

## Overview:



## 2 Creating sounds in space:

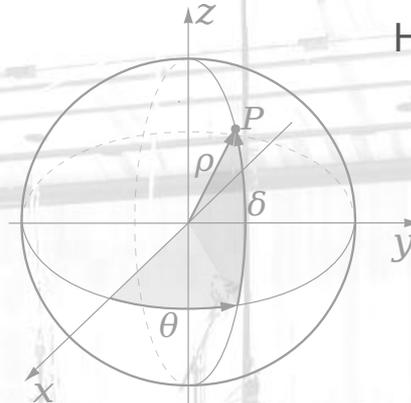
- straightforward yet effective: **“Put a loudspeaker there!”**
- generate psychoacoustic cues: **stereophony/VBAP**
- recreate ear signals: **binaural synthesis**
- recreate a soundfield in the listening space: **wave field synthesis, higher-order Ambisonics**

## 1 Psychoacoustics of spatial hearing

Why do we localize sounds?

How do we localize sounds?

- in front
- behind us
- above
- distance



**...and what it buys us  
in actual practice.**

about me:

**Jörn Nettingsmeier**

**<joern.nettingsmeier@tonmeister.de>**

**freelance sound designer and event engineer  
from Amsterdam, Netherlands**

**Focus: spatial audio and Ambisonics**

**certified event engineer ("Meister für Veranstaltungstechnik Bühne/Studio")**

**VDT board member and head of sound reinforcement dept.**

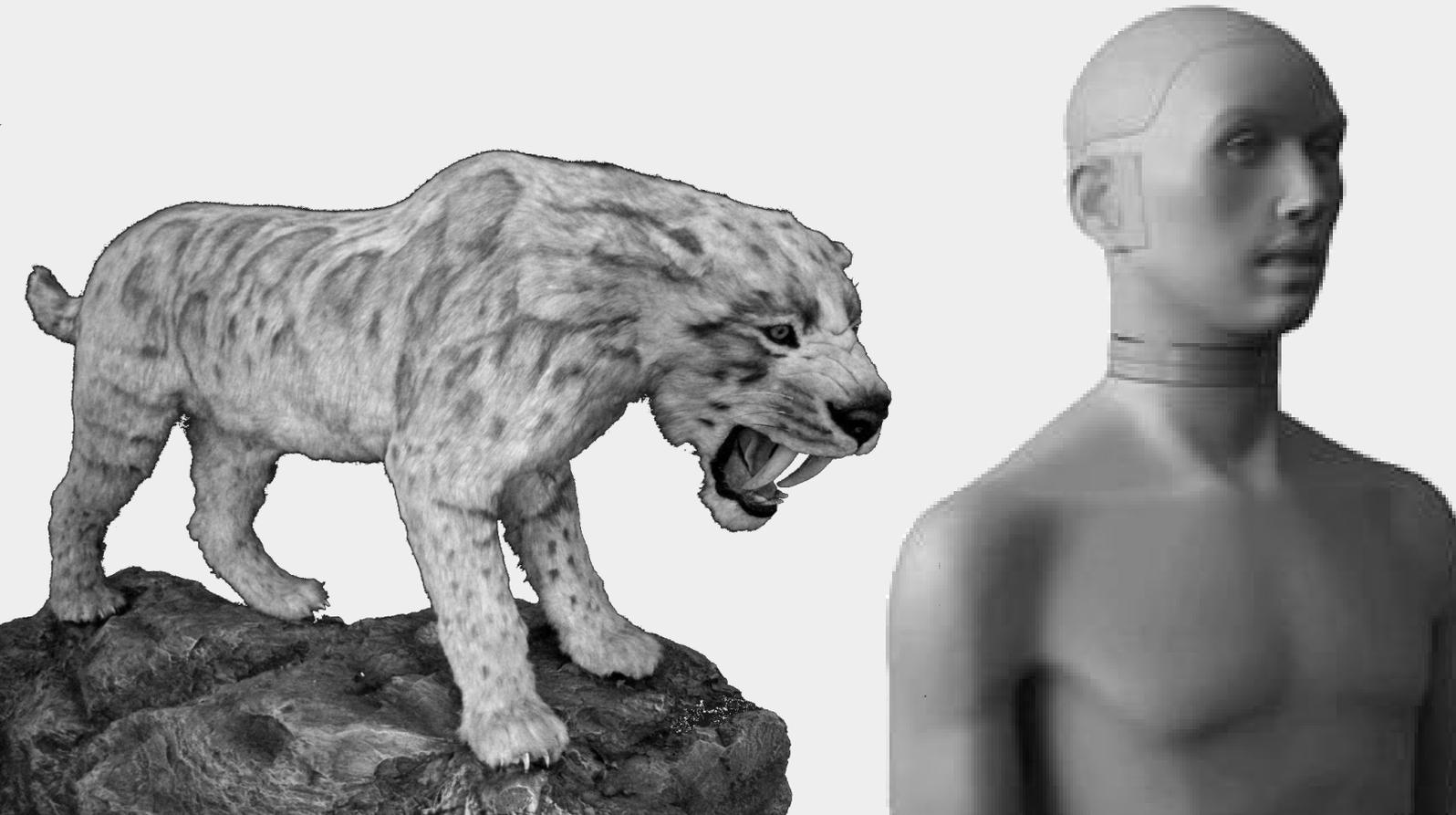
**AES member**

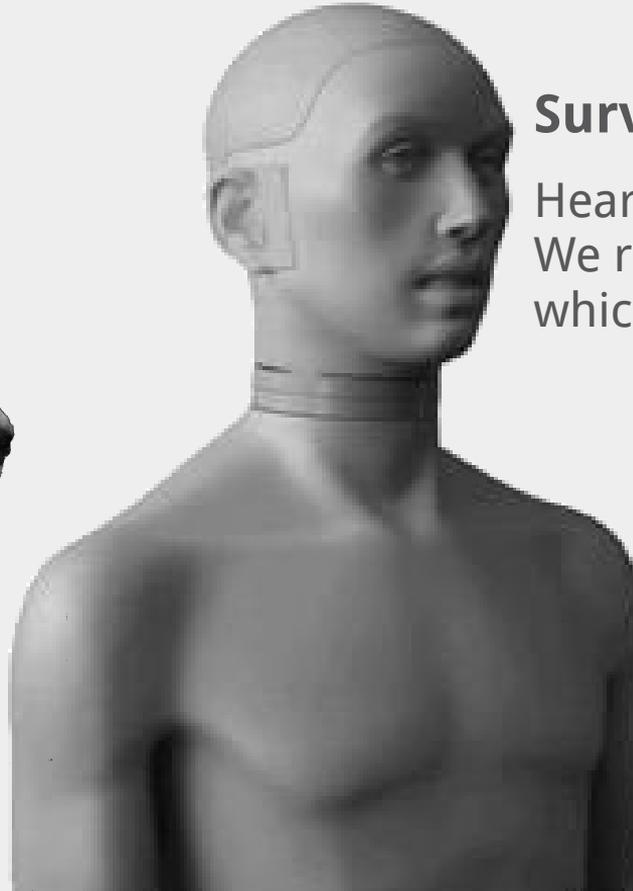
**VPT member**

# Why do we localize sounds?



# Why do we localize sounds?





## Survival.

Hearing covers all directions.  
We receive warnings of that  
which we **cannot see yet.**

- a predator?
- a car at a crossing?

# Why do we localize sounds?





## Orientation

We hear

- the characteristics of the ground we walk on
- the size of the room we're in
- the location of walls and obstacles.

# Why do we localize sounds?



# Why do we localize sounds?



## Social life.

We can call our friends before they can see us.

We can follow a conversation in a loud place by listening in a particular direction.

We can warn others, and receive warnings.

# Why do we localize sounds?



**Staying sane.**

We cannot shut our ears: it's impossible *not to listen.*

Directional hearing allows us to focus on what matters, even in a noisy environment.

# Why do we localize sounds?



## Creativity.

Spatial hearing is a low-level means of survival.

It connects to strong emotion.

A sound can tell us we just escaped being eaten.

Or that we're about to be.

# Why do we localize sounds?



**Creativity.**

Good **localization** keeps the focus on stage and screen.

**Envelopment** creates emotion that transcends the boundaries of the stage.

**Surround effects** put us in the center of the story.

**Spatialisation** allows for greater scene complexity.

# How do we localize sounds?



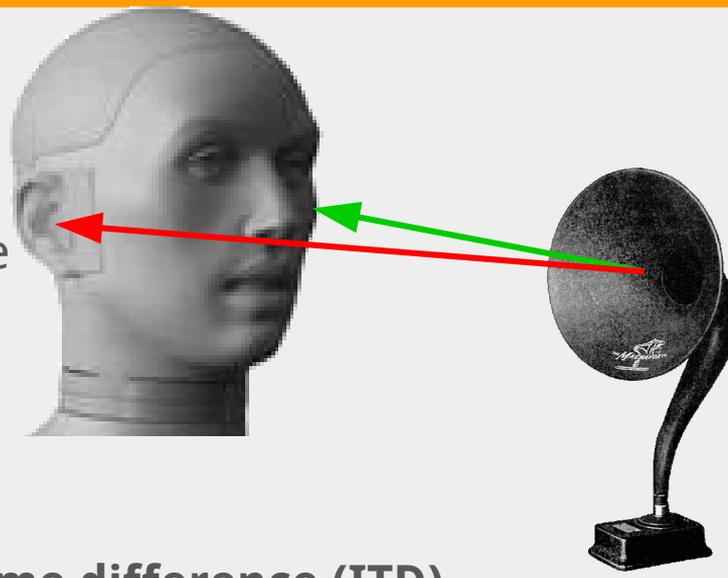
# How we localize sounds in front:

A sound from straight ahead reaches both ears at the same time, with the same level.

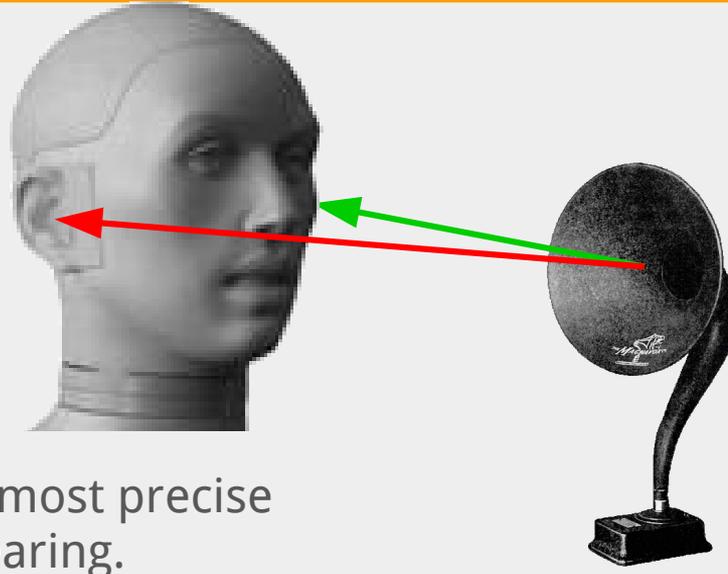
A sound from one side reaches the opposite ear a little bit later:

There is an **inter-aural time difference (ITD)**.

And the opposite ear is in the acoustic shadow of the head, where the sound is softer. We observe an **inter-aural level difference (ILD)**.

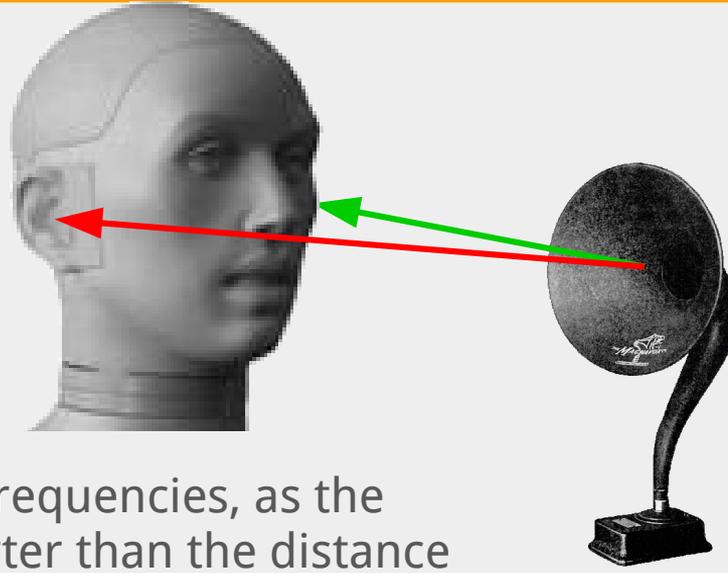


# How we localize sounds in front:



The time difference (**ITD**) and level difference (**ILD**) between our ears are the most precise **spatial cues** of human hearing.

With their help, our brain is able to localize sounds extremely accurately, with up to  $2^\circ$  of horizontal angular resolution.



The time difference (**ITD**) results in a phase difference. But at higher frequencies, as the wavelength becomes shorter than the distance between the ears. We don't know if the time difference is one wavelength, two wavelengths, or more. So the brain discards the information at high frequencies.

The inverse holds for **ILDs**: an acoustic shadow only happens at high frequencies. Low sounds (=long waves) are diffracted around the head and do not result in much level difference. **ITD and ILD complement each other.**

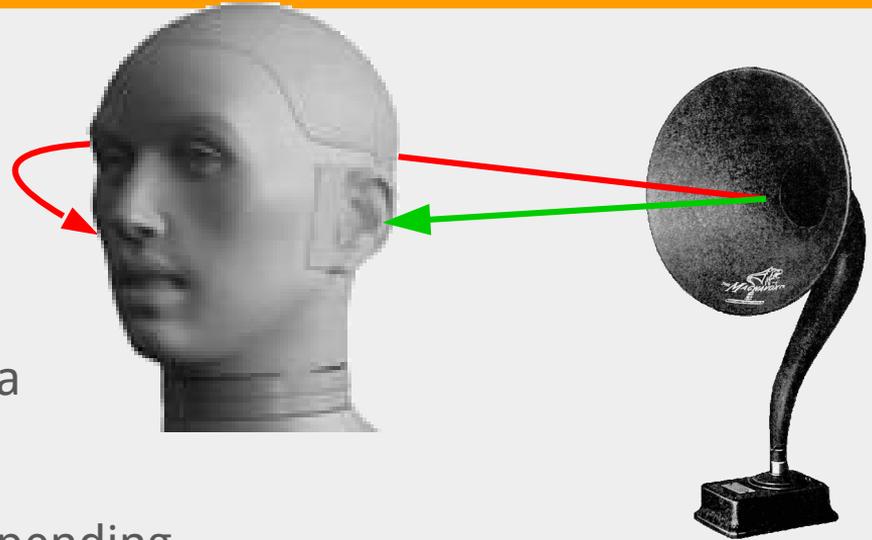
A sound from behind produces the exact same time and level differences (**ILD** und **ITD**) as one from the front.

We can tell them apart, because our head, torso, and outer ear function as a **directional equalizer**.

They change the sound's **spectrum** depending on where the sound came from.

