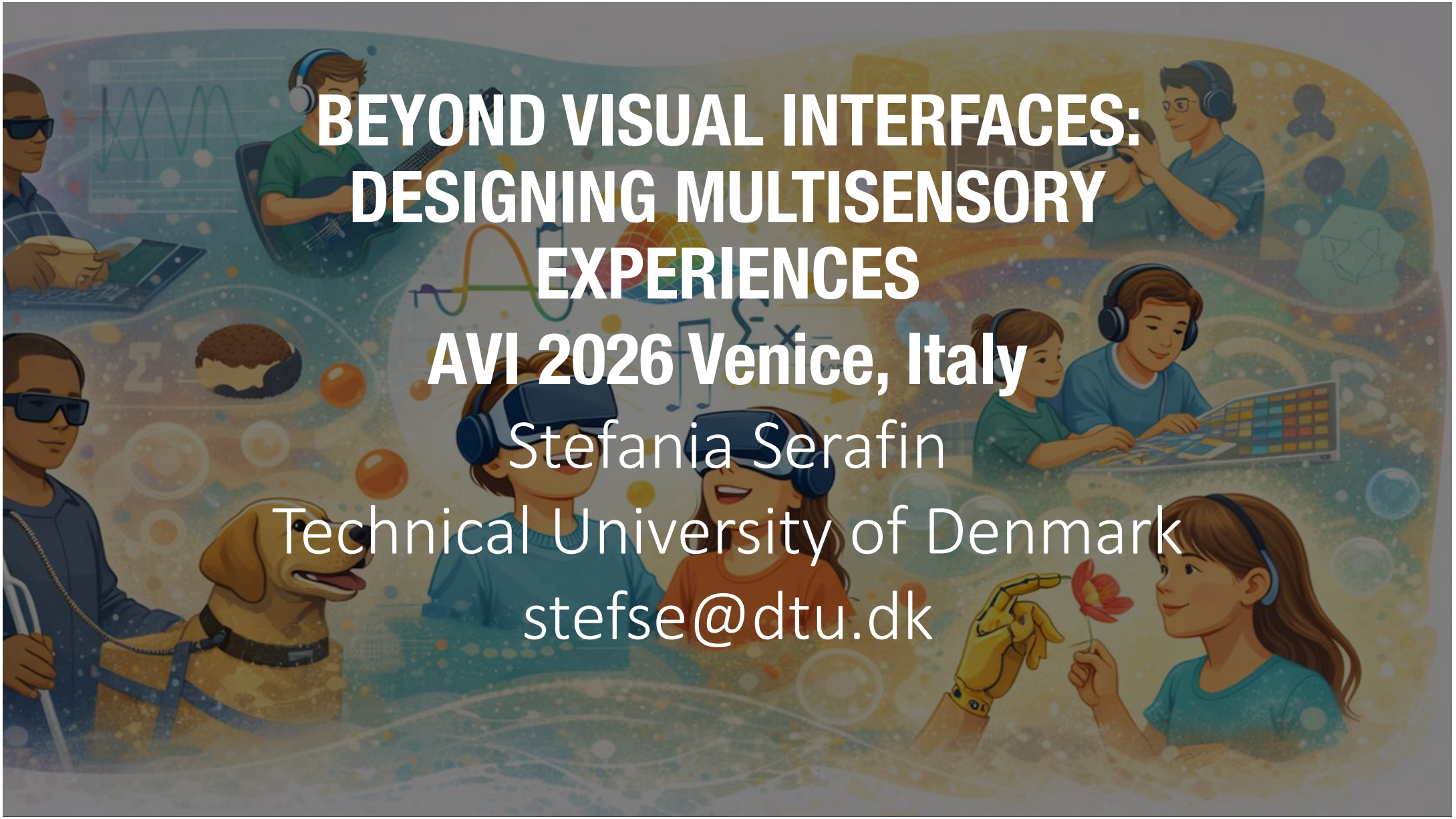


BEYOND VISUAL INTERFACES: DESIGNING MULTISENSORY EXPERIENCES

AVI 2026 Venice, Italy

Stefania Serafin
Technical University of Denmark
stefse@dtu.dk

WHAT IS YOUR FAVOURITE MULTISENSORY EXPERIENCE IN 2026?


REALITY
IS BETTER
THAN
VIRTUALITY


DTU
Danmarks Tekniske Universitet
Technical University of Denmark


MELAB

HOW MANY SENSES DO WE HAVE?


VISUAL
AUDITORY
TACTILE
OLFACTORY
GUSTATORY

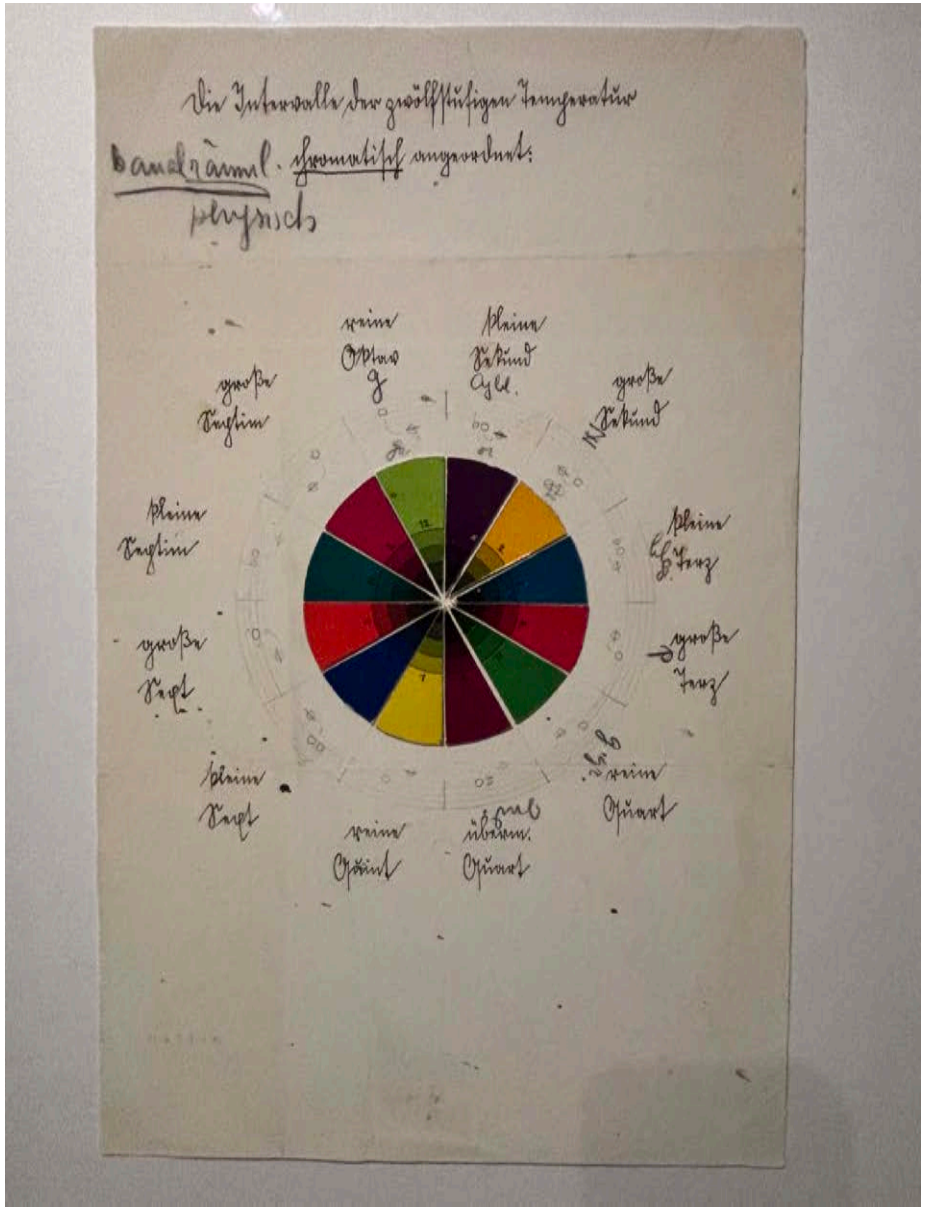
There are many opinions about how many senses we have

