Working hypotheses:

3D audio is the future.

3D audio is hard.

3D audio marketing is often dubious.

You cannot fool physics.

You can however fool people's ears.

Good 3D audio is a matter of the right trade-offs and compromises for each application.

Therefore, we must understand the mechanisms at work in the products we are using.

Understanding 3D audio: fundamental techniques

My name is **Jörn Nettingsmeier**, 42 years old, living in Amsterdam, The Netherlands.
I work as a live and recording sound engineer and sound designer, focusing on surround sound.

Furthermore, I lead the live sound and event technology department of the German sound engineering association, **Verband Deutscher Tonmeister** (VDT).

It's an honor to contribute to the International Audio Training Course, thanks to **ProLight + Sound Shanghai** and to **VPLT** for the invitation.

And most importantly, **thank you** for joining this presentation!

To begin, let's reflect on

0. What spatial hearing is good for

Topics:

Every solution on the market has to deal with

1. The psychoacoustics of spatial hearing

Every solution on the market will use one or a combination of several

2. Fundamental techniques

All of those techniques will come with

3. Different production paradigms

Finally, we will look at

4. 3D audio in practice

What spatial hearing is good for:

What spatial hearing is good for:

What spatial hearing is good for:

Survival.

Hearing covers all directions, not just the front.

Hearing warns us of that which we **cannot see yet**.

- a predator?
- a car at a crossing?

What spatial hearing is good for:

Orientation.

We hear

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

What spatial hearing is good for:

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 urgent information.

spatial hearing is good for:

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.

What spatial hearing is good for:

Creativity.

Spatial hearing is a basic **survival skill**.

Therefore, it is connected to **intense emotions**.

A gentle brushing sound from behind can tell us our lover is near by.

Or that we are about to be eaten by a tiger.

What spatial hearing is good for:

Creativity.

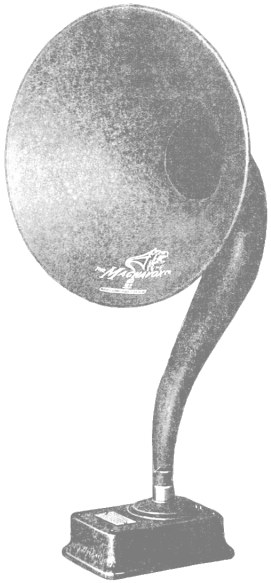
Good **localization** keeps the focus on stage and screen.

Envelopment creates emotion that transcends the boundaries of the stage.

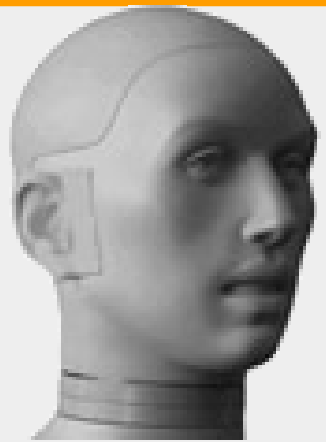
Surround effects put us in the center of the story.

Spatialisation allows for greater scene complexity.

1. The psychoacoustics of spatial hearing



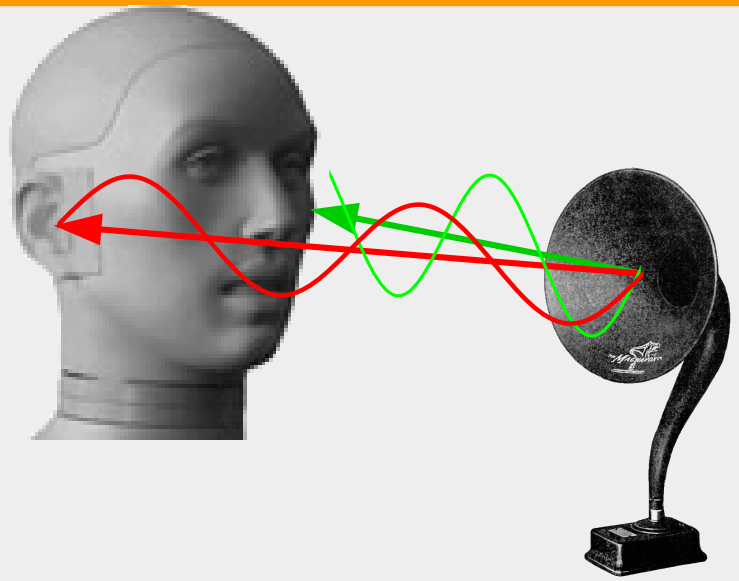
How we localize sounds.



How we localize sounds in front:

A sound from the left reaches the left ear first, and the right ear a little bit later:

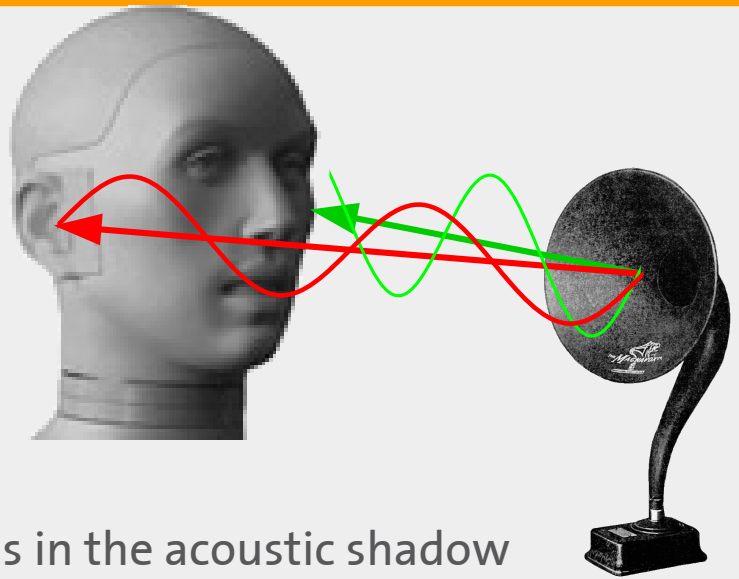
There is an **inter-aural time difference (ITD)**.



How we localize sounds in front:

A sound from the left reaches the left ear first, and the right ear a little bit later:

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



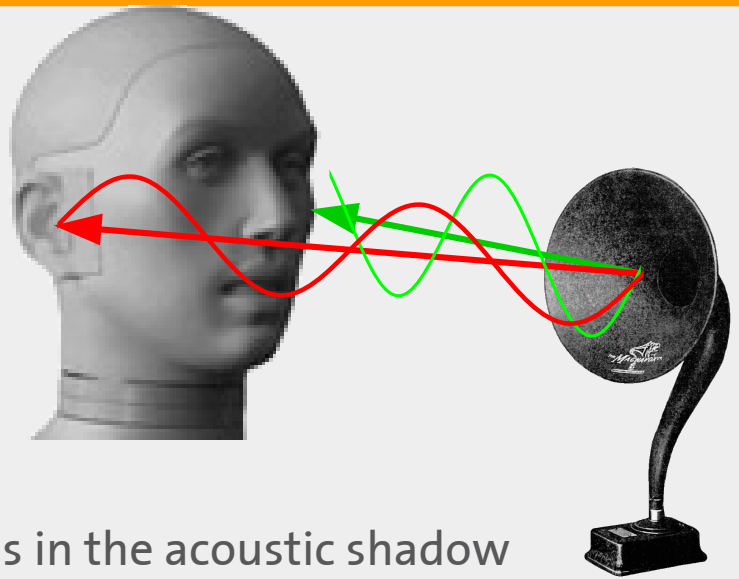
Additionally, the right ear is in the acoustic shadow of the head, where the sound is softer:

We observe an **inter-aural level difference (ILD)**.

How we localize sounds in front:

A sound from the left reaches the left ear first, and the right ear a little bit later:

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

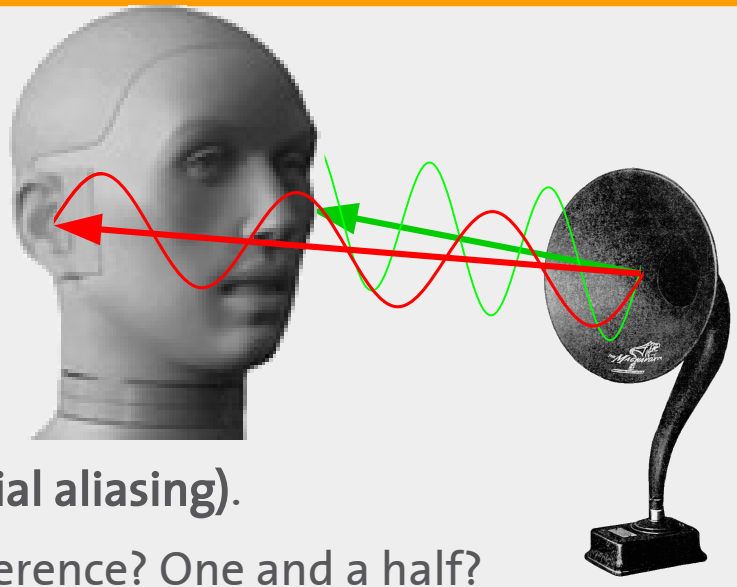


Additionally, the right ear is in the acoustic shadow of the head, where the sound is softer:

We observe an **inter-aural level difference (ILD)**.

These **spatial cues** are very precise:
with their help, we can localise sounds to an accuracy of **2 degrees**.

How we localize sounds in front:

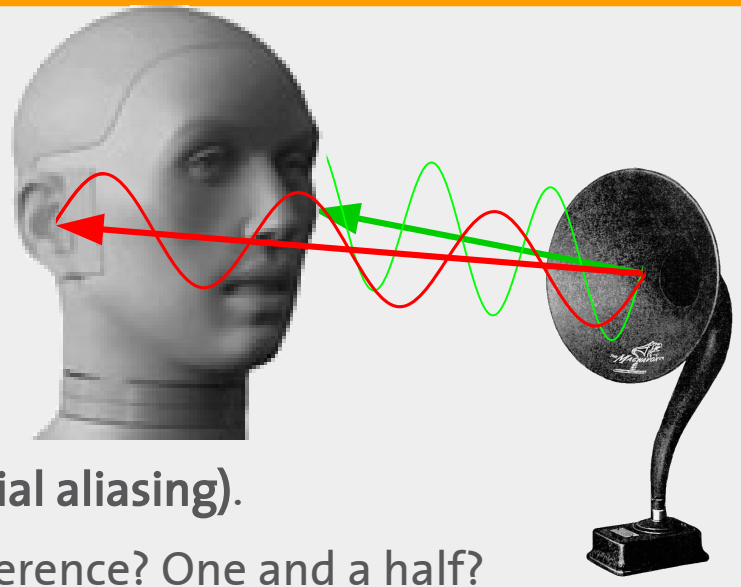


We perceive the **ITD** as a **phase difference** between the ears.

At wavelenghts shorter than twice ear distance, this phase information becomes ambiguous **(spatial aliasing)**.

Is it half a wavelength difference? One and a half? Two and a half? We do not know. So the brain discards **ITD** information above 700Hz.

How we localize sounds in front:



We perceive the **ITD** as a **phase difference** between the ears.

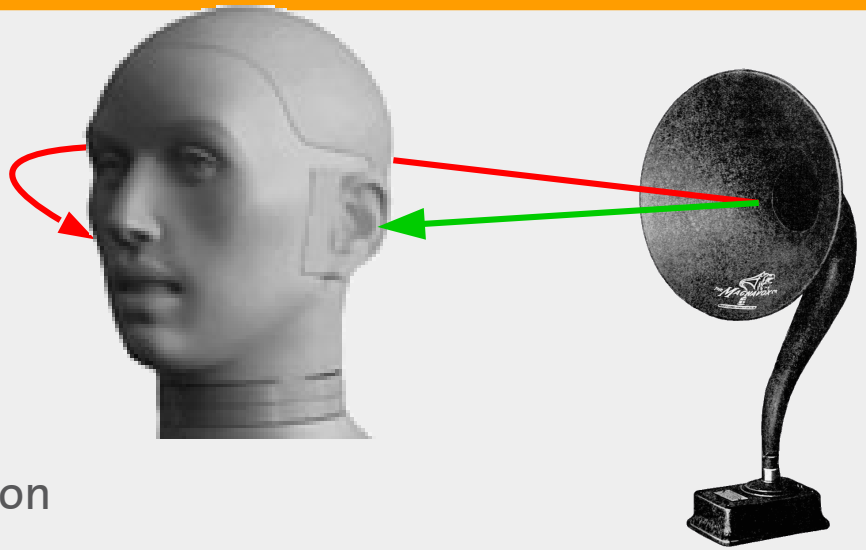
At wavelenghts shorter than twice ear distance, this phase information becomes ambiguous **(spatial aliasing)**.

Is it half a wavelength difference? One and a half? Two and a half? We do not know. So the brain discards **ITD** information above 700Hz.

The inverse holds for **ILD**s: only shorter wavelengths are affected by the head shadow. Long waves diffract around the head and do not result in much level difference.

ITD and ILD complement each other in different frequency bands.

How we localize sound behind us:

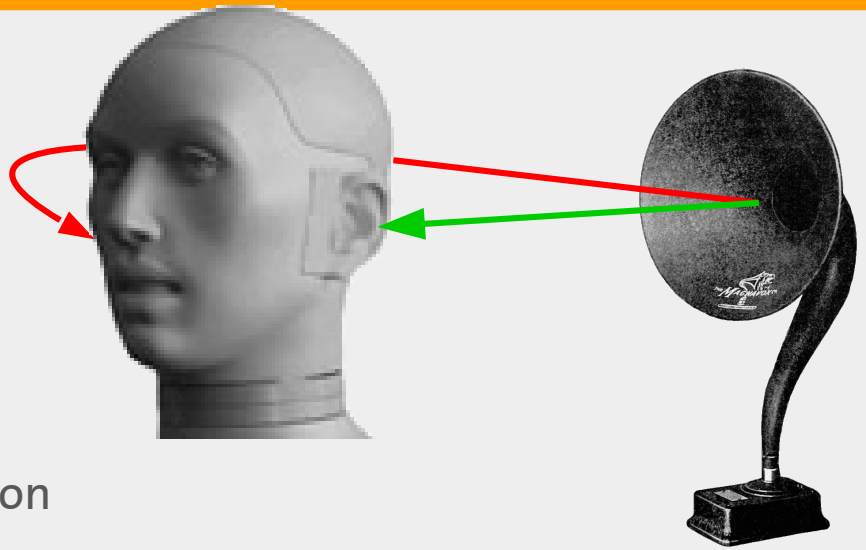


A sound from behind produces the exact same time and level differences (**ILD** und **ITD**) as one from the front.

We can tell them apart, because our head, torso, and outer ear function as a **directional equalizer,** changing the sound's spectral balance depending on the position of the source.

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

How we localize sound behind us:



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

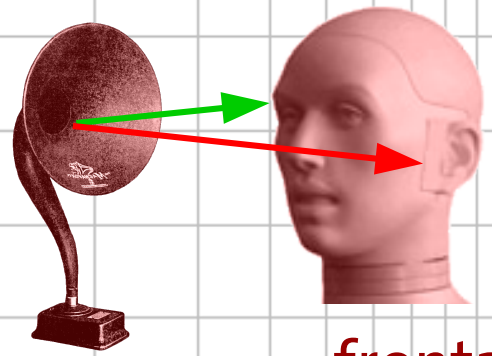
We can tell them apart, because our head, torso, and outer ear function as a **directional equalizer,** changing the sound's spectral balance depending on the position of the source.

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

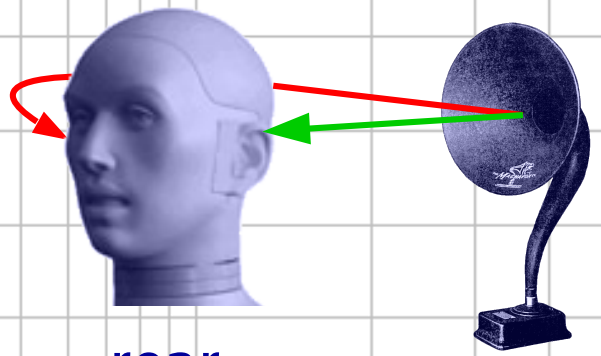
We are not consciously aware of this **substantial coloration!** But it is possible to measure it by placing microphones in the ear canal:

The ear filter curves of the author
(also known as **head-related transfer functions, HRTFs**)
for sources from the front and 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.



frontal



rear

100
95
90
85
80
75
70
65
60
55
50
45

11 20 30 40 50 60 80 100 200 300 400 600 800 1k 2k 3k 4k 5k 6k 8k 10k 26.5k Hz

1: vorn 1/12 80.9 dB NEW 2: hinten 1/12 57.4 dB NEW

How we localize sounds vertically:

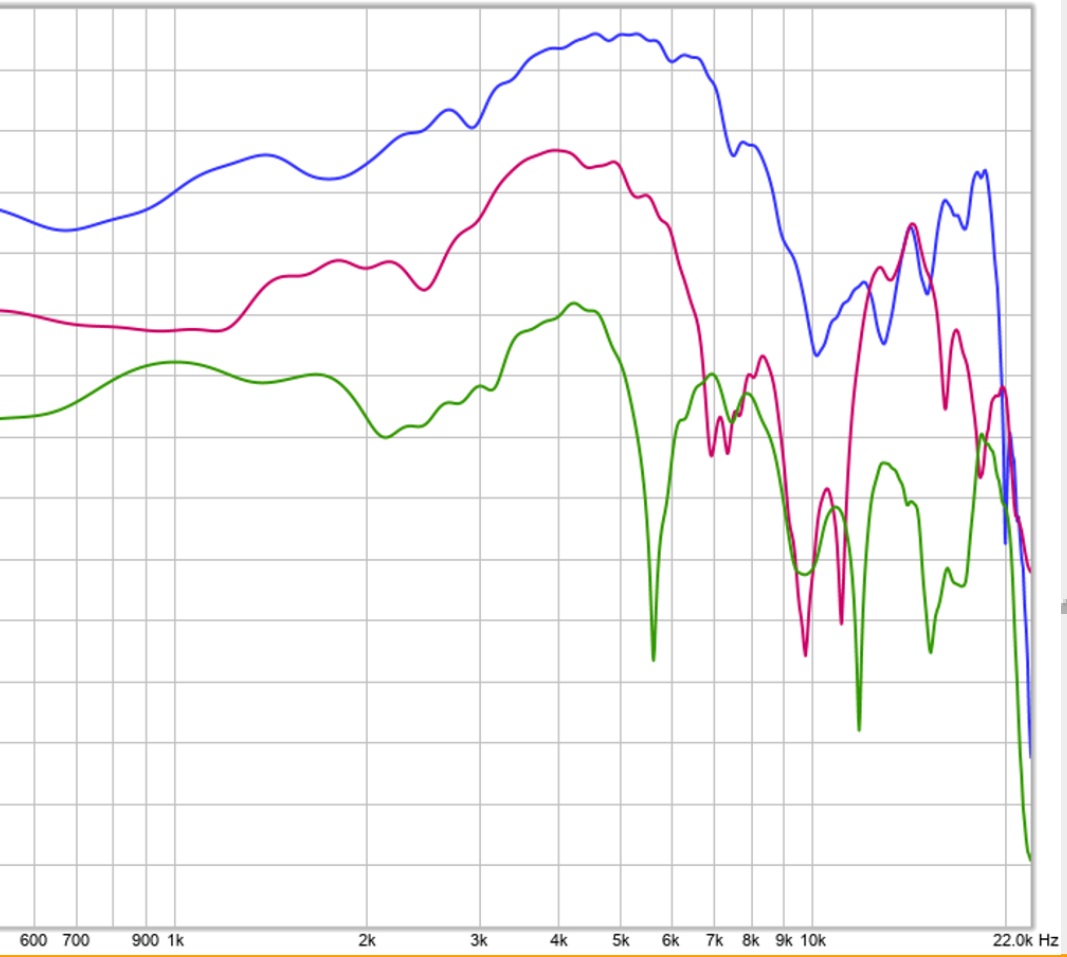
As a sound source moves up and down, there is no change in **ILD** and **ITD**.