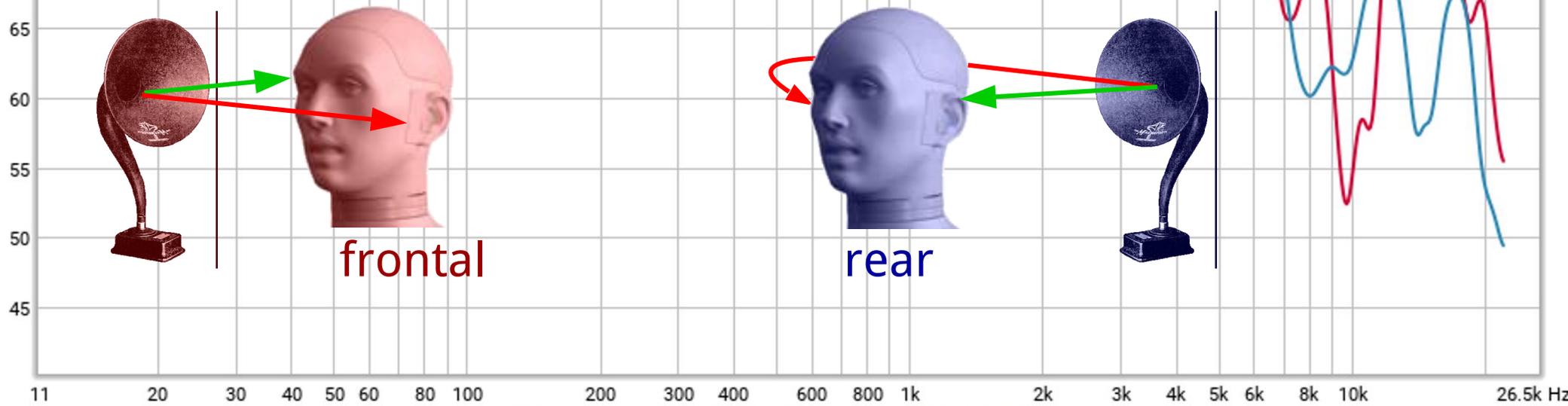
Our brain interprets these **spectral cues** as directional information.

We are not consciously aware of this substantial coloration. But it is possible to measure it by placing microphones in the ear canal:



The ear filter curves (also known as **head-related transfer functions, HRTFs**) of the author, for a source from ahead and another from the rear, into the left ear.

The right-ear curve will be quite similar, due to the strong (although not quite perfect) symmetry of the head.



# How we localize sounds vertically:

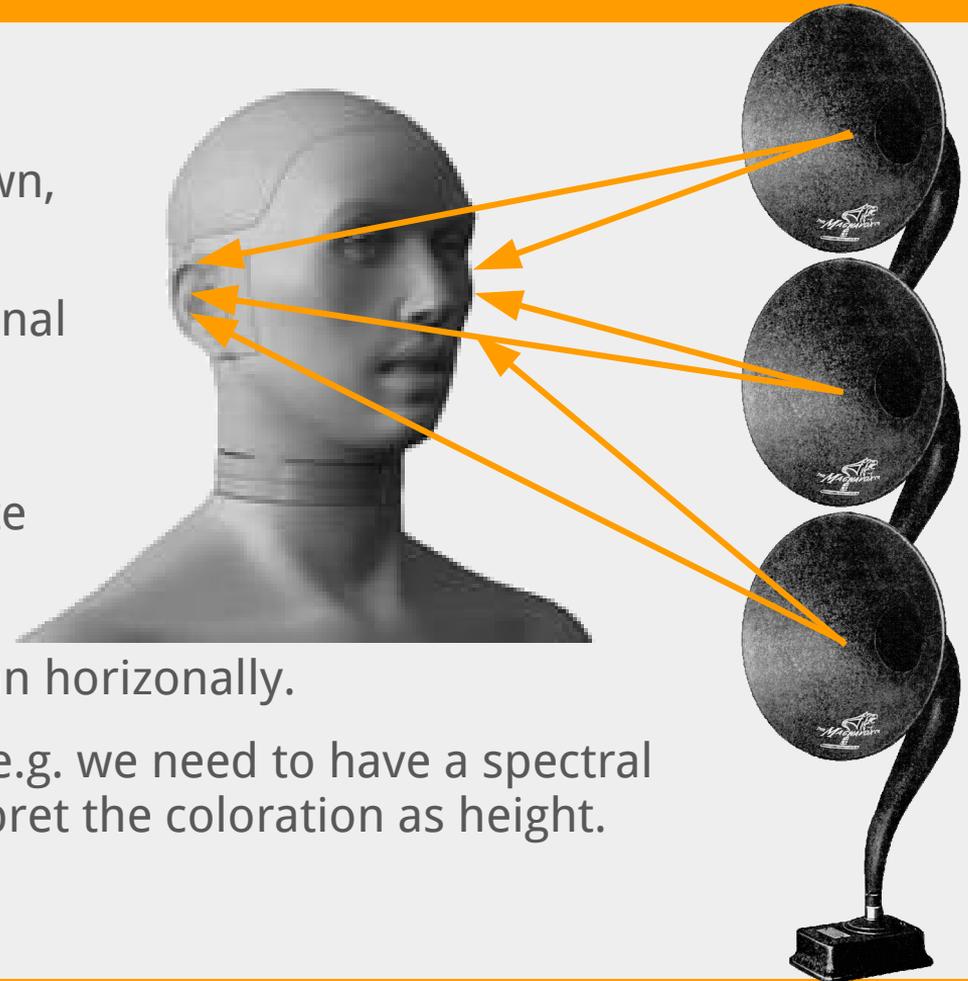
As a sound source moves up and down, there is no change in **ILD** and **ITD**.

Again, we have to rely on the directional equalizer of our outer ear.

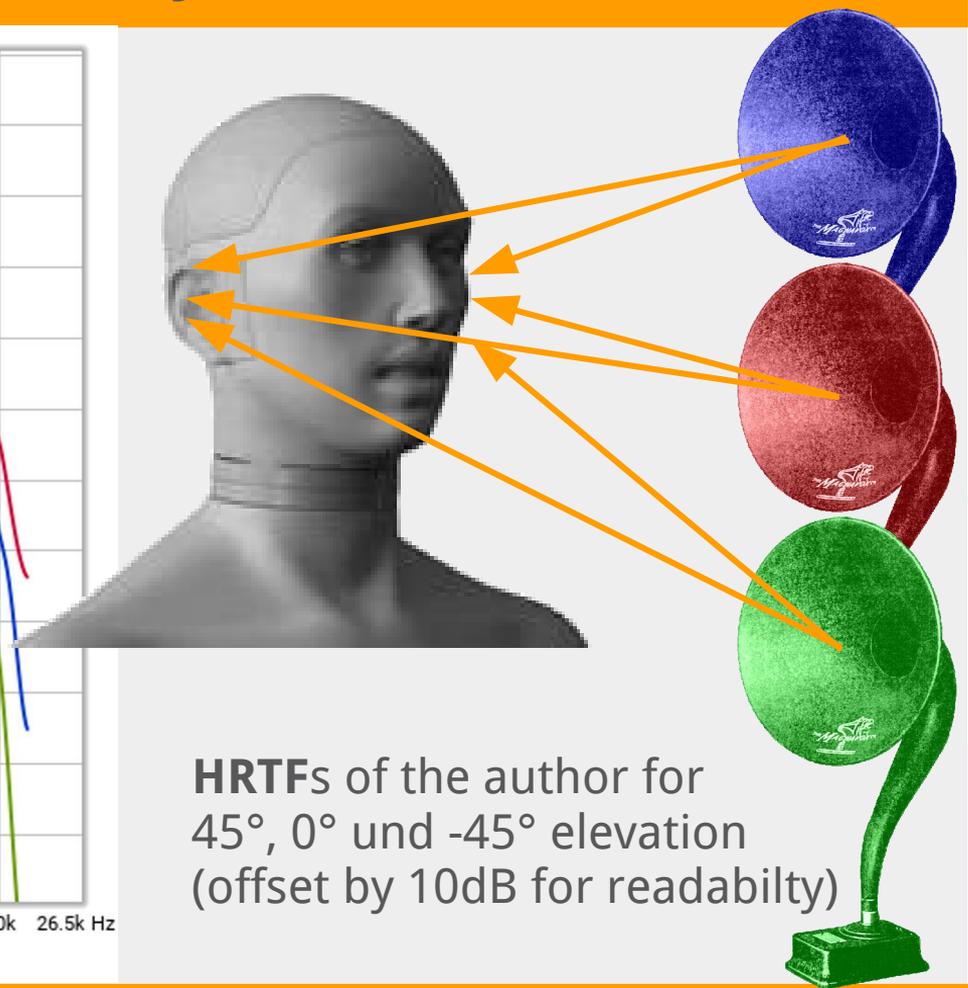
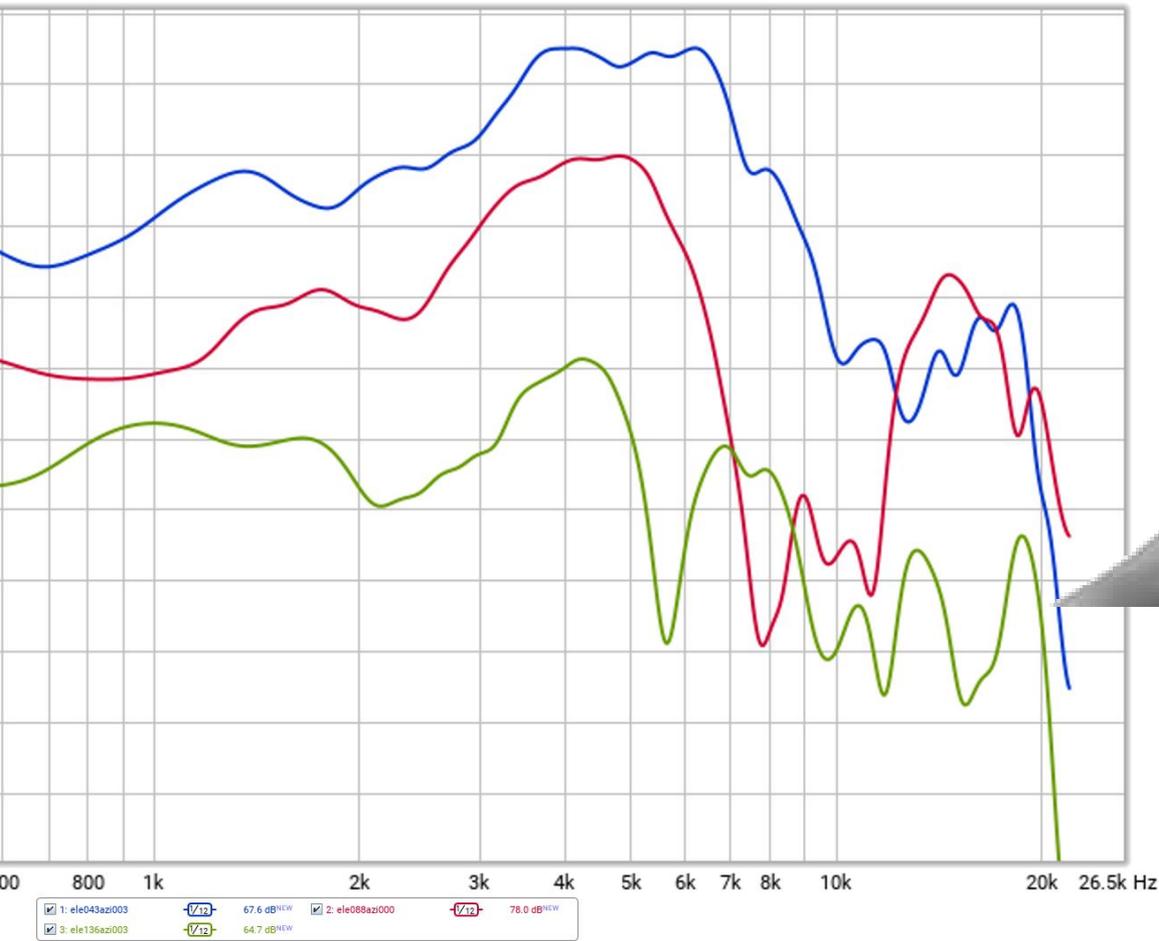
The **spectral cues** for height are quite subtle and ambiguous.

That's why our vertical localisation is a lot less precise and dependable than horizontally.

We need to know a sound very well (e.g. we need to have a spectral reference) in order to correctly interpret the coloration as height.



# How we localize sounds vertically:



**HRTFs of the author for 45°, 0° und -45° elevation (offset by 10dB for readability)**

# How we perceive distance:

We only have a direct way to perceive distance at very close range:

Like any directional microphone, our ears have some degree of **near-field effect**:  
sounds up close have more bass.

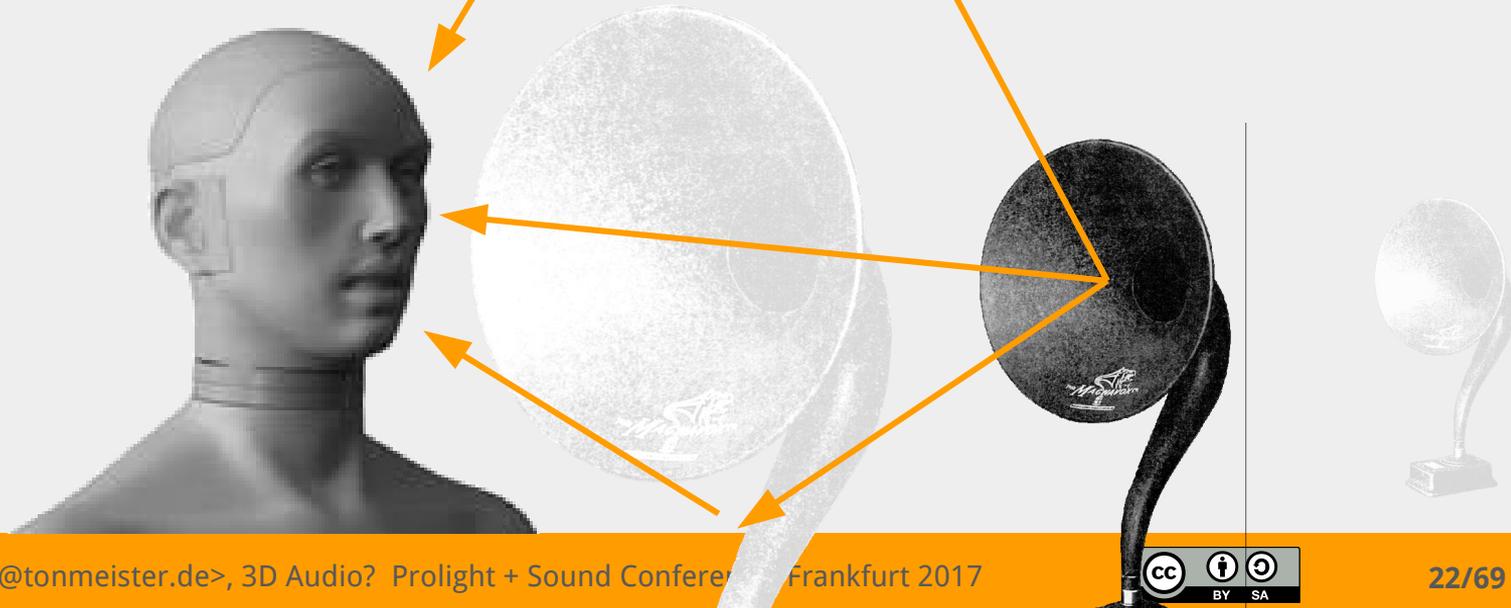


# How we perceive distance:

We only have a direct way to perceive distance at very close range:

Like any directional microphone, our ears have some degree of **near-field effect**: sounds up close have more bass.

