DTrees trained with:							
pure HRTF white noise and bee							
Tested on:							
bee sound	0,04	81,78	16,19	2,49	23,92	17,42	48,10%
white noise	0,92	115,11	36,14	9,60	45,22	32,92	47,26%
Number of positions	711						

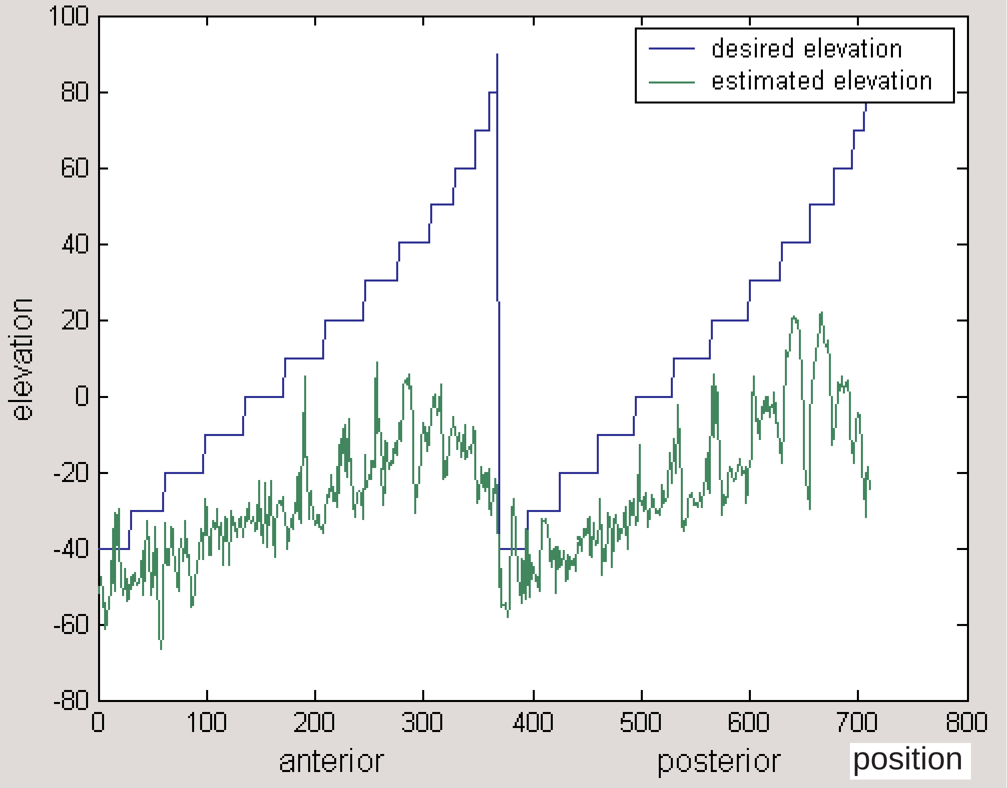
3.6.1 Interpretation

The results above the red line show, that, compared to the results below the red line, estimation accuracy is highest in a 90° cone anterior and posterior to the head, just as expected from theory and personal experience. While listening to test sounds over a headphone at positions around the head, it was much harder for the sounds that were positioned within the cone of confusion to estimate their correct position.The results also show that a certain knowledge about the shape of the signal is indeed required to achieve a decent increase of accuracy. It should also be pointed out that by testing near the mid sagittal plane and using broadband signals like noise overall front / back destinction accuracy significantly benefits from this signal characteristic.
It is remarkable that this even seems to be true for the Dtree that was solely trained with a pure HRTF.
Since the result that is marked green in the table is significantly better than the corresponding result in the other testset (below the red line), we take a look at both Matlab plots to see what they look like.



Matlab plot of the good result

As expected, quite large errors occur around 90 degrees since less information is available compared to 0 degree elevation. (also refer to section 2.3.4 Head-Related Transfer Functions for the psychoaccoustical explanation, the HRTF in the median / sagittal plane).



Matlab plot of the not so good result

Although the result does not look really good compared to the other result, they both share some simililarities. While enough information is available at low degrees of elevation, at higher degrees, the accuracy becomes very low, because probably the Dtrees become literally confused by signals that are lateral to the head, overall perfomance is quite poor.