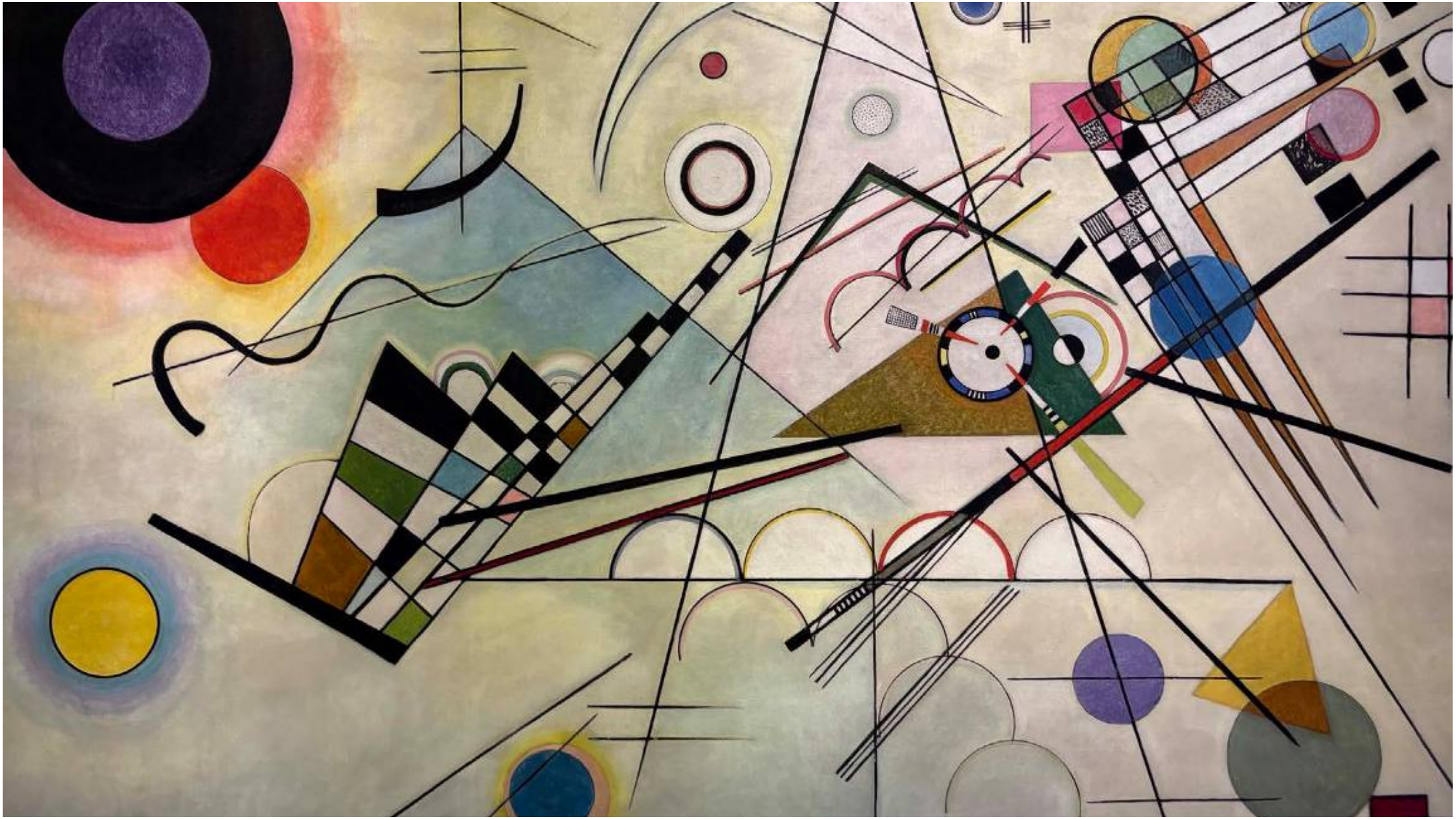
SENSORY MODALITY	Conservative	Accepted	Radical
Vision	■		
Light		■	■
Colour		■	
Red			■
Green			■
Blue			■
Hearing	■	■	■
Smell	■	■	
2000 or more receptor types			■
Taste	■		
Sweet		■	■
Salt		■	■
Sour		■	■
Bitter		■	■
Umami			■
Touch	■	■	
Light touch			■
Pressure			■
Pain	■	■	
Cutaneous			■
Somatic			■
Visceral			■
Mechanoreception	■		
Balance	■	■	
Rotational acceleration			■
Linear acceleration			■
Proprioception – joint position		■	■
Kinaesthesis		■	
Muscle stretch – Golgi tendon organs			■
Muscle stretch – muscle spindles			■
Temperature	■		
Heat		■	■
Cold		■	■
Interoceptors			
Blood pressure	■	■	
Arterial blood pressure			■
Central venous blood pressure			■
Head blood temperature			■
Blood oxygen content		■	■
Cerebrospinal fluid pH		■	■
Plasma osmotic pressure (thirst?)		■	■
Artery-vein blood glucose difference (hunger?)		■	■
Lung inflation		■	■
Bladder stretch			■
Full stomach			■
TOTAL	10	21	33

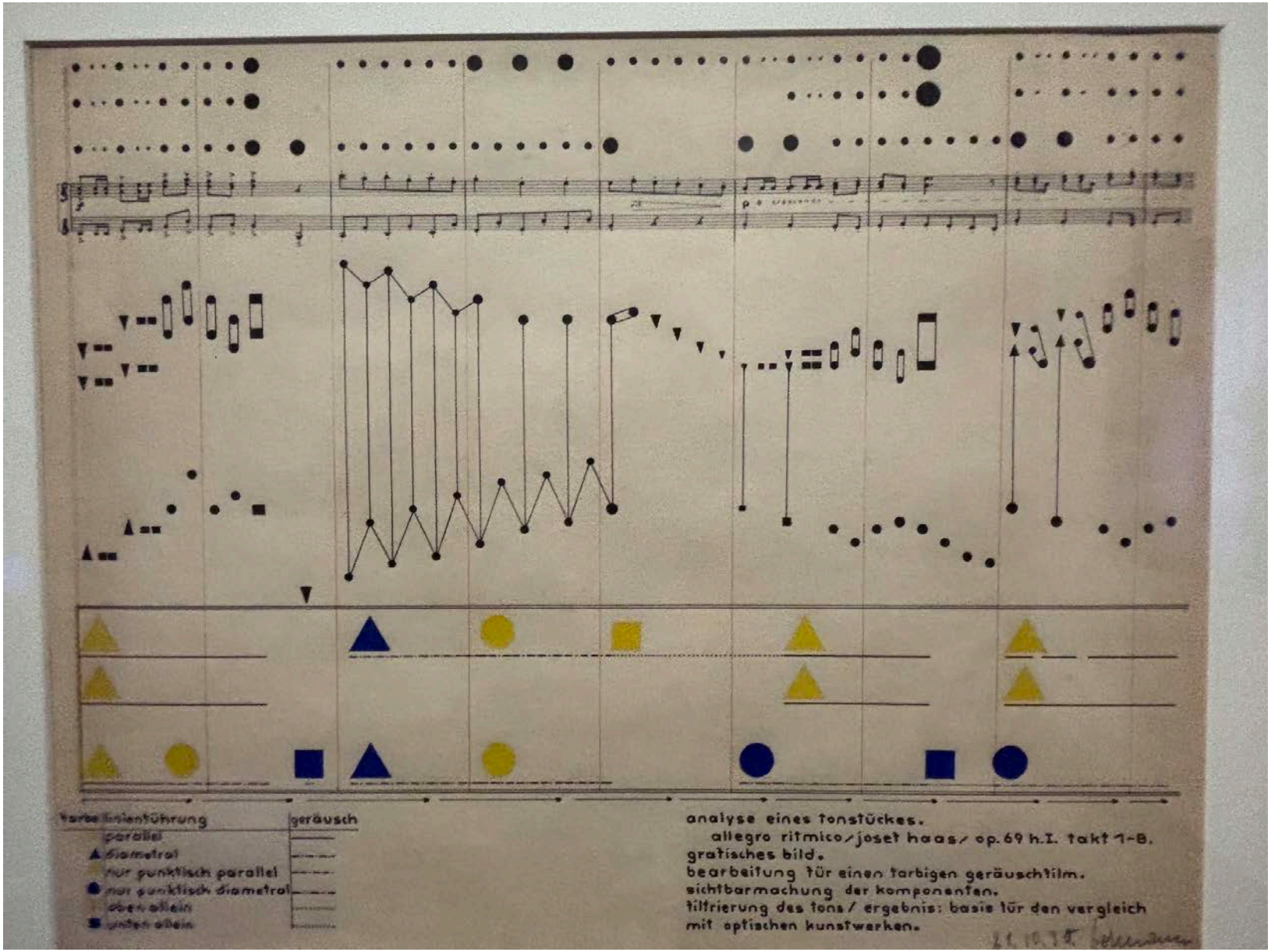
New Scientist article, 29 Jan 2005 by Bruce Durie