In enclosed spaces, we can estimate distance from the **pattern of early reflections** and the **ratio of direct to reverberated sound**.



# How we perceive distance:

We only have a direct way to perceive distance at very close range:

Like any directional microphone, our ears have some degree of **near-field effect**: sounds up close have more bass.

In enclosed spaces, we can estimate distance from the **pattern of early reflections** and the **ratio of direct to reverberated sound**.

The level attenuation that follows from the **1/r law** is another cue, provided we have a good reference of the original level.



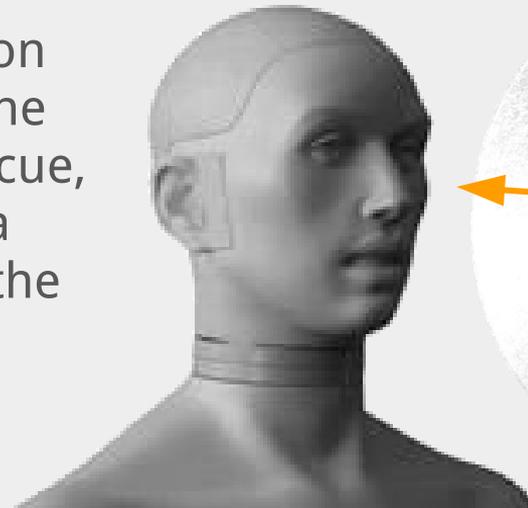
# How we perceive distance:

We only have a direct way to perceive distance at very close range:

Like any directional microphone, our ears have some degree of **near-field effect**: sounds up close have more bass.

In enclosed spaces, we can estimate distance from the **pattern of early reflections** and the **ratio of direct to reverberated sound**.

The level attenuation that follows from the **1/r law** is another cue, provided we have a good reference of the original level.



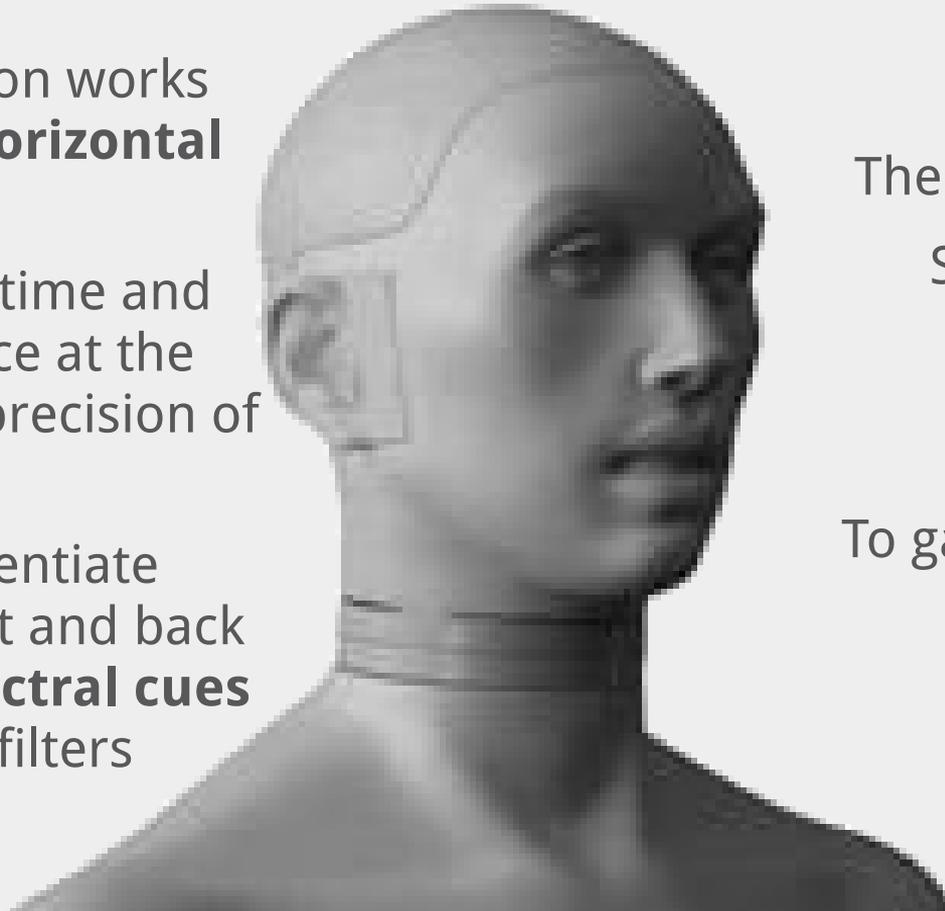
Finally, air absorbs more sound energy at higher frequencies. This leads to selective damping of the treble range over long distances (mostly relevant in the open).



Our localization works best on the **horizontal plane**.

**ILD** and **ITD** (time and level difference at the ears) have a precision of up to 2°.

We can differentiate between front and back thanks to **spectral cues** from our ear filters (**HRTFs**).



The same applies to height.

Spectral cues are subtle and ambiguous. That's why our height localization is worse.

To gauge distance, we need **reverberation**.

# Creating sounds in space:

*“If you want sound to come from a particular place,  
put a loudspeaker there!”*

- Curtis Roads

*“If you want sound to come from a particular place,  
put a loudspeaker there!”*

- Curtis Roads

- simple yet effective
- all psychoacoustic cues are naturally correct
- we have perfect localisation of that loudspeaker
- usually the tool of choice in theatre

*“If you want sound to come from a particular place,  
put a loudspeaker there!”*

- Curtis Roads

But:

- who wants to hear a loudspeaker?
- what if we need sounds from many directions?
- what if a sound needs to move?

## Alan Blumlein, 1930s



**Stereophony** attempts to recreate the **ILD** and **ITD** localization cues between **pairs of loudspeakers**.

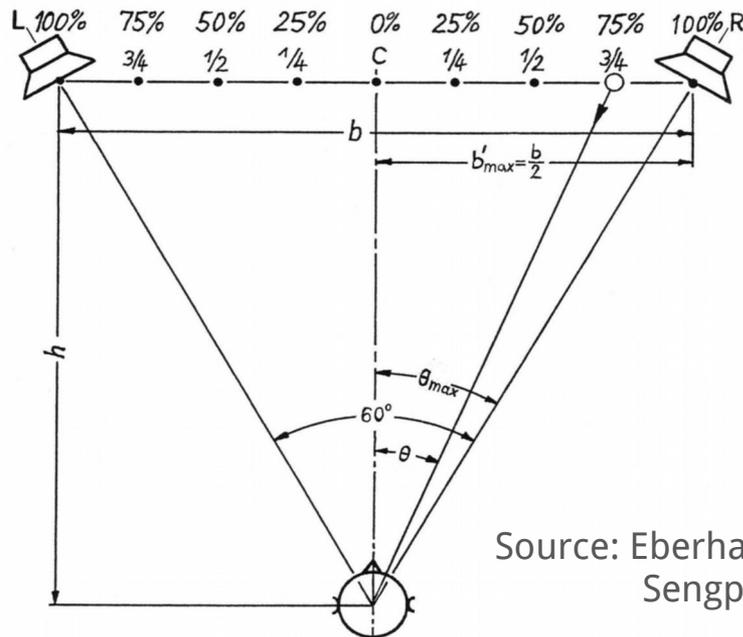
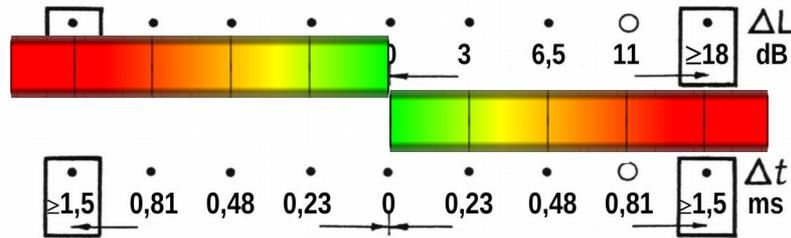
It can produce **phantom sources**, sounds that seem to originate between those loudspeaker pairs.

Stereo is not limited to two channels!  
Most channel-based formats (5.1, 7.1, 22.2, Auro-3D etc.) employ stereophonic localization techniques, at least in the front.



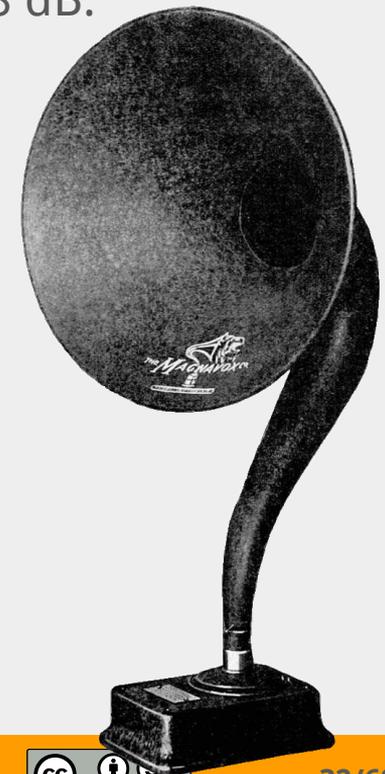
# Creating sounds in space: stereophony

ITD happens in a  
**time window** of  
about 1.5 ms.

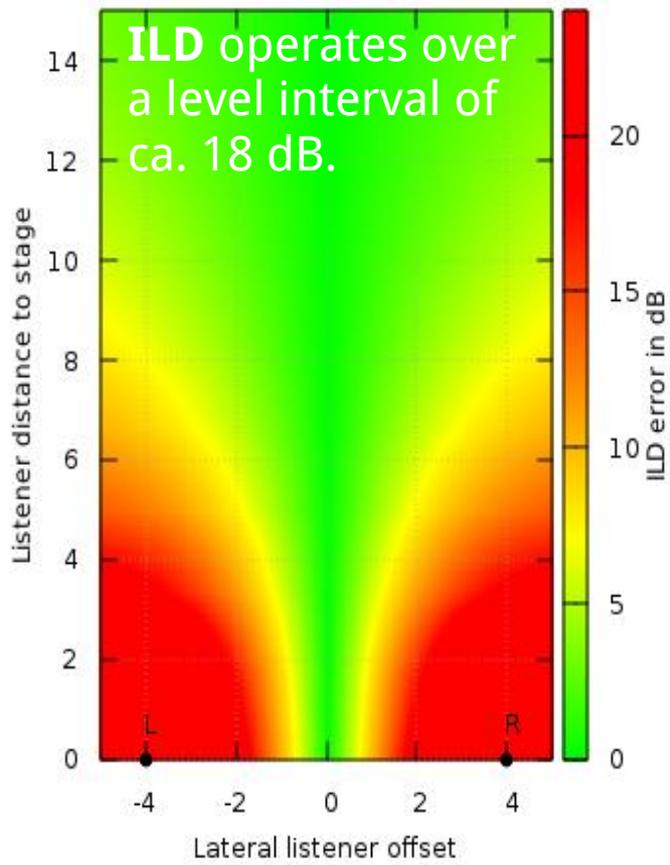
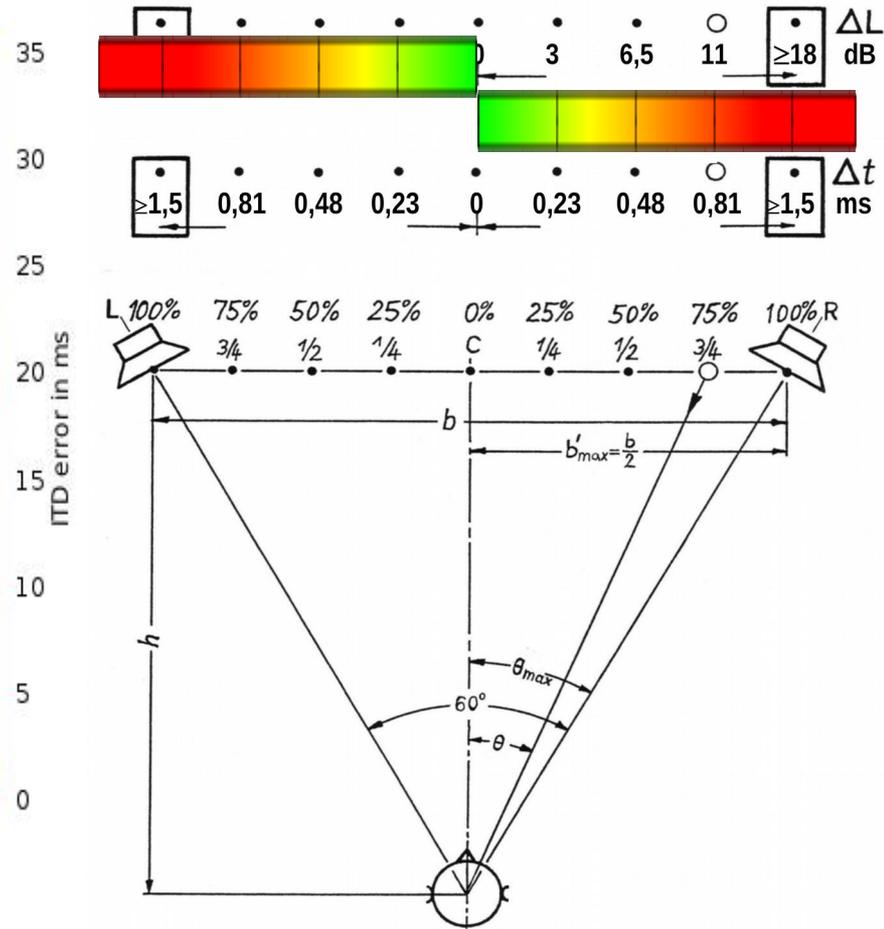
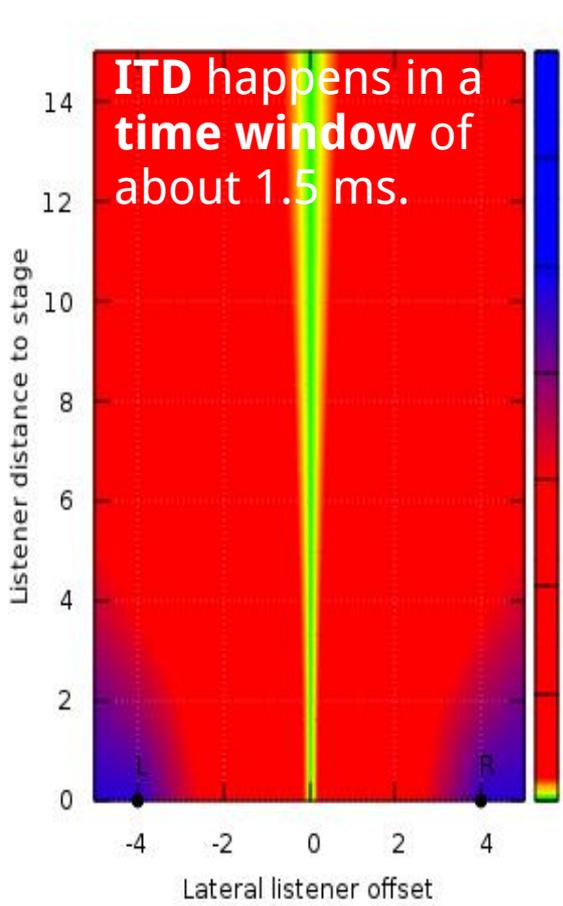


Source: Eberhard  
Sengpiel.

