

Virtual & Spatial Sound

かそうくうかん りったいおんきょう

仮想空間と立体音響

Virtueller und räumlicher Klang

Michael Cohen; 公園マイケル

Prof. Emeritus, University of Aizu; 名誉教授、会津大学

Adjunct Prof., Higashi Nippon International University; 東日本国際大学

Assoc. Editor, Presence: Virtual and Augmented Reality

Developer, Sonic — audio education web site

Organizing Committee, CoG: Conf. on Games, Aug. 2–5, 2027, Aizu

Follow along live:

mcohen.info/spatial-media/Sonic



What we will do

ほんじつ なが

本日の流れ

Ablauf des Seminars

- 1 • **Sound** What sound physically is, and the one number that anchors everything: $A_4 = 440$ Hz. ▶ tuning fork
- 2 • **Hearing** Air, ear, brain — three different places where a mix happens. ▶ bone conduction
- 3 • **Sound in media** The working vocabulary: voice, effects, music, diegesis, interface. Eight families, thirty-two terms.
- 4 • **Space** Channel, scene, and object encodings; reverberation, convolution, directivity. ▶ three demos
- 5 • **Control** Narrowcasting: who hears whom, and how to say it precisely. ▶ demo
- 6 • **Making it** Stems, workflow, and what AI generators are now good for. ▶ rhythm playground

Live demonstrations are marked ▶. Everything on the Sonic site runs in a browser — open it on your own phone or laptop and follow along.

Sonic runs in the browser

うご
ブラウザだけで動く

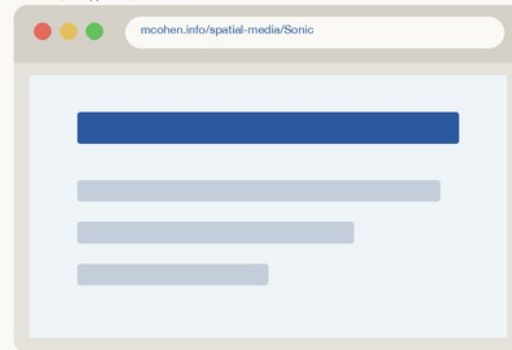
Läuft im Browser

Sonic runs in the browser

Web Audio is part of the browser, not an add-on — so every demo in this seminar runs on whatever you already have in your hand.

One URL

no install, no app store, no account



Same page, same code, same audio graph. What changes is only the screen it is laid out for —

and, for the spatial demos, whether you are on headphones. Bring headphones.

Nothing to install and nothing to agree to, which is the point: a seminar demo that needs a download is a demo half the room will not see. Every link in this deck opens the same way.

Twirleds: your phone as the controller

トワールド ワールドズ

Das Handy als Steuergerät



Your phone already has the sensors.

Gyroscope, accelerometer and magnetometer at 60 Hz.
Yaw, pitch and roll become control parameters, sent over OSC.

What it drives.

Hue lights, panoramic browsers, 3D scenes — and spatial sound: sweep your hand to pan a source around the room.

Including narrowcasting.

Channels can be targeted individually — one phone mutes a zone while another steers it. The predicates from Part Five, on a device everyone here is holding.

Free · no ads · no account · bilingual EN/ 日本語



apps.apple.com/app/twirleds



play.google.com — jp.ac.u_aizu.Twirleds



Mobile development

see Prof. Philipp Krieter;
フィリップ クリーター教授

HSD — this summer school

Built at the University of Aizu Spatial Media Group, and used to drive the installations this seminar draws on.

mcohen.info/spatial-media/Twirleds

PART ONE

Sound

おと

音

Der Schall

Before any of the design vocabulary makes sense, it helps to agree on what is physically travelling through the room.

Sound is a pressure wave

おと あつりょく なみ

音は圧力の波

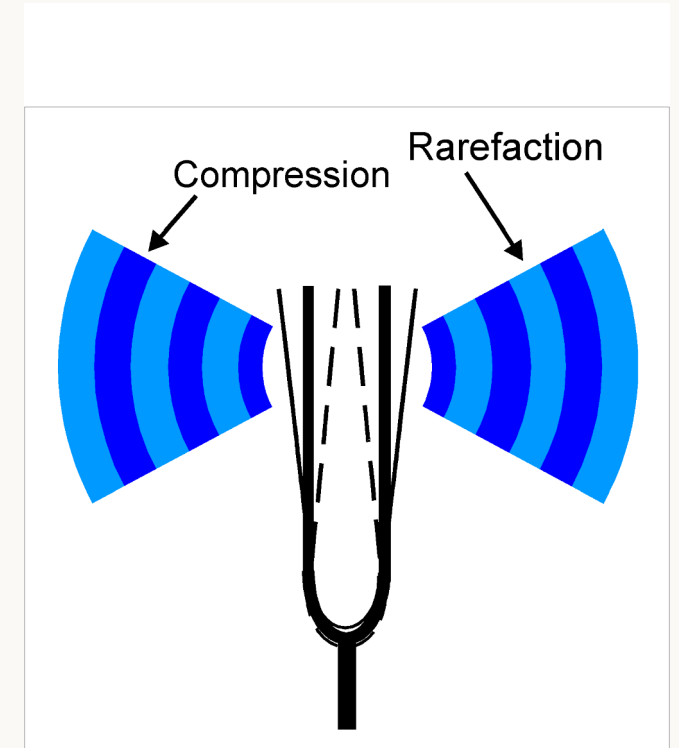
Schall ist eine Druckwelle

Not wind. Air molecules barely move — about 7.7×10^{-8} m, roughly 1/300 of the diameter of a hydrogen molecule. What travels is the disturbance, not the air.

Compression and rarefaction. Regions of slightly higher and slightly lower density chase each other outward at about 343 m/s.

Tiny deviations. Ordinary speech varies atmospheric pressure by around 0.00000001 %. Hearing is an extraordinarily sensitive instrument.

Three numbers describe it. How often the pressure cycles (frequency), how far it deviates (amplitude), and what mixture of frequencies is present (spectrum).



[Sound pressure vs. atmospheric pressure — Speaks, Intro. to Sound, Fig. 1-4](#)

LIVE DEMO

The tuning fork

おんさ
音叉
Die Stimmgabel

mcohen.info/spatial-media/Sonic/chapter1.html#ch1ToneGen

Explore Frequencies (Fig. 1.11)

What we will do

1. Strike a tine on the edge of a desk.
2. Hold it in the air: thin, quiet, and it fades fast.
3. Press the stem against the table: suddenly loud.
4. Open the tone generator and set it to 440 Hz.
5. Sound the fork and the tone together — same note, different timbre.

Listen for

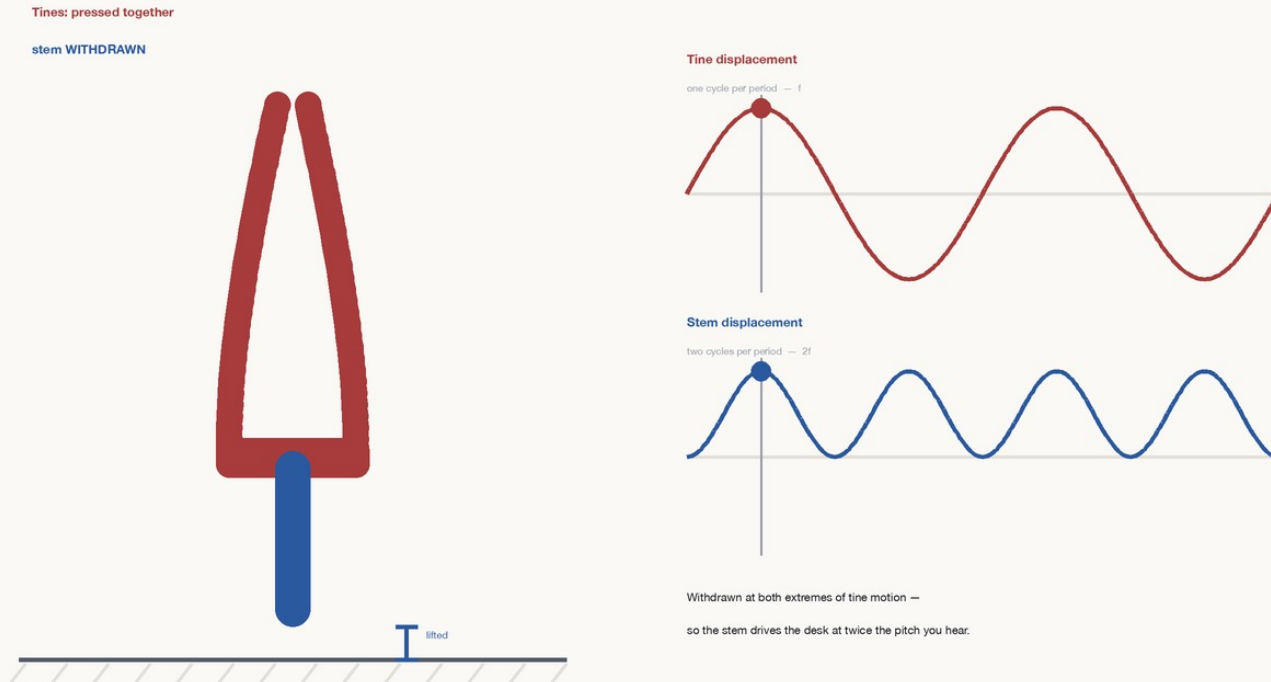
The fork alone moves very little air. The table is a bigger radiator, so it couples the same energy into the room — louder, and it dies sooner. Energy is conserved; loudness is not.