VISUAL
AUDITORY
TACTILE

MULTISENSORY ART





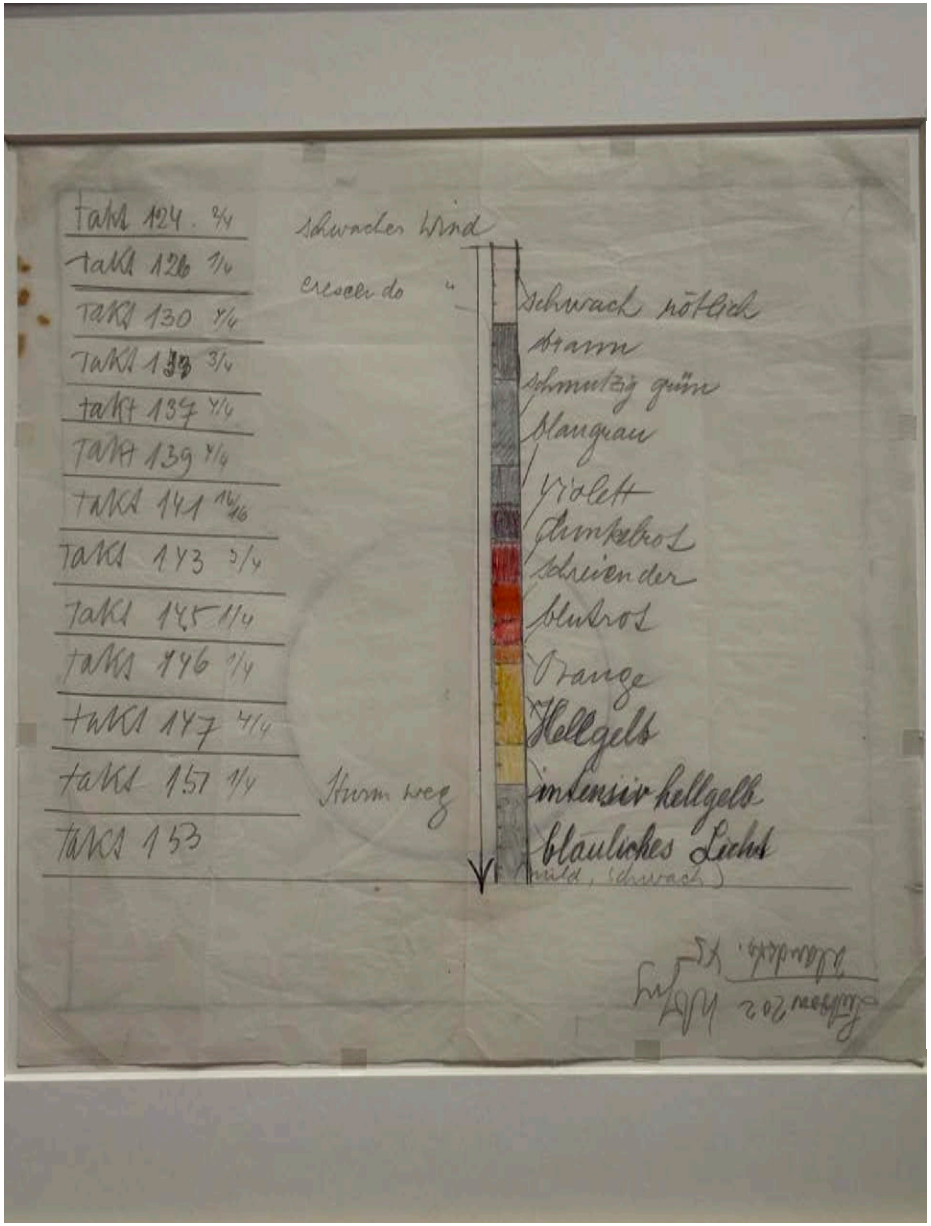


analyse eines tonstückes.
allegro ritmico / joset haas / op. 69 h.I. takt 1-8.
gratisches bild.
bearbeitung für einen farbigen geräuschfilm.
sichtbarmachung der komponenten.
filtrierung des tons / ergebnis: basis für den vergleich
mit optischen kunstwerken.

Die Intervalle der zwölfstufigen Tonskala

Dauerräume. schematisch angeordnet:

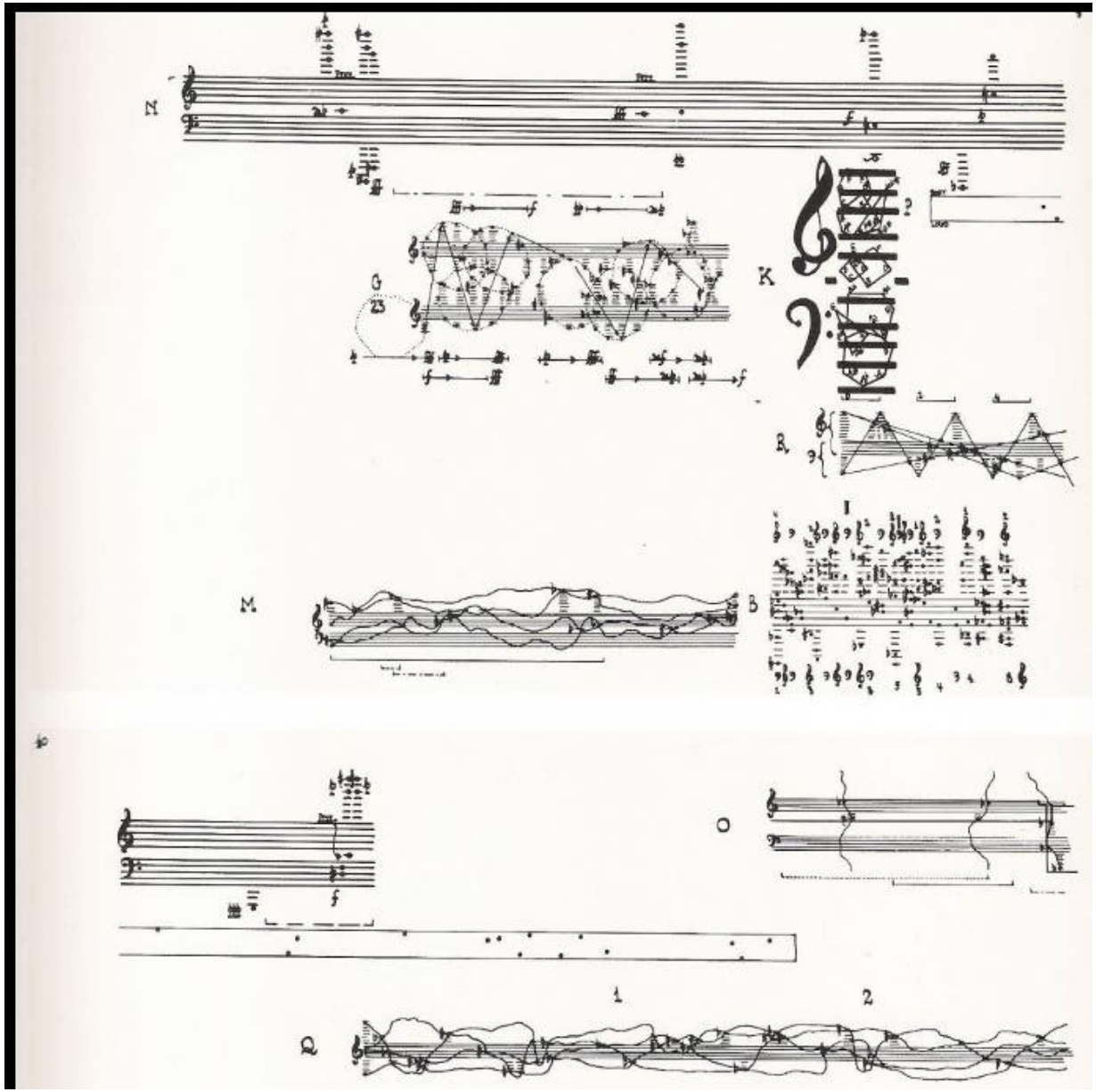
physisch



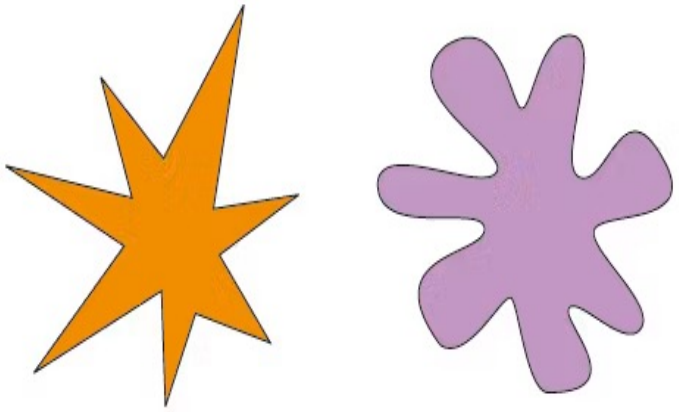
COLOR / CHARACTERISTIC / TONE

YELLOW	"warm," "cheeky and exciting," "disturbing for people," "typical earthly color," "compared with the mood of a person it could have the effect of representing madness in color [...] an attack of rage, blind madness, maniacal rage.	loud, sharp trumpets, high fanfares
BLUE	deep, inner, supernatural, peaceful "Sinking towards black, it has the overtone of a mourning that is not human." "typical heavenly color"	light blue: flute darker blue: cello darkest blue of all: organ
GREEN	mixture of yellow and blue. stillness, peace, but with hidden strength, passive "Green is like a fat, very healthy cow lying still and unmoving, only capable of chewing the cud, regarding the world with stupid dull eyes."	quiet, drawn-out, middle position violin
WHITE	"It is not a dead silence, but one pregnant with possibilities." "Harmony of silence",	"Harmony of silence", "pause that breaks temporarily the melody"
BLACK	"Not without possibilities [...] like an eternal silence, without future and hope." Extinguished, immovable.	"final pause, after which any continuation of the melody seems the dawn of another world"
GREY	mixture of white and black	soundlessness.
RED	alive, restless, confidently striving towards a goal, glowing, "manly maturity" Light warm red: strength, energy, joy; Vermilion: glowing passion, sure strength Light cold red: youthful, pure joy, young	"sound of a trumpet, strong, harsh" Fanfare, Tuba deep notes on the cello high, clear violin
BROWN	mixture of red + black dull, hard, inhibited	
ORANGE	mixture of red + yellow radiant, healthy, serious	middle range church bell, alto voice, "an alto violin, singing tone, largo"
VIOLET	mixture of red + blue "morbid, extinguished [...] sad"	english horn, shawm, bassoon

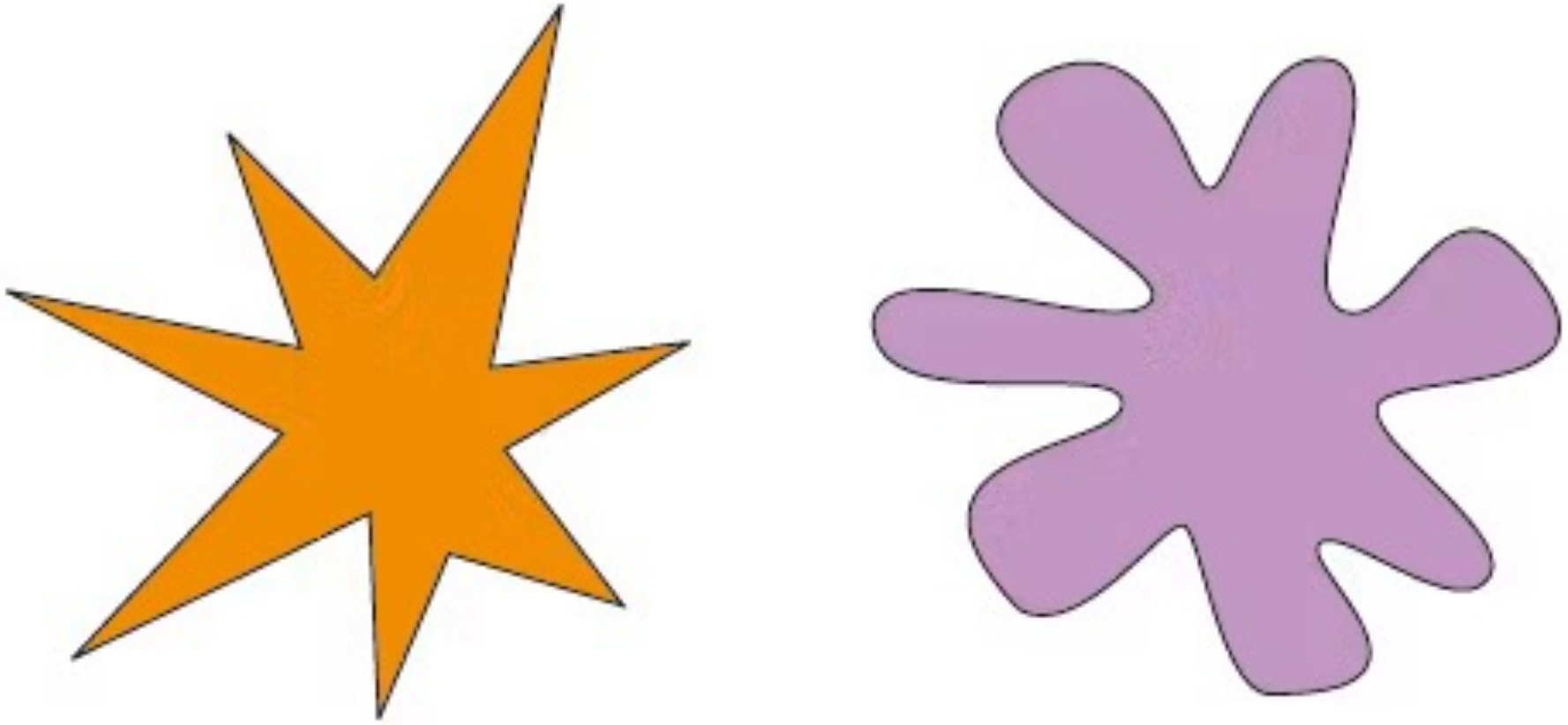
From Wassily Kandinksy, *Concerning the Spiritual in Art*, 1912


N
G
23
K
R
I
M
B
O
Q
1
2

CROSSMODAL CORRESPONDANCES



BOUBA KIKI OR TAKETE MALUMA?