ILD operates over  
a level interval of  
ca. 18 dB.



# Creating sounds in space: stereophony

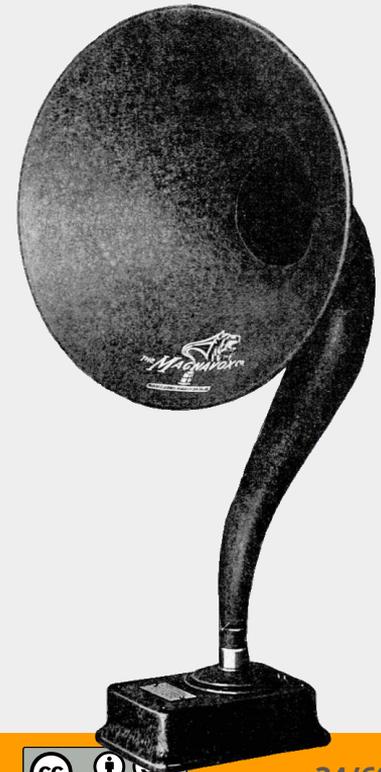
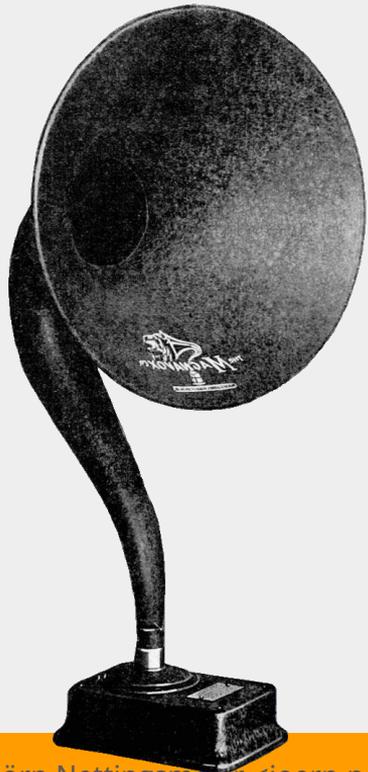


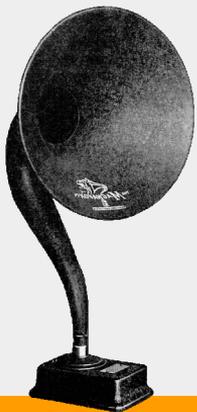
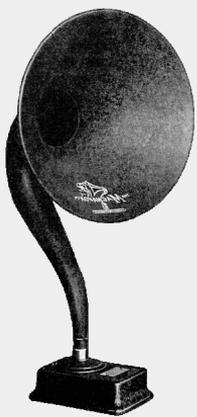
## advantages:

- very good timbre reproduction (no coloration)
- efficient: low bandwidth and hardware requirements

## disadvantages:

- unstable localization on the sides, no height
  - moving sources “stick” to speakers, variable tone color and apparent source width
- very small listening area, particularly troublesome in PA situations

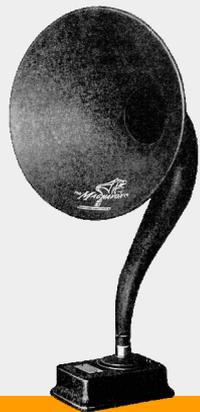
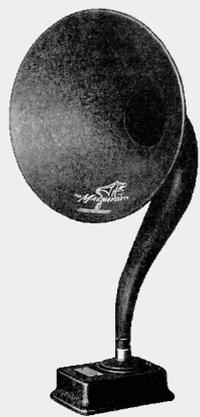




Horizontal stereo can be augmented  
with discrete **height channels**  
(NHK 22.2, Auro-3D).

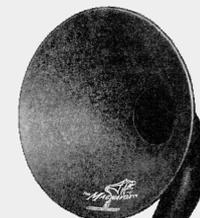
Vertical localisation however is bad –  
no **ILD** and **ITD**, incorrect **spectral cues**:  
no vertical phantom sources.

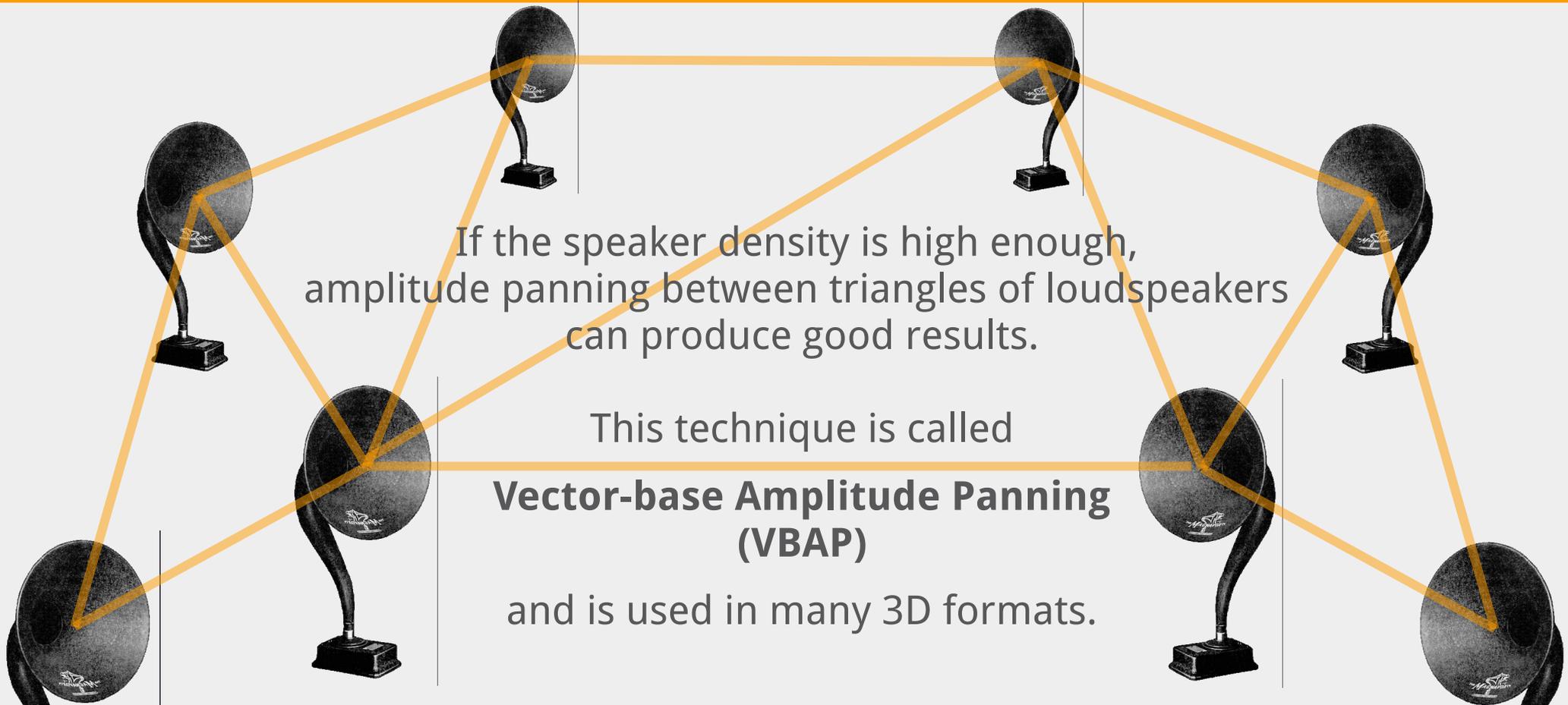
But: expanded listening area,  
tone colour and spatial impression appear  
more natural and convincing.





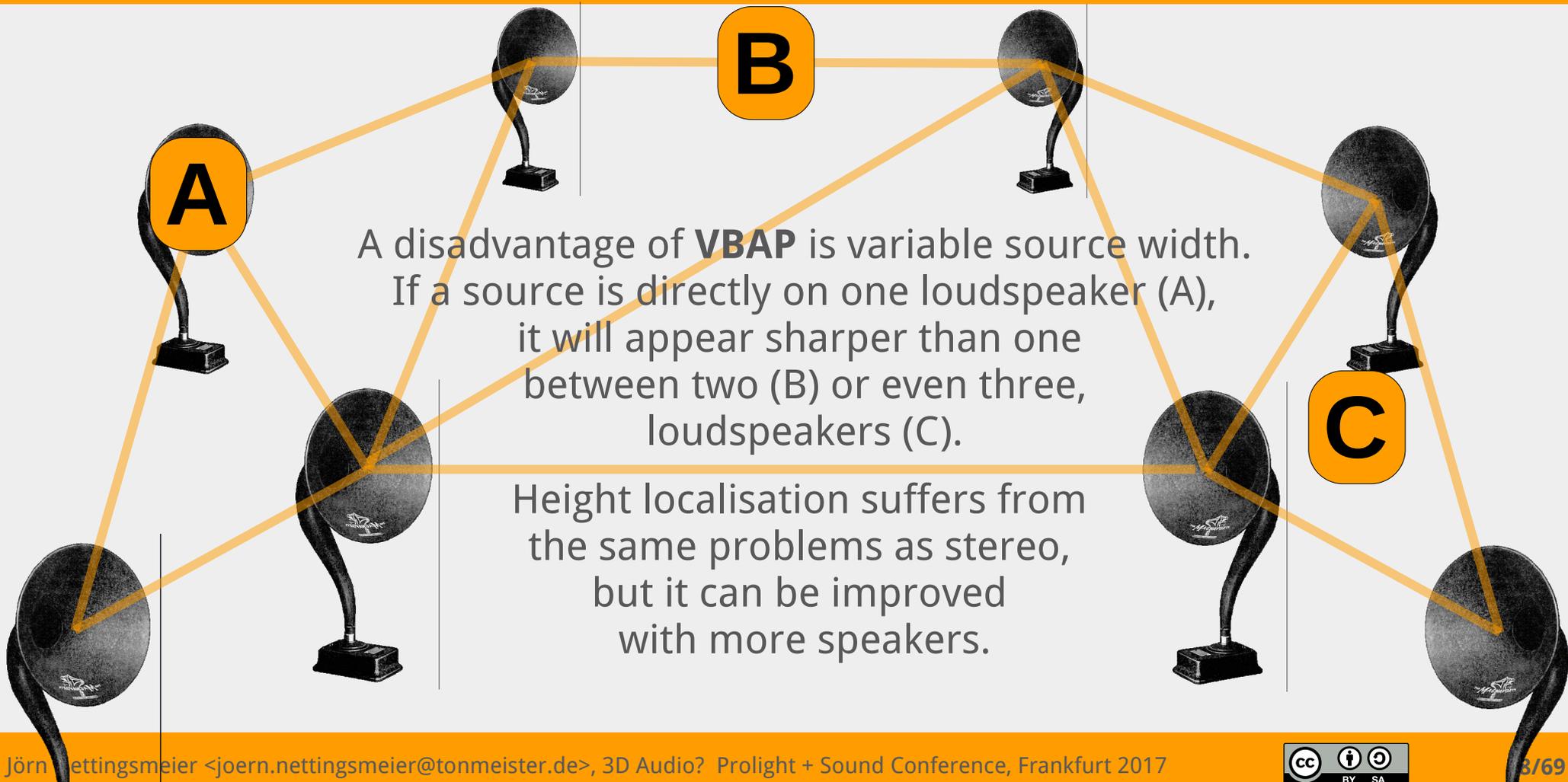
Ville Pulkki, 1997



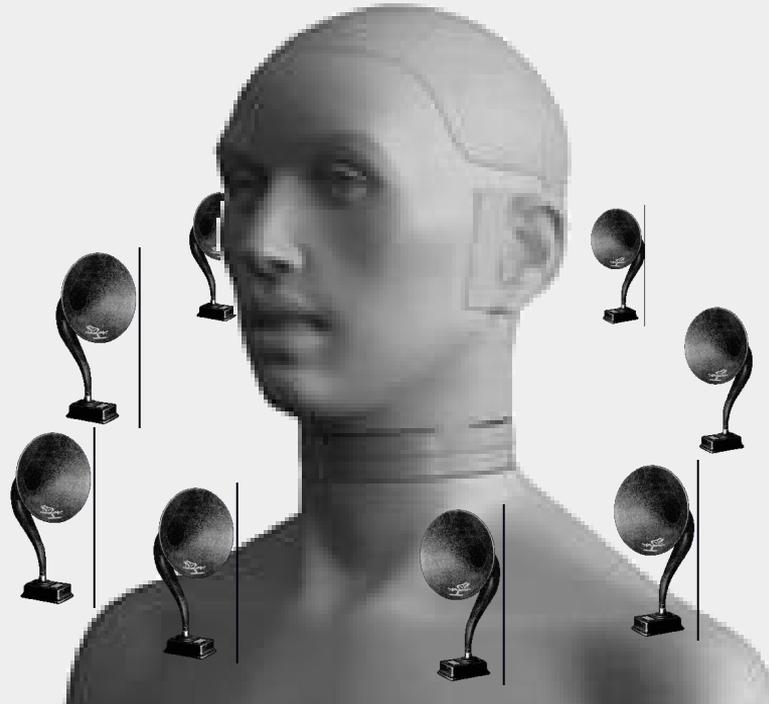


If the speaker density is high enough, amplitude panning between triangles of loudspeakers can produce good results.

This technique is called  
**Vector-base Amplitude Panning (VBAP)**  
and is used in many 3D formats.



## Harvey Fletcher et al., 1927

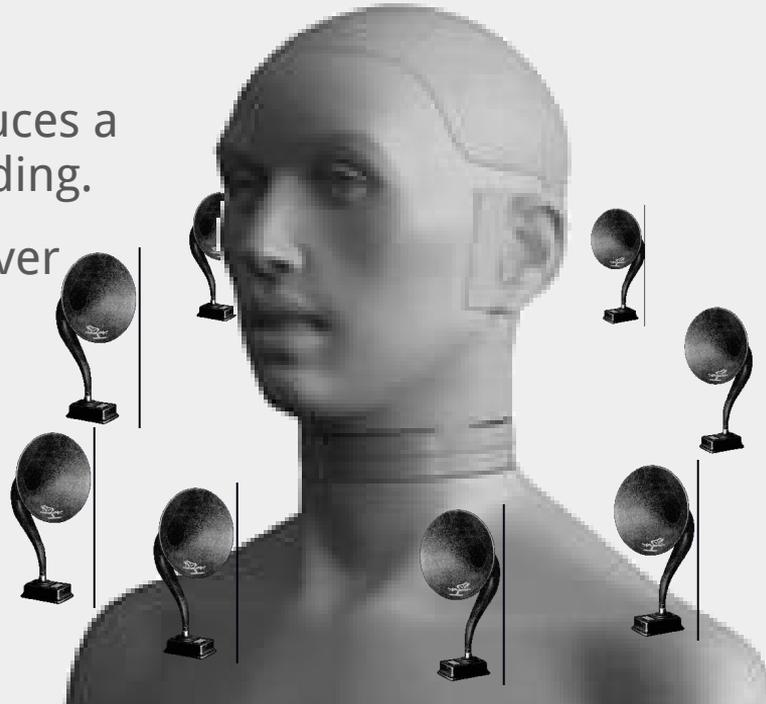


At the eardrums, the entire spatial information of a sound field is reduced to two channels.