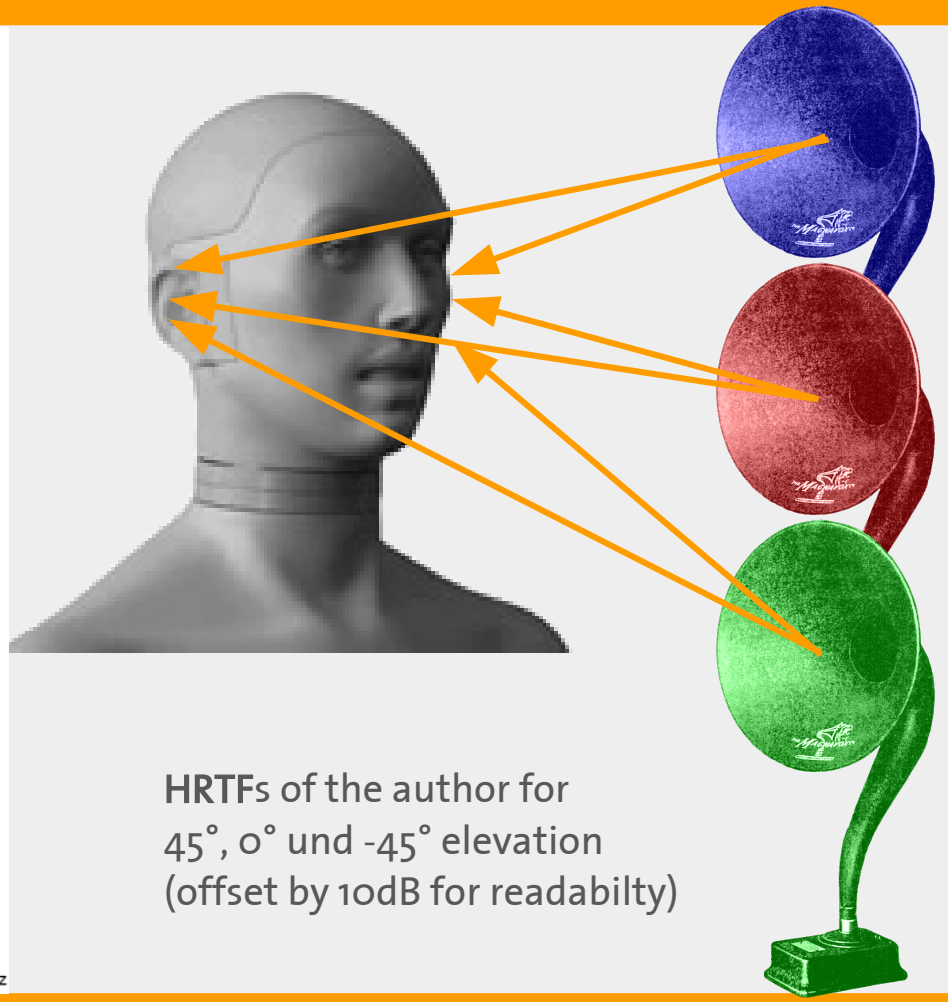
Again, we have to rely on the **directional equalizer** of our outer ear.

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

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

How we localize sounds vertically:





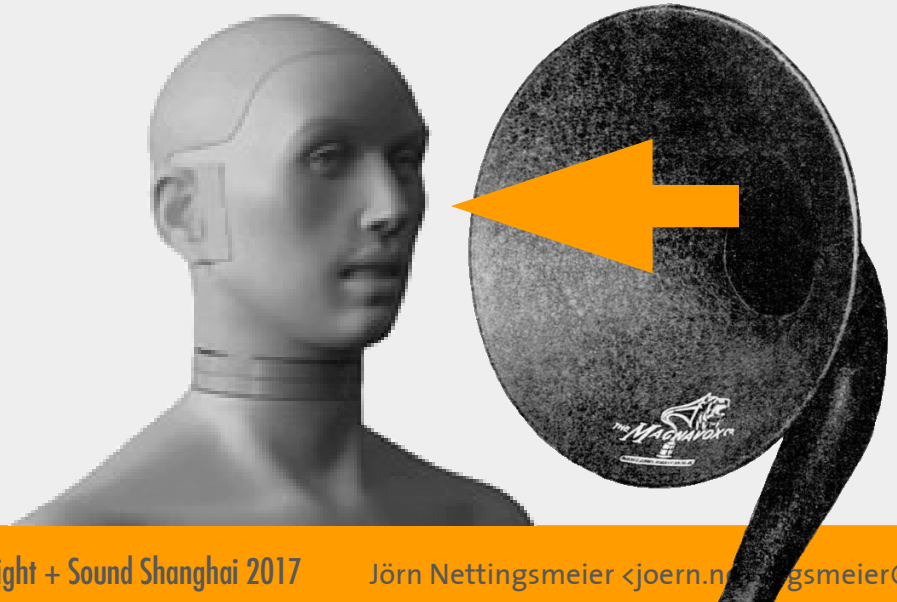
HRTFs of the author for
45°, 0° und -45° elevation
(offset by 10dB for readabilty)

How we perceive distance:

How we perceive distance:

Only at very close range can we perceive distance directly:

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

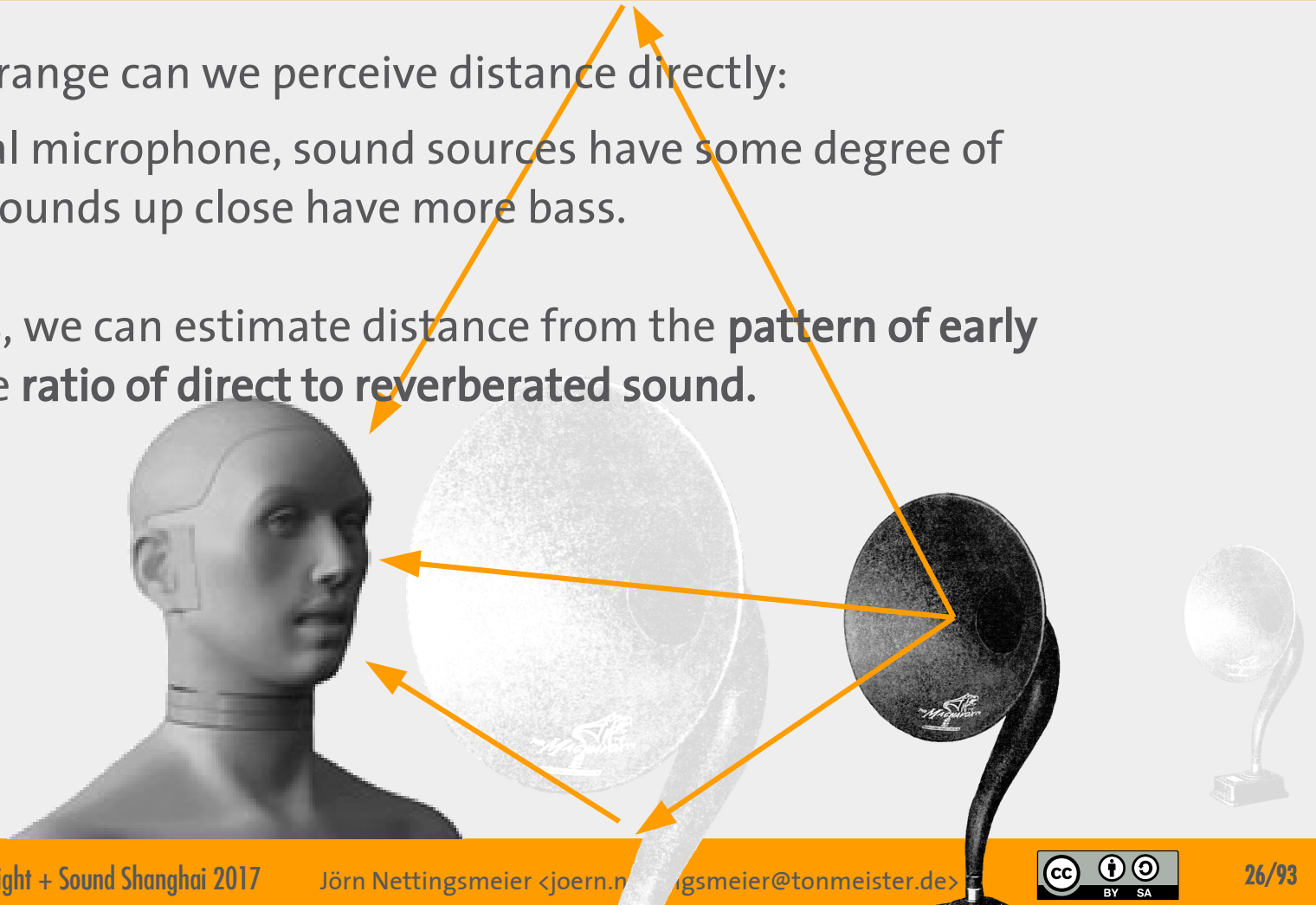


How we perceive distance:

Only at very close range can we perceive distance directly:

Like any directional microphone, sound sources 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:

Only at very close range can we perceive distance directly:

Like any directional microphone, sound sources 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.

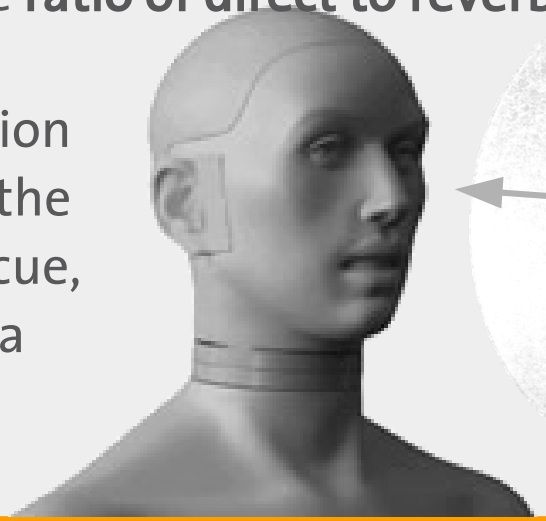
How we perceive distance:

Only at very close range can we perceive distance directly:

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

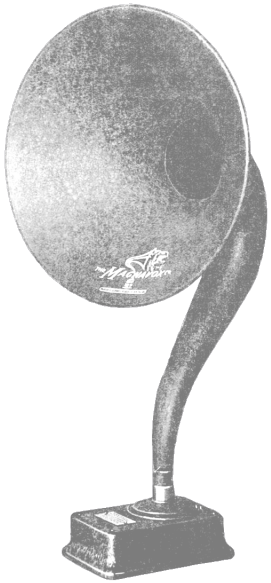
In enclosed spaces, we can estimate distance from the **pattern of early reflections** and the **ratio of direct to reverberated sound.**

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



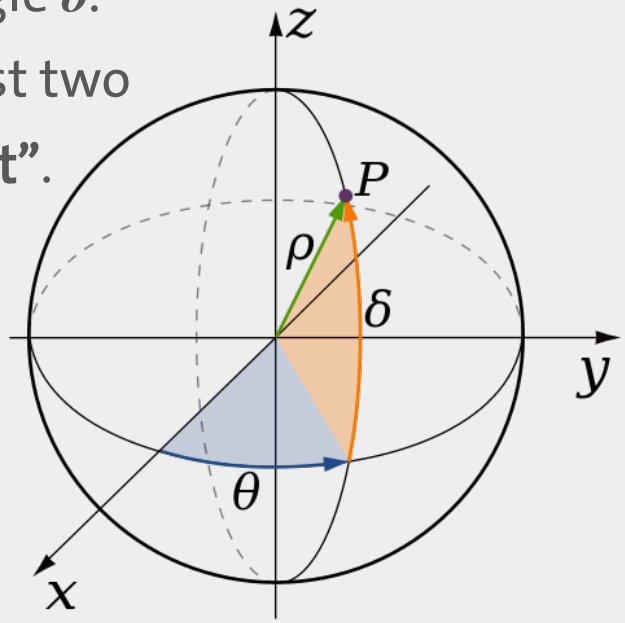
Finally, **air absorbs** more sound energy at **higher frequencies**. Very far-away sources sound less bright.

2. Fundamental 3D techniques



To begin with: what do we mean by **“3D”**?

Most systems can place **sounds on a sphere around the listener**. That gives two degrees of freedom: horizontal angle θ, and elevation angle δ.
That means all fundamental “3D” techniques have at most two degrees of freedom. In marketing, 3D means: **“with height”**.

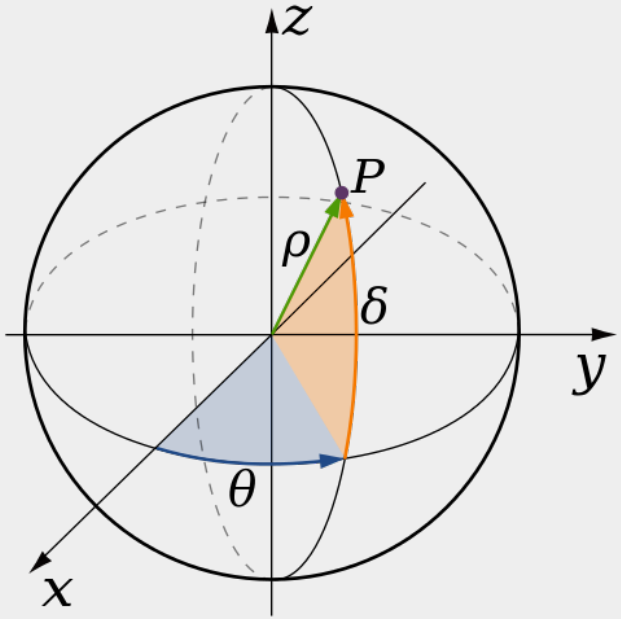


We can add distance cues, and then we get closer to real 3D. But **no system can produce sounds inside the loudspeaker array** (with the exception of WFS, but only under controlled conditions).

So: Do we do real 3D? No, but we get close.

Some vendors even try to market “4D”.
“4D” is not “more spatial than 3D”
“4D” is nonsense.

If someone tries to sell you 5D, run for cover.



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

- Curtis Roads

Fundamental techniques:

"If you want sound to come from a particular place, put a loudspeaker there!"

- Curtis Roads

- simple but 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

Fundamental techniques: stereophony

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

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

Stereo is not limited to two channels.

All channel-based formats (5.1, 7.1, 22.2, Auro-3D etc.) use stereophonic localization.

Fundamental techniques: stereophony

Problem: **stereophony is fragile**.

It has a very small **sweet spot**.

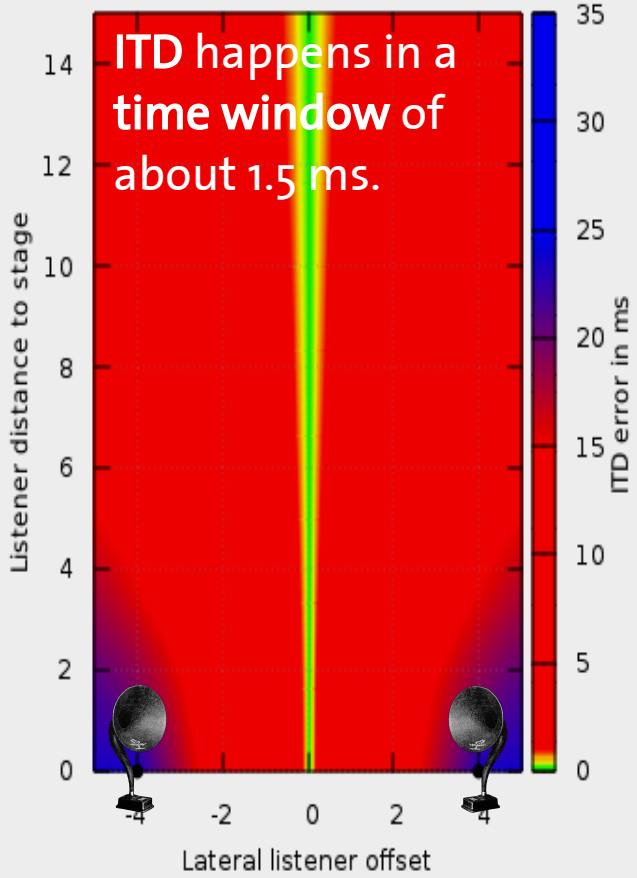
Fundamental techniques: stereophony

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

Let's look at a live concert
or cinema example
with a stage/screen 8 metres wide:

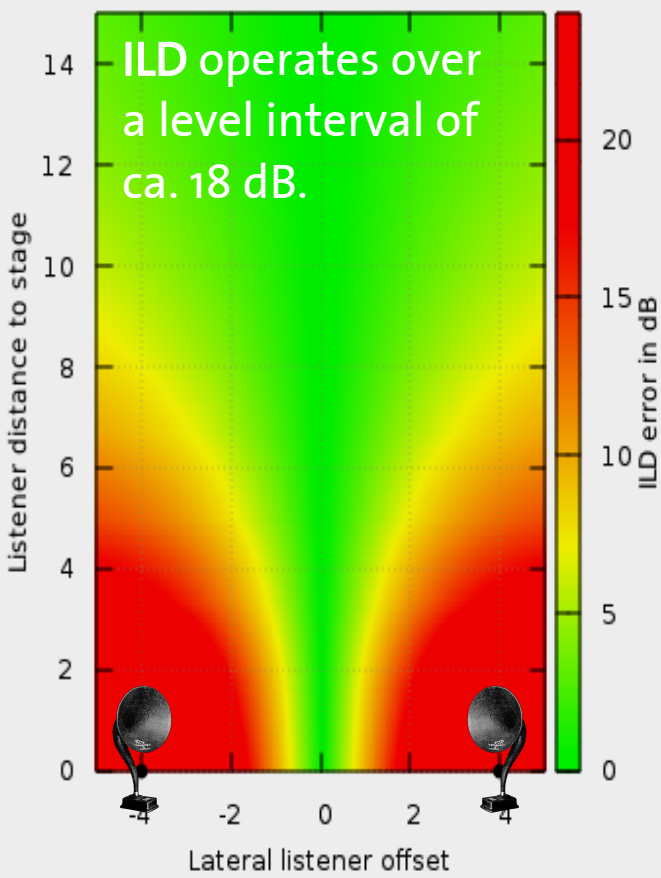
ILD operates over a level interval of ca. 18 dB.