At 440 against 441 you hear one beat per second: the two waves drift in and out of step.

What the fork is actually doing

おんさ うご いそう
音叉の動きと位相

Was die Stimmgabel wirklich tut



The stem's travel is exaggerated here by roughly a thousandfold — in a real fork it is microns. It is also why the stem is what you press to the desk: tiny motion, but enough force to drive a large radiator.

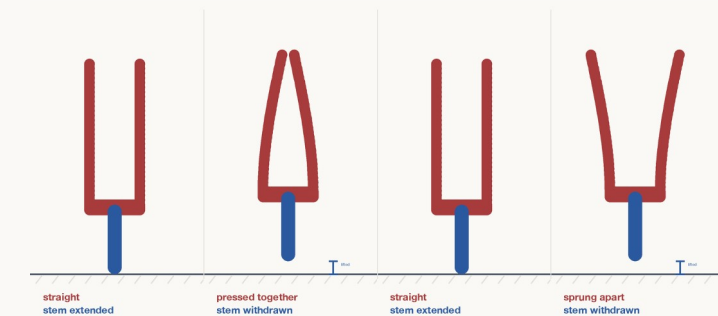
The tines: one cycle per period.

Struck, they bend toward and away from each other — 440 times a second for an A_4 fork.

The stem: two.

Bending shortens a tine at both ends of the swing alike, and the fork's centre of mass cannot move — so the yoke and stem ride up instead.

Together — withdrawn. Apart — withdrawn again. Extended only in passing.



LIVE DEMO

Seeing sound: the spectrogram

スペクトログラム

Das Spektrogramm

mcohen.info/spatial-media/Sonic/chapter7.html#spectrogram-waterfall-live

Live Spectrogram Waterfall (Fig. 7.8)

What we will do

1. Set Source to the live microphone — the browser will ask permission.
2. Each painted polyline is one FFT frame; newer frames sit in front.
3. Whistle into it: one bright ridge, moving with your pitch.
4. Sound the fork at it — a single line is what a pure tone looks like.
5. Change the FFT size, 256 up to 2048, and watch what it costs.

Listen for

The fork gives you the cleanest thing a spectrogram can draw: one horizontal line at 440 Hz. Speech and music are a moving texture by comparison.

FFT size is the time–frequency trade-off made adjustable: a longer window resolves frequency more finely but reacts more slowly. You cannot have both.

Frequency and pitch

しゅうはすう おんこう

周波数と音高

Frequenz und Tonhöhe

Frequency

Frequenz

しゅうはすう

周波数

Cycles per second, in hertz. A physical property of the wave.

Octave

Oktave

オクターブ

Double the frequency. $A_3 = 220$, $A_4 = 440$, $A_5 = 880$ Hz — equal musical steps, but a doubling each time.

Pitch

Tonhöhe

おんこう

音高

How high or low it sounds. A perception — related to frequency, but not identical to it.

Concert A

Kammerton

きじゅんおん

基準音

$A_4 = 440$ Hz by modern convention. In Mozart's day A was about 421 Hz; in Handel's, 422.5. The standard is a choice, not a law of nature.

LIVE DEMO

The siren outside this building

サイレン

Die Sirene

mcohen.info/spatial-media/Sonic/siren.html?d.siren=fc:756,fm:1.00,dev:252,mw:sawtooth-down

Sirens & FM — opens on the recorded siren

What we will do

1. Open the link — it arrives already tuned to the recorded siren.
2. “▶ Start audio”. That is the alert we heard here at HSD.
3. Check the FM plane: the dot sits at $f_m = 1$ Hz, far left.
4. Run the collapse: the rate sweeps 0.2 to 800 Hz over 18 seconds.
5. Somewhere near 20 Hz the movement stops being movement.

Listen for

A siren is frequency modulation running slow. Carrier 756 Hz, deviation 252 Hz: the pitch falls a full octave, 1008 down to 504 Hz, once a second, then snaps back in a few tens of milliseconds — an inverted sawtooth, which is why it descends rather than climbs.

Those numbers were measured off a recording of this siren, not chosen to sound right — and they travel in the link itself.

The endless staircase

むげんおんかい

無限音階

Die endlose Tonleiter

mcohen.info/spatial-media/Sonic/shepard.html

Shepard Tones

What we will do

1. Start the tone and let it run for half a minute.
2. It appears to rise without ever getting higher.
3. Turn on “Spatialization” — it orbits your head. Headphones.
4. Turn on “Show spectrogram” — octaves as a rising constellation.
5. Reverse to descending: the fall is endless, the orbit reverses.

Listen for

Pitch is not frequency. The Shepard tone stacks octaves under a fixed spectral envelope, so every partial rises while the envelope does not move — and hearing reports a rise that never arrives.

The spectrogram shows it: partials climb while the bright band they cross stays put — Escher’s staircase for the ear.

Bone conduction

こつでんどう

骨伝導

Knochenleitung

Two paths to the cochlea. Air conduction goes outer ear → eardrum → ossicles → cochlea. Bone conduction skips all of that: the skull itself vibrates and drives the inner ear directly.

Try it now. Strike the fork, then touch the stem to your mastoid bone, just behind the ear. You hear it with the ear canal completely blocked.

Why your recorded voice sounds wrong. When you speak you hear yourself through both paths at once, and bone conduction favours low frequencies. A recording gives you only the air path — thinner, higher, and disconcertingly unfamiliar.

Used in practice. Bone-conduction headphones leave the ear canal open for traffic and conversation; the Rinne and Weber tuning-fork tests use the two paths to separate conductive from sensorineural hearing loss.

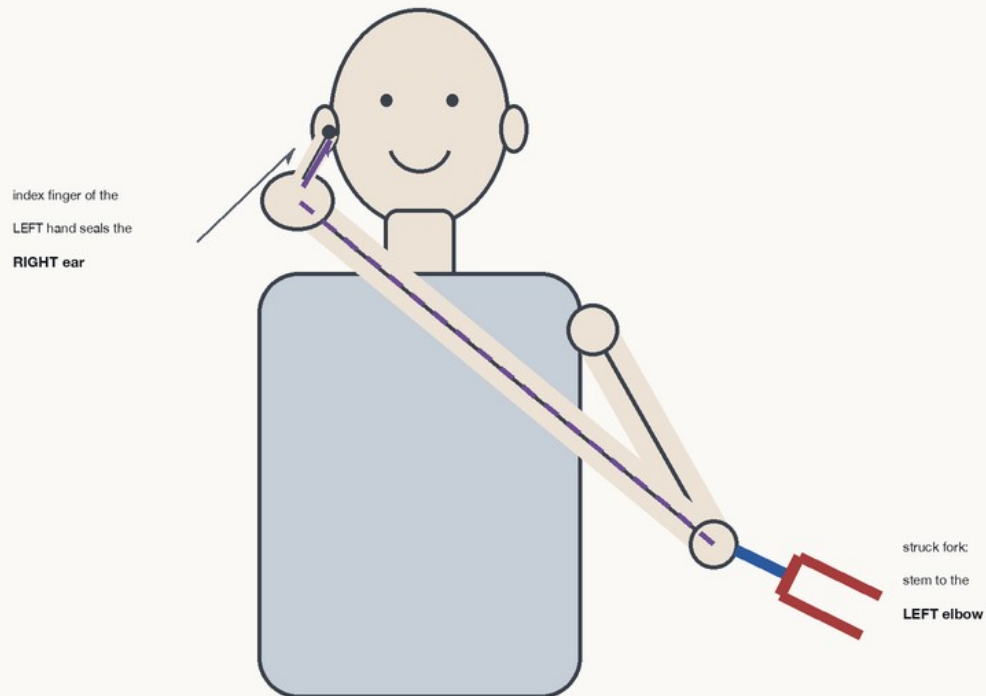
Two things to try with the fork

こつでんどう えんしゅう

骨伝導の演習

Zwei Versuche

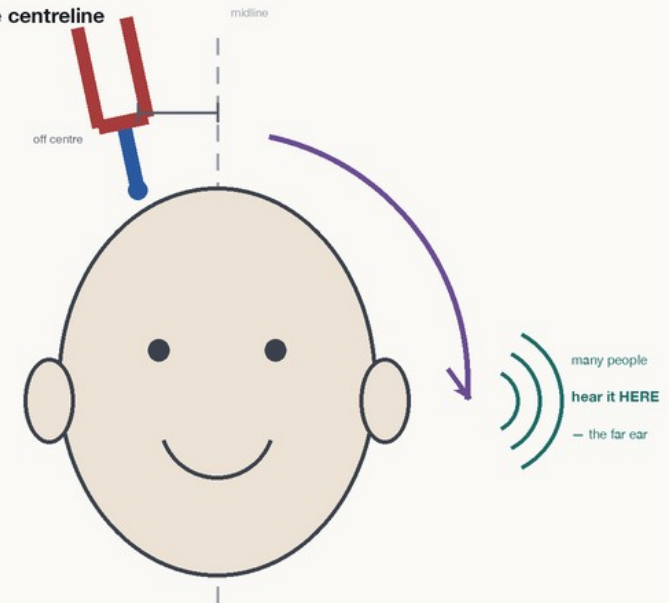
A. Up the arm — and in through your own finger



A. Up the arm — and in through your own finger.

Index finger of one hand into the opposite ear; kick the fork and press its stem to the elbow of that arm.