1929, Wolfgang Köhler

Matching sounds to shapes: Evidence of the Bouba-Kiki effect in naïve baby chicks

[1*]Maria Loconsole, [2]Silvia Benavides-Varela, [1]Lucia Regolin

[1]Department of General Psychology, University of Padova, Padova, Italy

[2]Department of Developmental Psychology and Socialisation, University of Padova, Padova, Italy

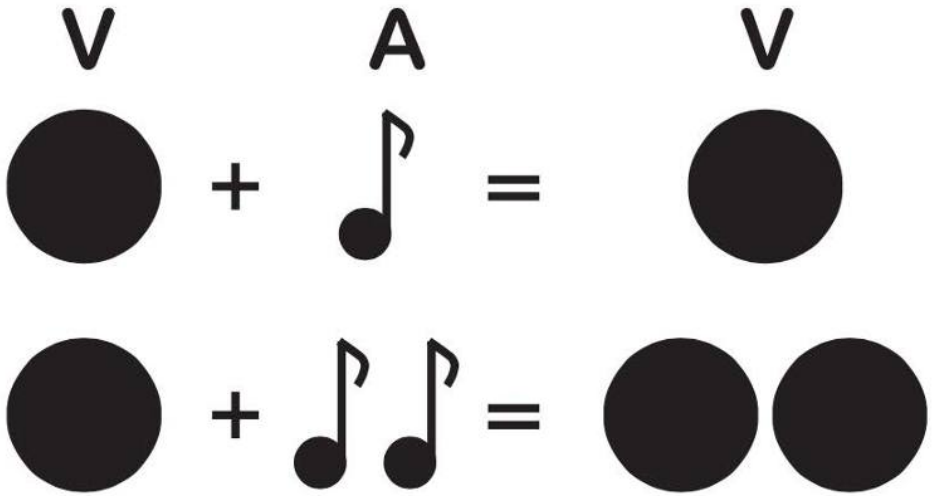
* Corresponding author: Maria Loconsole
Email: maria.loconsole@unipd.it

Keywords: Sound-symbolism; Crossmodal associations; Bouba-Kiki effect; Domestic chicken; Predispositions

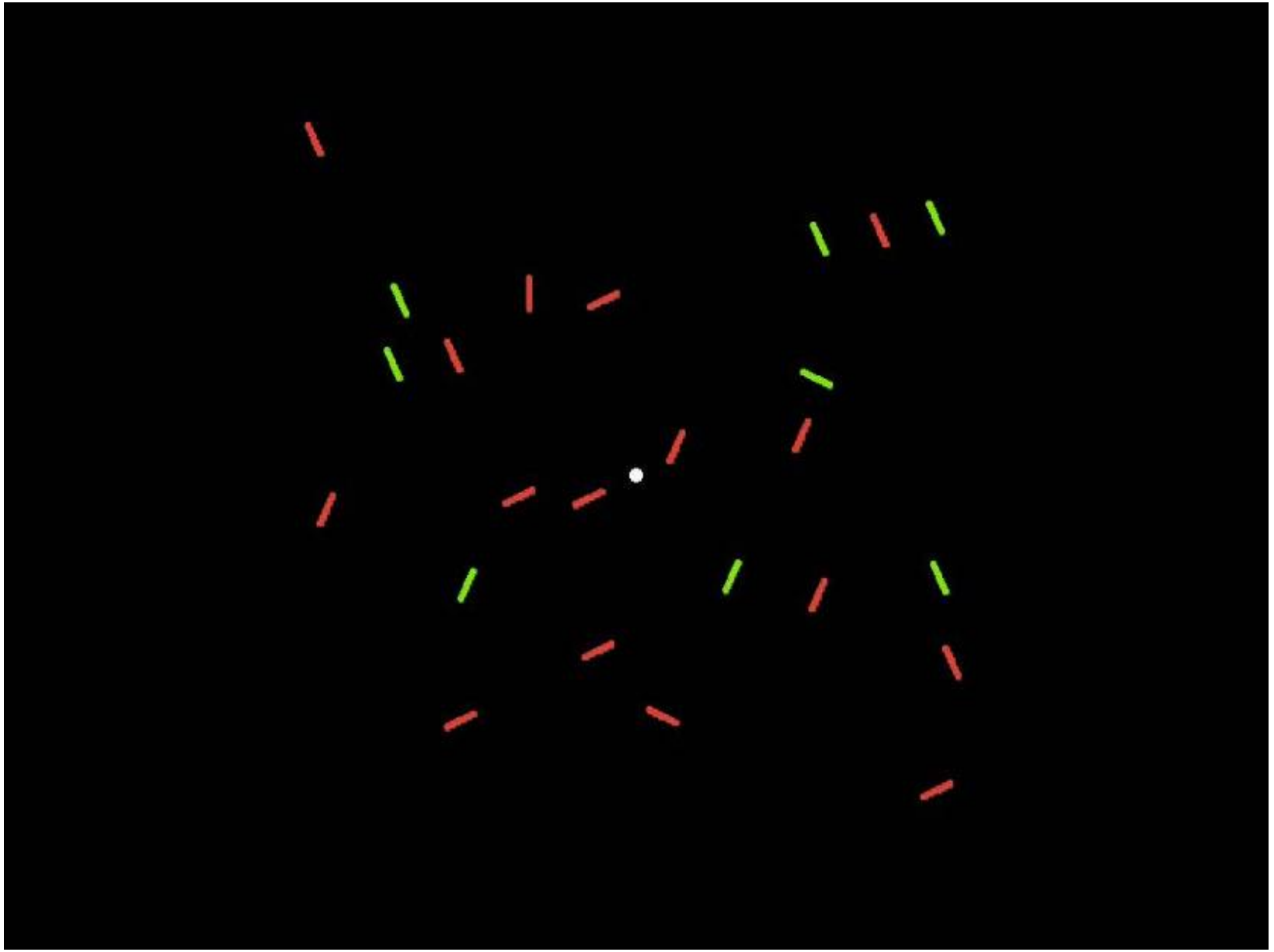
You are about to be FLASHED with black DISKS

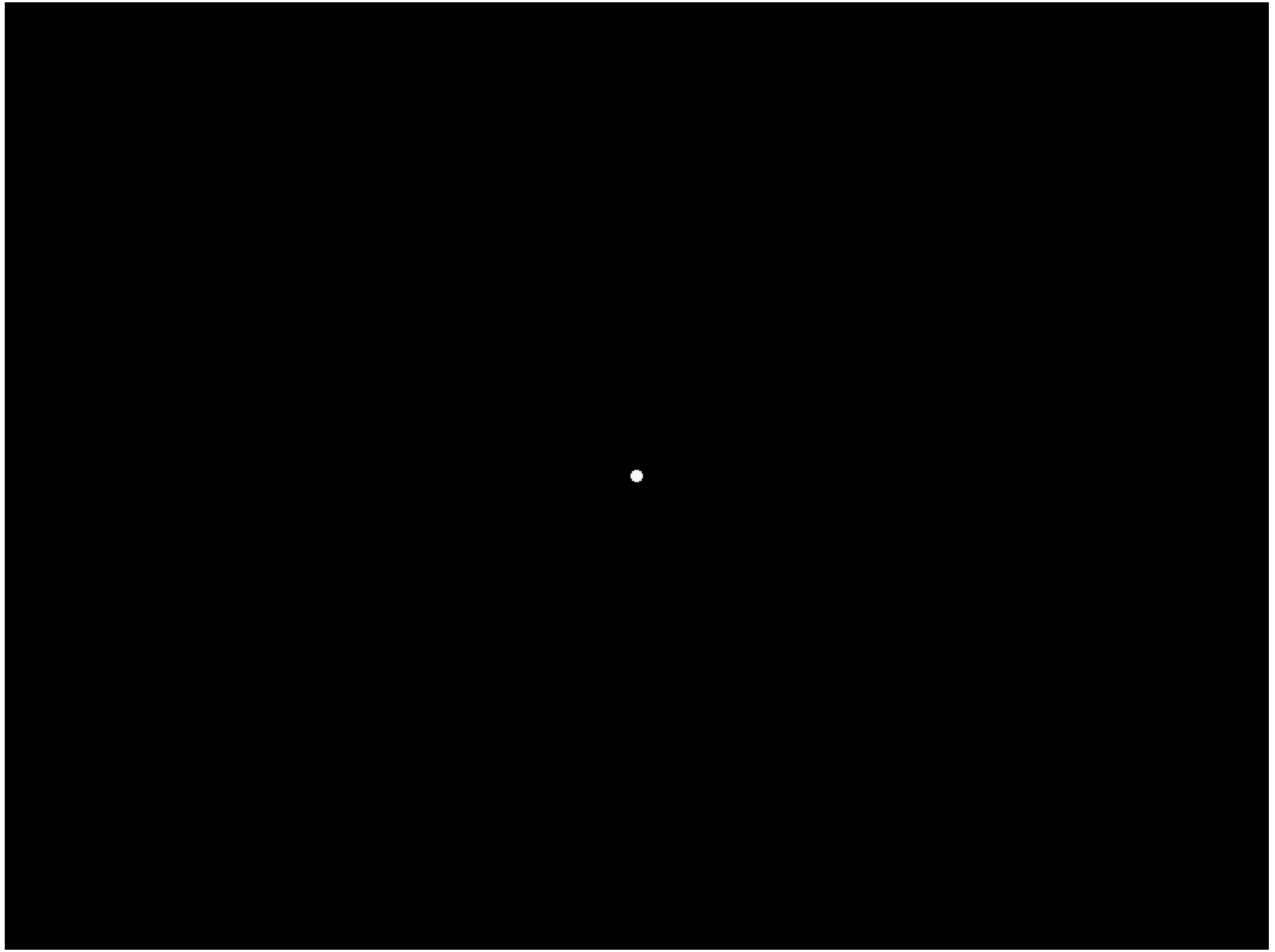
How many times does each disk flash?