Fundamental techniques: stereophony



Only people in the green area will get correct localisation from ITD (left) or ILD (right).

People in the red area will hear all sound coming only from the nearest speaker.



To address this problem, an additional center speaker can be added.

But great care must be taken to have enough **decorrelation** between pairs of speakers to avoid colouration.

The L–C and the C–R pair can have high correlation.

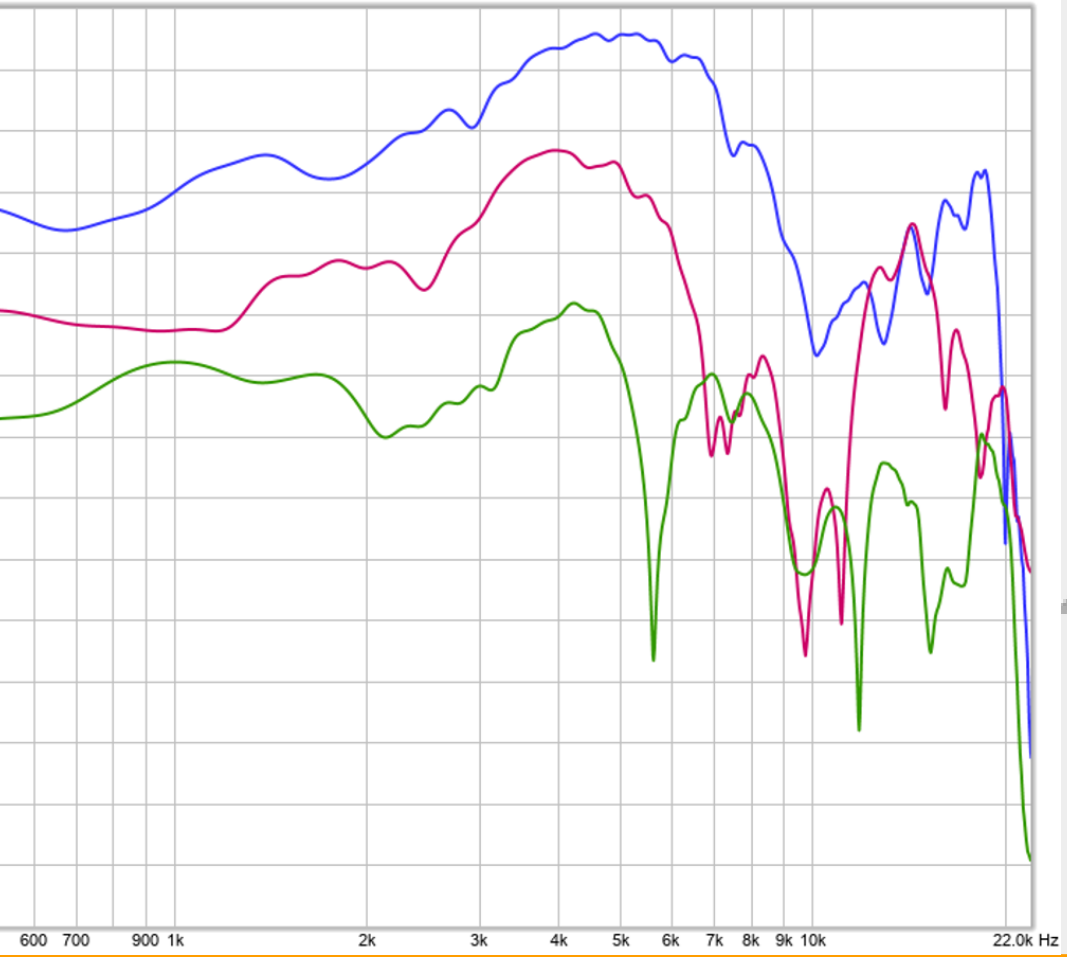
But L and R must not, which requires **carefully placed** main microphones with **high directivity**.

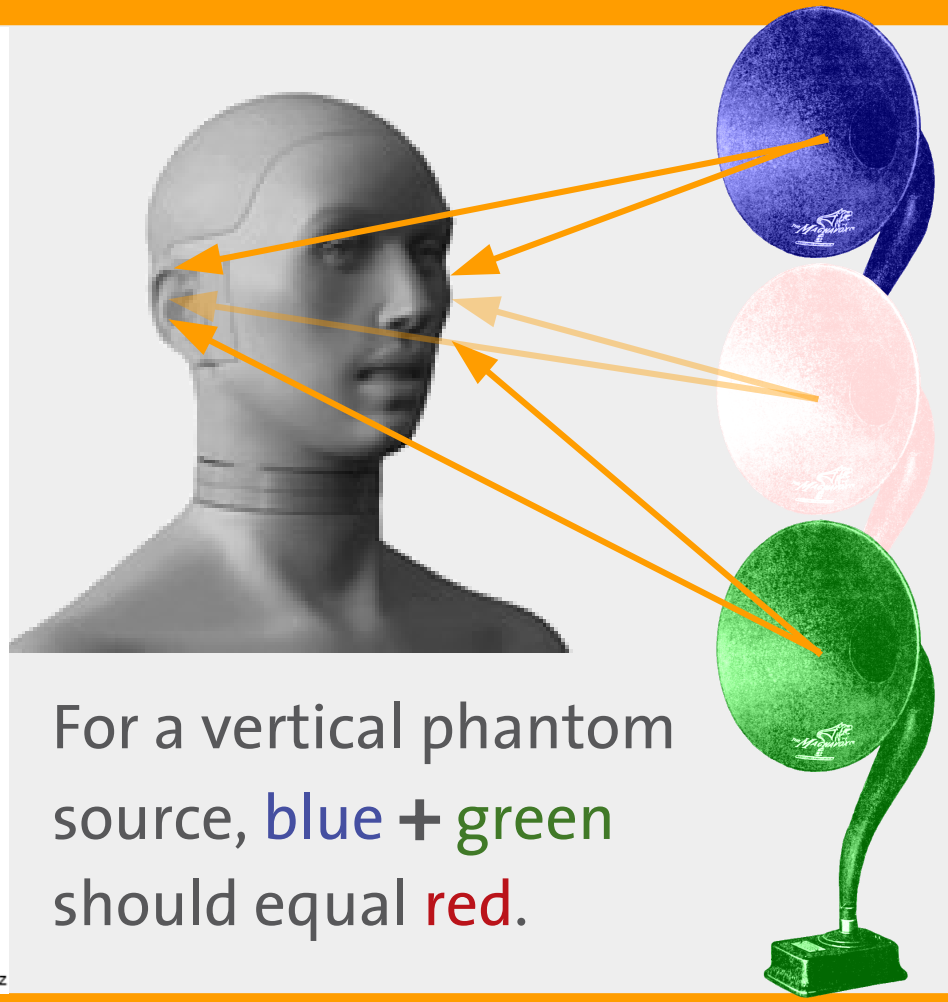
Fundamental techniques: stereophony

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

Vertical localisation however is bad – we don't have **ILD** and **ITD cues**, and the **spectral cues** are wrong:

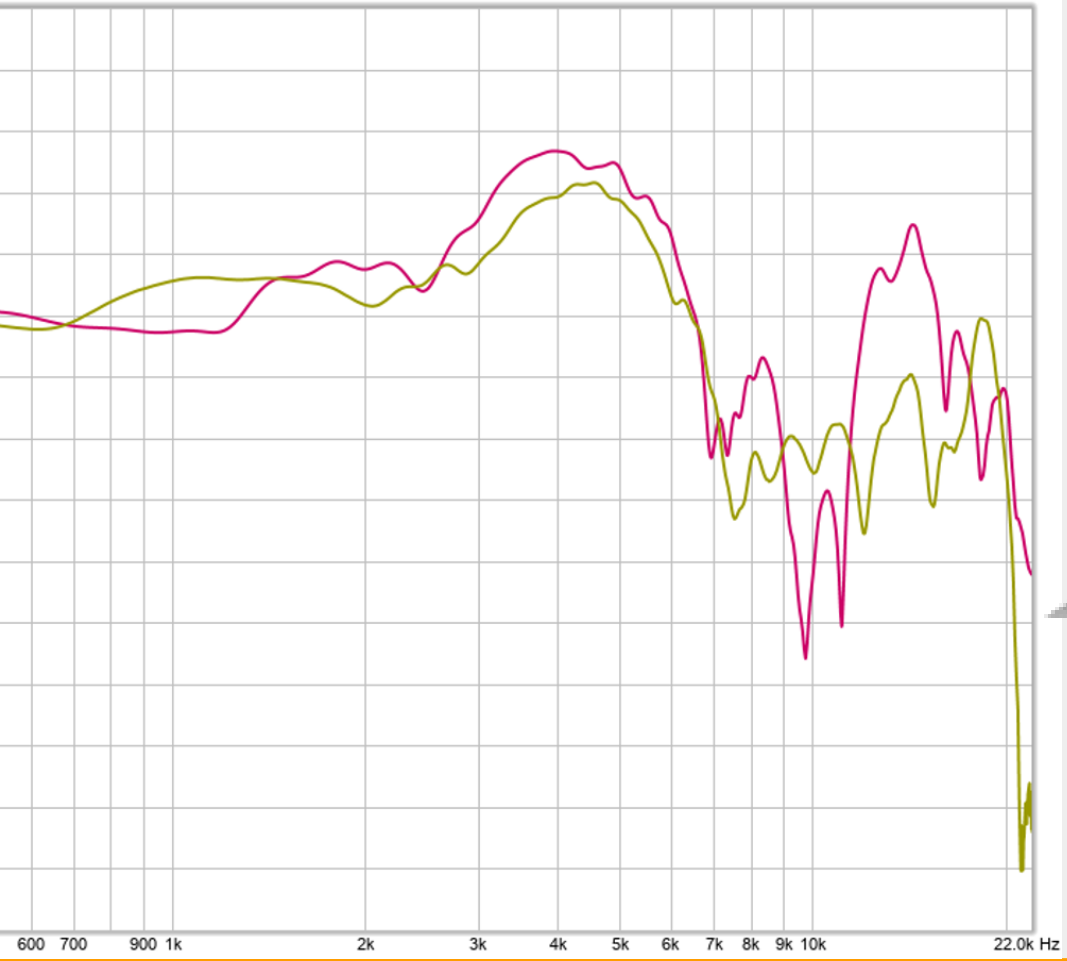
Fundamental techniques: stereophony

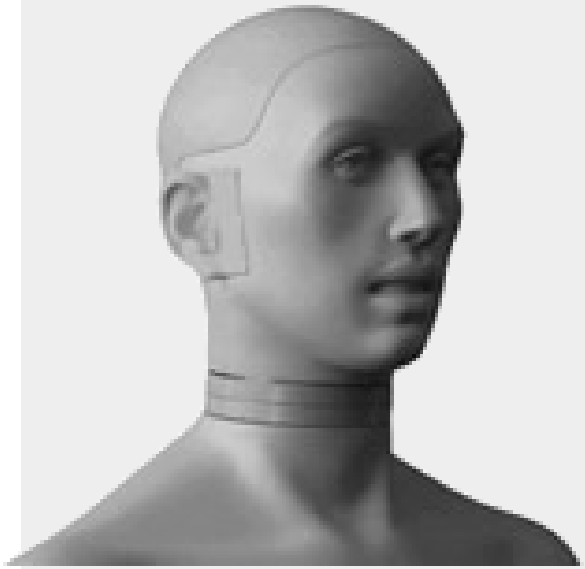




For a vertical phantom source, blue + green should equal red.

Fundamental techniques: stereophony





?

The light brown curve
is what we get instead.

Fundamental techniques: stereophony

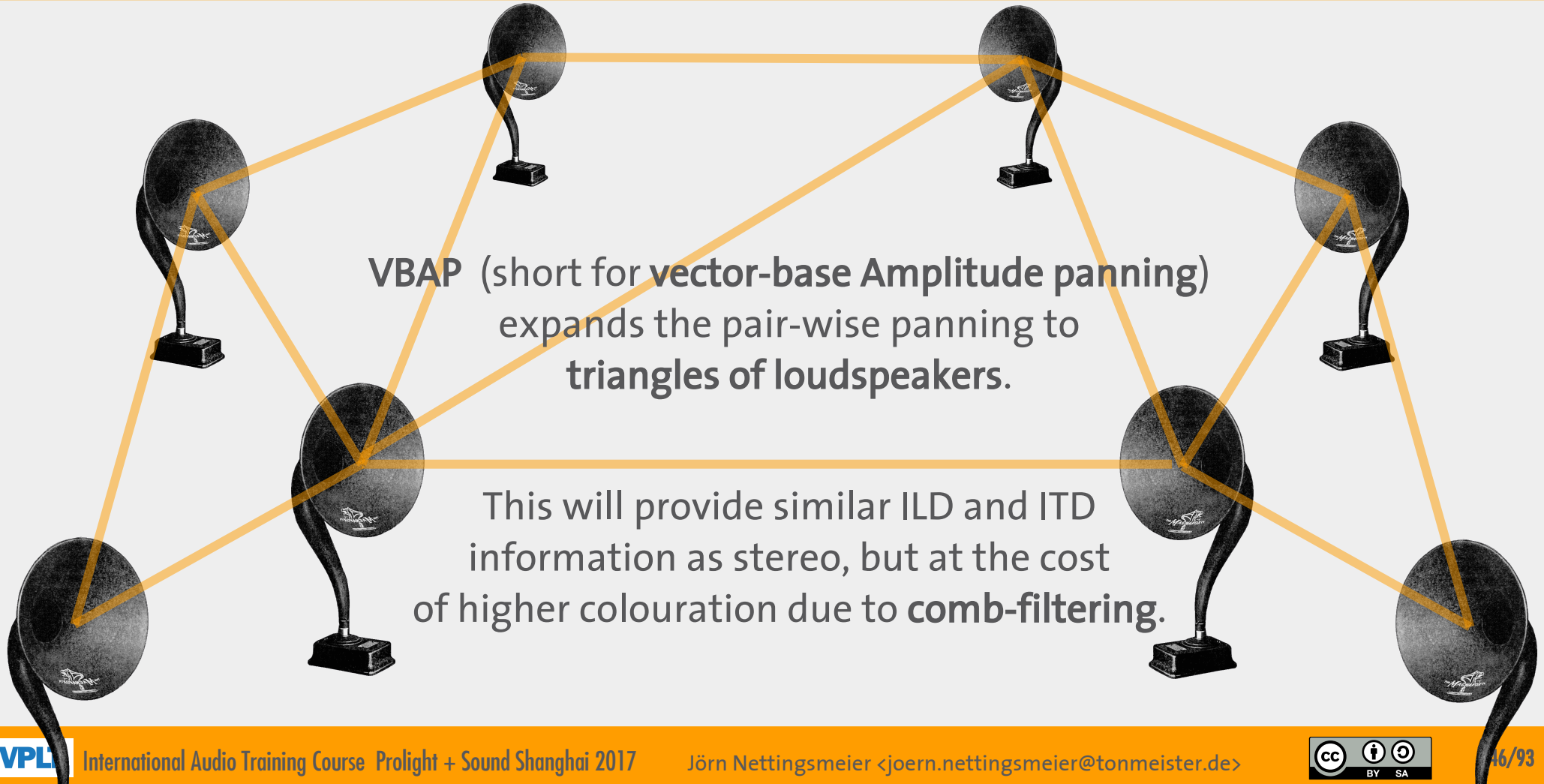
But there are large benefits to with-height stereophony
despite bad vertical localisation:

The useful **listening area becomes larger**,
and **tone colour and spatial impression** (envelopment)
appear more natural and convincing.

To avoid **colouration** due to **comb-filtering**,
the height information must either be sufficiently
decorrelated from the horizontal channels, or co-incident.