B. On the crown, off the centreline



B. On the crown, off the centreline.

Ring the fork to the top of the head, off the midline. Ask which ear hears it — for many, the far one.

PART TWO

Hearing

ちょうしゅ

聴取

Das Hören

The same pressure wave is mixed three times before you are aware of it — once in the air, once in the ear, once in the brain.

Air, ear, brain

くうき みみ のう
空気・耳・脳

Luft, Ohr, Gehirn

Air — acoustics

Luft — Akustik

くうき おんきょう
空気・音響

Physics. Propagation, reflection, absorption, interference. Objective, measurable, the same for everyone in the room.

Brain — cognition

Gehirn — Kognition

のう にんち
脳・認知

Auditory scene analysis. Grouping partials into objects, separating voices, deciding what to attend to and what to ignore.

Ear — sensation

Ohr — Empfindung

みみ かんかく
耳・感覚

Physiology and psychoacoustics. Critical bands, masking, binaural differences. Where physical intensity becomes perceived loudness.

Why it matters

Warum es zählt

せつけいじょう いみ
設計上の意味

A designer works on all three at once. Adding decibels is acoustics; being heard is cognition — and they often disagree.

[After Cohen & Martens, “Spatial Soundscape Superposition and Multimodal Interaction,” in The Technology of Binaural Understanding, Springer 2020, ch. 13.](#)

PART THREE

Sound in media

メディアなか おとの中の音

Klang in den Medien

A working vocabulary. Eight families of sound, thirty-two terms — the words a team actually uses to divide up the work.

Eight families, one mix

やっ そう ひと こんごう
八つの層、一つの混合

Acht Ebenen, eine Mischung

Voice

こえ

声

Stimme

Everything a person says, from scripted narration to the crowd you never see.

Effects

こうかおん

効果音

Geräusche

What objects and surfaces do — the evidence that the world is physical.

Music

おんがく

音楽

Musik

Scored sound that shapes how a moment feels rather than what it contains.

Frame

わくぐ

枠組み

Rahmen

Whether a sound exists in the world, or only for the person experiencing it.

Interface

そうさおん

操作音

Interface

Sound whose job is to tell the user what the system just did.

Space

くうかん

空間

Raum

The cues that put a sound at a place relative to the listener.

Practice

じつむ

実務

Praxis

The decisions that keep many simultaneous sounds intelligible.

Accessibility

だいたいしゅだん

代替手段

Zugänglichkeit

Every sound-carried meaning needs a route that does not require hearing.

Voice

こえ

声

Stimme

Everything a person says, from scripted narration to the crowd you never see.

Narration

Erzählerstimme

ナレーション

An unseen voice addressing the audience directly, from outside the story.

Walla

Hintergrundgemurmel

ガヤ

Unintelligible crowd murmur that gives a space its population.

Dialogue

Dialog

せりふ

台詞

Speech between characters inside the world, heard as they hear it.

ADR / looping

Nachsynchronisation

アフレコ

Dialogue re-recorded afterwards and matched to picture.

Effects

こうかおん

効果音

Geräusche

What objects and surfaces do — the evidence that the world is physical.

Hard SFX

Soundeffekte

こうかおん

効果音

Discrete designed events: a door, a weapon, a machine, an explosion.

Ambience

Atmo

かんきょうおん

環境音

The continuous bed of a place: wind, traffic, a forest, a server room.

Foley

Geräuschemacherei

フォーリー

Human-scale sounds performed live to picture: footsteps, cloth, handling.

Room tone

Raumton

ルーム トーン

The specific near-silence of one room; without it a space sounds dead.

Music

おんがく

音楽

Musik

Scored sound that shapes how a moment feels rather than what it contains.

Score / BGM

Filmmusik

げきばん

劇伴

Composed underscore carrying the emotional read of a scene.

Stinger

Akzent

スティンガー

A short musical hit marking a discovery, a failure, or an arrival.

Adaptive music

Interaktive Musik

おんがく

インタラクティブ音楽

Score that re-arranges itself in response to what the player does.

Leitmotif

Leitmotiv

しどうどうき

示導動機

A recurring figure tied to a character, place, or idea.

Frame

わくぐ
枠組み

Rahmen

Whether a sound exists in the world, or only for the person experiencing it.

Diegetic

diegetisch

ものがたりないおん

物語内音

Sourced inside the story world; the characters could hear it too.

Trans-diegetic

transdiegetisch

えっきょうおん

越境音

A sound that crosses the boundary — score revealed to be a car radio.

Non-diegetic

nichtdiegetisch

ものがたりがいおん

物語外音

Heard only by the audience: score, narration, most interface sound.

Fourth wall

Vierte Wand

だいし かべ

第四の壁

The boundary itself; breaking it means the world addresses you.

Interface

そうさおん

操作音

Interface

Sound whose job is to tell the user what the system just did.

UI signal

Bedienklang

そうさおん

操作音

A tone tied to a control: press, toggle, error, confirm.

Auditory icon

Auditives Icon

ちょうかく

聴覚アイコン

A sound resembling its referent: paper for a file, a latch for a lock.

Earcon

Hörsymbol

イヤコン

An abstract motif learned by convention, like a ringtone for an event.

Feedback

Rückmeldung

かくにんおん

確認音

Confirmation that an input registered — critical where there is no haptic.

Space

くうかん

空間

Raum

The cues that put a sound at a place relative to the listener.

Spatial audio

Raumklang

りったいおんきょう

立体音響

Rendering a source at a position in three dimensions around the head.

Occlusion

Verdeckung

しゃへい

遮蔽

What a wall does to a sound behind it — muffled, not merely quieter.

HRTF

Außenhörübertragungsfunktion

とうぶでんたつかんすう

頭部伝達関数

The filtering a head and ear impose, used to synthesise direction.

Reverb zone

Hallzone

ざんきょうりょういき

残響領域

A region with its own reflection character: tunnel, hall, open field.

Practice

じつむ
実務

Praxis

The decisions that keep many simultaneous sounds intelligible.

Ducking

Absenkung

おんりょうたいひ
音量退避

Lowering one layer automatically so another stays intelligible.

Loudness

Lautheit

ラウドネス

Perceived level over time, measured in LUFS rather than peaks.

Voice limiting

Stimmenbegrenzung

はつおんすうせいげん
発音数制限

Capping simultaneous sources, and choosing which ones survive.

Silence

Stille

むおん
無音

A designed absence — the loudest tool available, and the least used.

Accessibility

だいたいしゅだん

代替手段

Zugänglichkeit

Every sound-carried meaning needs a route that does not require hearing.

Subtitles

Untertitel

じまく

字幕

Spoken words as text; dialogue only.

Audio description

Audiodeskription

おんせいかいせつ

音声解説

A spoken account of what is visible, for people who cannot see it.

Captions

Erweiterte Untertitel

ちょうかくじまく

聴覚字幕

Dialogue plus the non-speech sound that carries meaning.

Visual substitute

Visuelle Entsprechung

しかくだいたい

視覚代替

A directional indicator standing in for a sound cue.

PART FOUR

Space

くうかん

空間

Der Raum

Everything so far assumed sound simply arrives. Now: where it comes from, how the room changes it, and how we reproduce that.

How is spatial sound modelled?

くうかんおんきょう さんほうしき

空間音響の三方式

Wie wird räumlicher Klang modelliert?

Channel-based

Kanalbasiert

ほうしき

チャンネル方式

Signals tied to fixed speaker positions — stereo, 5.1, 7.1. Simple and robust, but the mix assumes one listening geometry.

Scene-based

Szenenbasiert

ほうしき

シーン方式

The whole sound field captured or encoded at a point, as in Ambisonics. Rotate the listener and the field rotates with them.

Object-based

Objektbasiert

ほうしき

オブジェクト方式

Each sound kept as a separate stream plus its position; the renderer decides how to play it on whatever speakers are actually present.

Not three ways of storing a file: three ways of modelling the sound itself, and the choice runs through storage, transmission, control and rendering alike. What really differs is when the spatial decision is made — channel-based decides in the studio, scene-based at capture, object-based defers it to playback, which is why it alone survives a change of speakers.

Channel-based: what the numbers mean

ほうしき しょうげん
チャンネル方式の諸元

Kanalbasiert: die Kenngrößen

Width

Breite

すう
チャンネル数

How many channels — 2 for stereo, 6 for 5.1.

Sampling rate

Abtastrate

ひょうほんかしゅうはすう

標本化周波数

Time resolution: CD audio at 44.1 kHz reproduces up to 22.05 kHz.

Depth

Auflösung

りょうしか すう
量子化ビット数

Amplitude resolution: 16 bits $\approx$ 96 dB of dynamic range.