4 Discussion

The results show, that the approach of using Adaptive Logic Networks for elevation estimation and front / back classification from two monoaural correlograms provides quite good results for a single broadband sound source, if the tests are limited to a 90 degree cone anterior and posterior to the head.

Again, it should be pointed out again that due to lacking an actual KEMAR for testing and because the head related sounds were produced by convolving soundsamples with HRTF-measurements, the results are likely to be slightly better as if they would be under real life conditions.
Since the HRTF measurements are taken from a real KEMAR, and the differences are expected to be caused by imperfections like e.g. floor echo, the accuracy is expected to not seriously degrade in an actual anechoic chamber. However, this is a subject to further research.

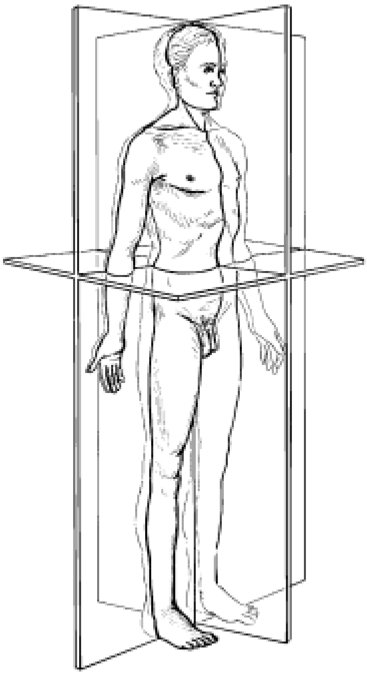
Assuming that enough information is provided by the training data, to increase accuracy and the robustness of the system, it is important to identify and separate the most important inputs from the least important inputs. Because abs. importance of each input value is available from a protocol file that is generated after training, it is easy to use this information to eleminate unneccissary data from the training set.
Another method could be, since training is fast, to use a Genetic Algorithm to identify the most important columns The RMSE after evaluating the generated Dtree should provide good feedback for the GA.

To build a system that is able to generalize well, it is important to provide good training data that includes enough information from broadband signals. E.g. to localize, speech well, words with long duration fricatives are likely to provide more localization information, since long duration fricatives are broadband signals.
Because estimation accuracy depends on room characteristics, it might be possible to use an ALN to roughly determine the room characteristics (room, hall, outside...) This approach, a form of basic auditory scene analysis, could then be used to choose an estimator, that is suited best to cope with environmental problems. This should not be confused with a clever sound locator mentioned in section 2.5.10 !
Moreover, utilizing reinforcement learning might also be advantageous.

Appendix A

Anatomical Terminology

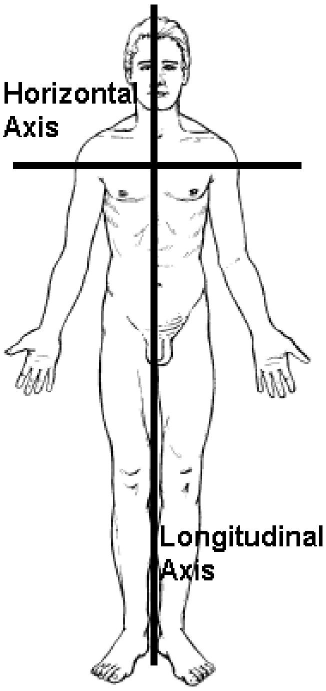
Anatomical planes



The plane which divides the body into a right and a left half is known as the median or the saggital plane. Vertical planes at right angles to the saggital plane divides the body into a front and a back and are described as coronal or frontal planes. Finally the body is divided into upper and lower portions by horizontal or transverse planes which are at right angles to both the frontal and saggital planes.

These planes are used to describe movement, for example, when you turn your head from right to left then your nose moves in a horizontal plane.

Anatomical Axes



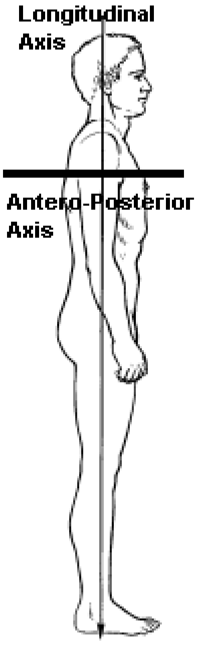
Anterior view

Horizontal (transverse) axis is perpendicular (at a right angle) to the longitudinal axis and runs from left to right.