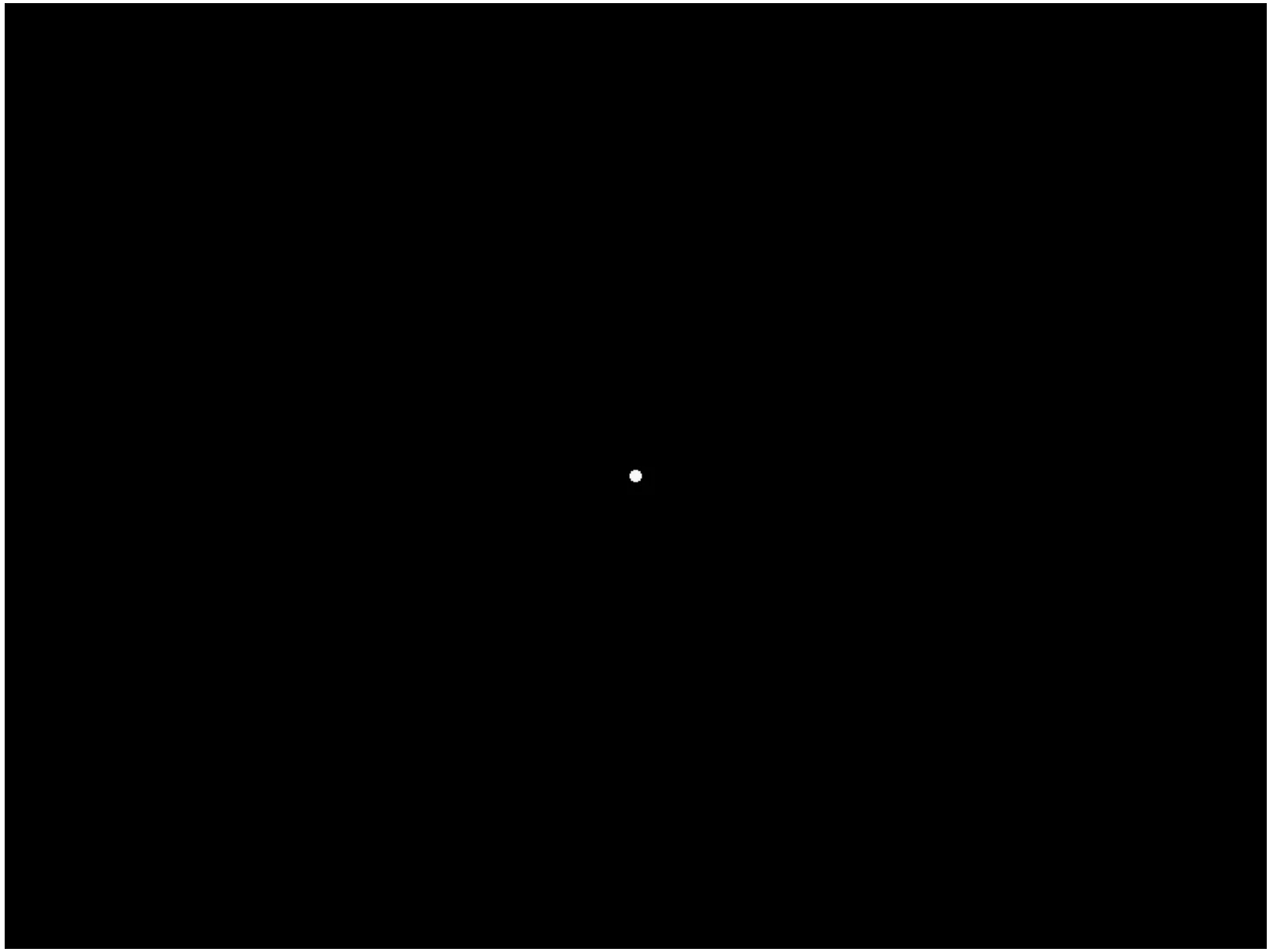
ILLUSIONARY FLASHES

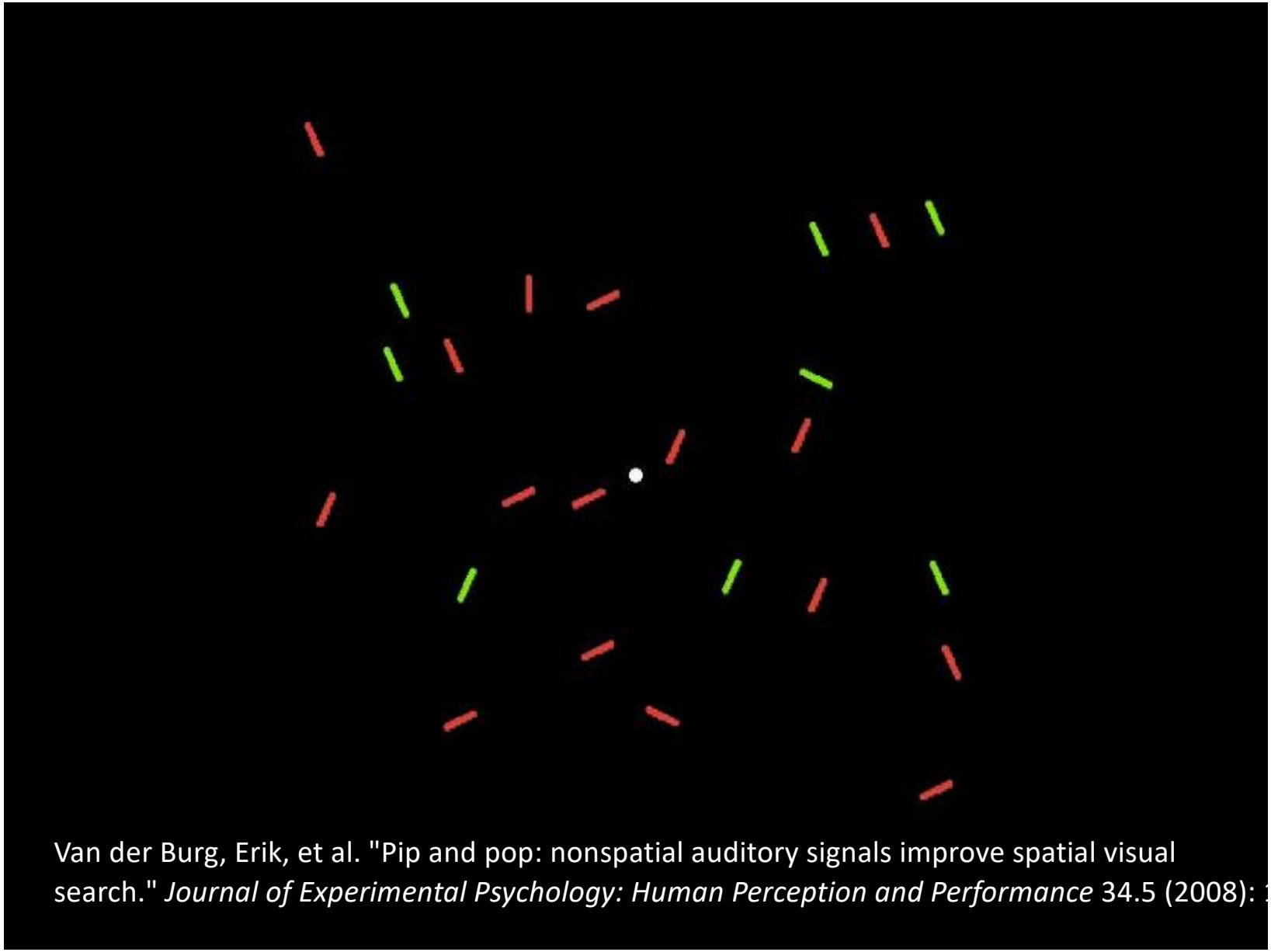


Shams, Ladan, Yukiyasu Kamitani, and Shinsuke Shimojo. "What you see is what you hear." *Nature* 408.6814 (2000): 788-788.