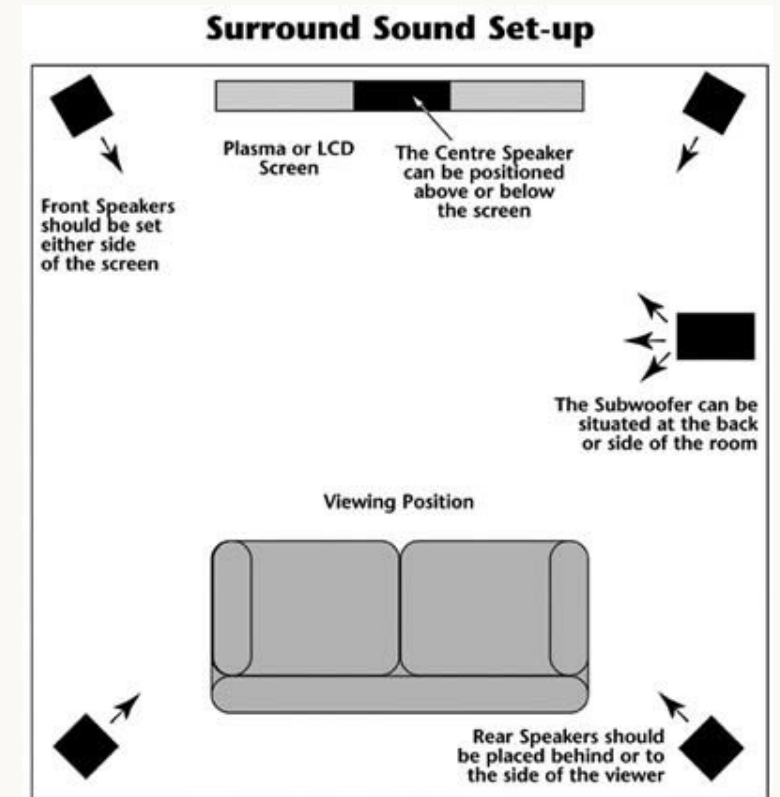
The “.1”

Der „.1“

ていおん

低音チャンネル

Not a channel count but a bandwidth ratio — the LFE carries only $\sim$ 120 Hz.



A 5.1 layout

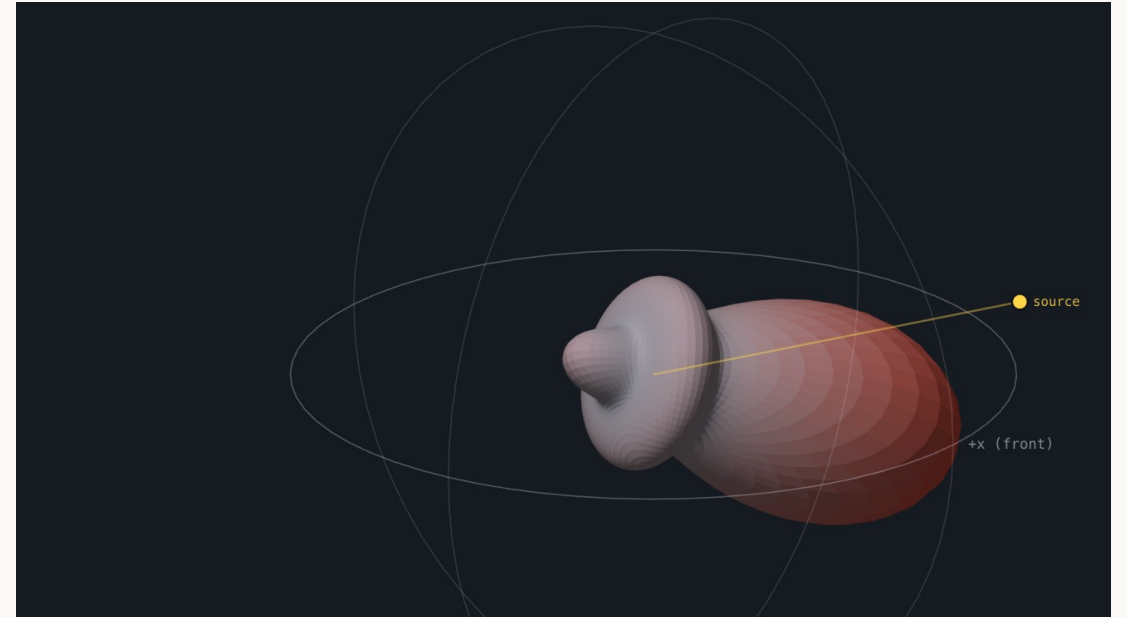
Scene-based: capturing a whole field

ほうしき しゅうろく
シーン方式の収録

Szenenbasiert: das ganze Feld



[Ambisonic microphone array — zylia.co](https://zylia.co)



[Sonic: Spherical Harmonics Explorer](#)

A first-order Ambisonic mic captures four channels that together describe the sound field at one point. Spherical harmonics are the basis functions used to write that field down — the explorer above lets you see them.

Object-based rendering in XR

XRにおけるオブジェクト方式^{ほうしき}

Objektbasiertes Rendering in XR

The scene is a list, not a mix. Each sound stays a separate stream carrying its own position, directivity and extent. Nothing is committed to a speaker until the moment of playback.

Which is what XR requires. The listener stands up and walks around. A channel mix assumed where their ears would be; an object scene simply re-renders from wherever they now are.

Every engine works this way. Unity, Unreal and WebXR all attach an audio source to a transform in the scene graph, then hand the list to a spatialiser that applies HRTF, distance and occlusion per frame.

And it degrades gracefully. The same scene renders to headphones, a phone speaker, or a 22.2 rig — because the spatial decision was never baked in.



Spatial media in AR/VR

Alaeddin Nassani;
アラディン ナッサーニ

LIVE DEMO

Spherical harmonics

きゅうめんちょうわかんすう

球面調和関数

Kugelflächenfunktionen

mcohen.info/spatial-media/Sonic/spherical-harmonics.html

Spherical Harmonics Explorer

What we will do

1. Start at order 0 — one lobe, no direction at all.
2. Step to order 1: three figure-of-eight patterns, one per axis.
3. Keep going and watch the lobes multiply.
4. Rotate the whole set and see the field turn with it.
5. “Play tone”, then drag the source: the level tracks the pattern.

Listen for

These are the basis functions an Ambisonic recording is written in.

Order 0 is one omni channel; first order adds three, and resolution improves with every order you can afford.

The gain readout dives as the source crosses a null: the pattern is a function of direction, and the level follows.

Gaussian splats: the same basis, seen

さんじげん

三次元ガウシアン スプラッティング

Gaussian Splats: dieselbe Basis, gesehen

mcohen.info/spatial-media/GaussianSplats/

Gaussian Splat Explorer

What we will do

1. Drag to orbit; click or tap a blob to select it.
2. Splat Anatomy: colour over direction — the acoustics picture, redrawn.
3. Degree cap: degree 0 is one coefficient, degree 3 is sixteen.
4. Orthonormal to SN3D (ambiX): coefficients rescale, the image does not.
5. Click one cell of the basis grid to render that harmonic alone.

Listen for

Split out of the Spherical Harmonics Explorer
— the same functions, read as view-
dependent colour instead of directional gain.

Degree 0 is the flat base colour and the
higher degrees are what changes with the
viewing angle: order 0 being a single omni
channel, said in graphics.

The same basis functions, twice

おな きていかんすう

同じ基底関数

Dieselben Basisfunktionen

In graphics — 3D Gaussian Splatting

In der Grafik — 3DGS

さんじげん

三次元ガウシアン スプラッティング

Each splat stores colour as spherical-harmonic coefficients, so its appearance changes with the direction you view it from. Order 0 is a flat colour; higher orders add view-dependent glint and sheen.

Same mathematics, different signal. A function on the sphere — colour as a function of view direction, or pressure as a function of arrival direction — is expanded in the same harmonic basis, and in both fields the order you can afford sets the angular resolution you get.

The neural-rendering parallel holds too: where NeRF and 3DGS learn a radiance field from photographs, neural acoustic fields learn room impulse responses from recordings — and both then render a novel viewpoint, or a novel listening point, that was never captured.

In audio — Higher-Order Ambisonics

Im Audio — HOA

こうじ

高次アンビソニックス

Each recording stores the sound field as spherical-harmonic coefficients, so what you hear changes with the direction you face. Order 0 is one omni channel; higher orders sharpen the image.



3D Gaussian Splatting

see Prof. Jens Herder;
イエンス ヘルダー教授

HSD — this summer school

One basis, two signals

おな きていかんすう ふた しんごう
同じ基底関数、二つの信号

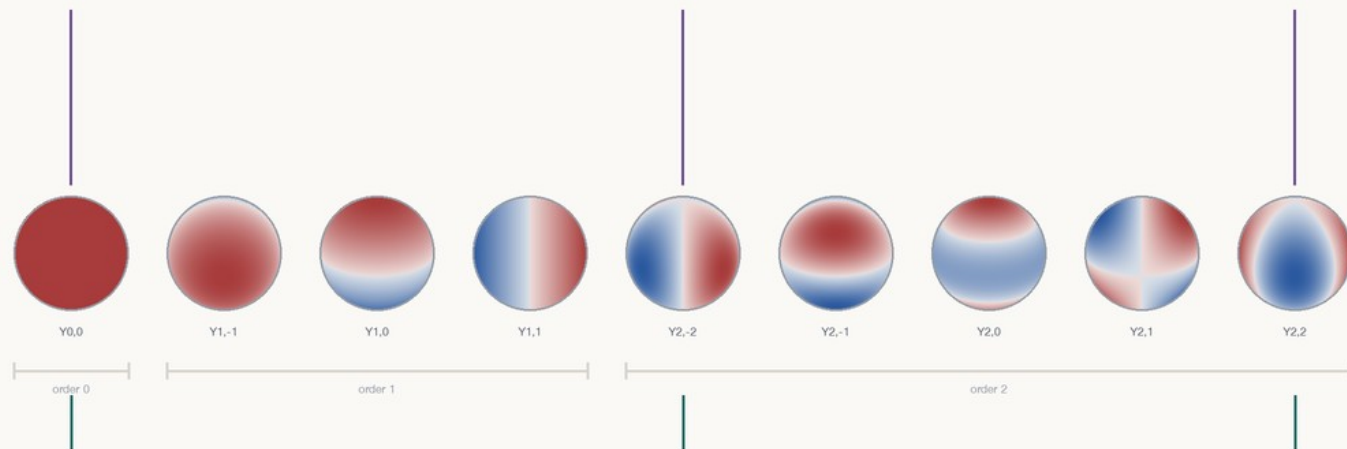
Eine Basis, zwei Signale

3D Gaussian Splatting

each splat stores COLOUR as a function of the direction you view it from

$$c(d) = \sum c(l,m) \cdot Y(l,m)(d)$$

coefficients are RGB triples; d = the direction the camera looks FROM. Order 0 alone is flat, unshaded colour.