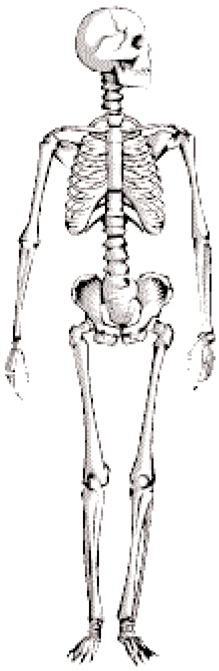
Longitudinal (vertical) axis, this is a vertical axis through the body in the upright position.

Lateral view

The saggital (antero-posterior) axis runs from front to back which is why it is also know as an antero-posterior axis. This axis is perpendicular (at right angles) to the other 2 axes



Details of some of the terms used in anatomy



Anterior - the front surface of the body or structure, the term is also used to describe the relationship of structures within the body. For example, the sternum is anterior to the heart so anterior also means towards the front. The heart will also have an anterior surface.

Posterior - the back surface of the body and towards the back of the body. For example, the spine is posterior to the heart and the heart has a posterior surface.

Superior - The upward surface. **Inferior** - The downward surface. For example, the top of the head is a superior surface, the soles of the feet an inferior surface.

Medial - Towards the middle or towards the median plane. **Lateral** - Away from the middle or away from the median plane

Dorsal - Towards the back. **Ventral** - Towards the front

Appendix B

The accompanying CD-Rom includes the following files:

Hearing examples: To get the sensation of direction, a headphone has to be used.

HRTF: The complete HRTF-measurements by [Gardner+Martin94].

Code:: Matlab code used used to precompute the test and training sets. Since the code uses a self explanatory makeFile, it is very easy to use. To evaluate the test and training samples you have to use evaulate.dll directly within Matlab or by using the performanceTest.m program.
performanceTest.m takes no inputs, the Dtrees used to evaluate the test data must be in the same directory where evaluate.dll resides. The test sets must be in the Matlab path. and evaluate.dll needs the dtr1032d.dll to run.

Dtrees and data sets: The DTrees can be found on the CD-Rom in the conv0, conv1 and conv2 directories. These are subdirectories of the directory called ”Precomputed datasets used for testing and training”

Toolboxes: The Matlab toolboxes directory does also include the AuditoryToolbox II by Malcolm Slaney that was used in this work.

Papers: The papers directory includes some very interesting papers.

The subdirectory called “KITANO Project” includes some very exciting videos about the project and a paper that further describes the work.
The dse.zip file in this directory implements a Dempster-Shafer Engine.
For more on the Dempster Shafer Theory see [Schafer89].

Bibliography

[David+Selfridge62]
David, E. E., Jr., and O. G. Selfridge (1962). "Eyes and ears for computers," Proc. IRE, Vol. 50, pp. 1093-1101. Describes the need for and the problems of machine perception;

[Gardner+Martin94]
http://sound.media.mit.edu/KEMAR.html

[Duda-98] a very well written and comprehensive tutorial about Sound Source Localization.
http://www.engr.sjsu.edu/~duda

Blauert [Bla97]
Blauert, J. (1997). Spatial Hearing (Revised Edition) (MIT Press, Cambridge, MA). The standard reference on the psychophysics of three-dimensional hearing. (See Mills (1972) for a shorter overview.)

[Mills72]
Mills, A. W. (1972). "Auditory localization," in J. V. Tobias, Ed., Foundations of Modern Auditory Theory, Vol. II, pp. 303-348 (Academic Press, NY). An excellent summary of the mechanisms and cues for sound localization.

[Web94]
Web94 WEBERS JOHANNES. Handbuch der Tonstudiotechnik. Franzis-Verlag GmbH, 6. Auflage, 1994

[Clifton87]
Clifton, R. K. (1987). "Breakdown of echo suppression in the precedence effect," J. Acoust. Soc. Am., Vol. 82, pp. 1834-1835. Demonstrates that the suppression of echoes is a dynamic process, probably related to the need for the auditory system to build a model of the acoustic environment; also demonstrates that Lindemann's model must be incomplete

[Han94]
Han, H. L. (1994). "Measuring a dummy head in search of pinna cues," J. Audio Eng. Soc., Vol. 42, pp. 15-37, January/February 1994.