Ville Pulkki, 1997

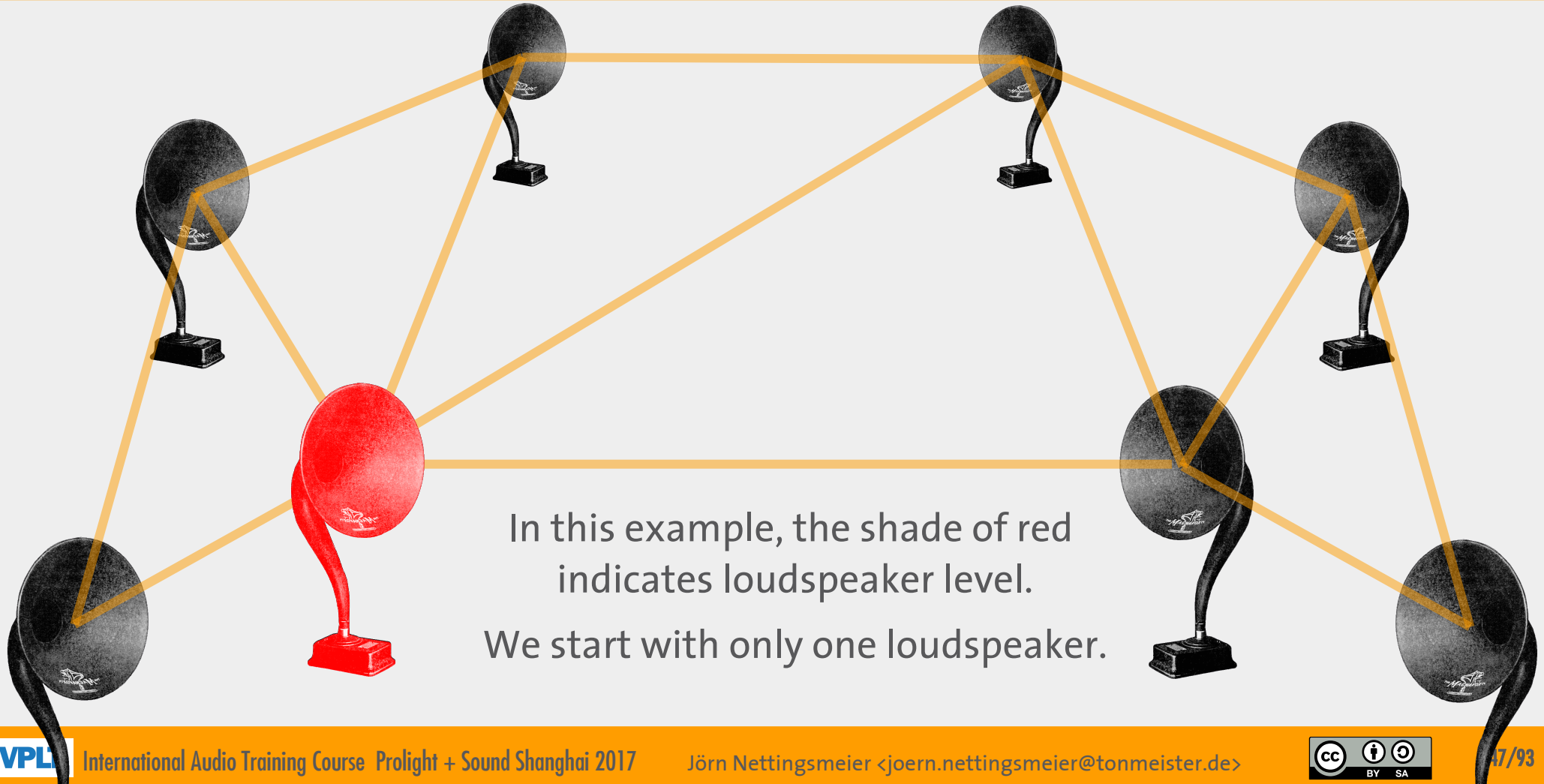
Fundamental techniques: VBAP



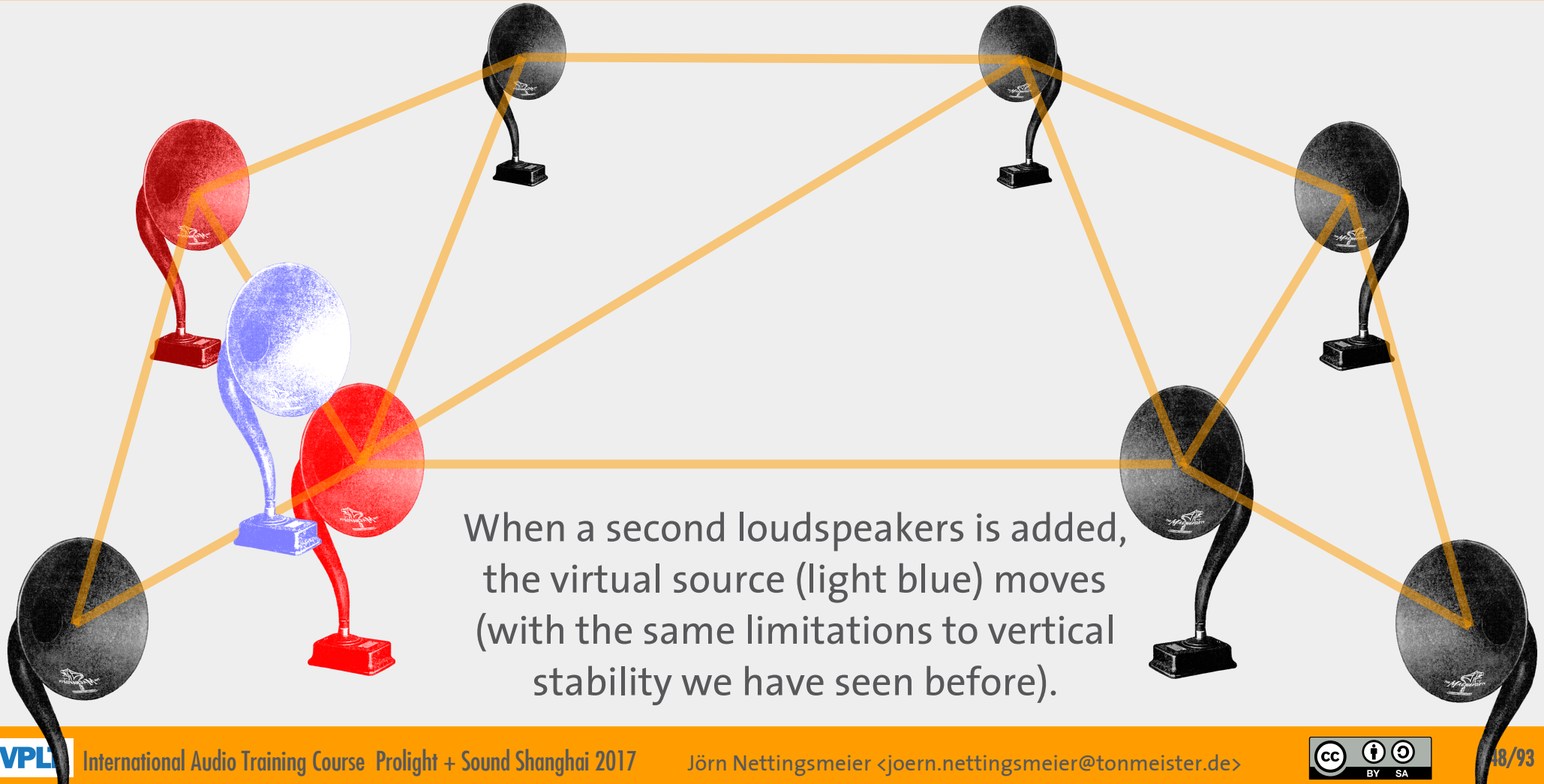
VBAP (short for **vector-base Amplitude panning**) expands the pair-wise panning to **triangles of loudspeakers**.

This will provide similar ILD and ITD information as stereo, but at the cost of higher colouration due to **comb-filtering**.

Fundamental techniques: VBAP

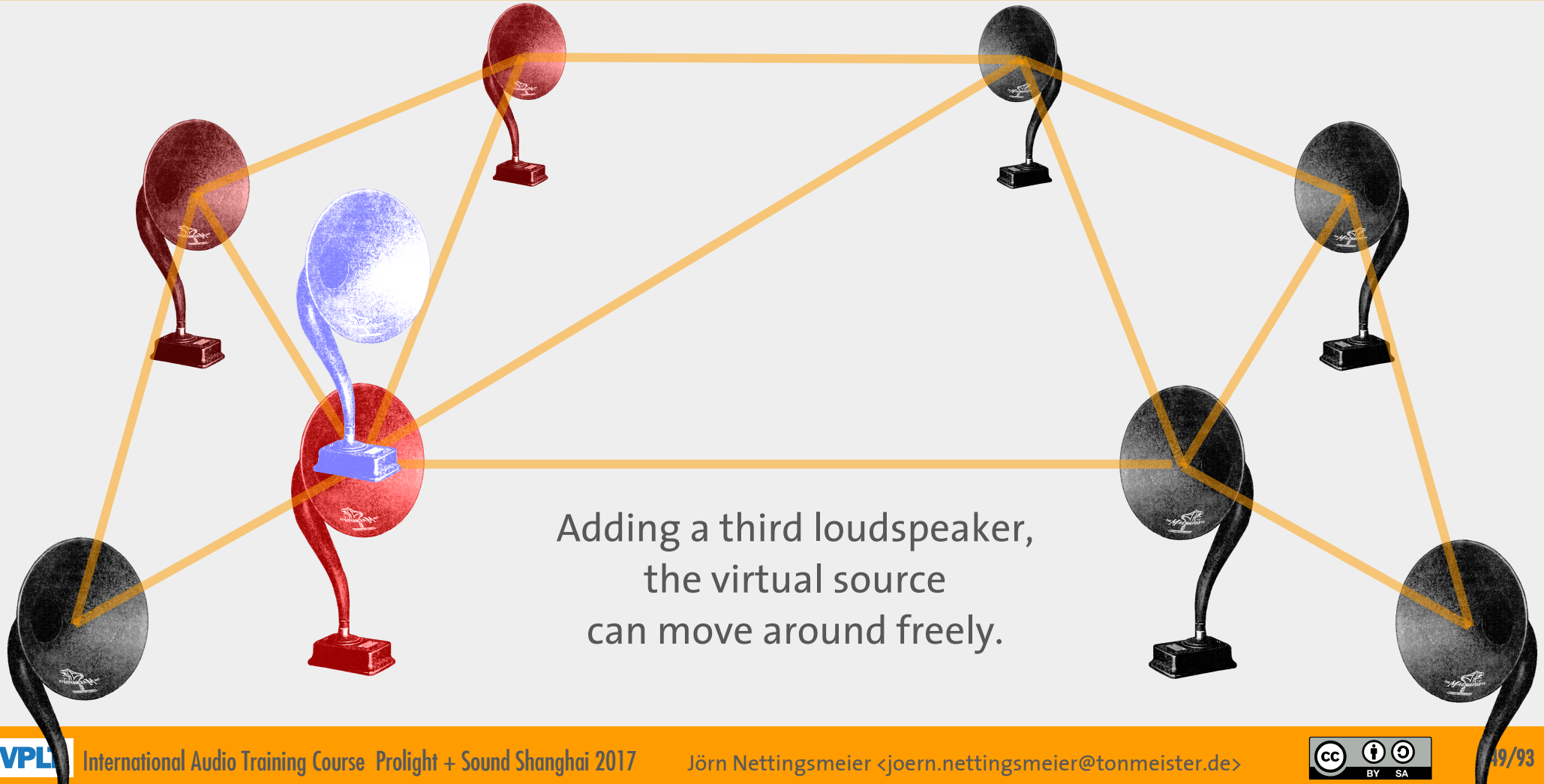


Fundamental techniques: VBAP



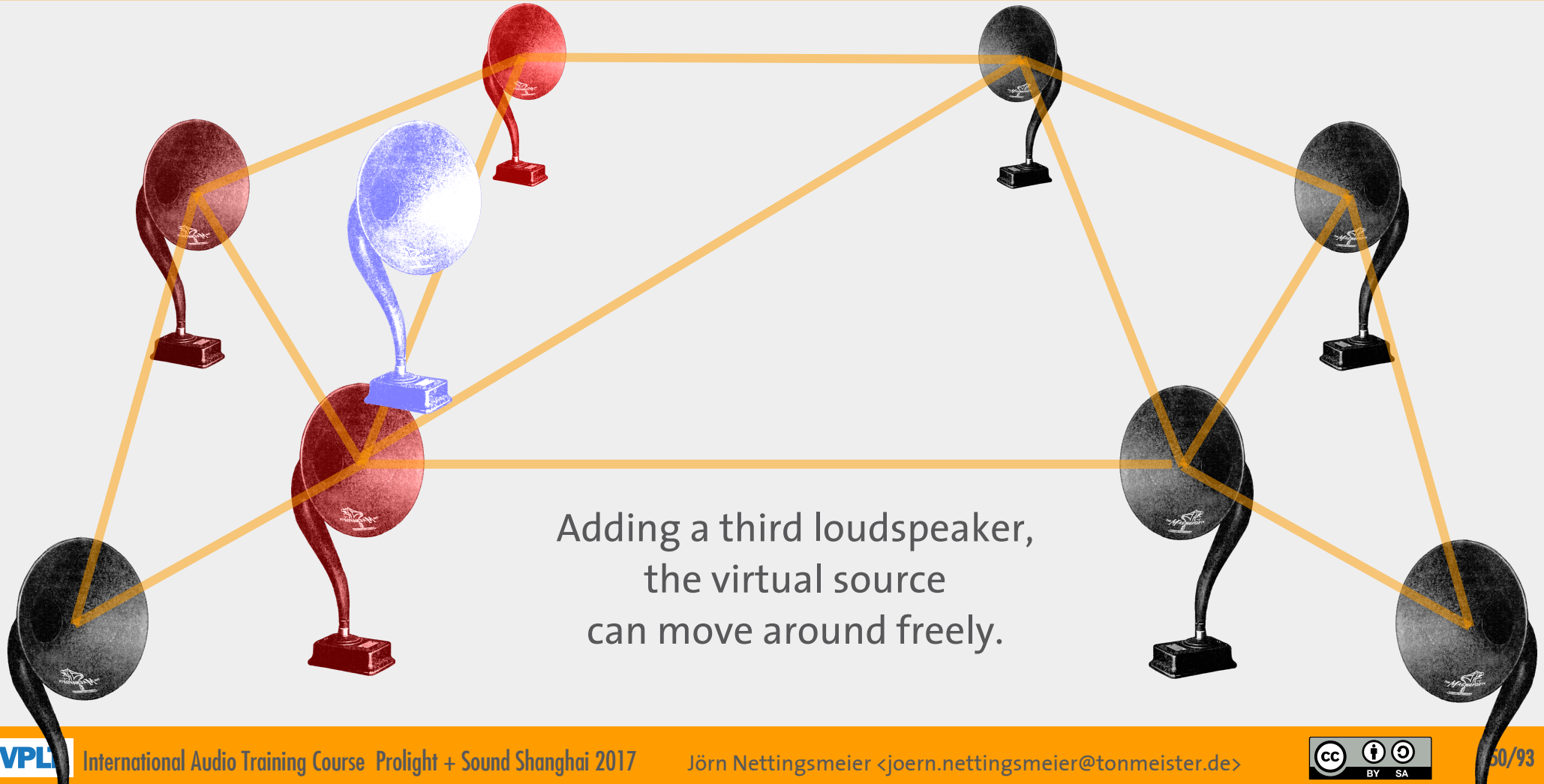
When a second loudspeakers is added, the virtual source (light blue) moves (with the same limitations to vertical stability we have seen before).

Fundamental techniques: VBAP



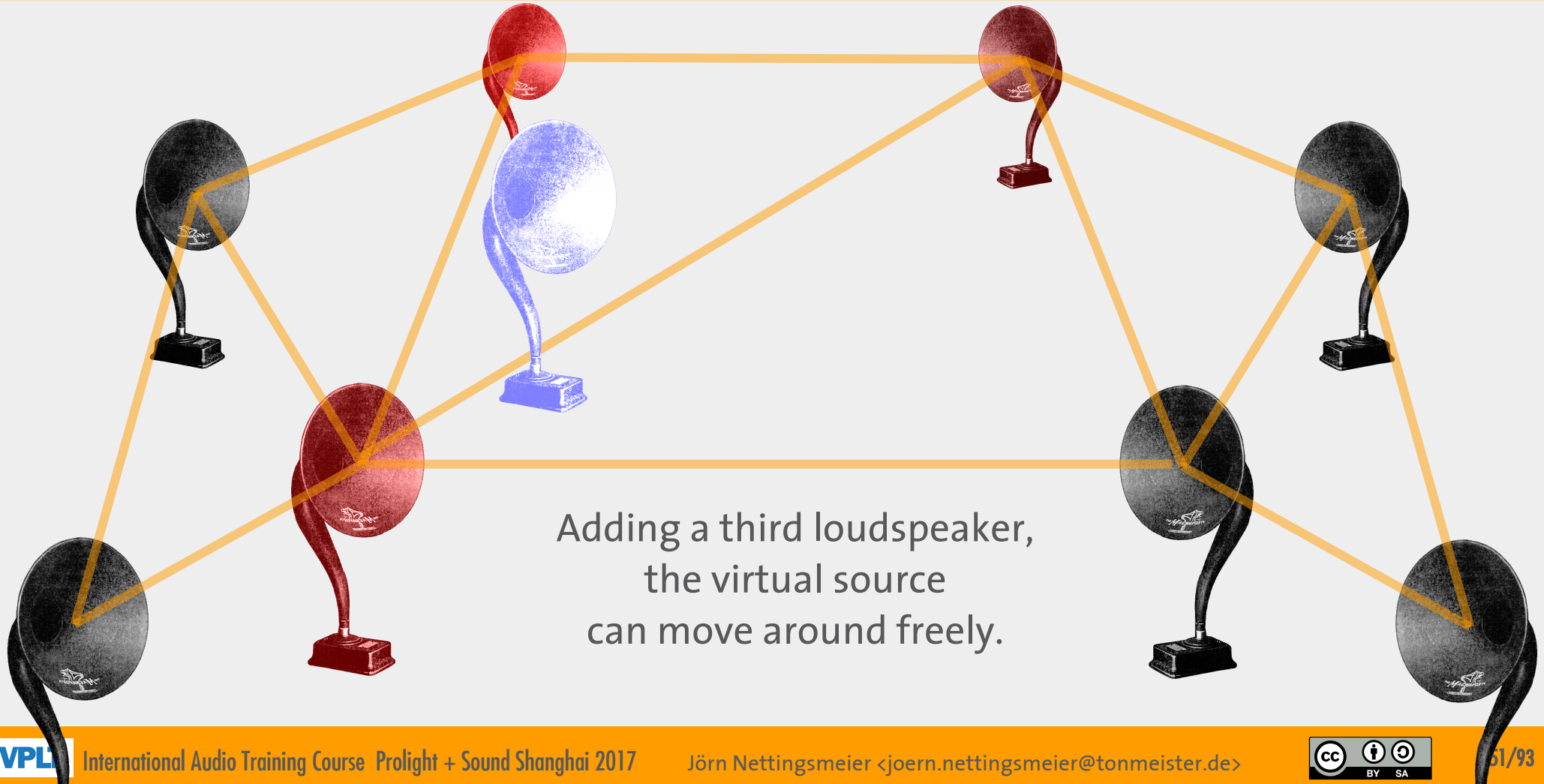
Adding a third loudspeaker,
the virtual source
can move around freely.

Fundamental techniques: VBAP



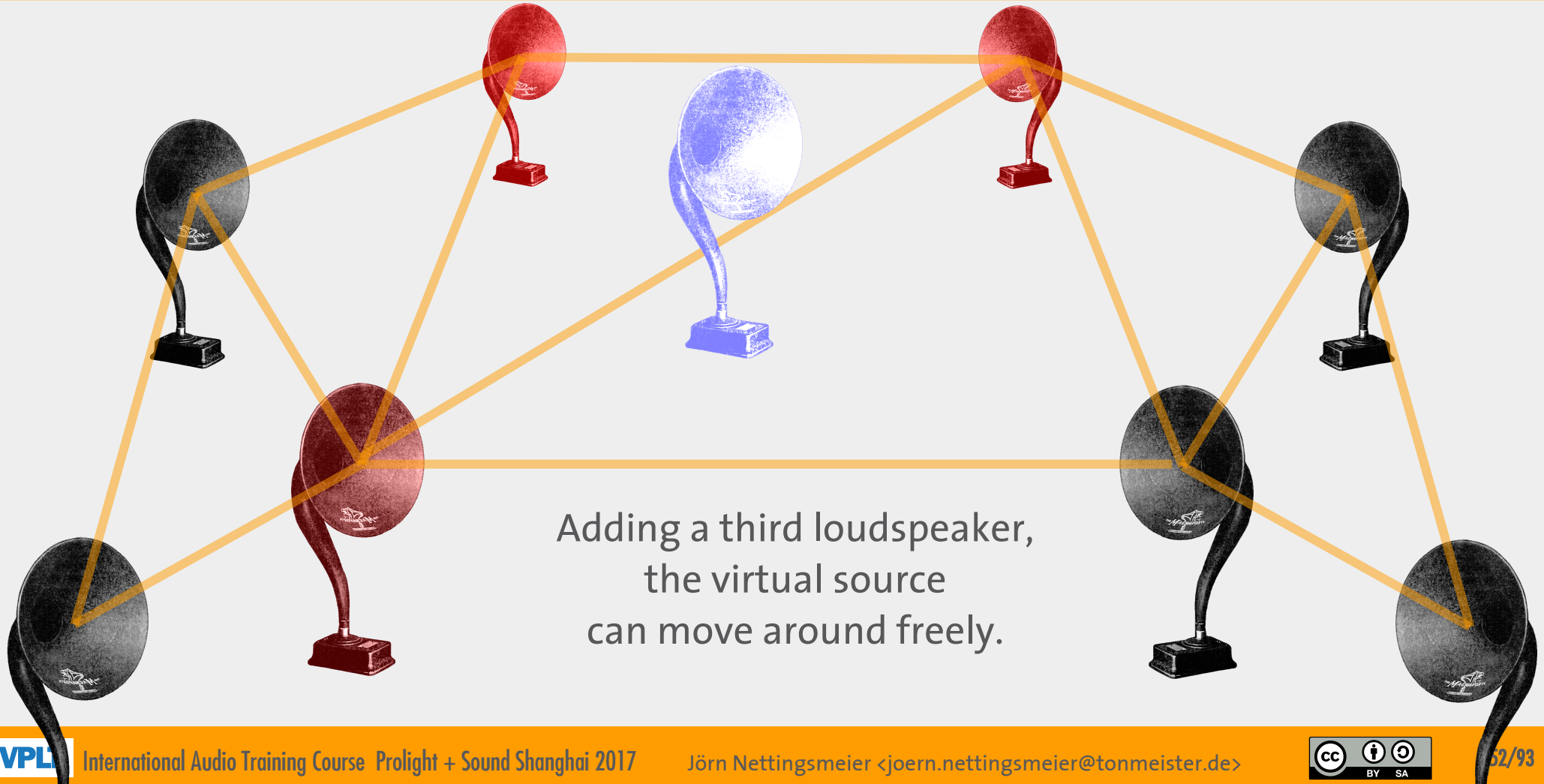
Adding a third loudspeaker,
the virtual source
can move around freely.