Higher-Order Ambisonics

one recording stores PRESSURE as a function of the direction it arrives from

$$p(d) = \sum b(l,m) \cdot Y(l,m)(d)$$

coefficients are audio signals; d = the direction the sound comes IN from. Order 0 alone is one omni channel.

What changes

what the coefficients weigh — an RGB triple per splat, or a whole signal per channel

what the direction d means — the way you look, or the way the sound came

What does not

the basis itself, its orthogonality, and the way a rotation of the scene becomes a matrix on the coefficients — rotate a splat cloud or steer a listener's head and it is the same linear algebra.

In both, the order you can afford is what sets the angular resolution you get.

Red where the function is positive, blue where it is negative. A first-order ambisonic microphone measures the first four of these directly; a splat stores a coefficient against each. Nobody had to agree on this — it is what the sphere makes available.

Relighting and re-reverberating are the same move

さいしょうめい さいざんきょう

再照明と再残響

Relighting und Nachhall-Ersatz

Neural relighting

Neuronal Relighting

さいしょうめい

ニューラル再照明

Factor a photograph into what the object is — albedo, normals, material — and what the light did. Keep the first, substitute the second, and the scene is lit anew.

BRDF ↔ directivity. How a surface reflects light by direction, and how a source radiates sound by direction — both are functions on a sphere. Your HRTF is the receiving half of the same idea.

Light probe ↔ reverb zone. Graphics samples illumination at points and interpolates; audio samples room response at points and interpolates. Same trick, same artefacts at the boundaries.

Novel view ↔ novel listening point. NeRF and 3DGS render a camera position that was never photographed; neural acoustic fields render a listening position that was never recorded.

And a link that runs the other way. Networks now predict a listener's individual HRTF from photographs of their ears — a vision model personalising the spatial audio.

Convolution / auralization

Auralisation

たた こ さいざんきょう

畳み込みによる再残響

Factor a recording into what the source is — the dry signal — and what the room did. Keep the first, substitute the second, and the voice is in a cathedral.



Neural rendering

see Prof. Yoshihiro Kanamori;

金森 由博 教授

Almost — but the dry sound is not the albedo

るいすい ていせい

類推の訂正

Fast — aber trocken ist nicht Albedo

Albedo is a reflectance: a ratio that scales light already arriving. The dry signal does not scale anything — it is what leaves the source. So it sits in the emission slot, and the acoustic albedo is somewhere else entirely: the absorption of the walls.

Light source, emission

Emission

ほうしゃ

放射

Dry signal — what the source puts into the room.

IES / goniometric profile

Abstrahlcharakteristik

しこうせい

指向性

Source directivity. Both are radiated power by direction.

Albedo, reflectance

Reflexionsgrad

はんしゃりつ

反射率

Absorption coefficient of each surface, per frequency band.

Camera response

Kameraantwort

とうぶでんたつかんすう

頭部伝達関数

HRTF — the receiver's own directional filter.

Direct lighting

Direktes Licht

ちよくせつおん

直接音

Direct sound: straight line, source to ear.

Global illumination

Globale Beleuchtung

こうぶざんきょう

後部残響

Late reverberation. Image-source method is mirror reflection; acoustic ray tracing is path tracing with a slower wave.

With that one swap the rest lines up: rendering equation and room acoustics are the same linear transport problem, and a game engine really is “lighting” its sounds.

Where the analogy breaks

るいすい げんかい

類推の限界

Grenzen der Analogie

No sharp shadows

Keine scharfen Schatten

おと かげ

音に影はできない

Light's wavelength is $\sim 0.5 \mu\text{m}$, so obstacles cast crisp silhouettes. Sound's is 2 cm to 17 m — comparable to the obstacle — so it diffracts around it. Occlusion is a frequency-dependent low-pass, not a silhouette. You hear round corners; you do not see round them.

A third difference follows from the second: because wavelength is comparable to room features, acoustics cannot use geometric optics everywhere. Wave solvers are needed at low frequencies and ray methods at high, so practical auralisation is hybrid in a way renderers rarely have to be.

The transient is the percept

Der Einschwingvorgang ist der Reiz

かとおうとう

ちかく

過渡応答こそが知覚

Light equilibrates in nanoseconds, so we only ever see the steady state. Sound takes milliseconds, so the arrival pattern is the experience. A room impulse response is the transient that graphics normally throws away — the analogue is transient rendering, not ordinary rendering.

Can you un-reverberate a recording?

ざんきょう と じょ

残響は取り除けるか

Lässt sich Nachhall entfernen?

Yes, partly — and it is the same ill-posed problem as recovering albedo from a photograph.

One equation, two unknowns. The microphone hears $y = s * h$: the dry source convolved with the room. Infinitely many pairs give the same y , exactly as infinitely many albedo-and-shading pairs give the same pixel. This is intrinsic decomposition, in air.

So both sides lean on priors. Graphics assumes albedo is piecewise constant and shading is smooth. Audio assumes speech is sparse in time–frequency and that a room decays smoothly and causally. Both fields now mostly learn the prior instead of writing it down.

More observations help. Multiple views or lighting conditions in graphics; multiple microphones in audio — multichannel methods such as weighted prediction error are the standard practical answer.

But it is estimation, not inversion. A room response is generally not minimum-phase, so no stable causal inverse filter exists; and wherever the room has a spectral null, that information is simply gone. You can estimate what was there. You cannot recover what the room discarded — just as you cannot read what a shadow hid.

Which is why we record dry.

Baking a room into a stem commits you to acoustics you cannot undo — the audio equivalent of painting highlights into a texture and then trying to relight it.

Reverberation

ざんきょう

残響

Nachhall

mcohen.info/spatial-media/Sonic/chapter7.html#reverb-playground

Reverb playground

What we will do

1. Start dry, then raise the room size.
2. Separate the three arrivals: direct sound, early reflections, late tail.
3. Change absorption and watch RT60 change.
4. Compare a small tiled room with a large hall.

Listen for

Reverberation is not an effect added to a sound — it is the room answering. Early reflections tell you the size and shape; the late tail tells you how absorbent the surfaces are.

RT60: the time for the sound to decay by 60 dB.

Direct, early, late

ちよくせつおん しょきはんしゃ こうぶざんきょう

直接音・初期反射・後部残響

Direktschall, frühe Reflexionen, Nachhall

Direct sound

Direktschall

ちよくせつおん

直接音

Straight line, source to ear. Tells you the direction.

Early reflections

Frühe Reflexionen

しょきはんしゃおん

初期反射音

First bounces, ~5–80 ms. Tell you the size and shape of the room.

Late reverberation

Nachhall

こうぶざんきょう

後部残響

Dense, directionless tail. Tells you how absorbent the surfaces are.

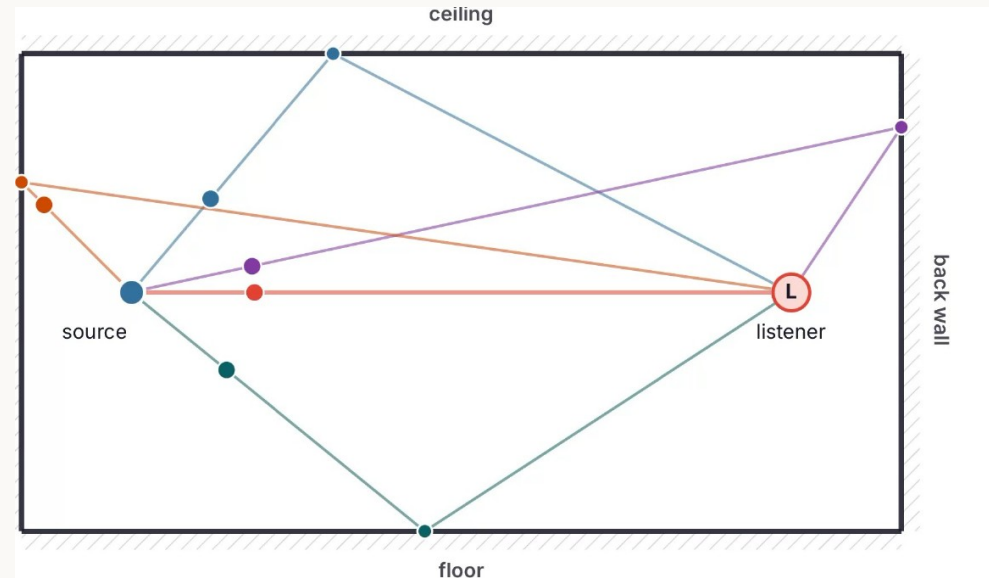
RT60

Nachhallzeit

ざんきょうじかん

残響時間

Time to decay by 60 dB. A concert hall ≈ 2 s; a studio ≈ 0.3 s.