[Angel+Algazi+Duda2002]
On the design of canonical sound localization environments
Eric J. Angel V. Ralph Algazi, and Richard O. Duda
Audio Engineering Society Convention Paper Presented at the 113th Convention 2002 October 5–8 Los Angeles, California, USA

[Dic87]
Dic87 DICKREITER MICHAEL Handbuch der Tonstudiotechnik. K.G.Saur München, 5. Auflage, 1987

[Bracewell 86]
Bracewell, R. N. (1986). The Fourier Transform and its Applications, 2nd ed. (McGraw-Hill, New York). A standard reference for Fourier analysis.

[Oppenheim89]
Oppenheim, A. V. and R. W. Schafer (1989). Discrete-Time Signal Processing (Prentice-Hall, Englewood Cliffs, NJ). A standard and widely respected textbook on digital signal processing (DSP).

[Schafer89]
http://yoda.cis.temple.edu:8080/UGAIWWW/lectures/dempster.html#4

[Begault94]
Begault, D. (1994). 3-D Sound for Virtual Reality and Multimedia (Academic Press, Boston, MA, 1994). A clear and comprehensive presentation of 3-D audio principles and current technology.

[Wright96]
Wright, M. (1996). "3-D Audio: Above and behind you or moving low left to right," EDN, pp. 87-88,90,92,94,96,99, June. A trade publication that accurately summarizes current commercialproducts and technical trends

[Zemlin97]
Zemlin, W.R., Speech and Hearing Science: Anatomy and physiology, 4th edition, Allyn and Bacon, 1997 (QP306.Z4/1998) (Starting page 435)

[Jeffress48]
Jeffress, L. A. (1948). "A place theory of sound localization," J. Comp. Physiol. Psychol., Vol. 41, pp. 35-39 (1948). The classic cross-correlation model of horizontal localization, proposed long before neural correlation circuits were identified in the Medial Superior Olive (MSO); primarily of historical interest.

[CFRL02]
http://oto.wustl.edu/cochlea/

[NT97]
http://thalamus.wustl.edu/course/

[Allen85]
Allen, J. B. (1985). "Cochlear modeling," IEEE ASSP Magazine , Vol. 2, pp. 3-29. A good survey of alternative models of the cochlea.

[Lyon82 Lyon88]
Lyon, R. F., (1982). "A computational model of filtering, detection and compression in the cochlea," ICASSP 82 (Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing, May, 1982). Describes an important filter-bank model of the cochlea. See also Lyon (1988) and Slaney (1988).

[Lyon88 Mead88]
Lyon, R. F., and C. Mead (1988). "An analog electronic cochlea," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. 36, pp. 1119-1134. Describes further development of Lyon's filter-bank model of the cochlea. See also Lyon (1982) and Slaney (1988).

[Slaney88]
Slaney, M. (1988). "Lyon's Cochlear Model," Apple Technical Report No. 13, Advanced Technology Group, Apple Computer, Inc., Cupertino, CA. Describes and includes a C program for Lyon's filter-bank cochlear model.

[Lyon90]
Richard F. Lyon, "Automatic Gain Control in Cochlear Mechanics", The Mechanics and Biophysics of Hearing, P. Dallos et al., eds., pp. 395-402, Springer-Verlag, 1990.

[Fourcin70]
Fourcin AJ (1970). In Frequency Analysis and Peroidicity Detection in Hearing, ed R Plomp and GF Smoorenburg (Sijthoff, The Netherlands). underpins the perception of these pitches.

[Fourcin58]
Fourcin AJ (1958). SDRE report #1126

[Akeroyd98] [Summerfield98]
Akeroyd MA and Summerfield AQ (1999) "A fully temporal account of the perception of dichotic pitches," Br. J. Audiol., 33(2), 106-107.