Fundamental techniques: VBAP



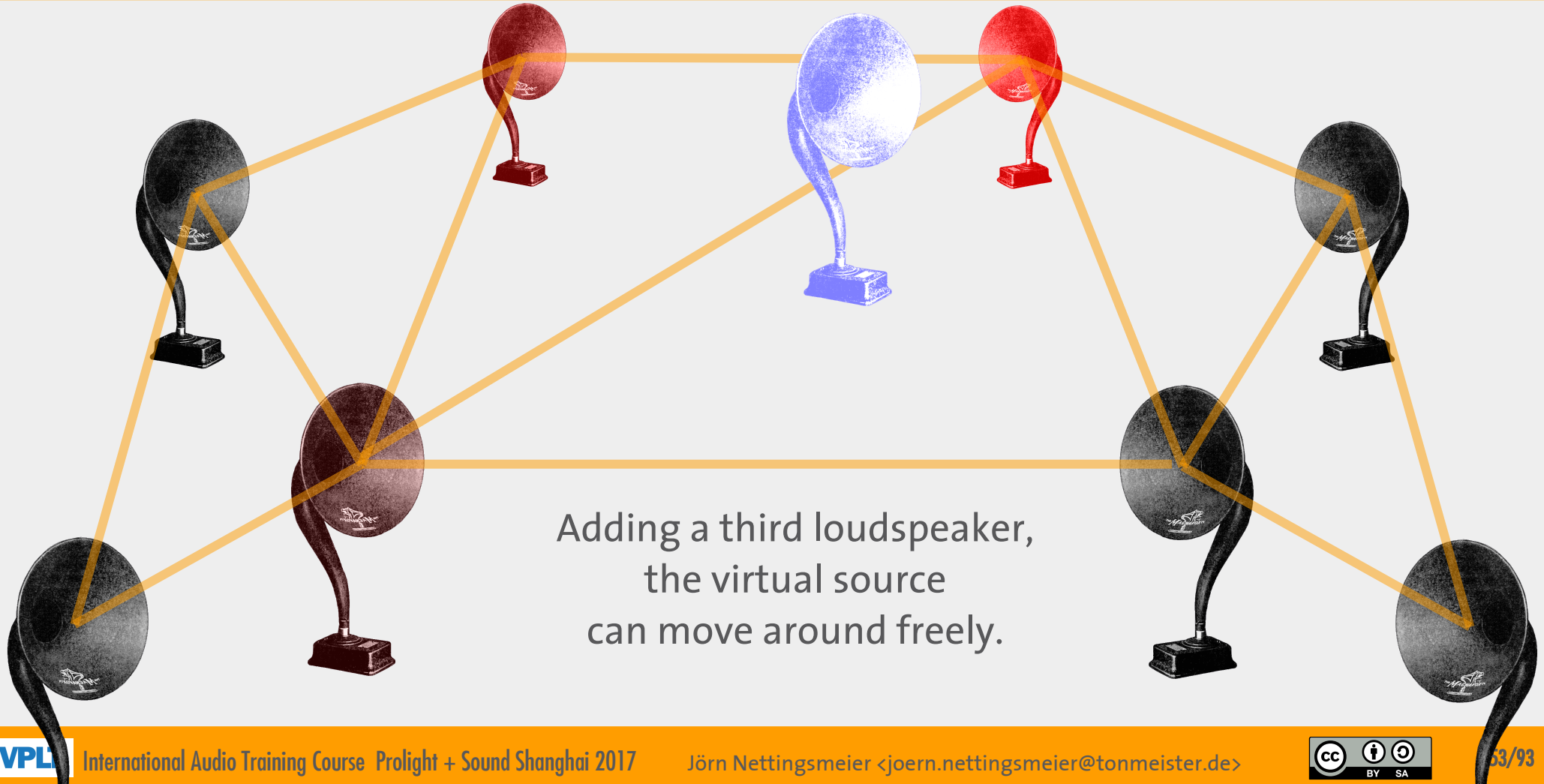
Adding a third loudspeaker,
the virtual source
can move around freely.

Fundamental techniques: VBAP



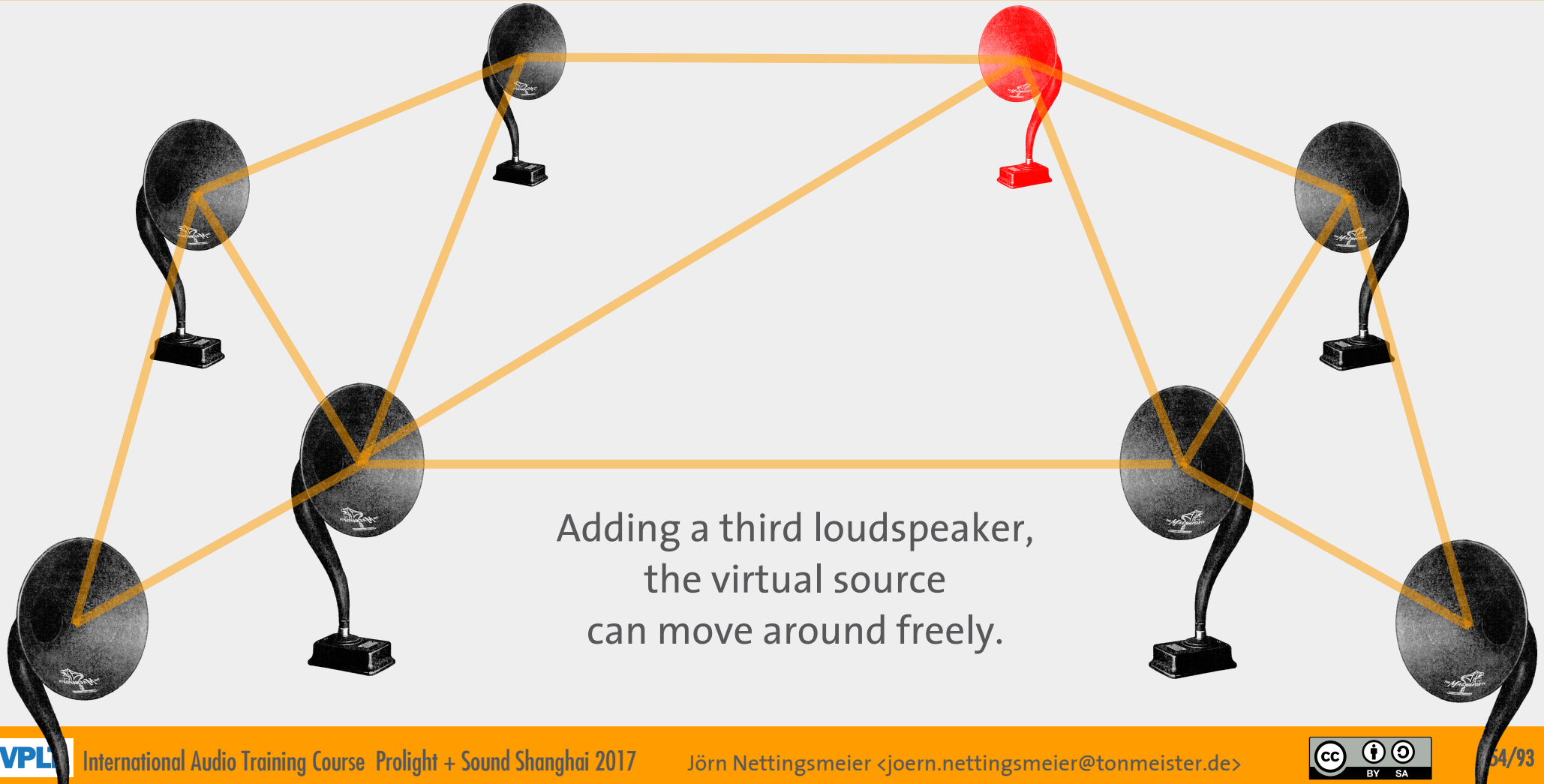
Adding a third loudspeaker,
the virtual source
can move around freely.

Fundamental techniques: VBAP



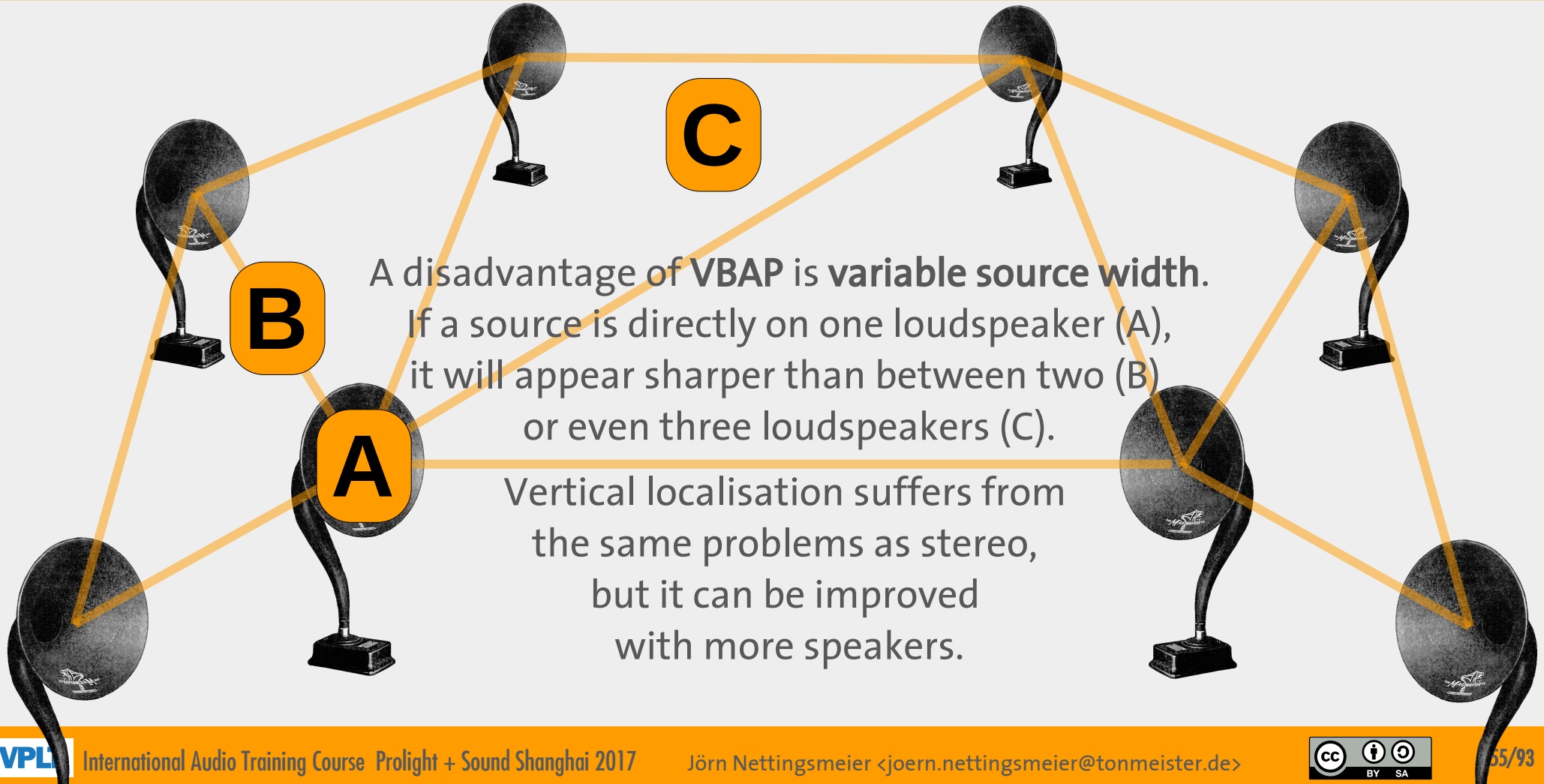
Adding a third loudspeaker,
the virtual source
can move around freely.

Fundamental techniques: VBAP



Adding a third loudspeaker,
the virtual source
can move around freely.

Fundamental techniques: VBAP



A disadvantage of **VBAP** is **variable source width**.
If a source is directly on one loudspeaker (A),
it will appear sharper than between two (B)
or even three loudspeakers (C).

Vertical localisation suffers from
the same problems as stereo,
but it can be improved
with more speakers.

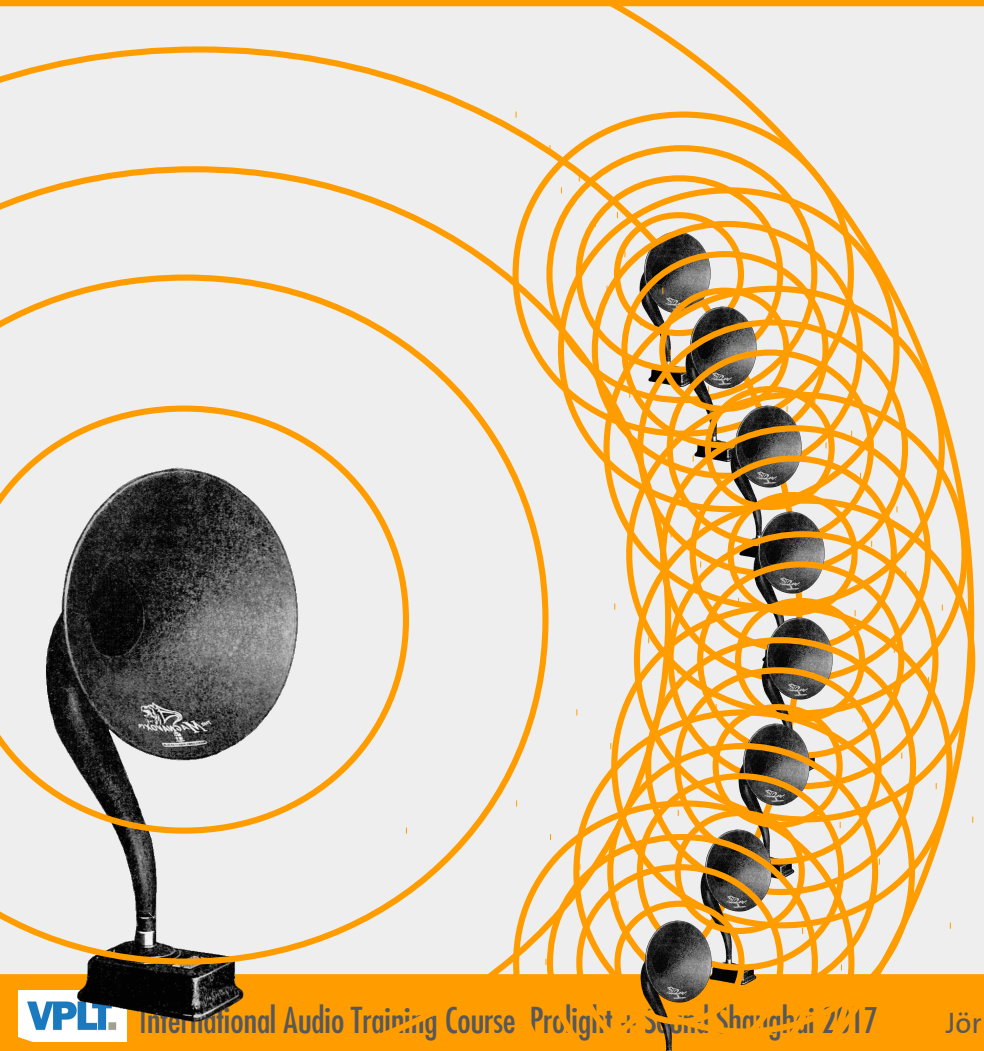
Berkhout and de Vries, 1989-1993

Fundamental techniques: wave field synthesis

Huygens' principle:

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

Fundamental techniques: wave field synthesis

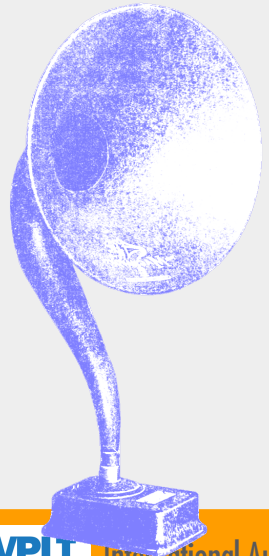


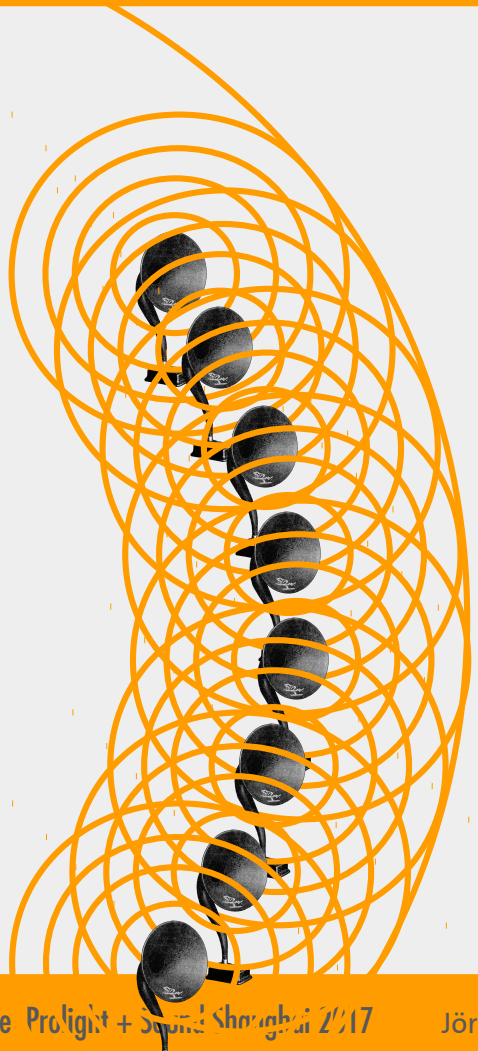
Huygens' principle:

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 line of loudspeakers working together.