A **dummy head** with artificial outer ears can record this information.

This approach produces a **head-related** recording.

In theory, it can deliver perfect 3D with only two channels.

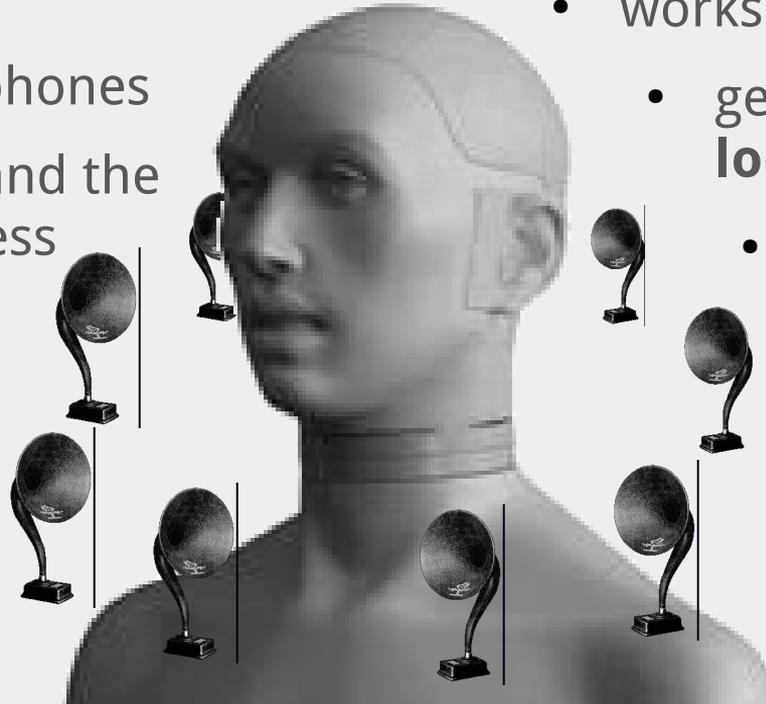


Alternatively, a mono signal recorded “dry” can be **convolved** with the left- and right-ear **HRTFs** for the desired direction.

The result at the ear drums is exactly the same as a sound actually coming from there.

## advantages:

- only two transmission channels
- works over headphones
- **HRTFs** are short and the convolution process is efficient and easy on the CPU.



## disadvantages:

- works only over headphones
  - generic HRTFs cause **in-head locatedness** and coloration.
  - stable localization requires either **head tracking** or mechanically fixing the listener's head

## advantages:

- only two transmission channels
- works over headphones
- **HRTFs** are short and the convolution process is efficient and easy on the CPU.

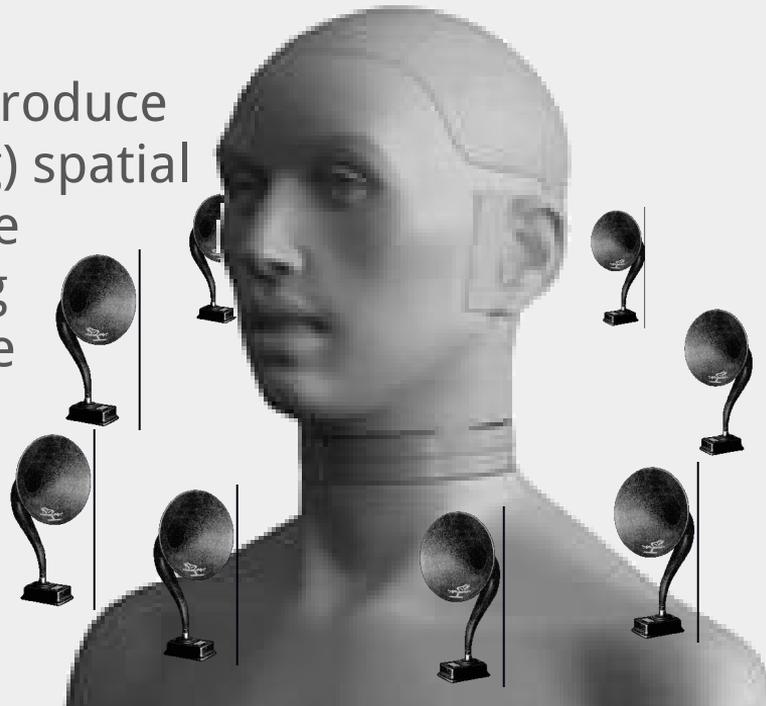
## disadvantages:

- works only over headphones
- generic HRTFs cause **in-head** localization and coloration.
- stable localization requires either **head tracking** or mechanically fixing the listener's head

Every smartphone is a 3D renderer

Binaural signals can be reproduced over loudspeakers. But now each ear signal reaches both ears. This can be suppressed using **crosstalk cancellation (CTC)**.

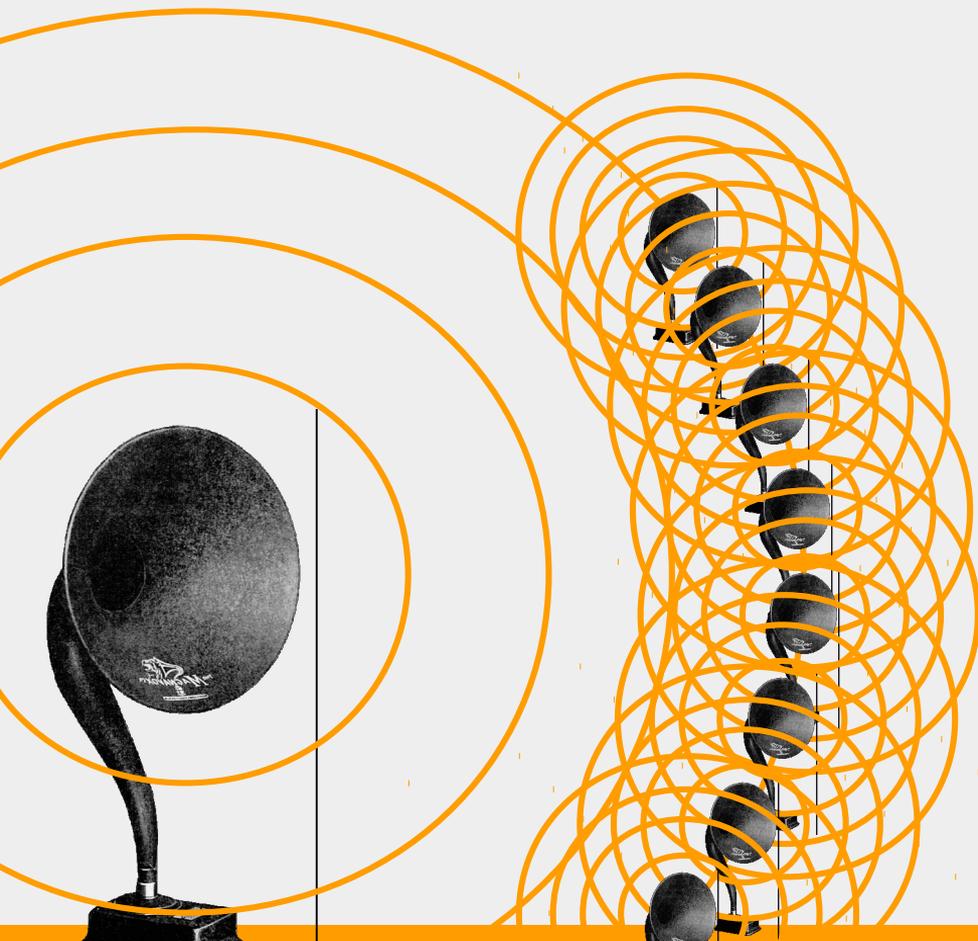
**CTC binaural** can produce nice (and surprising) spatial impressions, but the required processing can result in audible coloration.



CTC only works exactly on the axis between the two loudspeakers.

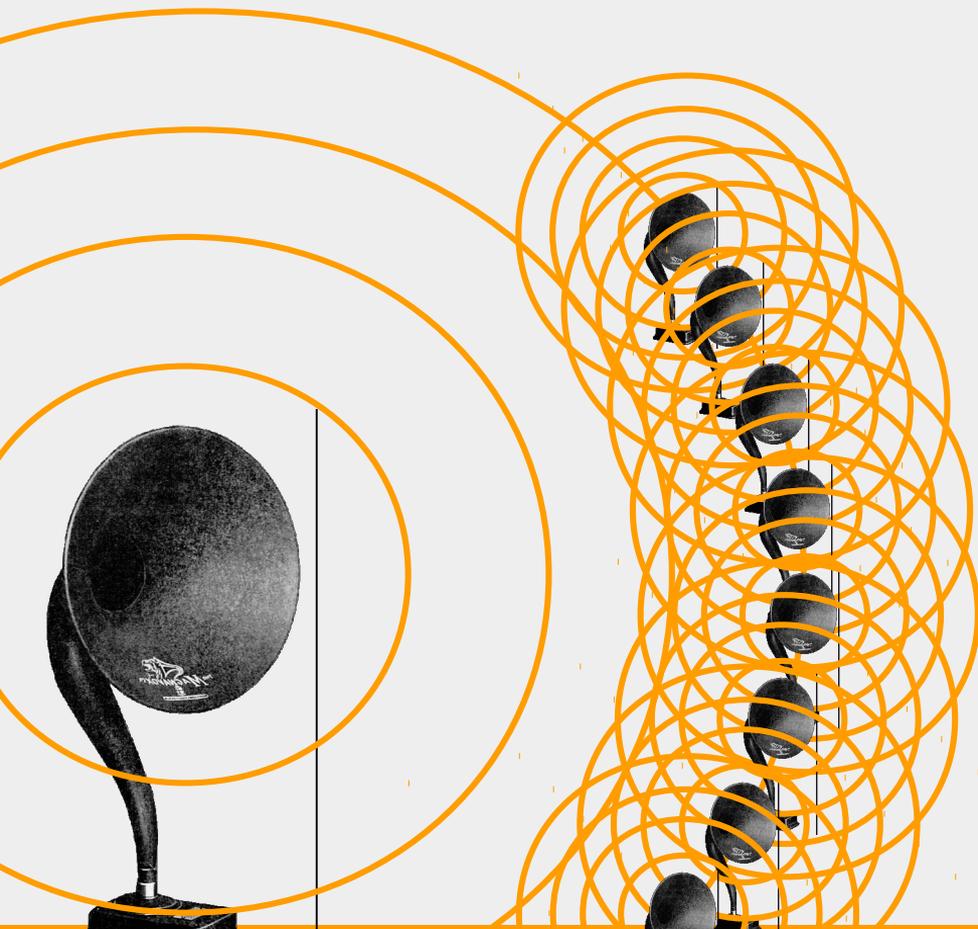


## Berkhout and de Vries, 1989-1993



Huygens' principle:

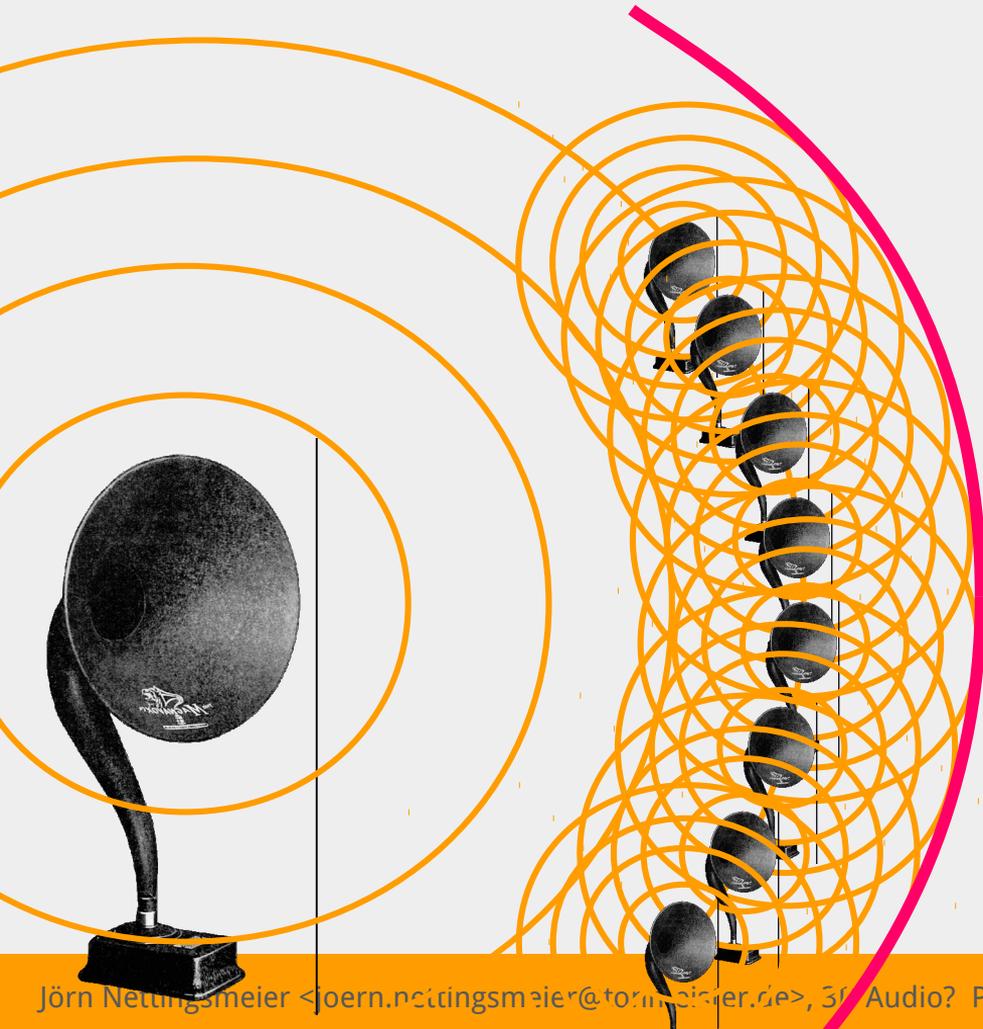
*Every wave front can be understood as a superposition of an infinite number of **elementary waves**.*



Huygens' principle:

*Every wave front can be understood as a superposition of an infinite number of elementary waves.*

Or: any sound source at any distance can be simulated by a tight row of loudspeakers.