Gaik [Gaik93]
Gaik, W. (1993). "Combined evaluation of interaural time and intensity differences: Psychoacoustical results and computer modeling," J. Acoust. Soc. Am., Vol. 94, pp. 98-110. An extension of Lindemann's model to include interaural intensity effects; the very large interaural differences shown in the data raise questions about the results

[Lindemann86]
Lindemann, W. (1986). "Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals," J. Acoust. Soc. Am., Vol. 80, pp. 1608-1622; "Extension of a binaural cross-correlation model by contralateral inhibition. II. The law of the first wave Front," J. Acoust. Soc. Am., Vol. 80, pp. 1623-1630. An important paper that presents an appealing modification of the classic Jeffress cross-correlation model to account for the precedence effect. But see Clifton (1987).

[Middlebrooks+Green91]
Middlebrooks, J. C., and D. M. Green (1991). "Sound localization by human listeners," Annu. Rev. Psychol., Vol. 42, pp. 135-159 (1991). An excellent review of the abilities of people to localize sound. Highly recommended.

[Carlile+King94]
CARLILE, S. AND KING, A. J. Monaural and binaural spectrum level cues in the ferret: acoustics and the neural representation of auditory space. J. Neurophysiol.71: 785-801, 1994.

[CARLILE+Pralong94]
CARLILE, S. AND PRALONG, D. The location-dependent nature of perceptually salient features of the human head-related transfer functions. J. Acoust. Soc. Am. 95: 3445-3459, 1994.

[OLDFIELD+PARKER84]
OLDFIELD, S. R. AND PARKER, S.P.A. Acuity of sound localisation: a topography of auditory space. I. Normal hearing conditions. Perception 13: 581-600, 1984b.

[PALMER+KING85]
PALMER, A. R. AND KING, A. J. A monaural space map in the guinea-pig superior colliculus. Hear. Res. 17: 267-280, 1985.

[Schnupp98]
SCHNUPP, J.W.H., KING, A. J., AND CARLILE, S. Altered spectral localization cues disrupt the development of the auditory space map in the superior colliculus of the ferret. J. Neurophysiol. 79: 1053-1069, 1998.

[Reed+Blum90]
Reed, M. C., and J. J. Blum (1990). "A model for the computation and encoding of azimuthal information by the lateral superior olive," J. Acoust. Soc. Am., Vol. 88, pp. 1442-1453. A complex model that attempts to account for both the functionality and the detailed neurophysiological characteristics of the lateral superior olive (LSO), where interaural intensity differences seem to be determined.

[Duda95+Chau95]
Chau, W. and R. O. Duda (1995). "Combined Monaural and Binaural Localization of Sound Sources," Proc. Twenty-Ninth Asilomar Conference on Signals, Systems, and Computers (Asilomar, CA, November 1995). Extends Chau's MS work by computing probability distributions for the location every 10 ms and averaging these distributions over time.

[Lim+Duda94]
Lim, C., and R. O. Duda (1994). "Estimating the azimuth and elevation of a sound source from the output of a cochlear model," in Proc. 28th Asilomar Conf. on Signals, Systems and Computers (Asilomar, CA). Extends Duda (1993) by replacing FFT's by a cochlear model, showing that high spectral resolution at high frequencise is unnecessary.

[Stern88]
Stern, R. M., Jr. (1988). "An overview of models of binaural perception," in Proc. 1988 National Research Council CHABA Symposium (Washington, DC). A useful survey paper directed at models that attempt to explain all known psychoacoustic phenomena.

[Wightman+Kistler93], [Hartmann95]
Wightman, F. L., and D. J. Kistler (1993). "Factors affecting relative importance of sound localization cues," Conference on Binaural and Spatial Hearing (Dayton, OH), September 9-12, 1993. Surveys psychoacoustic responses when interaural time and intensity cues are inconsistent.

[Tran+Tran93]
Tran, B. and T. Tran (1993). "A Sound Localization System Using Lyon's Cochlear Model and Lindemann's Cross-Correlation Model," Technical Report No. 5, NSF Grant No. IRI-9214233, Dept. of Elec. Engr., San Jose State Univ. (December, 1993).

[Bregman90]
Bregman, A. S. (1990). Auditory Scene Analysi (MIT Press, Cambridge, MA). A massive description of experiments by the author and his students on the factors that influence the formation and segregation of sound streams. The first and last chapters are readable by nonspecialists, but see Handel (1989) or Yost (1991) for an easier introduction.