[Sonic: reflection explorer — direct path, early reflections, and the room](#)

LIVE DEMO

Convolution

たたこ
畳み込み
Faltung

mcohen.info/spatial-media/Sonic/convolution.html

Convolution Explorer

What we will do

1. Play the dry signal — a clap or a short phrase.
2. Load an impulse response: a hall, a stairwell, a pipe.
3. Convolve, and hear the source placed in that space.
4. Swap the impulse response without touching the source.

Listen for

An impulse response is a recording of what a space does to a single instantaneous click — its acoustic fingerprint. Convolution applies that fingerprint to anything else.

This is how a voice recorded in a dry booth ends up in a cathedral.

Where the fingerprint comes from

しょきはんしゃ ざんきょう

初期反射と残響

Woher der Fingerabdruck kommt

mcohen.info/spatial-media/Sonic/chapter7.html#reflection-explorer

Reflection Explorer (Fig. 7.18)

What we will do

1. Drag the source (+) and the sink (–) anywhere in the room.
2. Choose how many wall bounces to include, then press Play.
3. One impulse per path: they leave together and arrive apart.
4. “Two ears” splits the sink into a head — one echogram per ear.
5. “Swap source & sink” — the rays reverse, the echogram does not.

Listen for

Each path is a straight line from a mirror image of the source: the echogram IS the impulse response.

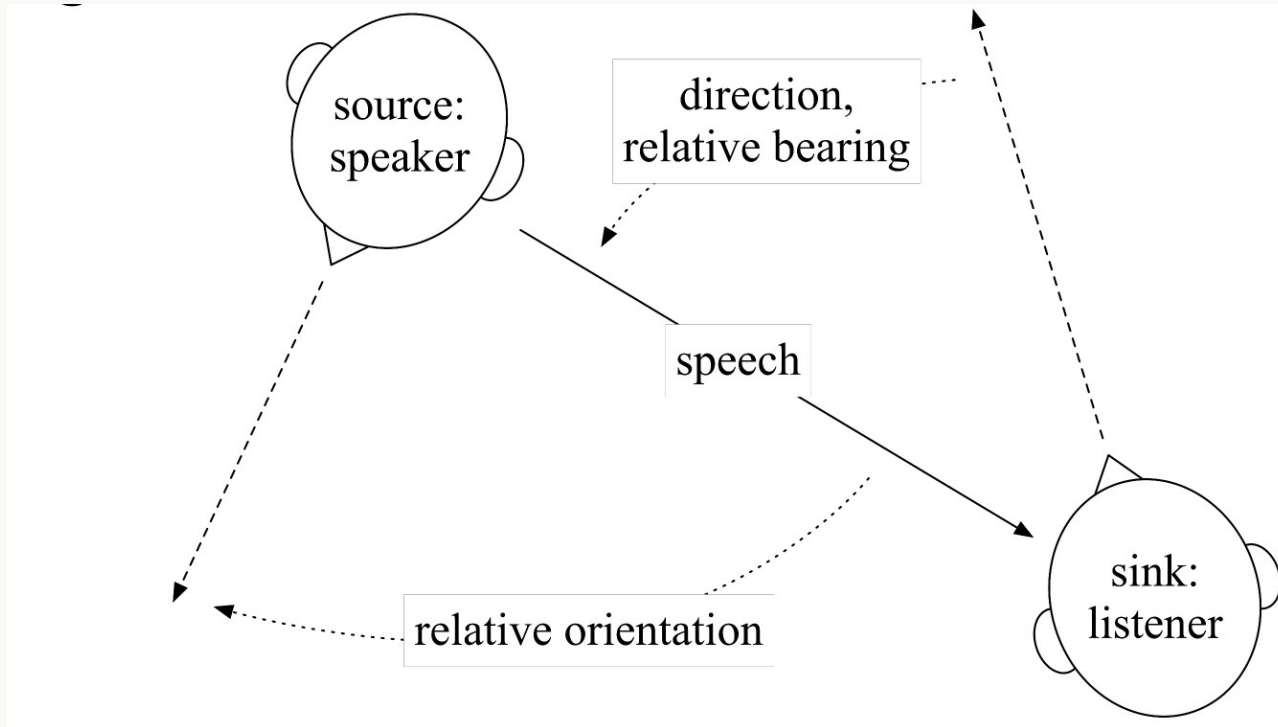
Two ears make it binaural — paired spikes give the interaural time and level differences.

Helmholtz reciprocity: swap source and sink and the echogram does not move.

Direction and orientation

ほうこう しこうせい
方向と指向性

Richtung und Ausrichtung



Bearing is heard; facing is inferred.

A listener localises where a source is from binaural cues — accurate to a degree or two straight ahead.

You can also tell which way it is aimed, but only because sources radiate directionally: turning away tilts the spectrum, and that tilt is the cue.

Both are social signals. Whether someone faces you is proxemic information, and a virtual environment that ignores it feels subtly wrong even when every level is correct.

[Directivity patterns are explored in the narrowcasting sandbox →](#)

LIVE DEMO

Spatial sound

りったいおんきょう

立体音響

Räumlicher Klang

mcohen.info/spatial-media/Sonic/chapter9.html#ch9Spatializer

Binaural Spatializer (Fig. 9.3)

What we will do

1. Headphones on — this one needs two ears.
2. Move the source in azimuth: left, front, right, behind.
3. Raise and lower elevation.
4. Pull the source away and listen to distance, not just level.
5. Enable head tracking and turn your head — the scene stays put.

Listen for

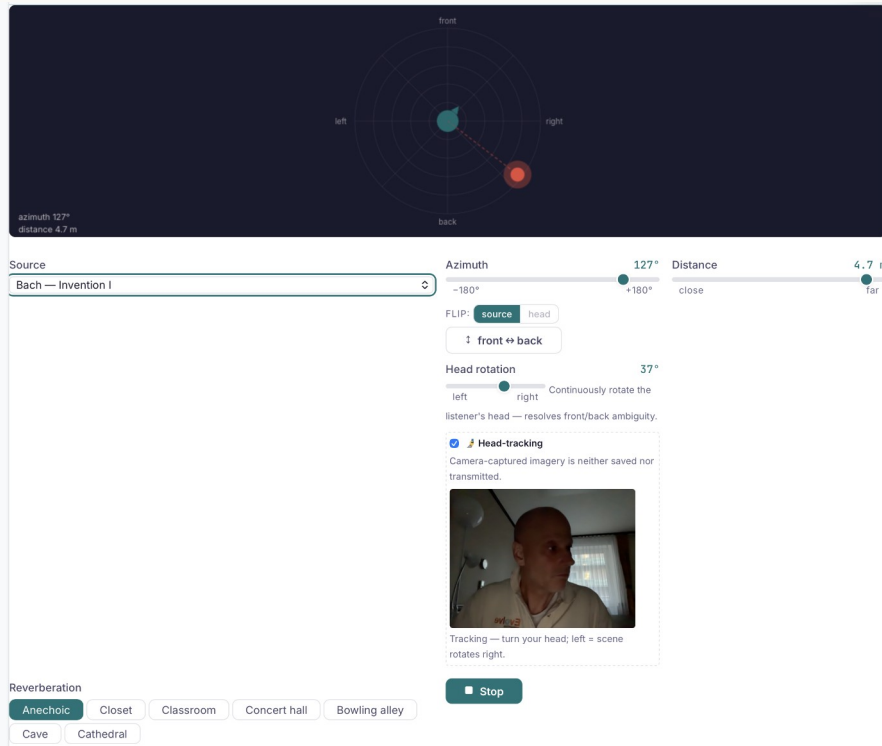
Two cues do most of the work: the sound reaches the near ear sooner (interaural time difference) and louder (interaural level difference). Your outer ears add a spectral signature that resolves front from back and gives elevation.

That signature is your HRTF.

Sound Park & the wider playground

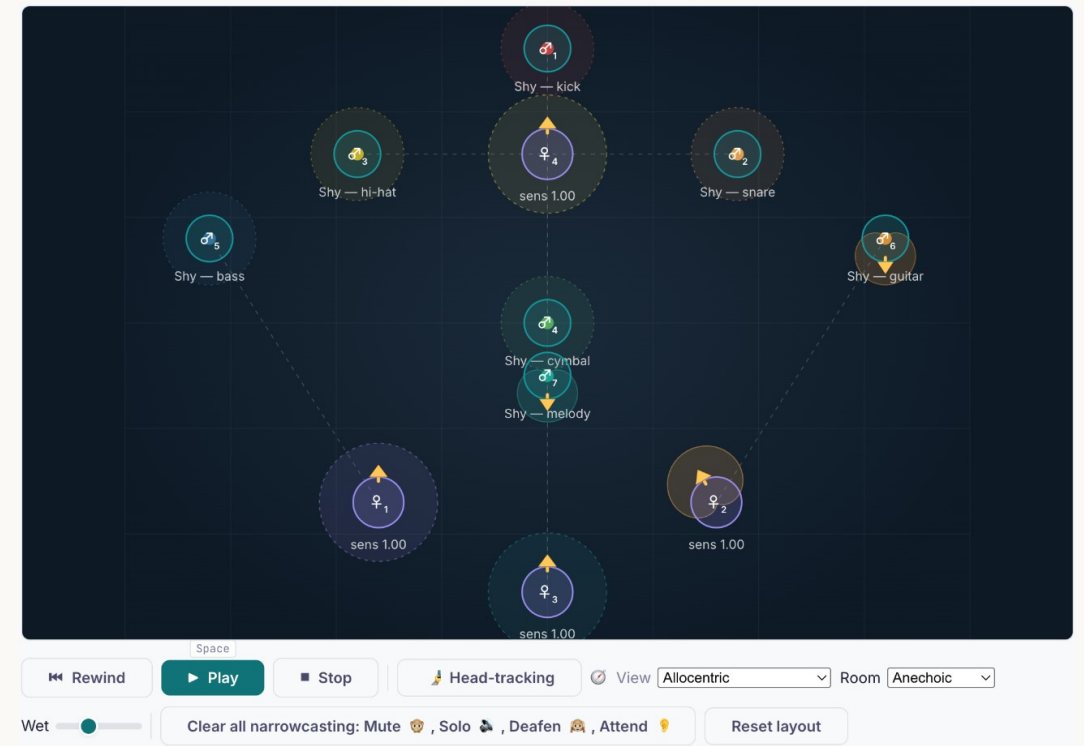
サウンド パーク

Klangpark



[Sonic: spatial sound — azimuth and range](#)

Sound Park puts several sources in one scene at once, so you can hear azimuth and distance interact — and hear how quickly a mix becomes unintelligible when everything is equally loud and equally close.