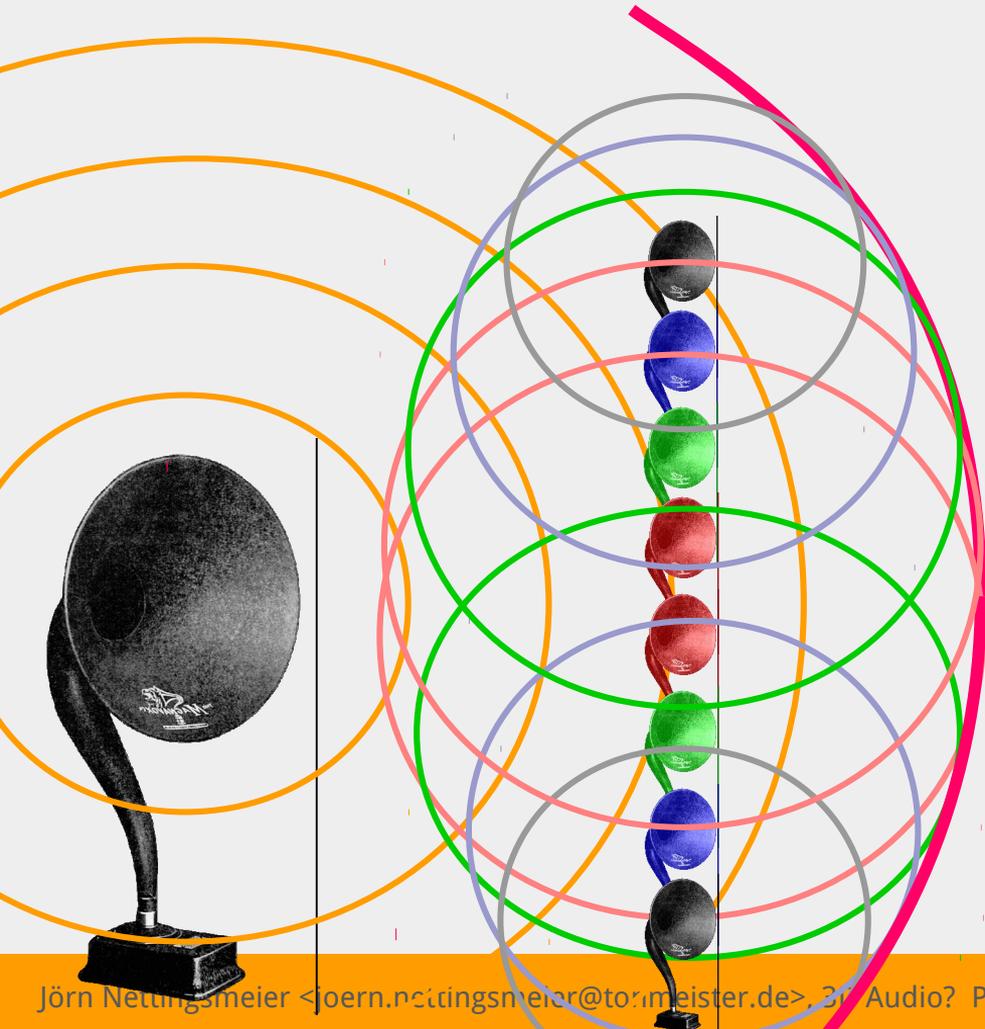


Huygens' principle:

*Every wave front can be understood as a superposition of an infinite number of **elementary waves**.*

Such a row of speakers can perfectly imitate the sound field of a source behind it, up to the **aliasing frequency**, which depends on the spacing of the speakers.

Practical systems start to alias above 1 or 2 kHz.



Huygens' principle:

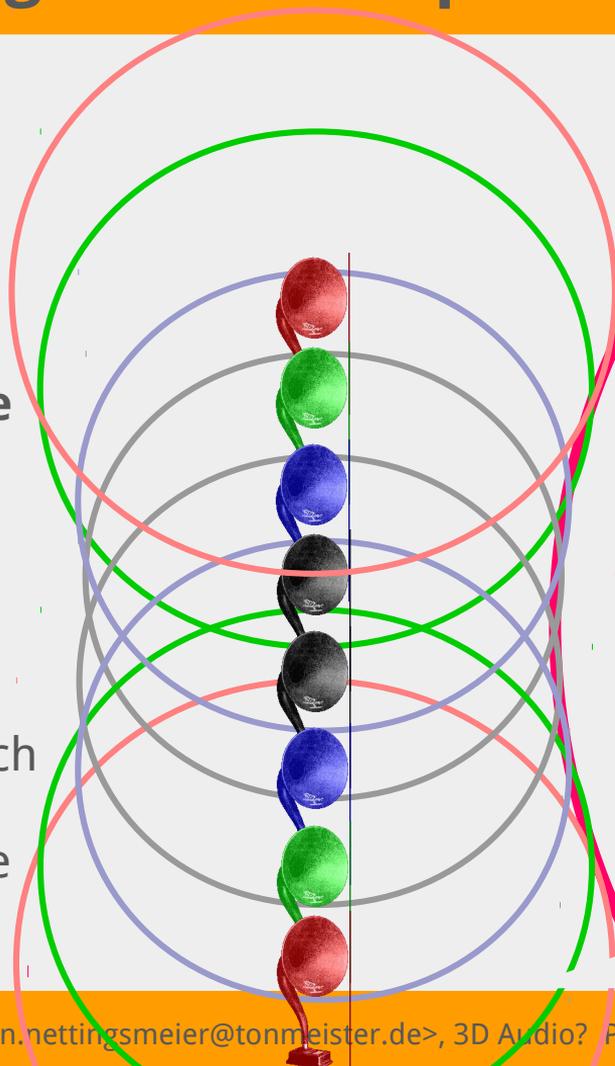
*Every wave front can be understood as a superposition of an infinite number of **elementary waves**.*

The speaker row does not need to follow the shape of the desired wave front.

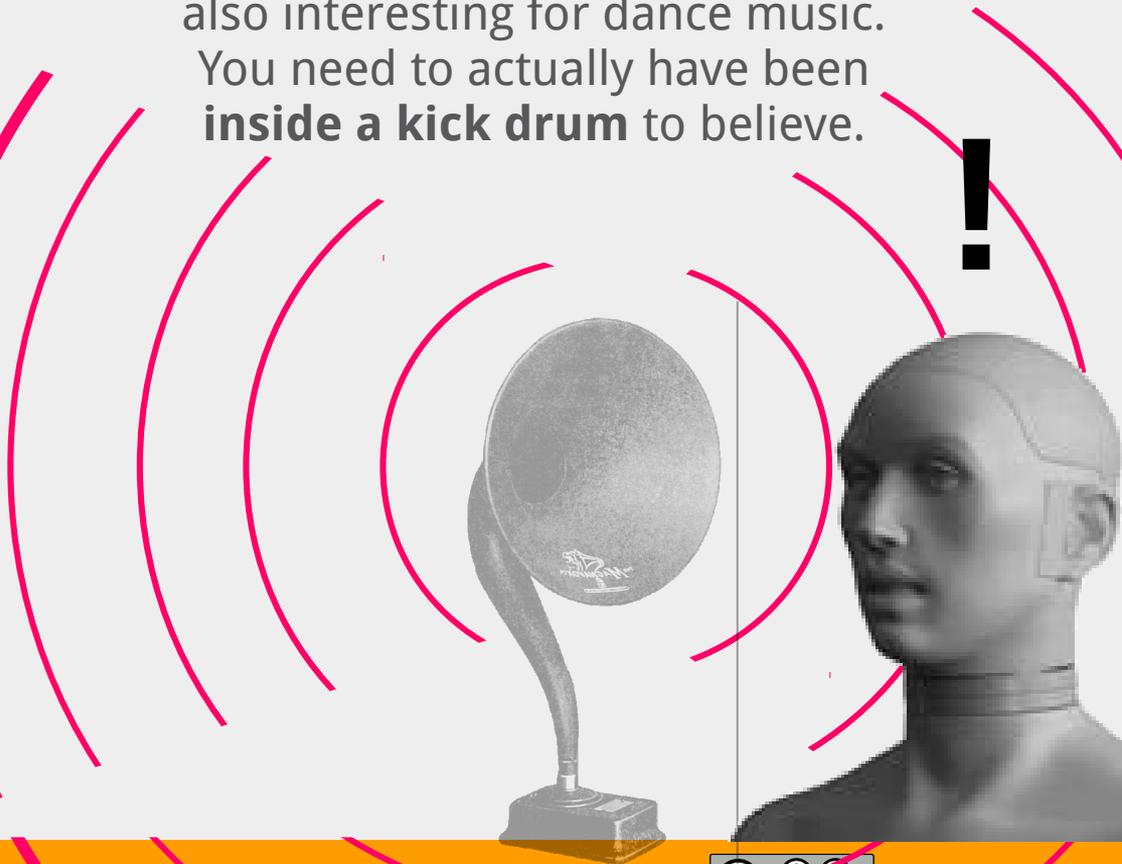
Using delays, any desired wave front can be synthesized.

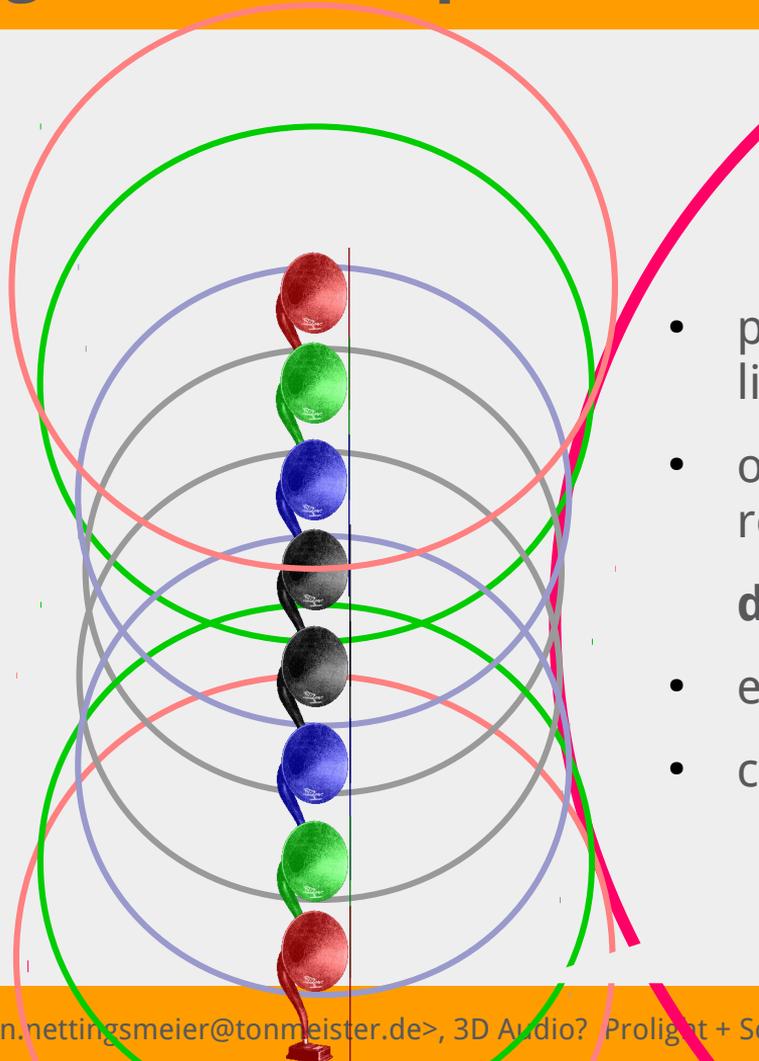
A neat trick:  
inverted delays  
produce a  
**focused source**  
inside the  
listening area.

The field  
converges in a  
single point in  
free space, which  
we perceive as  
the origin of the  
sound.



A favourite gimmick of theme parks, but  
also interesting for dance music.  
You need to actually have been  
**inside a kick drum** to believe.





## advantages:

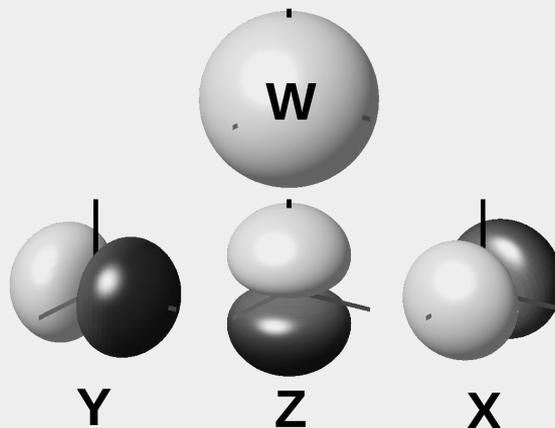
- large listening area, no “sweet spot”
- very stable localisation
- possibility of focused sources inside the listening area
- object-based mixing, content can easily be rendered to other formats

## disadvantages:

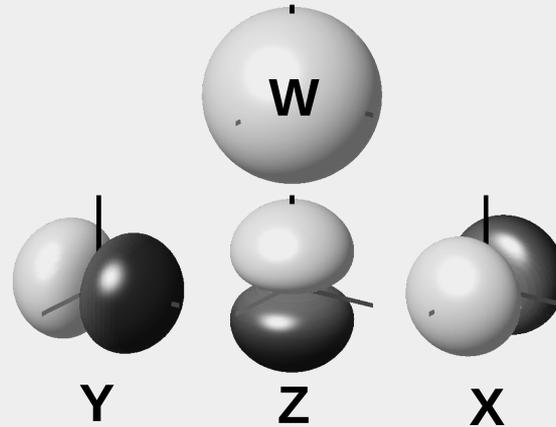
- extremely hardware intensive
- computationally expensive
  - audible aliasing artefacts
  - height localisation not feasible

# Creating sounds in space: wave field synthesis



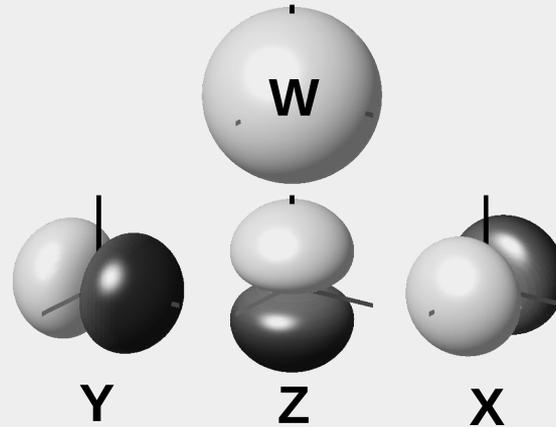


Michael Gerzon et al., 1975



Ambisonics deconstructs the sound field into directional components. In the most simple case, it is possible to use microphones that are placed as closely as possible (**co-incident**):

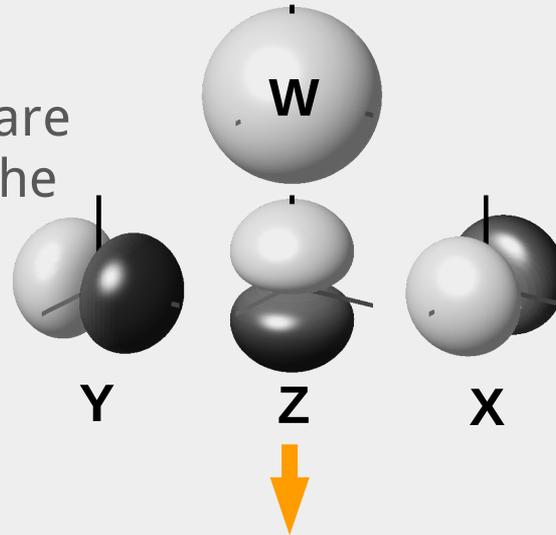
- an omni-directional mic records the sound pressure **W**;
- three fig8 mics record the sound pressure difference along the **X, Y** and **Z axes**.