Van der Burg, Erik, et al. "Pip and pop: nonspatial auditory signals improve spatial visual search." *Journal of Experimental Psychology: Human Perception and Performance* 34.5 (2008): 1

Missing The Point: An Exploration of How to Guide Users' Attention During Cinematic Virtual Reality

Lasse T. Nielsen, Matias B. Møller, Sune D. Hartmeyer, Troels C. M. Ljung*, Niels C. Nilsson, Rolf Nordahl, and Stefania Serafin†
Aalborg University Copenhagen

Abstract

Recent technological advances have brought Virtual Reality (VR) into the homes of consumers, and there is a growing interest in bringing cinematic experiences from the screen and into VR. However, cinematic VR limits filmmakers' ability to effectively guide the audience's attention. In this paper we present a taxonomy of approaches to guiding users' attention, and present a study comparing two such approaches with a control condition devoid of guidance. One approach guides users by controlling their body's orientation, and the other implicitly directs their attention by encouraging them to follow a firefly with their gaze. The results revealed interesting, albeit statistically insignificant, indications that assuming control of the user's action may negatively influence presence, whereas the firefly was perceived as significantly more helpful.

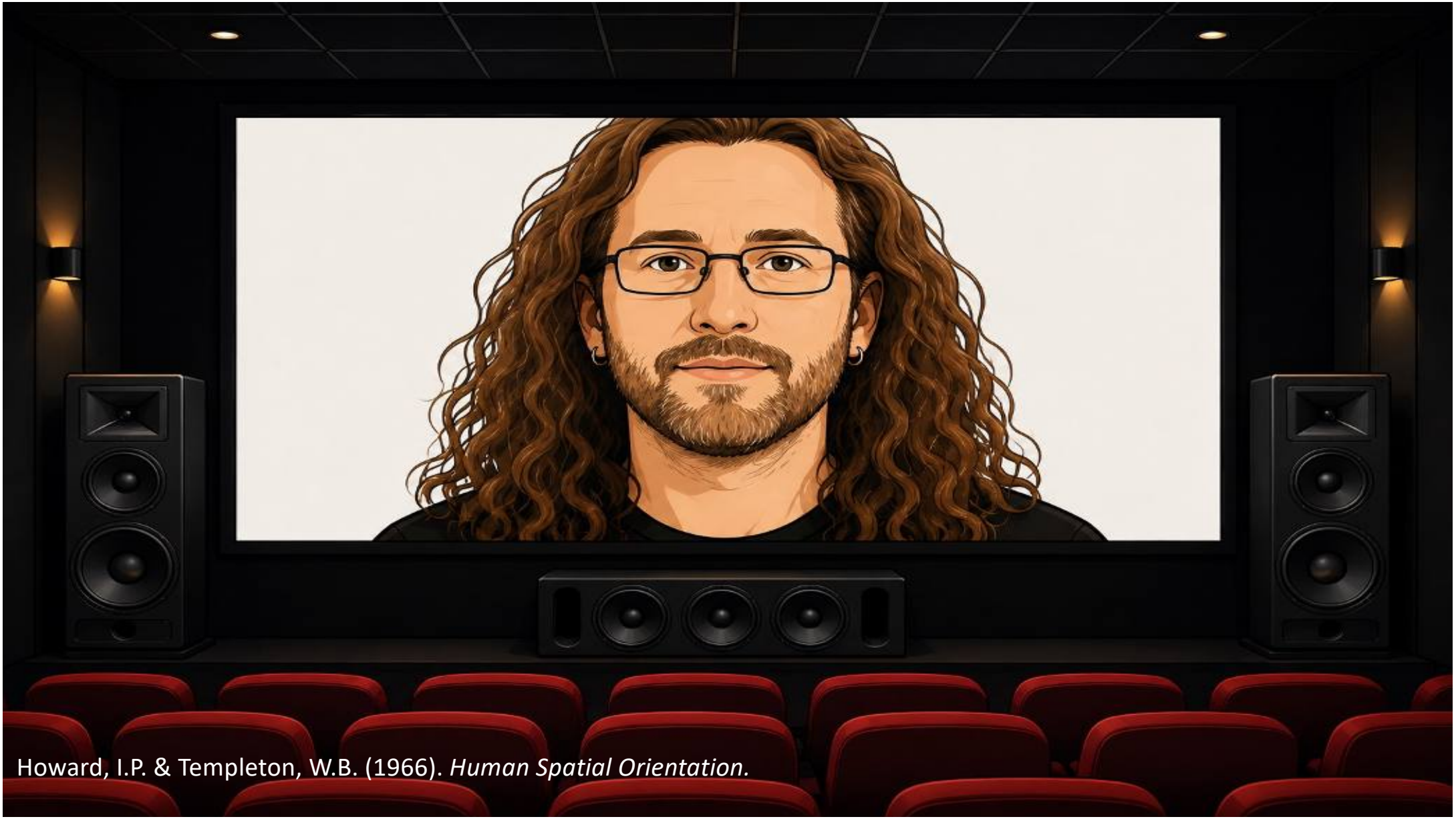
Keywords: virtual reality; film, attention; presence; cinematic VR

Concepts: •**Computing methodologies** → **Virtual reality;** •**Human-centered computing** → *Empirical studies in HCI;*

1 Introduction

Most contemporary filmmakers adhere to conventions that have simply by turning their heads, or even change the vantage point by moving around the scene, and it remains to be seen if editing can be used effectively in relation to cinematic VR without disorienting the audience. This inability to control the audience's attention naturally poses a problem when one aspires to create a coherent narrative since coherence is contingent upon "careful selection and presentation of actions whose causal and temporal relationships highlight an underlying plot" [Young 2000]. In relation to cinematic VR, the filmmaker can no longer rely on cinematography to show the audience the building blocks of the plot since the camera is identified with the user and no longer is under authoral control [Aylett and Louchart 2003]. In relation to pre-authored narratives this leaves the filmmaker with at least three different, albeit not mutually exclusive, options: 1) progression of the story is halted until the user's head or gaze direction makes it reasonable to assume that important events and objects have been observed; 2) the system dynamically presents events and objects within the user's field of view; and 3) the filmmaker uses cues to steer the user's attention towards relevant events and objects (e.g., using mise-en-scène and sound). The work documented in the current paper is focussed on the third option. Section 2 outlines a novel taxonomy of different categories of cues which can be used to guide users' attention during cinematic VR, and section 3 presents a user study aimed at exploring how two different types of cues influence the sensation of presence in the virtual environment (VE) and recollection of the scene. Finally,

McGurk, Harry, and John MacDonald. "Hearing lips and seeing voices." *Nature* 264.5588 (1976): 746-748.



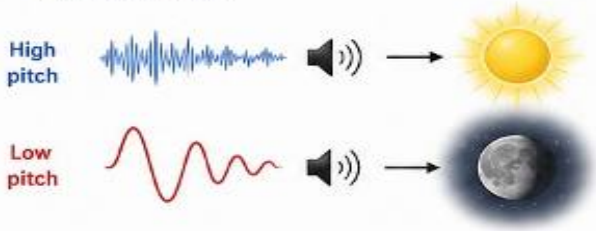
Howard, I.P. & Templeton, W.B. (1966). *Human Spatial Orientation*.