[Sonic: spatial sound playground](#)

PART FIVE

Narrowcasting

ナローキャスティング

Narrowcasting

Broadcast sends to everyone; unicast sends to one. Narrowcasting is the general case — and virtual environments need to say it precisely.

Broadcast, unicast, multicast, narrowcast

ほうそう こべつ たしや げんてい
放送・個別・多者・限定

Rundfunk bis Narrowcasting

Broadcast

Rundfunk

ほうそう

放送

One source, everybody receives.
No addressing at all.

Unicast

Punkt-zu-Punkt

こべつはいしん

個別配信

One source, one named recipient.
A telephone call.

Multicast

Mehrpunktverbindung

たちてんはいしん

多地点配信

One source, a subscribed group.
Membership is the only control —
you are in it or you are not.

Narrowcast

Narrowcasting

げんていはいしん

限定配信

Each source and each sink carries
its own predicate, so the decision
is per-pair and can change
moment to moment.

The same four operations, three times over — people have been narrowcasting by hand for a very long time.



三猿 Sanzarū — Nikkō Tōshō-gū, 17th c.



hand-tinted photograph, c. 1900



lenticular postcard — “think no evil”

Who hears whom

だれ だれ き

誰が誰を聞くか

Wer hört wen

mute

stummschalten

ミュート

Turn this source off for everyone.

select

auswählen

せんたく

選択

Turn everything else off — solo
this source.

deafen

taub schalten

ちょうしゅかいじょ

聴取解除

This listener stops receiving.

attend

zuhören

けいちょう

傾聴

This listener receives only the
chosen sources.

Two operations on each side

Sources can be muted (turned off) or selected (everything else turned off). Sinks can be deafened (stop receiving) or attended (receive only these). Exclusion is explicit; inclusion is exclusive.

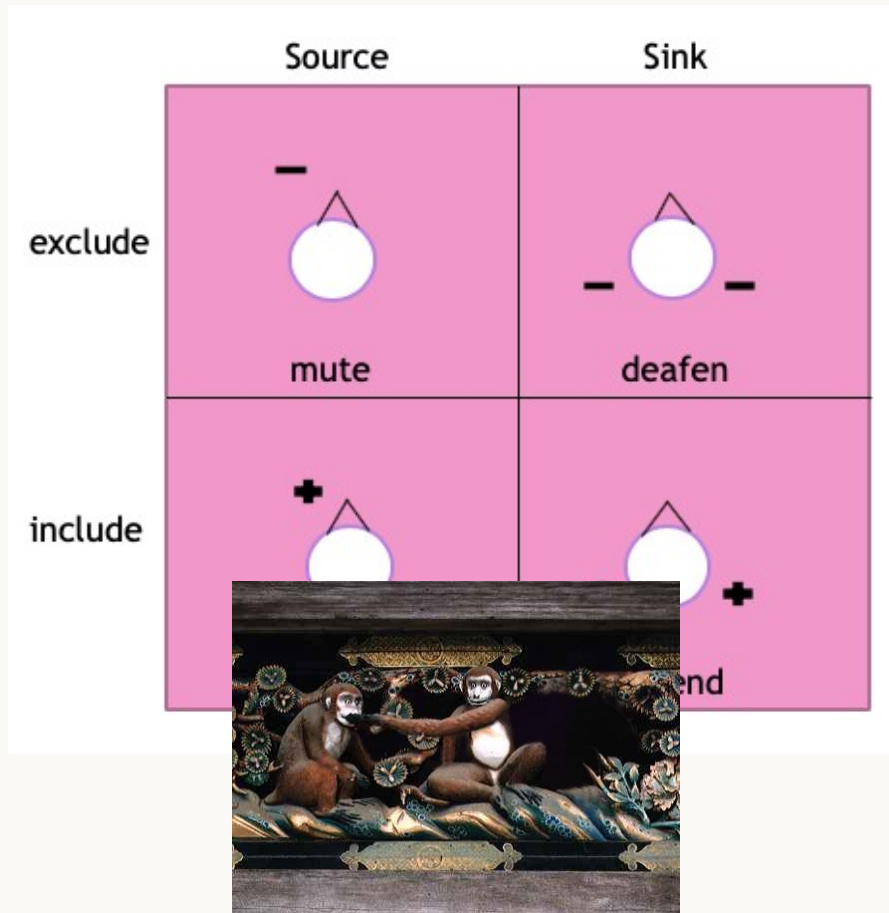
active(x) = $\neg$ exclude(x) $\wedge$ ($\exists$ y include(y) $\Rightarrow$ include(x))

An object is active if it has not been excluded, and — if anything at all has been explicitly included — it is one of the included. That single line covers mute/select for sources and deafen/attend for sinks.

The operations, as symbols

そうさ きごう
操作の記号

Die Operationen als Symbole



distal mute; 言わさせざる — silencing another

Michael Cohen; 公園マイケル · University of Aizu; 会津大学 · Higashi Nippon International University; 東日本国際大学 · HSD Summer School 2026

Two axes, four operations.

Across: is this a source (something heard) or a sink (someone listening)? Down: are we excluding, or including?

Exclusion is absolute — mute this, deafen that. Inclusion is exclusive: the moment anything is soloed or attended, everything not chosen falls silent.

The symbols are deliberately figurative: a minus by the mouth is mute, a minus by each ear is deafen. A user can read the state of the scene, without a legend.



distal deafen; 聞かせざる — preventing another from hearing

LIVE DEMO

Narrowcasting sandbox

えんしゅう
ナローキャスティング演習
Narrowcasting-Sandbox

mcohen.info/spatial-media/Sonic/narrowcasting.html

Narrowcasting Sandbox

What we will do

1. Start with several sources and several sinks audible.
2. Mute one source: it drops out for everyone.
3. Now select a different one — everything else falls silent.
4. Switch to the sink side: deafen, then attend.
5. Add directivity and hear the same controls behave spatially.

Listen for

The point is that presence and absence are both designed. A virtual meeting that cannot express “these three, privately” is missing something a physical room provides for free — you simply turn away and lower your voice.

PART SIX

Making it

せいさく

制作

Die Produktion

From raw material to a scene that a spatialiser can work with — and what the new generative tools change about the first step.

Prepare the stems

そざい じゅんび
素材の準備

Stems vorbereiten

Collect

Sammeln

そざい しゅうしゅう
素材収集

Record it, take it from a library, or synthesise it.

Keep it dry

Trocken lassen

ドライのまま

No baked-in reverb — the engine adds the space at run time.

“Dry” matters more than it sounds. If reverberation is baked into the file, the engine will add the room a second time — and the source will not move convincingly, because its reflections stay put.

Trim & normalise

Schneiden & normalisieren

せいおん せいぎか
整音・正規化

Remove leading and trailing silence; set a consistent peak level.

Export mono

Mono exportieren

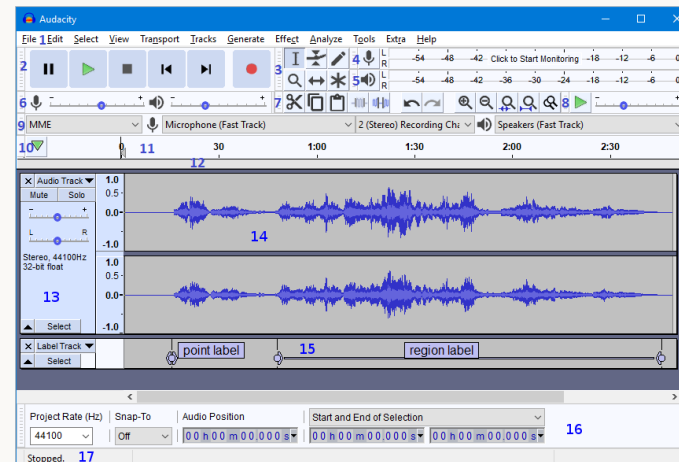
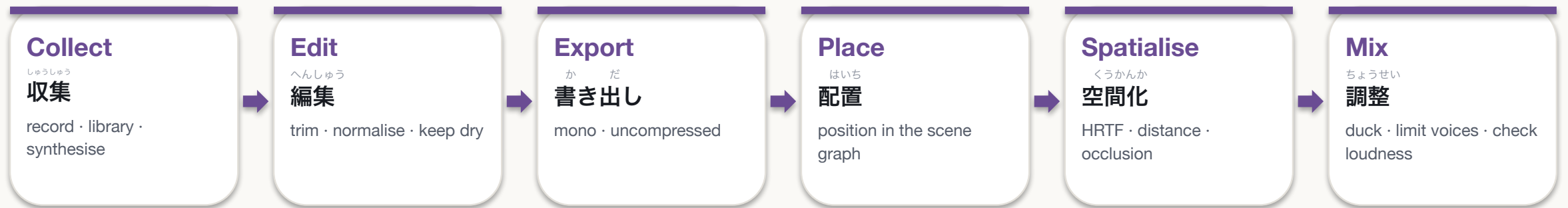
か だ
モノラル書き出し

A spatialiser needs one channel per source to place it in the scene.