Here, **W** is the mono component (directionless). The **X** difference signal means “front minus back”, the **Y** signal means “left minus right”, and **Z** means “up minus down”.

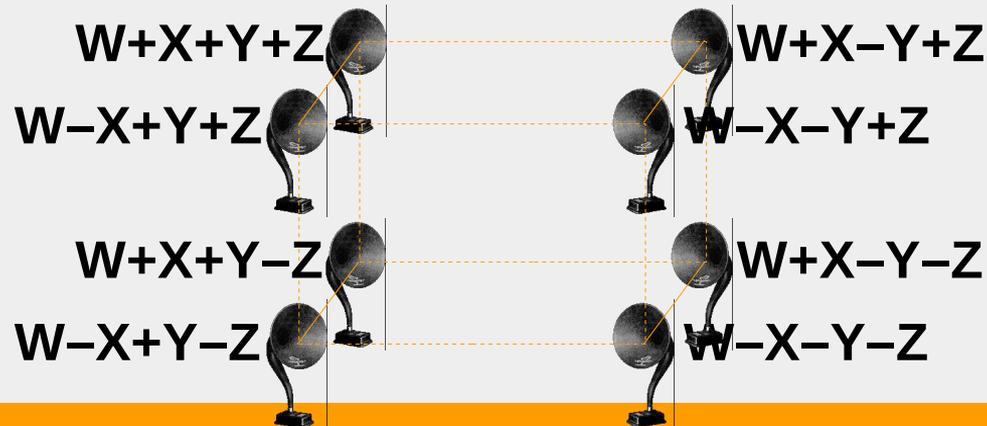
These 4 components are collectively called **B-Format**. They can be used to drive different loudspeaker arrangements with great flexibility:

The loudspeaker signals are simple combinations of the B-format components.

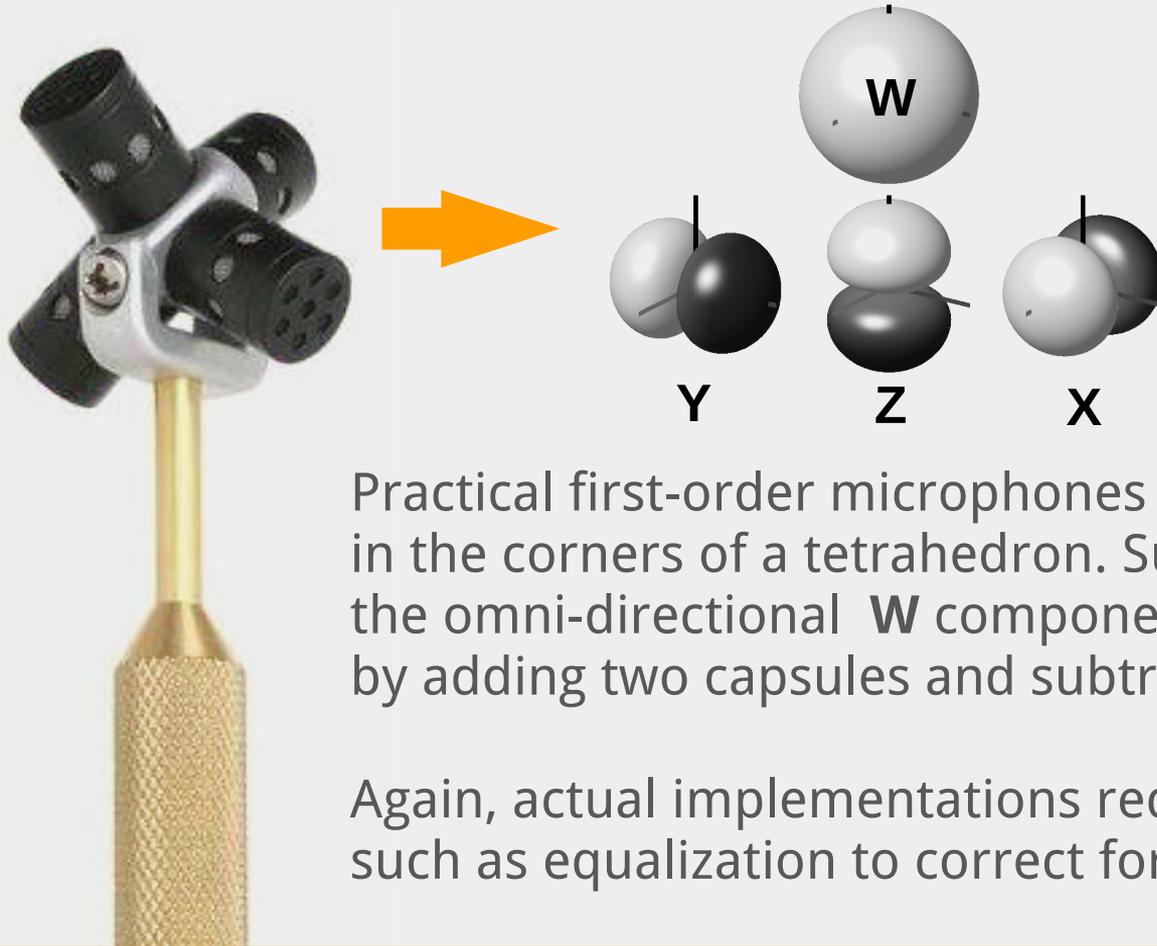


For best results, this matrix needs to be augmented with some psycho-acoustic optimizations.

Here is a simplified matrix for a cube of speakers:

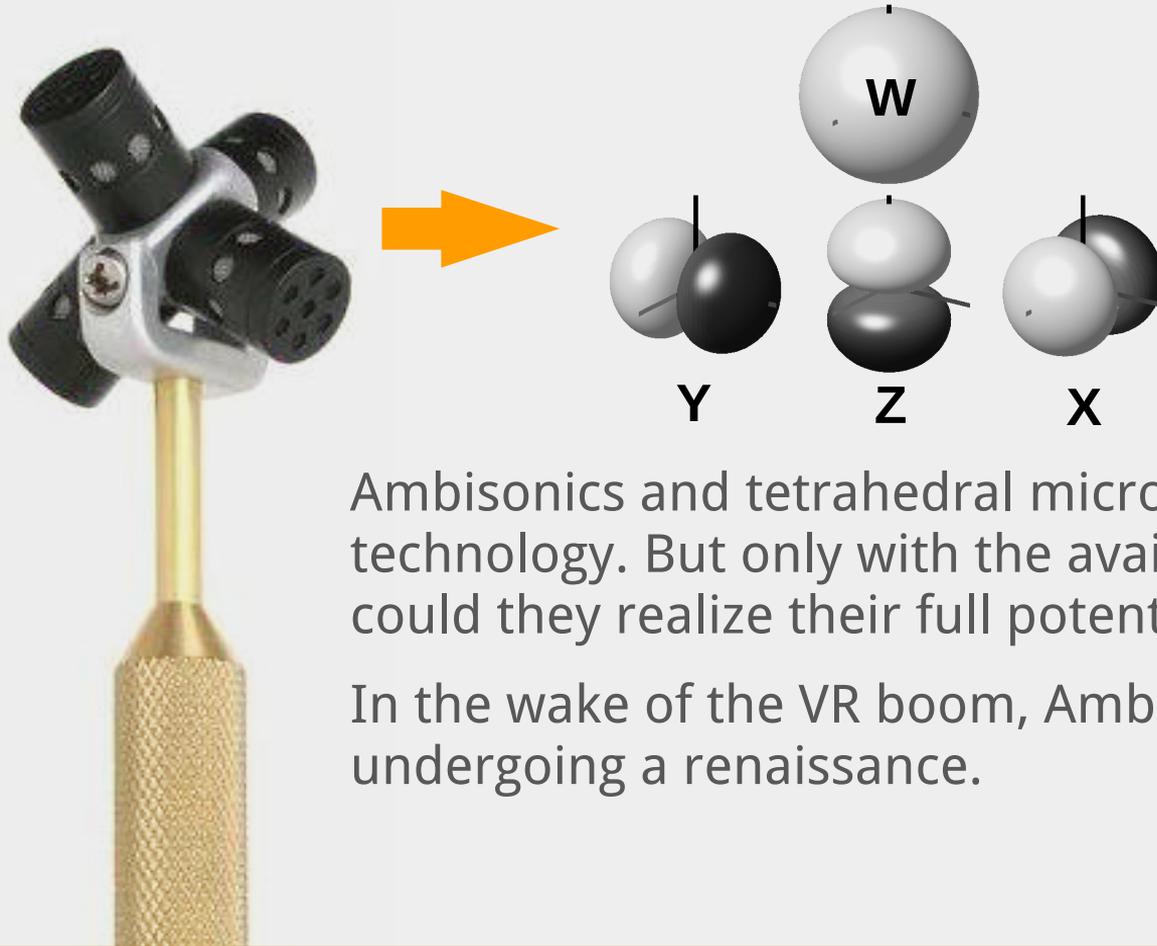


Other speaker layouts may need additional, angle-dependent gain factors.



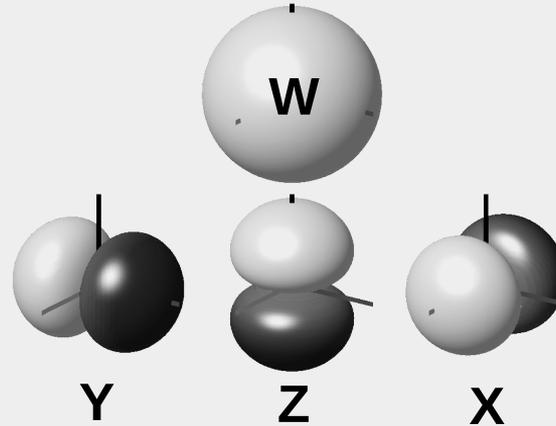
Practical first-order microphones are made from four cardioids in the corners of a tetrahedron. Summed together, they yield the omni-directional **W** component. The fig8s **XYZ** are derived by adding two capsules and subtracting the other two.

Again, actual implementations require additional optimizations such as equalization to correct for imperfect co-incidence.



Ambisonics and tetrahedral microphones are basically 70s technology. But only with the availability of digital technology could they realize their full potential.

In the wake of the VR boom, Ambisonic technology is currently undergoing a renaissance.



Even without a soundfield microphone, mono tracks (such as parts of a pop production) can be arranged in space using **Ambisonic panners**.

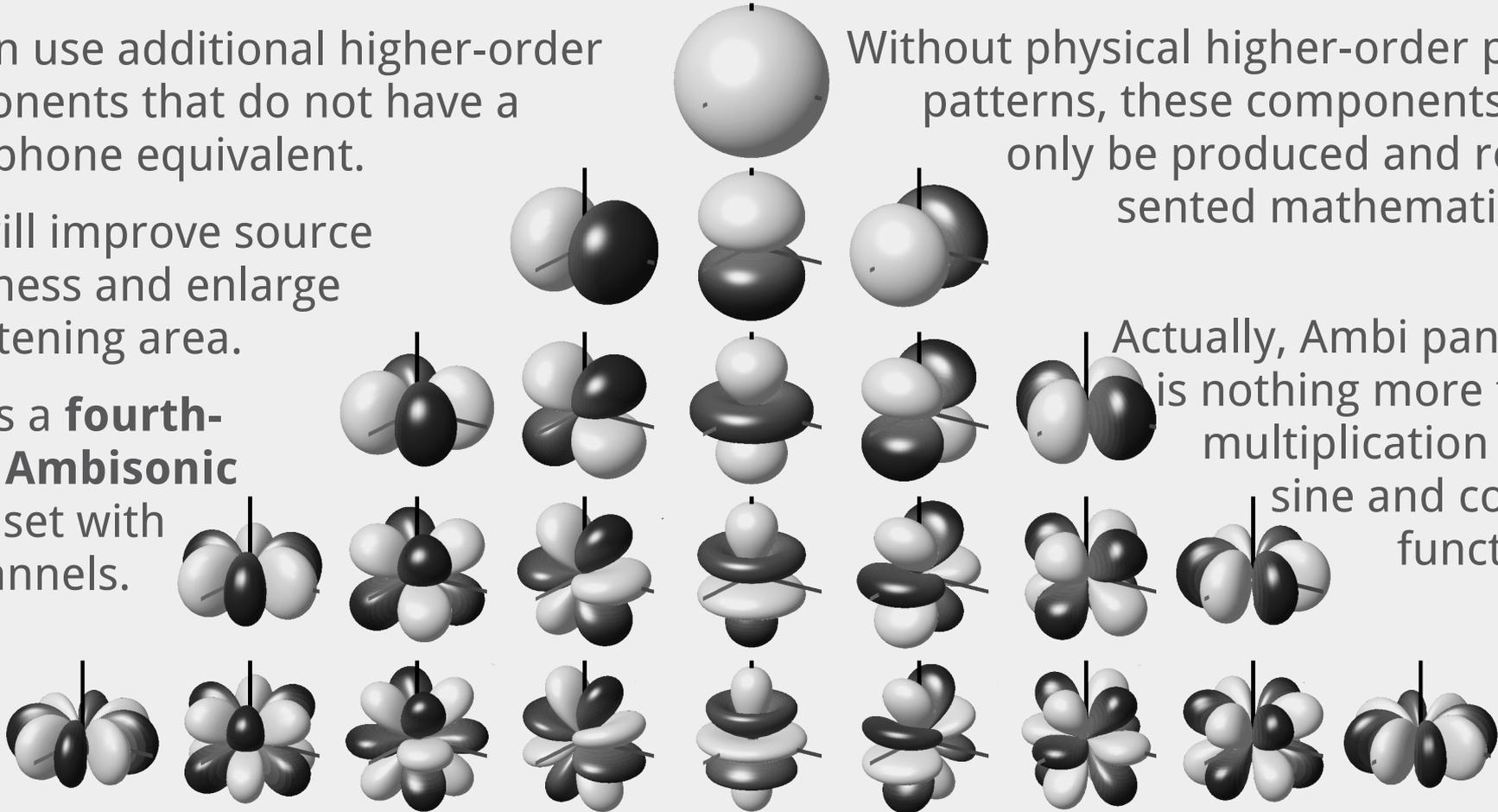
Such artificially created Ambisonic signal sets have a big advantage:

# Creating sounds in space: Ambisonics

We can use additional higher-order components that do not have a microphone equivalent.

This will improve source sharpness and enlarge the listening area.

Here is a **fourth-order Ambisonic** signal set with 25 channels.