[Handel89]
Handel, S. (1989). Listening (MIT Press, Cambridge, MA). A recommended general introduction to the psychology of hearing; includes a good summary chapter on neurophysiology

[Yost91]
Yost, W. A. (1991). "Auditory image perception and analysis: The basis for hearing," Hearing Research, Vol. 56, pp. 8-18.

[Slaney, 1995]
Slaney, M. (1995). "A critique of pure audition," Proceedings of the Computational Auditory Scene Analysis Workshop; 1995 International Joint Conference on Artificial Intelligence, Montreal, Canada (August 1995).

[Pickles88]
Pickles, J. O. (1988). An Introduction to the Physiology of Hearing (2nd Ed.) (Academic Press, London). Describes in detail most of what is known of the neurophysiological processes along the path from the cochlea to the auditory cortex. See Handel (1989) for an easier introduction.

[Akeroyd01]
A Binaural Cross-correlogram Toolbox for MATLAB February 2nd, 2001

[Lyon+CarverMead88]
Lyon, R. F., and C. Mead (1988). "An analog electronic cochlea," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. 36, pp. 1119-1134. Describes further development of Lyon's filter-bank model of the cochlea. See also Lyon (1982) and Slaney (1988).

[Zweig+Lipes+Pierce76]
Zweig76 - G. Zweig, R. Lipes and J. R. Pierce, "The cochlear compromise," JASA 59, pp. 975-982, 1976.

[Slaney+Lyon93]
ON THE IMPORTANCE OF TIME-A TEMPORAL REPRESENTATON OF SOUND Malcolm Slaney and Richard F. Lyon A TEMPORAL REPRESENTATION OF SOUND

[AppleTechnicalReport #13 1988]
Lyon's Cochlear Model Malcolm Slaney Advanced Technology Group Copyright © 1988 Apple Computer, Inc

[Slaney94]
AUDITORY MODEL INVERSION FOR SOUND SEPARATION Malcolm Slaney, Daniel Naar, and Richard F. Lyon Apple Computer, Inc., One InÞnite Loop, Cupertino, CA 95014 USA Reprinted from Proceedings of the ICASSP 94 Ñ 1994 International Conference on Acoustics, Speech, and Signal Processing, Adelaide, Australia, 19-22 April 1994.

[Chuck Lim94]
Chuck Lim, "A Sound Localization System Using Correlograms and Croscorrelograms from a Cochlear Model," Technical Report No. 9, NSF Grant No. IRI-9214233, Dept. of Elec. Engr., San Jose State Univ. (November, 1994).

[Breiman96]
Bias, Variance, and] Arcing Classifiers February, 1996 ;revised July, 1996

[Breiman99]
Using adaptive bagging to debias regressions February 1999

[Ear Anatomy]
http://www.earaces.com/

[Hiroshi+Okuno+Nakadai+Kitano2002]
Hiroshi G. Okuno, Kazuhiro Nakadai, Hiroaki Kitano: Real-time Sound Source Localization and Separation for Robot Audition Proc. of 7th International Conference on Spoken Language Processing (ICSLP-2002), to be appeared, Denver, USA, Sep. 2002.

[Zwiers+Opstal+Cruysberg2001]
Zwiers, M.P., Van Opstal, A.J. <../../People/OpstalvanAJ/index.html>, & Cruysberg, J.R.M. (2001). Two-dimensional sound-localization behavior of early-blind humans. Experimental Brain Research, 140 (2), 206-222. http://www.nici.kun.nl/Publications/2001/14713.html

[Armstrong2001]
http://www.dendronic.com/main.htm

[Mizumachi+Akagi98]
M. Mizumachi and M. Akagi, “Noise reduction by pairedmicrophones using spectral subtraction.,” Proc. of ICASSP-98, 1998, pp. 1113–1116, IEEE.

[Zwicker82]
The book Psychoakustik (in German) by E. Zwicker (1982) - which is out of print - contained a record with acoustic demonstrations.

Versicherung über Selbständigkeit

Hiermit versichere ich, daß ich die vorliegende Arbeit im Sinne der Prüfungsordnung nach §24(5) ohne fremde Hilfe selbständig verfaßt und nur die angegebenen Hilfsmittel benutzt habe.

______________________ ________________________________

Ort, Datum   Unterschrift des Studenten