Fundamental techniques: wave field synthesis





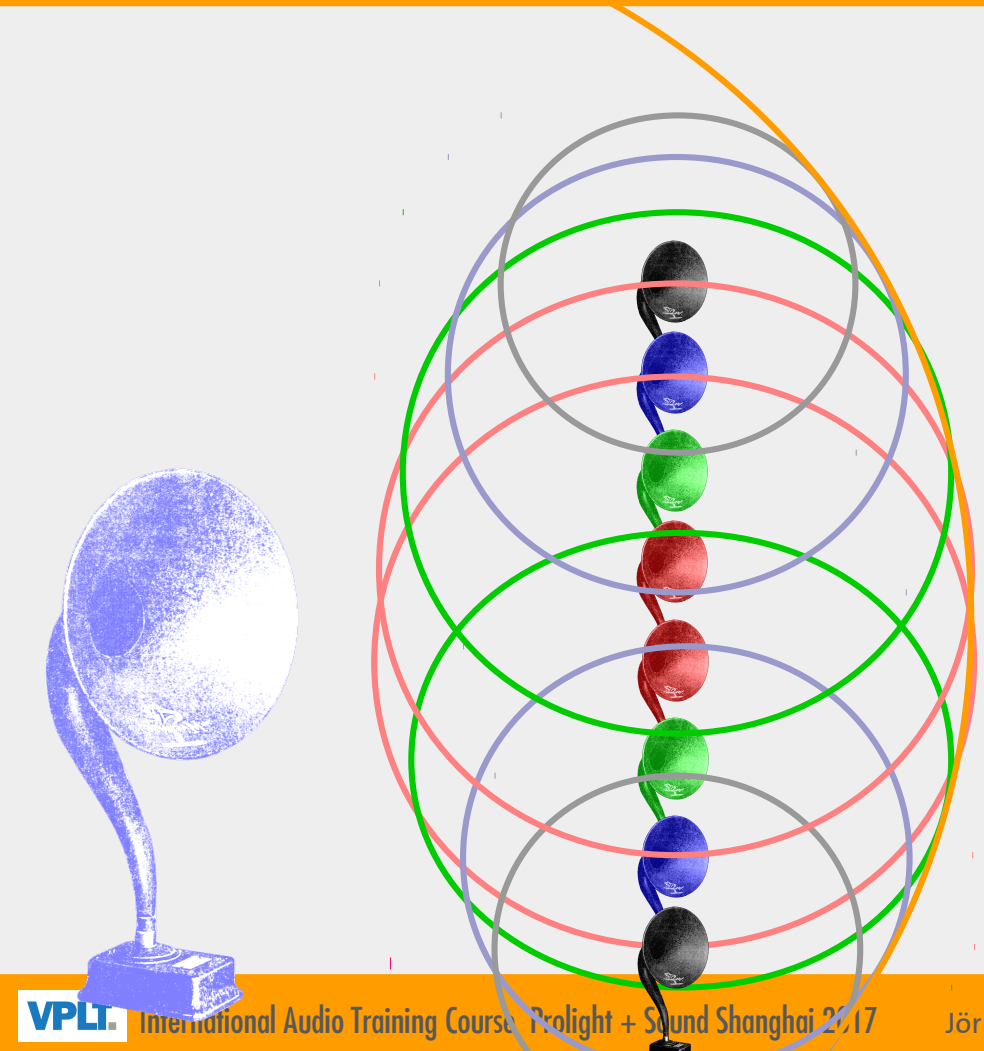
Huygens' principle:

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

Such a line 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.

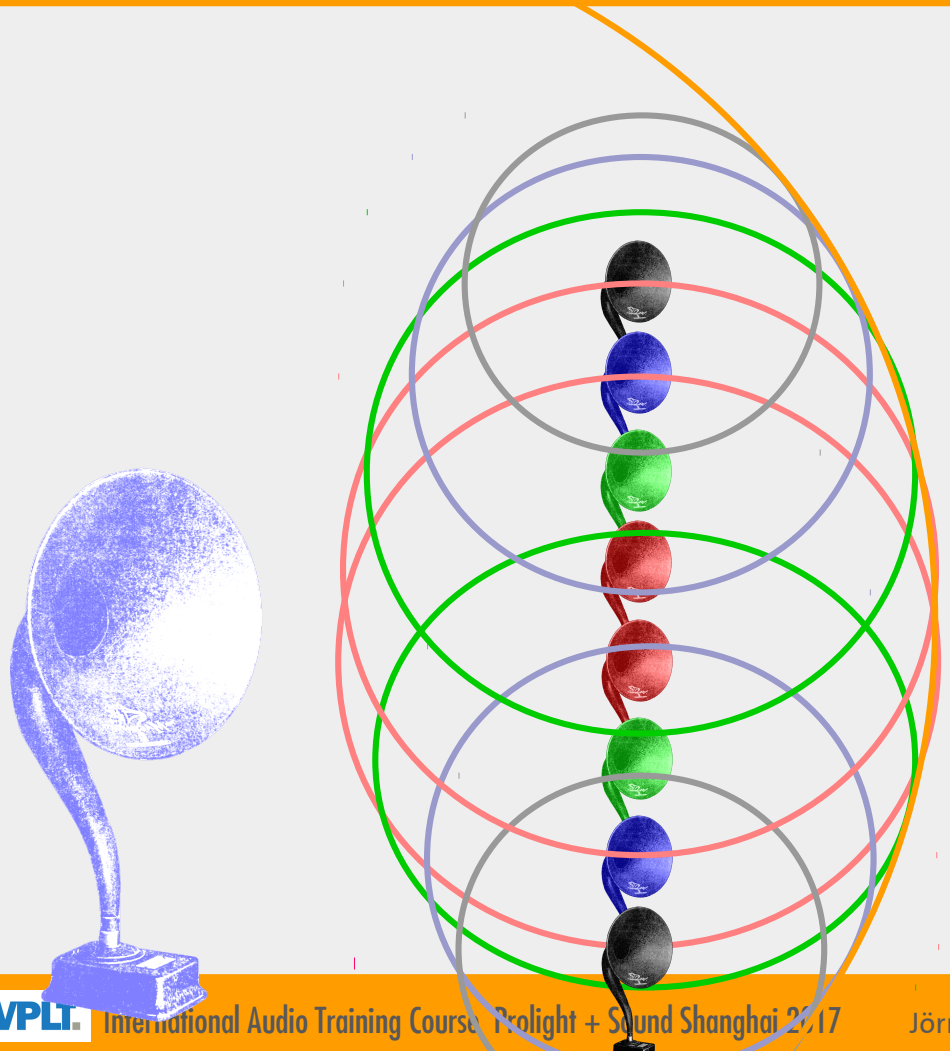
Fundamental techniques: wave field synthesis



The speakers do not have to follow the desired wave front.

Arbitrary shapes can be synthesized by using the appropriate delays for each speaker.

Fundamental techniques: wave field synthesis



The speakers do not have to follow the desired wave front.

Arbitrary shapes can be synthesized by using the appropriate delays for each speaker.

Because the soundfield is **physically correct,** WFS gives very robust localisation.

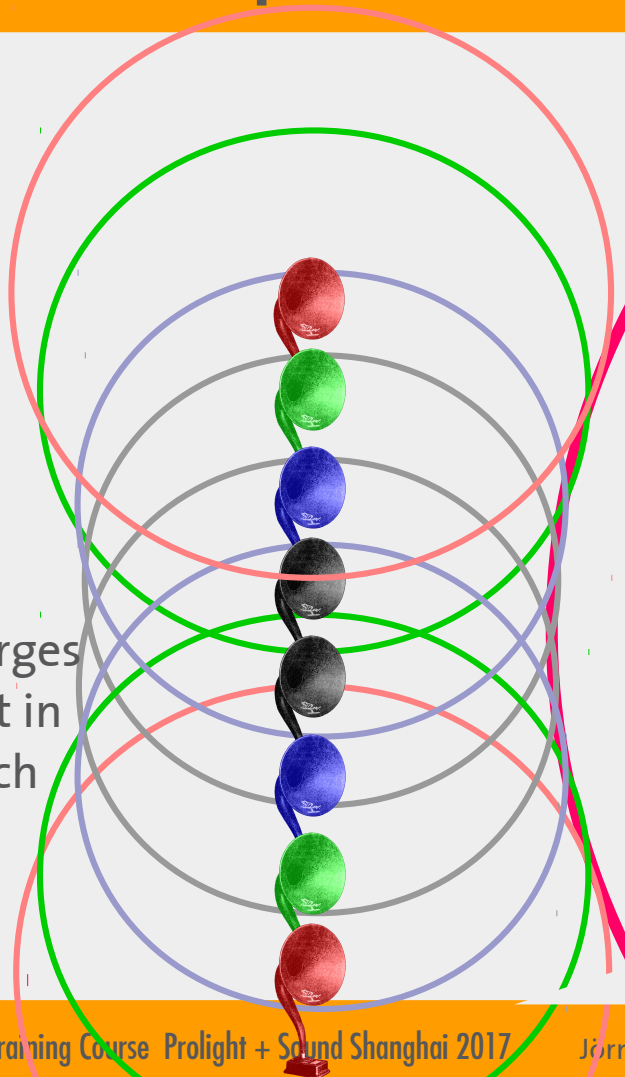
Since the speakers are coupling perfectly, there is **no comb filtering** and no colouration below the alias frequency. Above that frequency, special care must be taken to avoid artefacts.

Fundamental techniques: wave field synthesis

A neat trick:

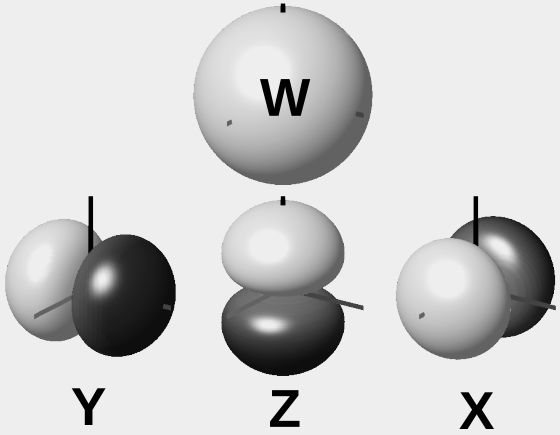
Invert the delays to produce a **focused source** inside the listening area.

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

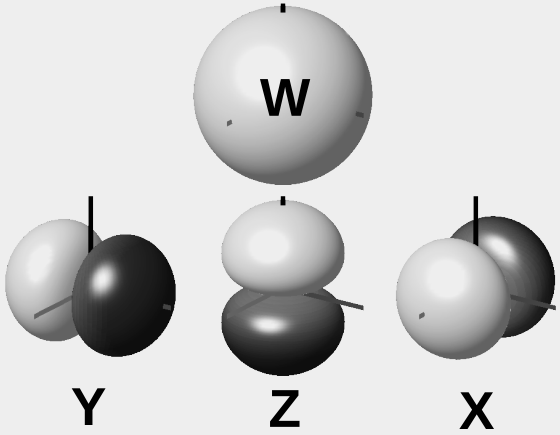


A favourite gimmick of theme parks: you can create virtual sources very close to the listener, i.e. “whisper in their ears”!

Fundamental techniques: Ambisonics



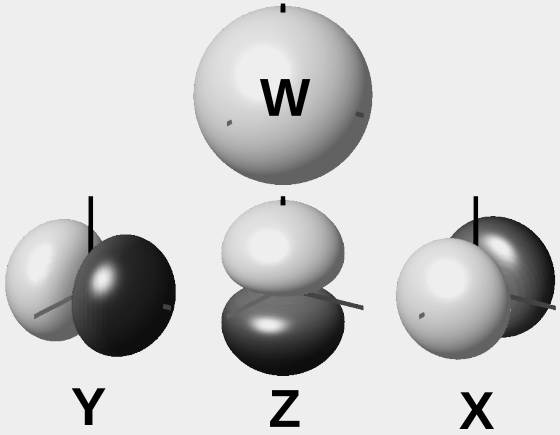
Michael Gerzon et al., 1975



Ambisonics splits the sound field into directional components.
In the basic first-order case, we can do that with microphones:

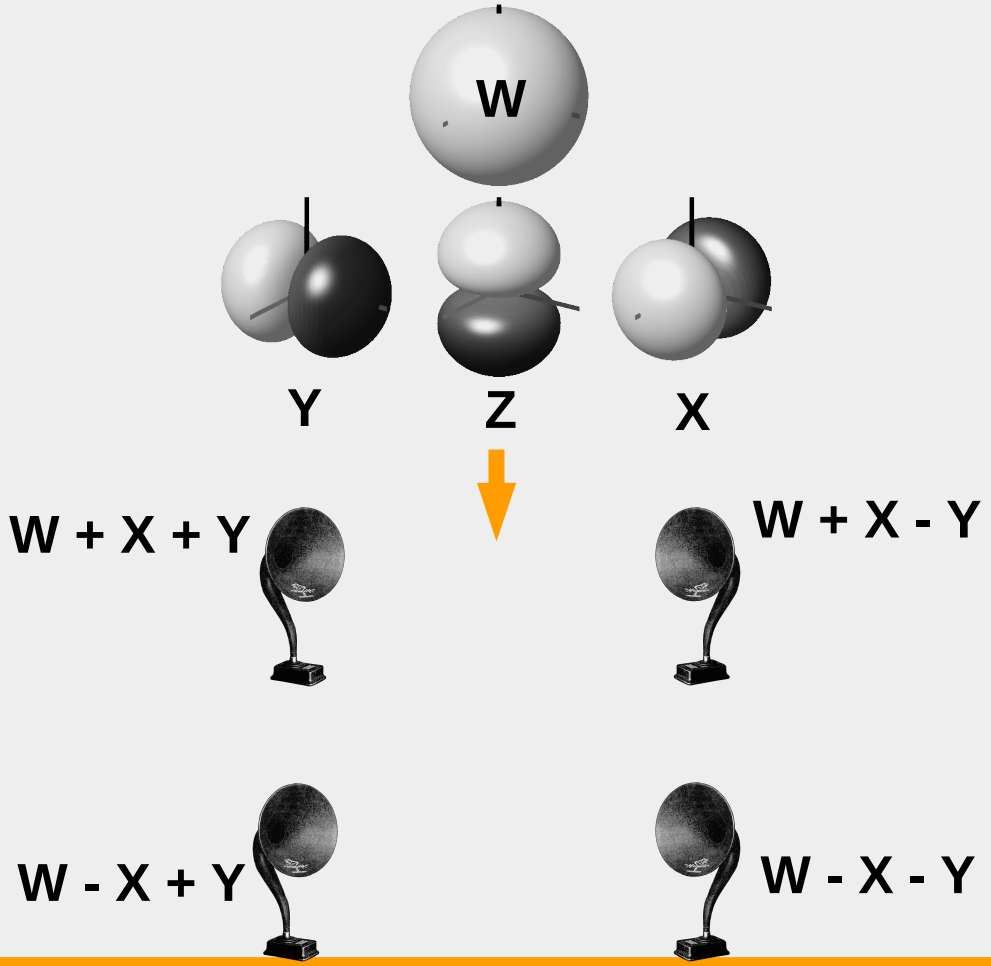
- an omni-directional mic captures the sound pressure **W**;
- three fig8 mics capture the sound pressure gradient along the **X**, **Y** and **Z** **axes.**

Fundamental techniques: Ambisonics



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

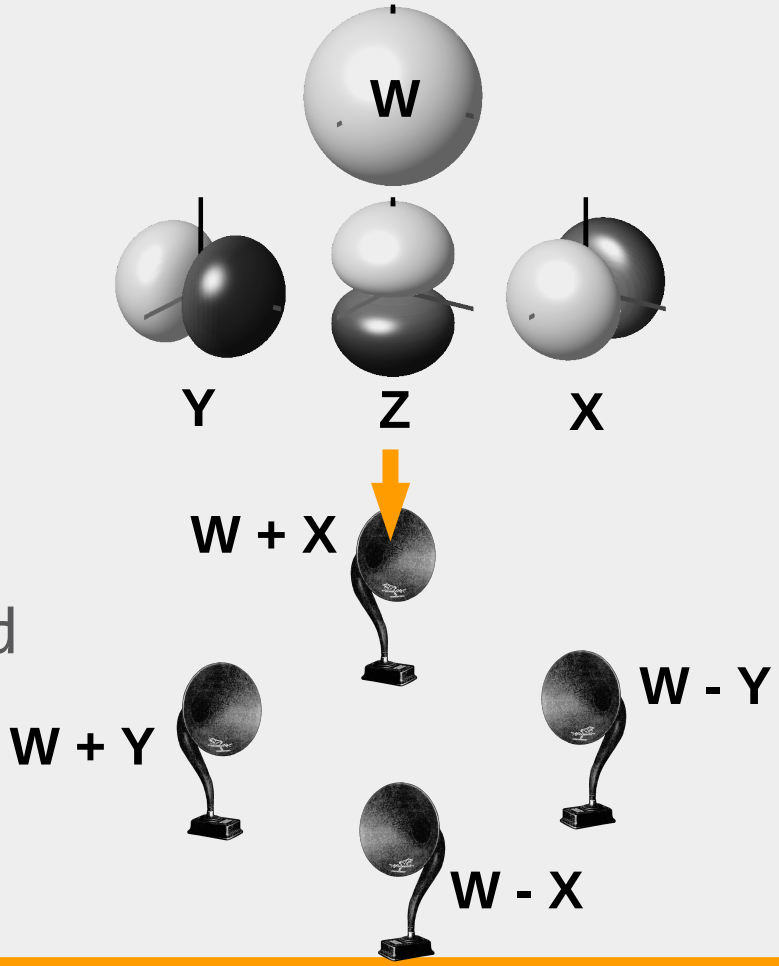
Here is a simplified **decoding matrix** for four speakers in a square:



\+ means mix

− means mix with inverted polarity

Fundamental techniques: Ambisonics



The content encoding does not change at all.