Without physical higher-order polar patterns, these components can only be produced and represented mathematically.

Actually, Ambi panning is nothing more than multiplication with sine and cosine functions.

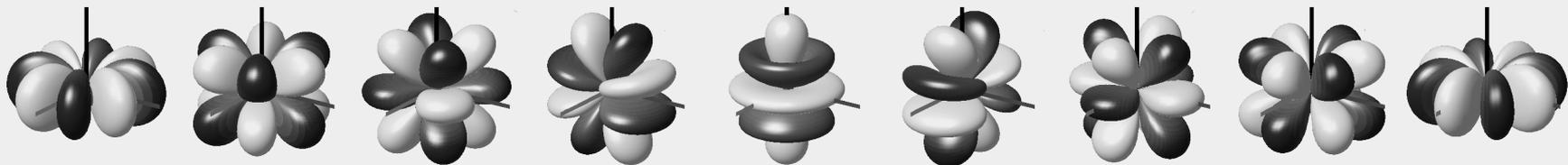


## advantages:

- simple 3D starts at 4 channels and 8 loudspeakers
- scalable, third order is sufficient for larger audiences
- the production process is independent of the targetted speaker layout
- flexible reproduction options
- can be reproduced binaurally, with decent stability even at first order
- simple rotatability is helpful for head-tracked binaural reproduction

## disadvantages:

- high correlation (=crosstalk) between speakers at low orders leads to phasing and coloration
- difficult to deploy in theatres and cinemas
- not currently directly supported by consumer equipment (but part of the MPEG-H spec)
- a little tricky to grasp
- troubleshooting playback systems is non-trivial

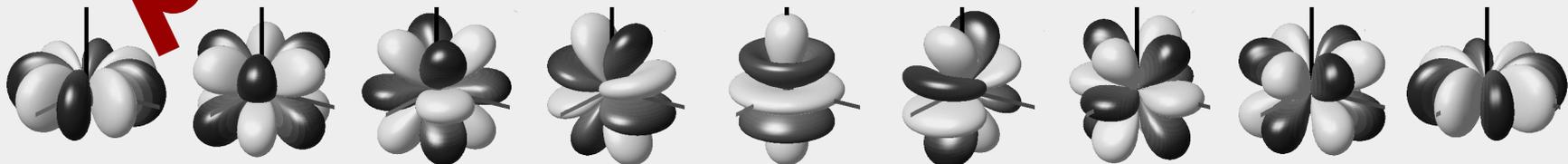


## advantages:

- simple 3D starts at 4 channels and 8 loudspeakers
- scalable, third order is sufficient for larger audiences
- the production process is independent of the targetted speaker layout
- flexible reproduction options
- can be reproduced binaurally with decent stability even at first order
- simple rotatability is helpful for head-tracked binaural reproduction

## disadvantages:

- high correlation (=crosstalk) between speakers at low orders leads to phasing and coloration
- difficult to deploy in theatres and cinemas
- not currently directly supported by consumer equipment (but part of the MPEG-H spec)
- a little tricky to grasp
- troubleshooting playback systems is non-trivial



Smart-phones

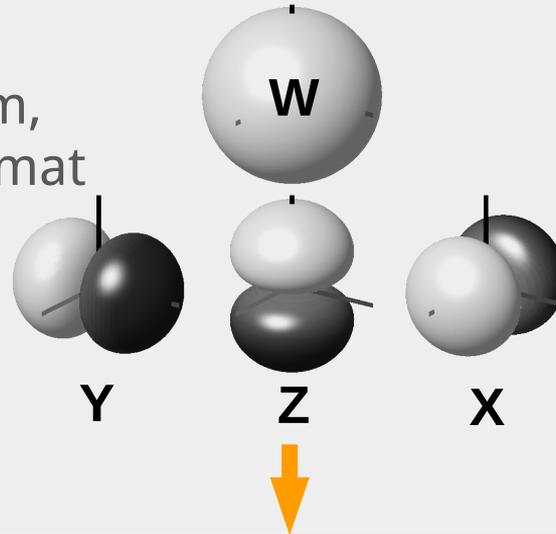
Thank you  
for your  
attention!

Unless I really botched  
the lecture, you will  
now have questions.

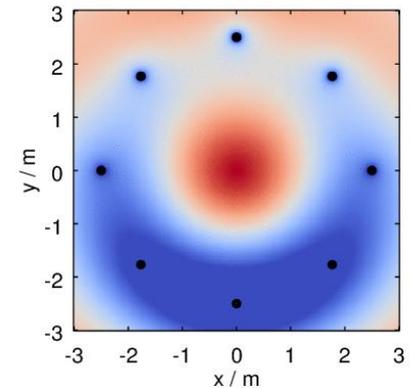
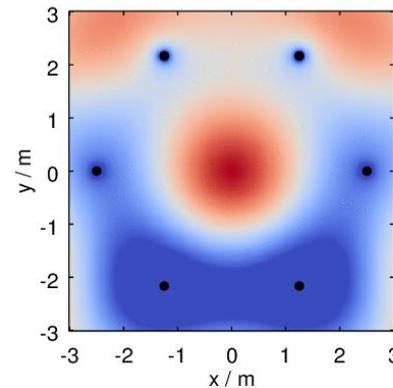
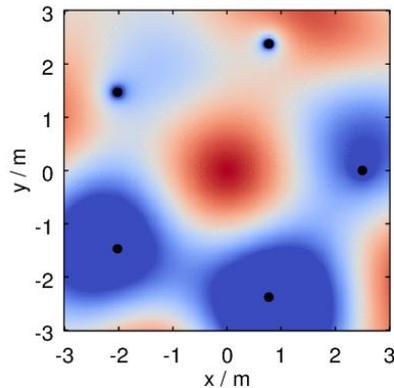
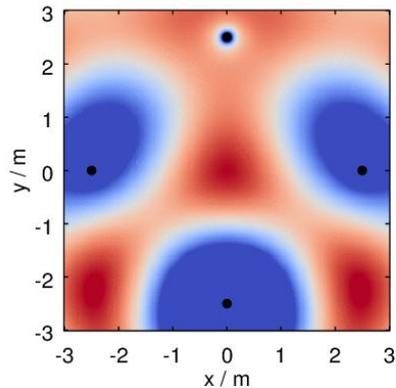
- [https://commons.wikimedia.org/wiki/File:Smilodon\\_Fatalis\\_by\\_Salvatore\\_RabitoNBG.png](https://commons.wikimedia.org/wiki/File:Smilodon_Fatalis_by_Salvatore_RabitoNBG.png)
- <https://commons.wikimedia.org/wiki/File:Maniki.jpg>
- [https://commons.wikimedia.org/wiki/File:Shanghai\\_Hongqiao\\_Railway\\_Station\\_-\\_platform\\_-\\_P1080122.JPG](https://commons.wikimedia.org/wiki/File:Shanghai_Hongqiao_Railway_Station_-_platform_-_P1080122.JPG)
- <https://commons.wikimedia.org/wiki/File:Alphah.jpg>
- <http://core-sound.com/TetraMic/1.php>

# Addendum 1: Ambisonics

A plane wave at 100 Hz,  
running from top to bottom,  
rendered in 1<sup>st</sup> order B-format  
on 4, 5, 6 and 8 speakers:



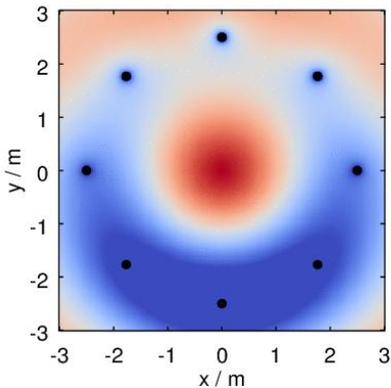
The shape of the sound  
field improves, but at the  
expense of increased  
phasing and coloration  
(think crosstalk!)



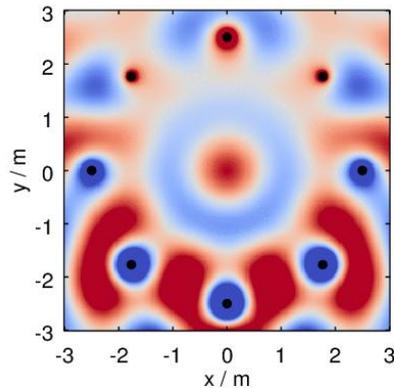
First-order Ambisonics has a very small **sweet spot** (usable listening area).

The volume in which the soundfield is reproduced physically accurately varies with frequency. Above 800 Hz, it is smaller than a human head.

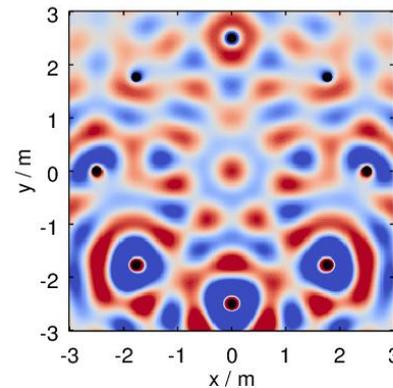
100 Hz



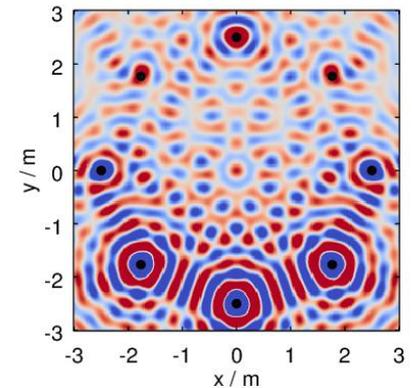
200 Hz



400 Hz



800 Hz



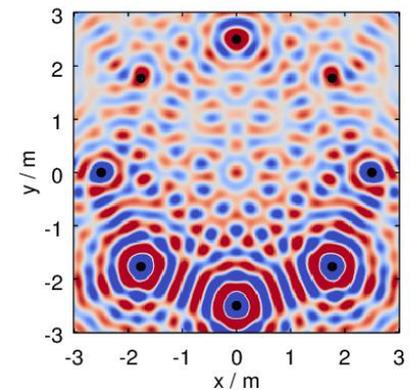
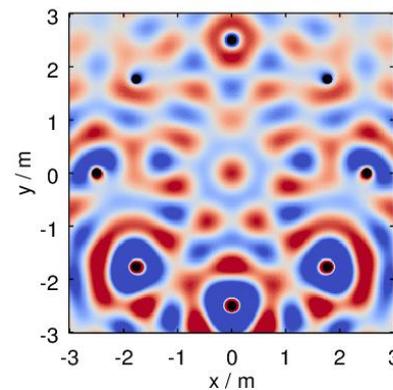
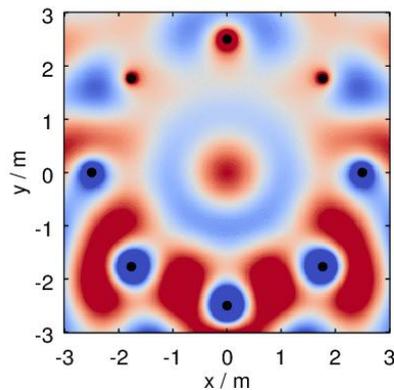
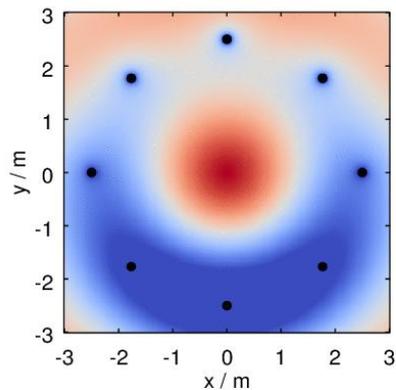
But: we don't really need physically correct reproduction everywhere, only at low frequencies (where **ITDs** matter). With a few psycho-acoustic optimizations in the decoder, we still get correct **ILDs** up to 4 kHz despite incomplete reproduction.

100 Hz

200 Hz

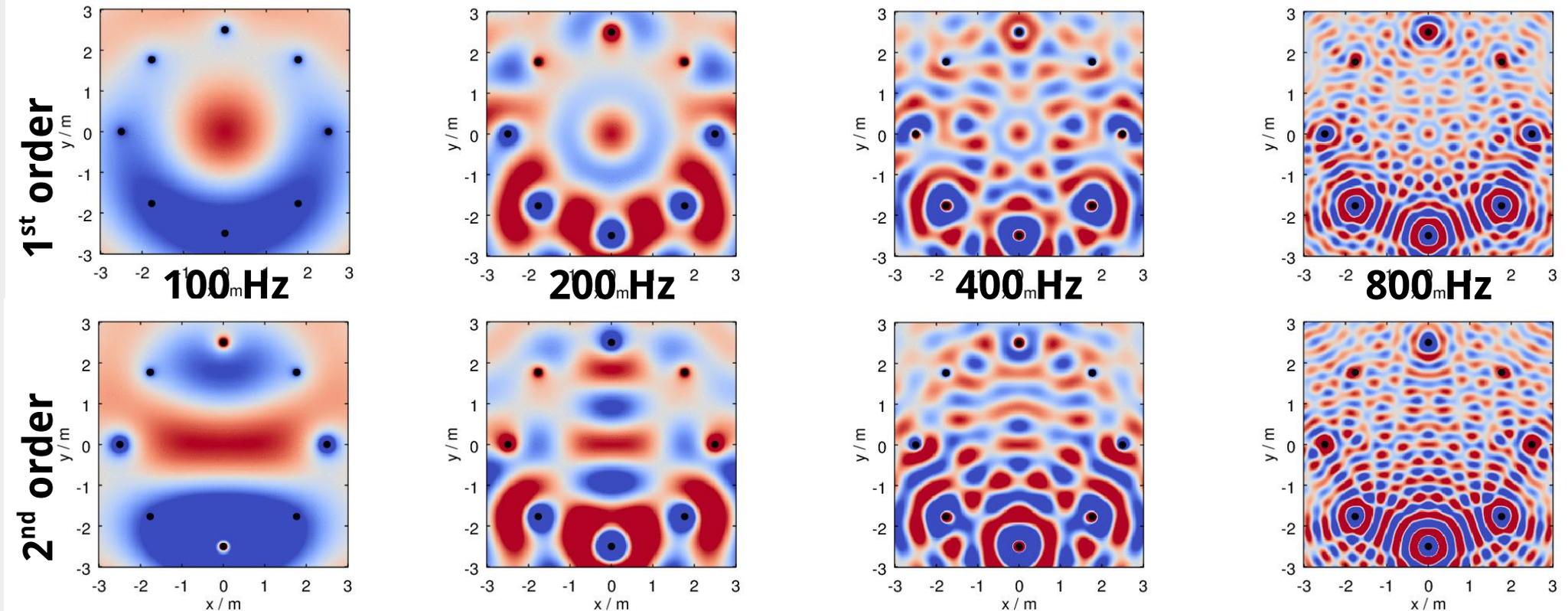
400 Hz

800 Hz



# Addendum 1: Ambisonics

Adding higher-order components reduces phasing artefacts, improves source sharpness and localization, and extends the usable listening area.



# Addendum 1: Ambisonics

Adding higher-order components reduces phasing artefacts, improves source sharpness and localization, and extends the usable listening area.