The workflow

ワークフロー

Der Arbeitsablauf



Audacity — trimming and normalising a stem

Generating sound with AI

おと せいせい
A I による音の生成

Klangerzeugung mit KI

Literal

Wörtlich

しゃじつてき

写実的

“Footsteps on wet pavement, leather soles.”
Names the physical event.

Abstract

Abstrakt

ちゅうしょうてき

抽象的

“Granular noise fading into a smooth sine.”
Names the signal, not a source.

Poetic

Poetisch

してき

詩的

“Wind arguing with bamboo in a moonlit garden.” Names a feeling; useful when you want something you could not have specified.

Tools

[ElevenLabs — elevenlabs.io/sound-effects](https://elevenlabs.io/sound-effects)

[Canva SFX — www.canva.com/features/ai-sound-effect-generator](https://www.canva.com/features/ai-sound-effect-generator)

[PopPop AI — poppop.ai/ai-sound-effect-generator](https://poppop.ai/ai-sound-effect-generator)

[Remusic — remusic.ai/ai-sound-effect-generator](https://remusic.ai/ai-sound-effect-generator)

[SFX Engine — sfxengine.com](https://sfxengine.com)

What they are good for

Placeholder and concept material, fast: something to put in the scene today so the interaction can be tested. Poetic prompts are genuinely useful here — they return things you would not have thought to specify.

Weakest where it matters most: a specific object, recorded consistently, in sync with a picture. That is still Foley.

LIVE DEMO

Rhythm Playground

リズム プレイグラウンド

Rhythmus-Spielplatz

mcohen.info/spatial-media/Sonic/rhythm-playground.html

Rhythm Playground

What we will do

1. Build a pattern on the wheel and let it loop.
2. Move the room slider — the pattern gets wetter.
3. Push tempo until the separate hits fuse into a pitch.
4. Compare with the Shepard tone: an illusion of endless rise.

Listen for

Rhythm and pitch are the same phenomenon at different rates. Below about 20 Hz you count events; above it you hear a note. The slider crossing that boundary is one of the more startling things you can demonstrate in a browser.

PART SEVEN

Access

アクセシビリティ

Zugänglichkeit

Every meaning carried by sound needs a second route. This is not an add-on at the end; it is a constraint on the design from the start.

If it matters, it needs a second route

おといがい けいろ
音以外の経路

Ein zweiter Weg

Subtitles

Untertitel

じまく
字幕

Spoken words as text; dialogue only.

Audio description

Audiodeskription

おんせいかいせつ
音声解説

A spoken account of what is visible, for people who cannot see it.

Captions

Erweiterte Untertitel

ちょうかくじまく
聴覚字幕

Dialogue plus the non-speech sound that carries meaning.

Visual substitute

Visuelle Entsprechung

しかくだいたい
視覚代替

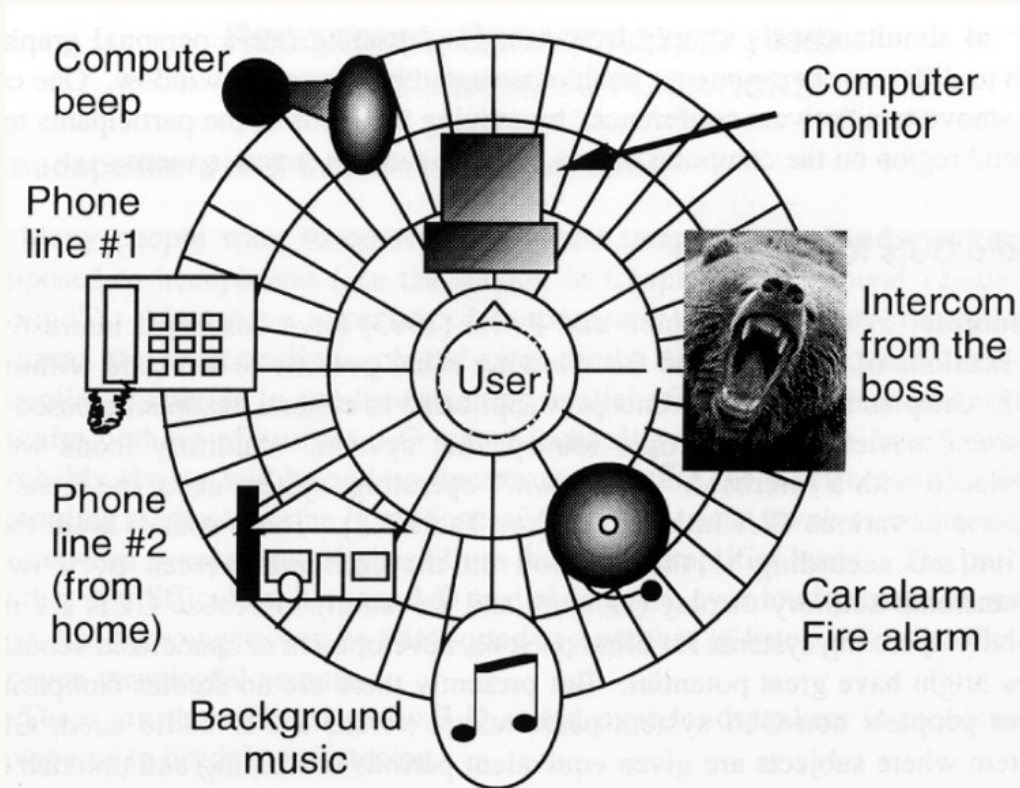
A directional indicator standing in for a sound cue.

Subtitles are not captions. Subtitles carry dialogue; captions also carry the door closing behind you. In a game where an approaching enemy is audible but not visible, that difference is the whole game.

Audio windowing

オーディオ ウィンドウ

Audio-Fenster



After Begault, 3D Sound for Virtual Reality and Multimedia

Windows, but for listening.

A graphical interface gives each task a frame you can move, resize, raise and close. Begault imagined the auditory equivalent: several soundscapes superimposed, each independently placed and attenuated.

Narrowcasting is the control surface for such a system — it is how you close an audio window without leaving the room.

This is also where accessibility and spatial audio meet: separating sources in space is one of the most effective ways to make speech intelligible in noise.

IN ONE MINUTE

Sound tells you where you are, what just happened, and how to feel about it.

おと いま なに お かん つた
音は、今どこにいるか、何が起こったか、どう感じるべきかを伝える。

Klang sagt Ihnen, wo Sie sind, was geschah und wie es sich anfühlt.

Three things worth keeping:

The room is not an effect — it is an answer. Reverberation and convolution are the same idea seen twice.

Space is a mixing tool. Separating sources by direction buys intelligibility that no equaliser can.

Design the silence and the absence too — who does not hear what is as much a decision as who does.

Everything is online

さんこうしりょう

参考資料

Alles online

Sonic — the whole site

mcohen.info/spatial-media/Sonic

Every demonstration, in the order they appear

[Sound check — selftest.html](#)

[The tuning fork — chapter1.html#ch1ToneGen](#)

[Seeing sound: the spectrogram — chapter7.html#spectrogram-waterfall-live](#)

[The siren outside this building — siren.html \(opens on the preset\)](#)

[The endless staircase — shepard.html](#)

[Convolution: the room as a filter — convolution.html](#)

[Spherical harmonics — spherical-harmonics.html](#)

[Gaussian splats: the same basis, seen — https://mcohen.info/spatial-media/GaussianSplats/](#)

[Reverberation — chapter7.html#reverb-playground](#)

[Convolution — convolution.html](#)

[Where the fingerprint comes from — chapter7.html#reflection-explorer](#)

[Spatial sound — chapter9.html#ch9Spatializer](#)

[Narrowcasting sandbox — narrowcasting.html](#)

[Rhythm Playground — rhythm-playground.html](#)

Also on Sonic

[Sound Park — sound-park.html](#)

[Delay / echo playground — chapter7.html#delay-echo-playground](#)

[Glossary — glossary.html](#)

[Keyboard & spectrogram — chapter1.html#ch1KeyboardDemo](#)

[Reverberation time & level — chapter7.html#reverberation-time-and-level](#)

[Beats — 440 vs 441 Hz — chapter1.html#ch1Beating](#)

Background reading

[Cohen & Martens, “Spatial Soundscape Superposition and Multimodal Interaction,” in The Technology of Binaural Understanding, Springer 2020, ch. 13.](#)

Begault, 3D Sound for Virtual Reality and Multimedia. ISBN 978-0120847358.

[Unity — importing audio files.](#)

[Slides and links: this PDF keeps every URL clickable. Questions afterwards: mcohen.info](#)