For a diamond-shaped layout, the matrix is even simpler:

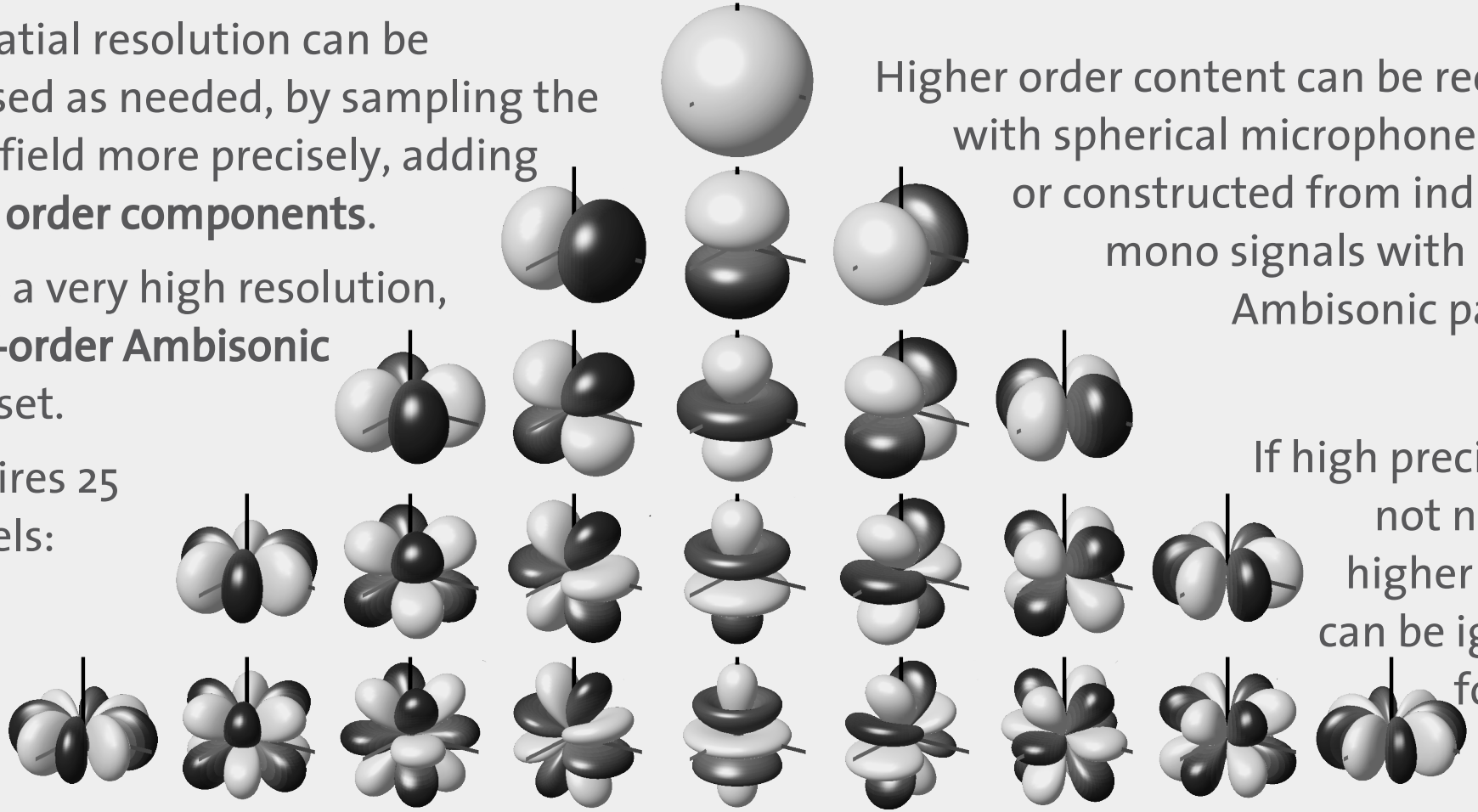
The same principle makes Ambisonics very easy to rotate.

Fundamental techniques: Ambisonics

The spatial resolution can be increased as needed, by sampling the sound field more precisely, adding **higher order components**.

Here is a very high resolution, **fourth-order Ambisonic** signal set.

It requires 25 channels:

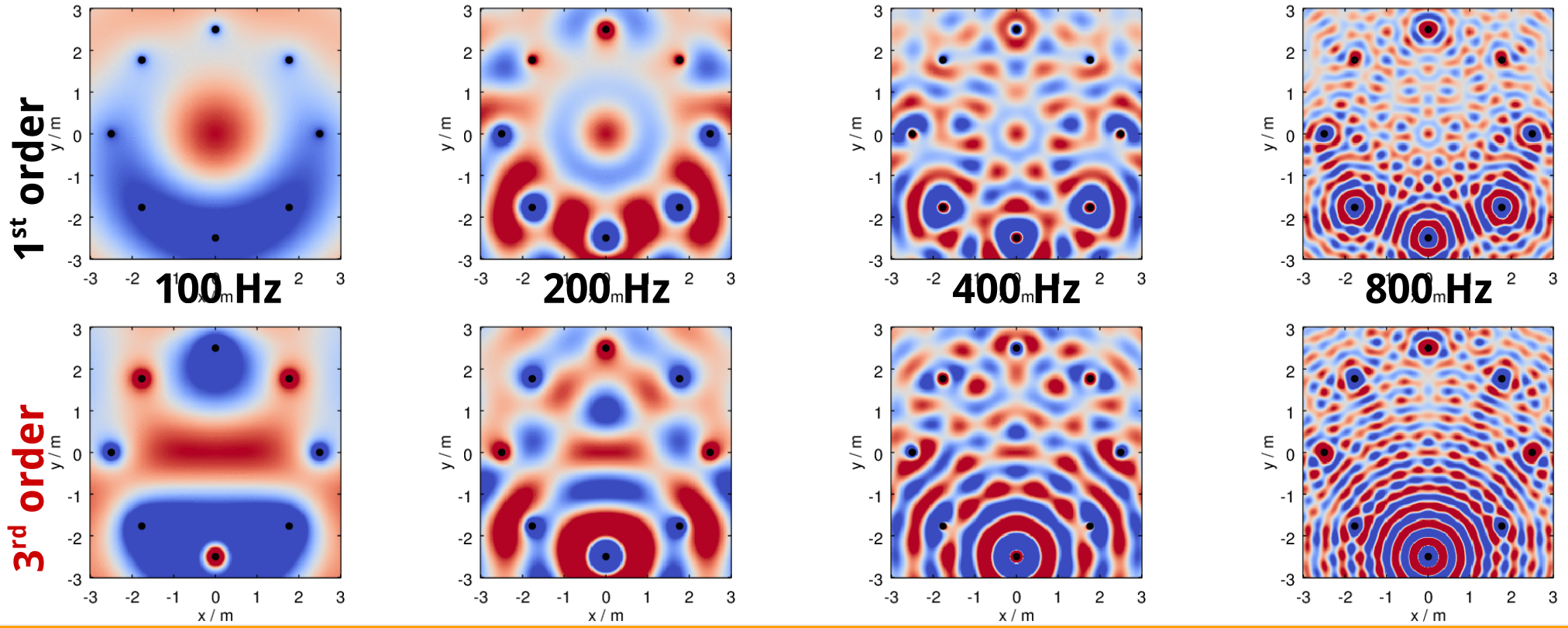


Higher order content can be recorded with spherical microphone arrays, or constructed from individual mono signals with simple Ambisonic panners.

If high precision is not needed, higher orders can be ignored for play-back.

Fundamental techniques: Ambisonics

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



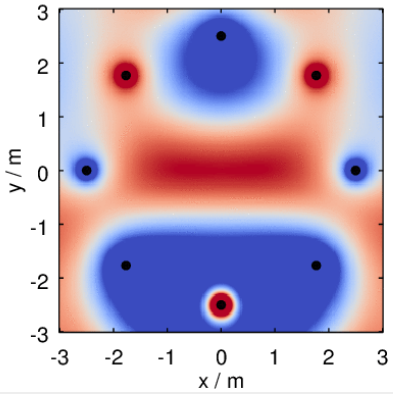
Fundamental techniques: Ambisonics

Ambisonics reproduces a **physically correct** sound field at low frequencies, where the human hearing needs it to obtain **correct ITD** information.

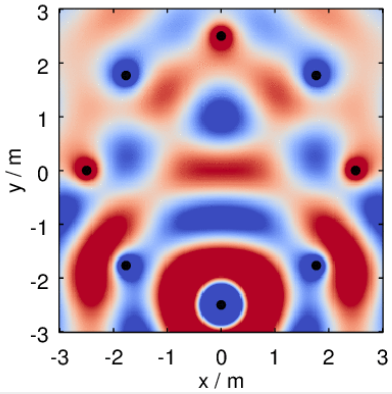
In the higher frequencies, the phase information is essentially randomized, but the level distribution still gives **correct ILD** information.

Ambisonics is therefore neatly optimized to the way we localize.

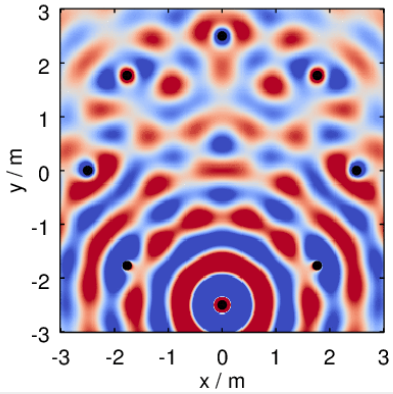
100 Hz



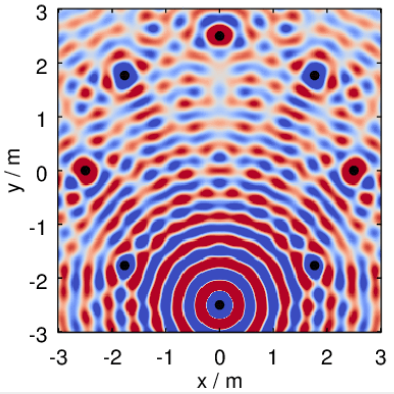
200 Hz



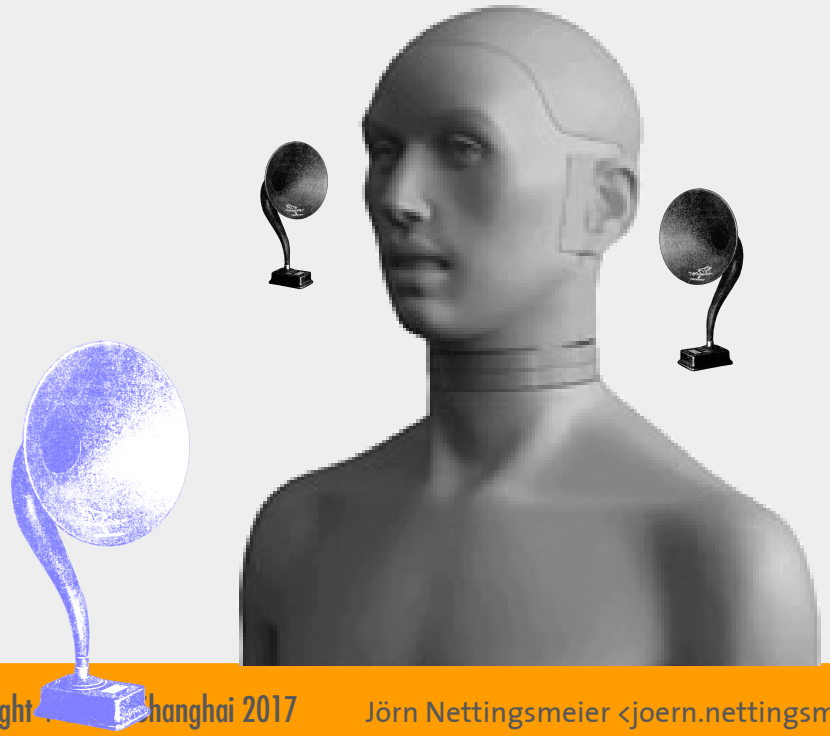
400 Hz



800 Hz



Harvey Fletcher et al., 1927



Fundamental techniques: Binaural synthesis

At the eardrums, the entire spatial information of a sound field is reduced to two signals. A **dummy head** with artificial outer ears can record this information. We can then play it back using **headphones**.

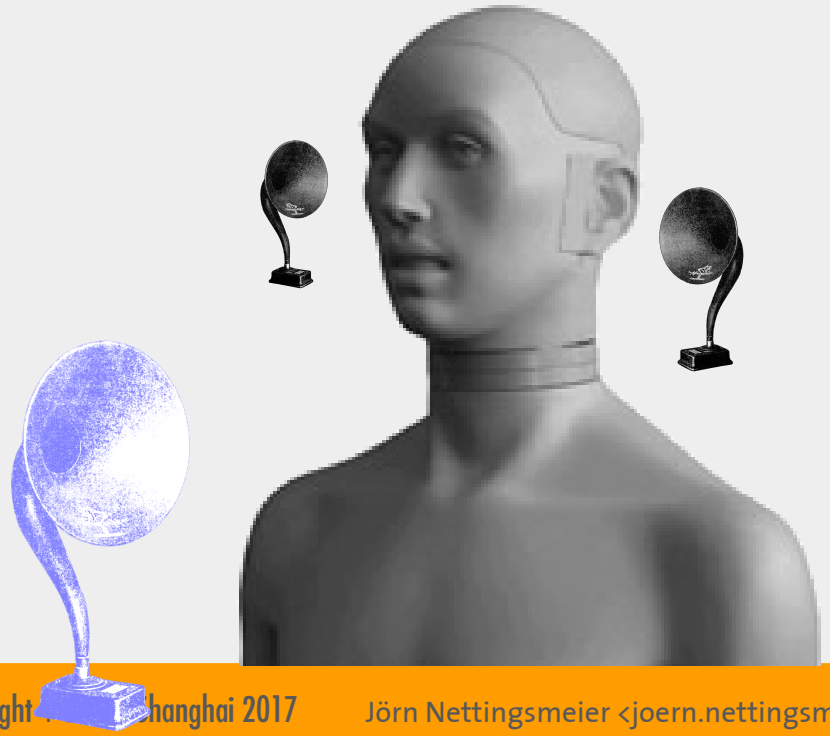
This approach produces a **head-related** recording. In theory, it can deliver perfect 3D from only two channels.

Alternatively, a mono signal 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 for a real sound source.

Fundamental techniques: Binaural synthesis

The illusion of binaural listening can be very good, once our brain begins to believe in it: we perceive sounds outside the head, at the correct distance, like real.

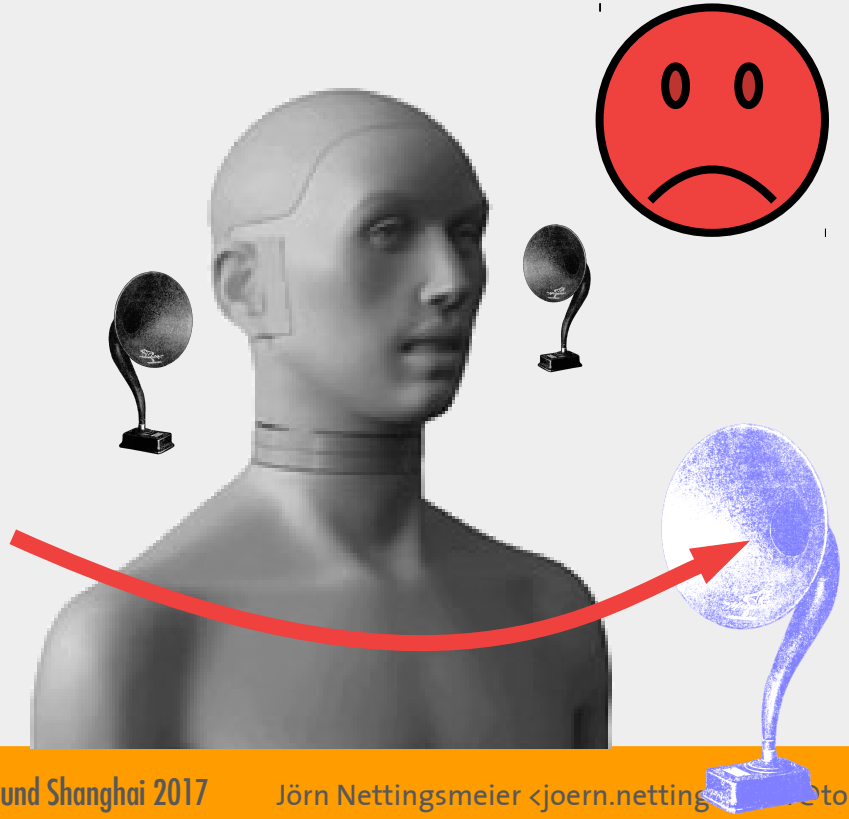


Fundamental techniques: Binaural synthesis

The illusion of binaural listening can be very good, once our brain begins to believe in it: we perceive sounds outside the head, at the correct distance, like real.

However, once we turn our head, the sources move, too.

At that moment, the illusion breaks down and we hear the sounds in our head again.

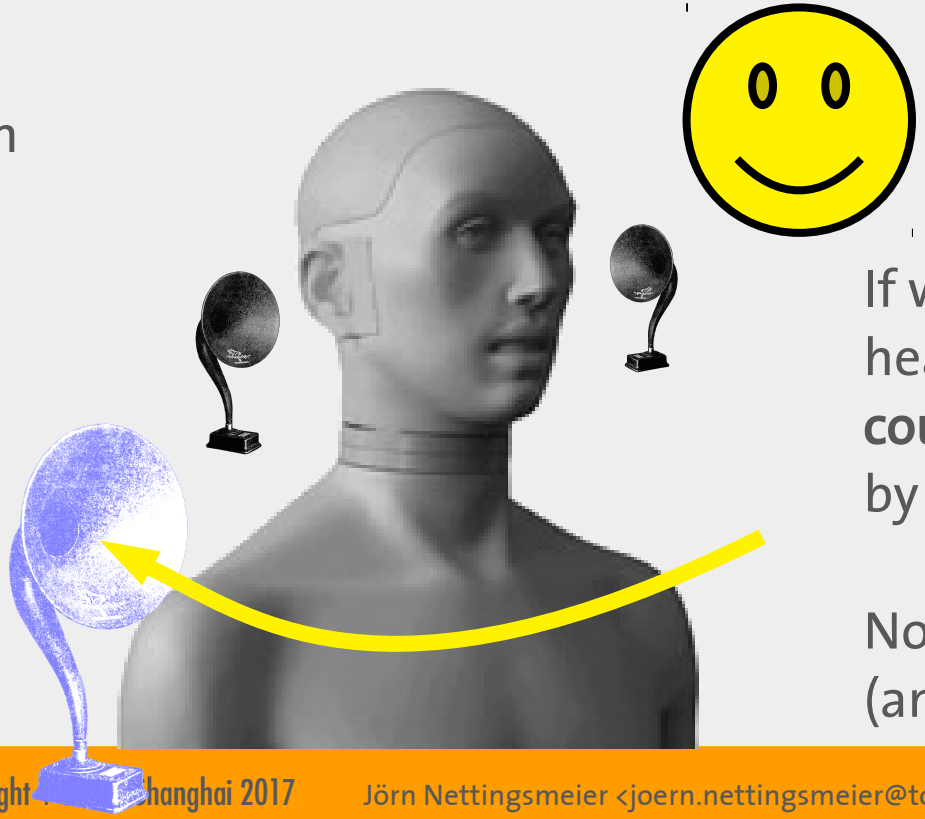


Fundamental techniques: Binaural synthesis

The illusion of binaural listening can be very good, once our brain begins to believe in it: we perceive sounds outside the head, at the correct distance, like real.

However, once we turn our head, the sources move, too.

At that moment, the illusion breaks down and we hear the sounds in our head again.



If we **track** the listener's head movements, we can **counter-rotate** the sound by applying new HRTFs.

Now the source location (and the illusion) is stable.

Fundamental techniques: Summary

technique	advantages	disadvantages	minimal 2D setup	minimal 3D setup
Stereo (incl. 5.1, 9.1, 22.2...)	• low hardware and cpu requirements • very natural tone colour reproduction • world-wide standard • one channel per speaker, easy to understand	• unstable localization on the sides and vertically • tone colour and apparent width of moving sources changes • very small listening area, bad for concert situations • not scalable - larger listening areas require remixing or an intrusive (and error-prone) upmixing algorithm • highly optimized to frontal direction	5 channels, 5 speakers (ITU 775)	9 channels, 9 speakers (ITU 2051)

Fundamental techniques: Summary

technique	advantages	disadvantages	minimal 2D setup	minimal 3D setup
VBAP	• low cpu requirements • easy to understand and set up • works reasonably well even on irregular speaker layouts • can be used for full 3D or sphere subsets • object-based, content can be played on arbitrary speaker layouts • works equally well in all directions	• variable colouration and source width for moving sources • needs a rendering engine for playback	3 speakers	~ 20 speakers, less if only upper hemisphere is used.

Fundamental techniques: Summary

technique	advantages	disadvantages	minimal 2D setup	minimal 3D setup
WFS	• close to perfect physical reproduction up to aliasing freq. • very convincing • arbitrarily large listening area • focused sources offer new creative possibilities • object-based, content can be played on arbitrary speaker layouts	• immense hardware and computational requirements • high investment threshold • needs careful tuning to avoid colouration above the aliasing frequency • needs a very powerful parallel rendering engine for playback	>16 for basic frontal beam-steering. >100 for basic horizontal surround	not feasible (other methods used for height)

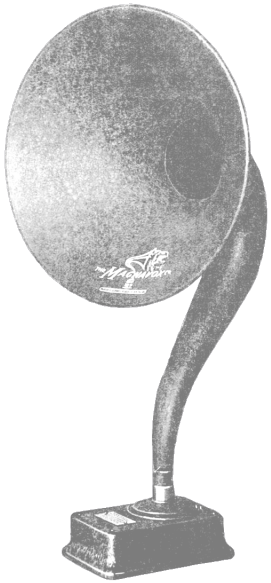
Fundamental techniques: Summary

technique	advantages	disadvantages	minimal 2D setup	minimal 3D setup
Higher-order Ambi-sonics	• good approximation of real sound field • no colour change when moving, good vertical localisation • efficient, scalable, low investment threshold • easy to rotate: good basis for headtracked binaural and VR • content independent of speaker layout • works equally well in all directions	• small sweet spot at low orders • low-order systems need careful tuning to avoid colouration and phasing effects • conceptually hard to understand • needs decoder for playback	3 channels, 6 speakers for home listening. 7 channels, 8 speakers for larger audiences	4 channels, 8 speakers for simple home listening. 16 ch. to 20 speakers for larger audiences, less if only upper half-sphere is used.

Fundamental techniques: Summary

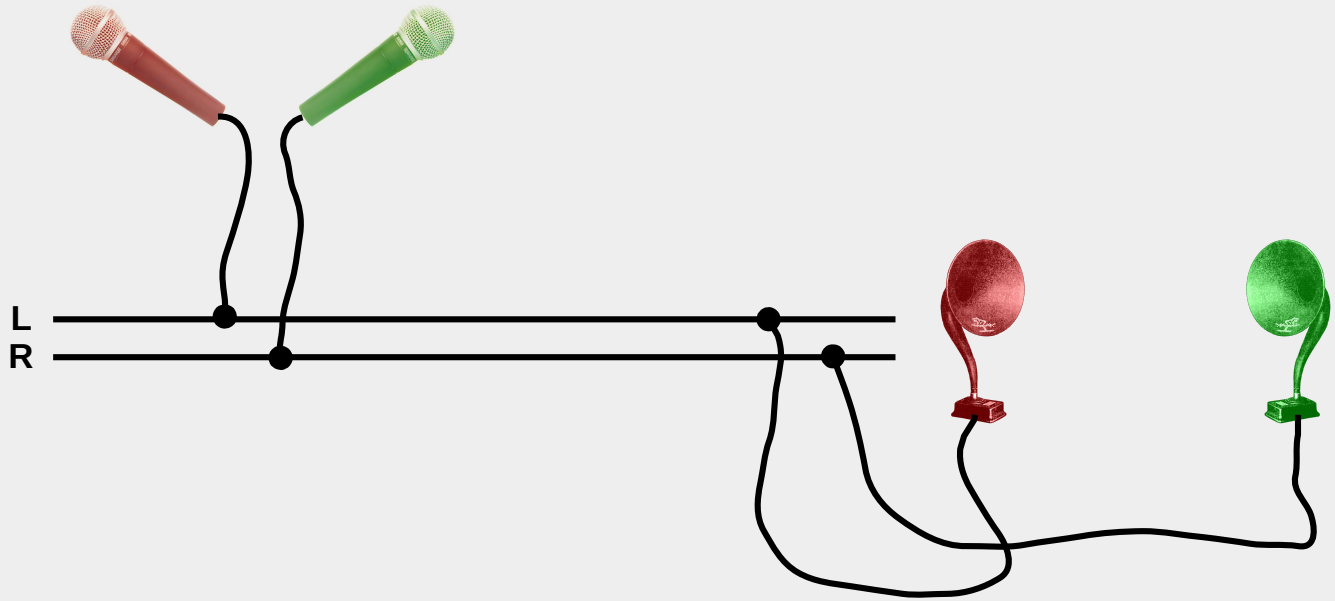
technique	advantages	disadvantages	minimal 2D setup	minimal 3D setup
binaural synth-esis	• very compact and affordable playback via headphones • optional isolation from ambient noise • low adoption thres-hold, can be played on smartphones • can be broadcast over 2 channels (radio drama)! • Ambisonic content suitable for other forms of playback	• needs head-tracking and at least 2D surround content for serious use • needs individually adjusted HRTFs for best effect • only possible via headphones, difficult to create a collective movie or concert experience • results differ between individuals, difficult QA	a pair of head-phones, optional tracking device and renderer	a pair of head-phones, optional tracking device and renderer

3. Production paradigms



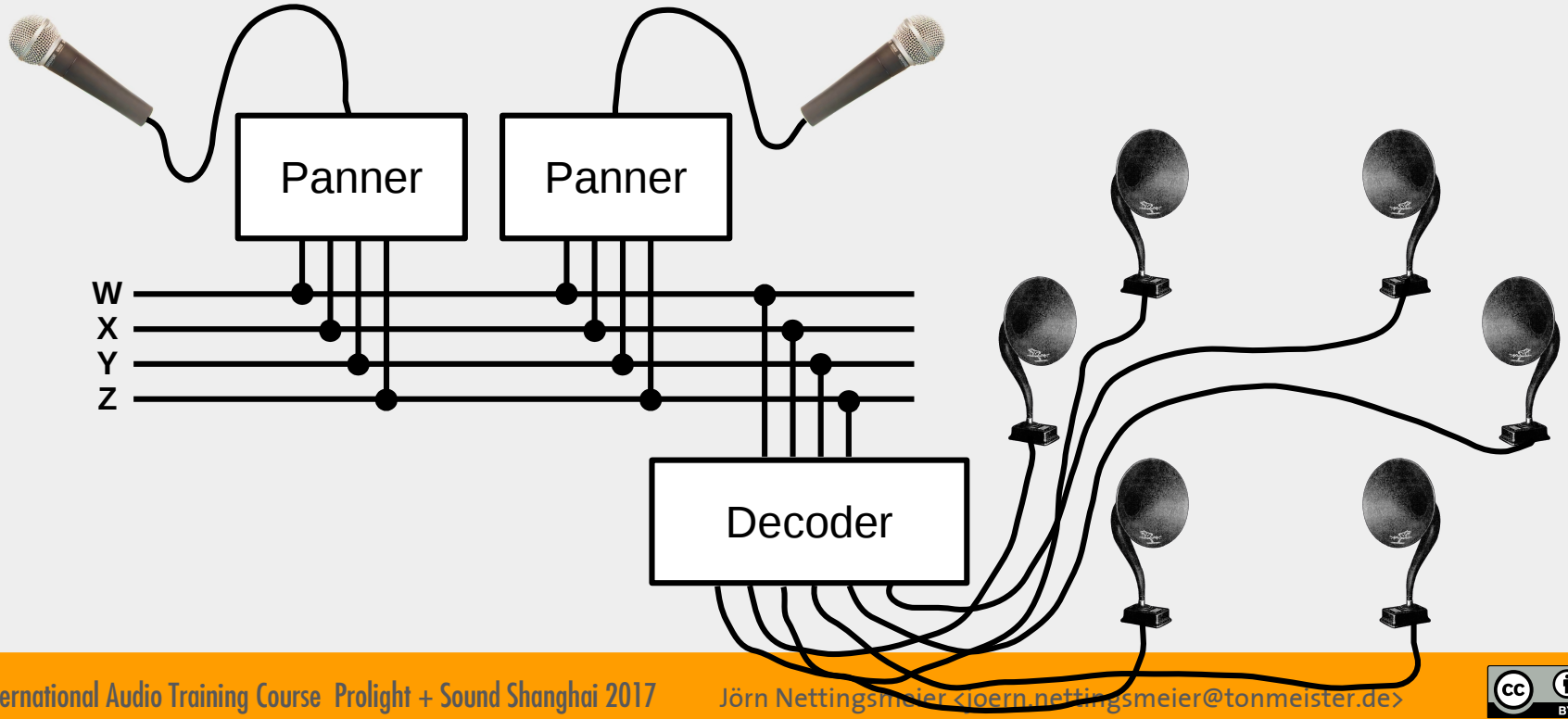
Production paradigms: channel-based

Traditionally, content was produced to **channels**, one per loudspeaker.
This is still used by stereophonic methods.



Production paradigms: scene-based

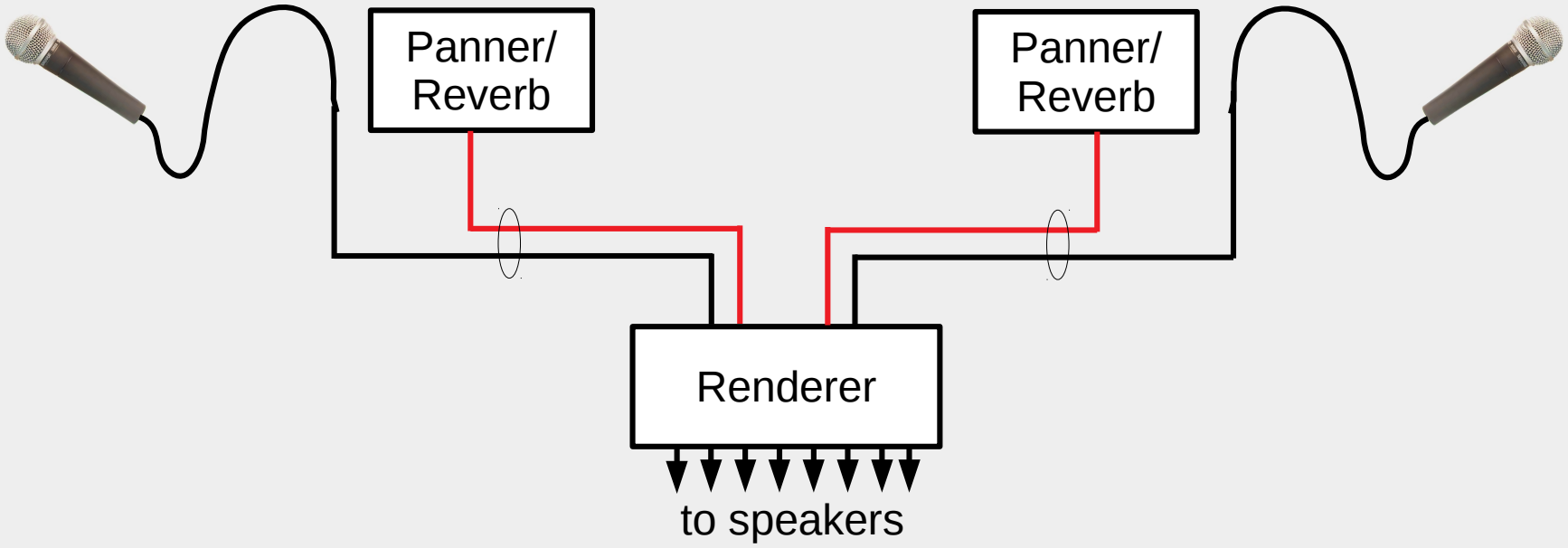
Ambisonics works with a set of channels reproducing an acoustic **scene.**
There is still a mixing bus, but the channels no longer have a fixed relationship to loudspeakers.



Production paradigms: object-based

VBAP and Wavefield synthesis use an **object-based** approach:

mono signals are stored together with **metadata** describing their direction, distance, motion and the reverberation characteristics of the scene.



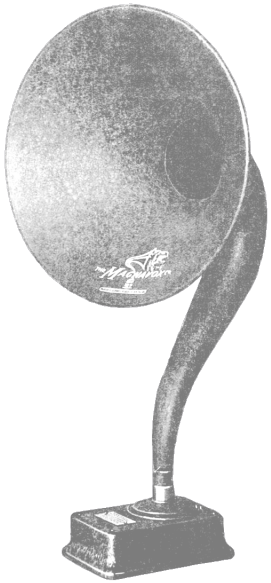
Production paradigms

Most mixing consoles and digital audio workstations still assume the **channel-based** paradigm.

Scene-based Ambisonic production requires very wide mixing buses for higher orders (with 16, 25, or 36 channels), which are not supported by mainstream products. Alternatives exist.

Object-based production on standard equipment requires workarounds: mixers and workstations feed each signal separately into the external rendering engine, using direct outs.
Panners no longer create signal outputs, they only produce positional metadata used to control the rendering engine.

4. 3D audio in practice



Many techniques complement each other well:

Ambisonics is a very convenient format for head-tracked binaural reproduction, because it is so easy to rotate.

A widely used **Ambisonics decoder internally uses VBAP** in its rendering stage (“AllRad”, Zotter and Frank 2012). VBAP is also used as a simple 9.1 panner.

Some **WFS systems borrow the height speakers from 9.1 to improve spaciousness.**

It is possible to **mimic HRTFs** with EQ to give a sense of height over loudspeakers.

3D audio in practice

Content made for one method can often be reproduced via another:

Ambisonic, WFS or binaural systems can create **virtual speakers** to enable **sterephonic N.1** playback.

Audio objects created for one system (for example WFS) can in theory be rendered on another object based system such as VBAP, or Ambisonics.

All types of content can be rendered **binaurally.** This is a great option for pre-production, to reduce the time required in expensively equipped mixing studios.

3D audio in practice

The **object-based** approach to content production offers the highest flexibility.

However, since there are no widely implemented data exchange formats for audio objects, there is a danger of **vendor lock-in**. Moving your content from one platform to another is still a substantial investment requiring extensive manual work if not a complete re-mix.

The **MPEG-H** standard promises to solve that problem.

If you are targetting online VR or gaming, the method of choice today is **Ambisonics**, possibly in combination with audio objects.

3D audio in practice

Because of the strong advantages and disadvantages of each method, commercial offerings will usually use a **combination of fundamental techniques**.

Because of the strong advantages and disadvantages of each method, commercial offerings will usually use a **combination of fundamental techniques**.

Example:

A system with a loudspeaker spacing of 1 metre may use wavefield synthesis to a little over 100 Hz.

Above that, it might use higher-order Ambisonics or VBAP.

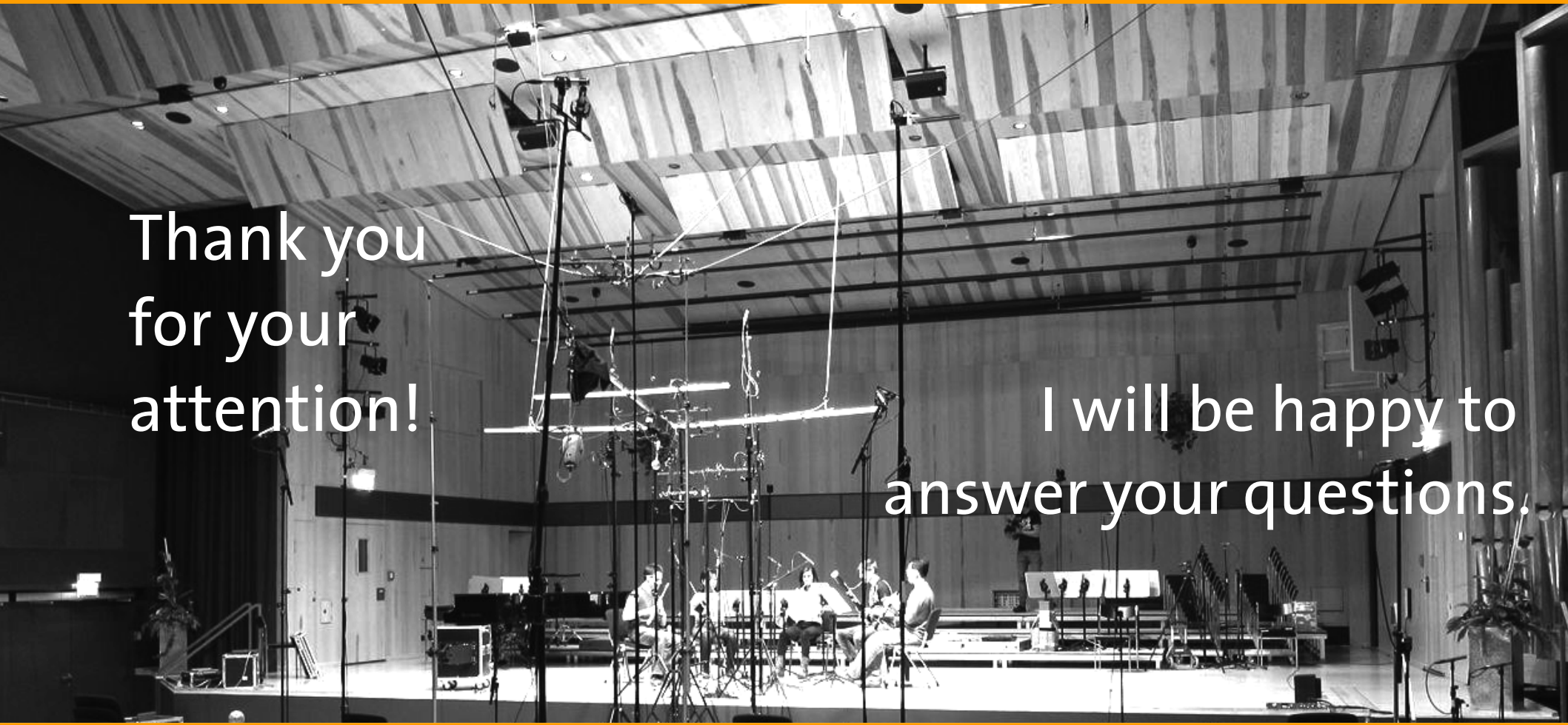
To avoid phasing problems at high frequencies, it could use a “nearest neighbour speaker” approach for frequencies above, say, 4 kHz.

3D audio in practice

No vendor in the market does anything fundamentally new.

The quality of a 3D solution is not determined by revolutionary secrets.

Rather, it results from carefully **selecting** methods **and tuning** them for best subjective results, and from offering an efficient, inspiring workflow.


Thank you
for your
attention!
I will be happy to
answer your questions.