1. PITCH ↔ SIZE

High-pitched sounds are associated with small objects; low-pitched sounds with large objects.

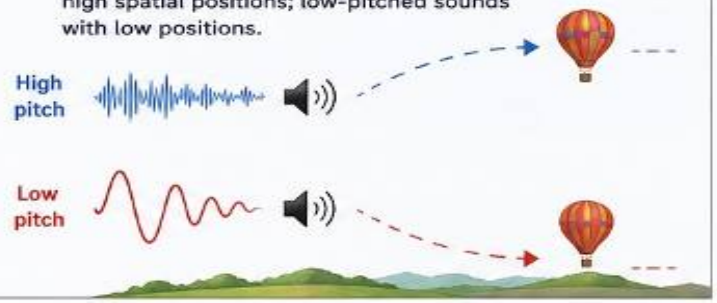
2. PITCH ↔ BRIGHTNESS (LIGHTNESS)

High-pitched sounds are matched with bright or light stimuli; low-pitched sounds with dark stimuli.



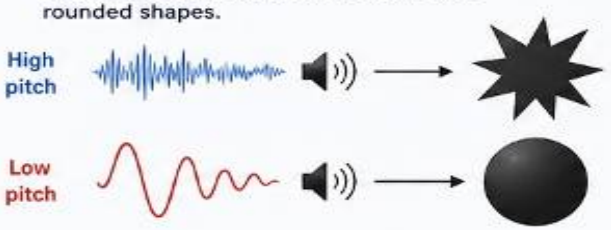
3. PITCH ↔ SPATIAL ELEVATION

High-pitched sounds are associated with high spatial positions; low-pitched sounds with low positions.



4. PITCH ↔ SHAPE

High-pitched sounds are associated with angular shapes; low-pitched sounds with rounded shapes.



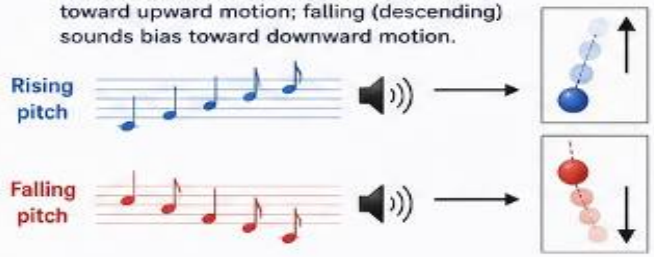
5. SOUND SYMBOLISM: BOUBA / KIKI EFFECT

People consistently match invented words with shapes.

Most people match "bouba" with rounded shapes and "kiki" with spiky shapes.

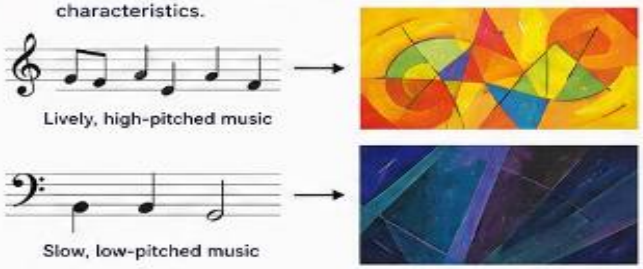
6. PITCH ↔ MOTION DIRECTION

Rising (ascending) sounds bias perception toward upward motion; falling (descending) sounds bias toward downward motion.



7. MUSIC ↔ VISUAL ART / COLOR

People reliably match music with paintings or colors and visual forms with musical characteristics.



Lively, high-pitched music

Slow, low-pitched music

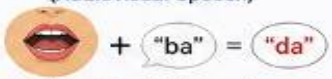
8. TASTE ↔ SOUND

Tastes and flavors are associated with particular sound characteristics.

Flavors can also be matched with musical characteristics (e.g., brighter vs. darker timbre).

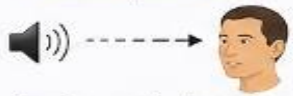
9. CLASSIC MULTISENSORY ILLUSIONS & EFFECTS

McGurk Effect (Audiovisual Speech)



What we hear is influenced by what we see.

Ventriloquism Effect



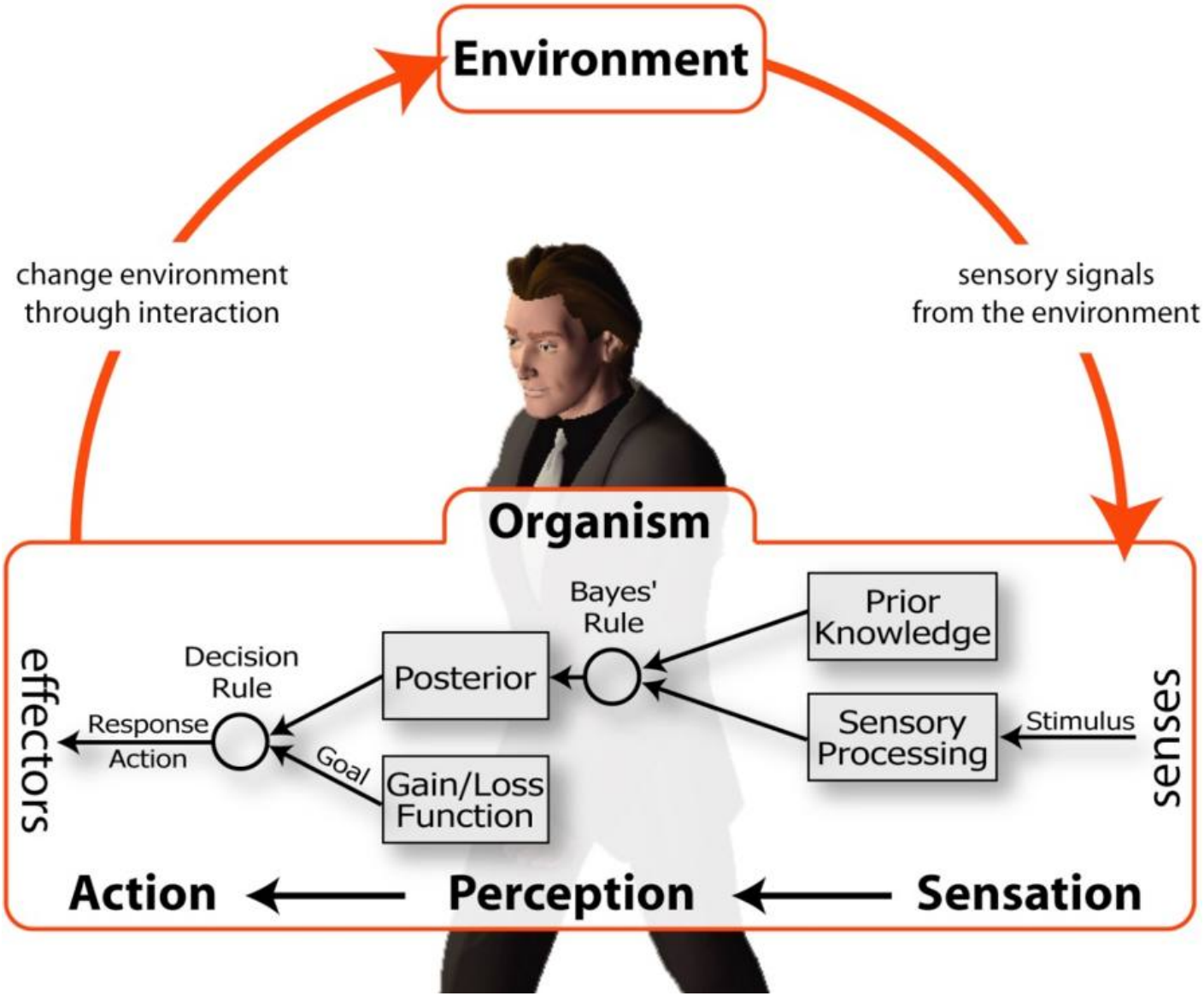
Sounds are perceived as coming from the visual location.

Pip-and-Pop Effect

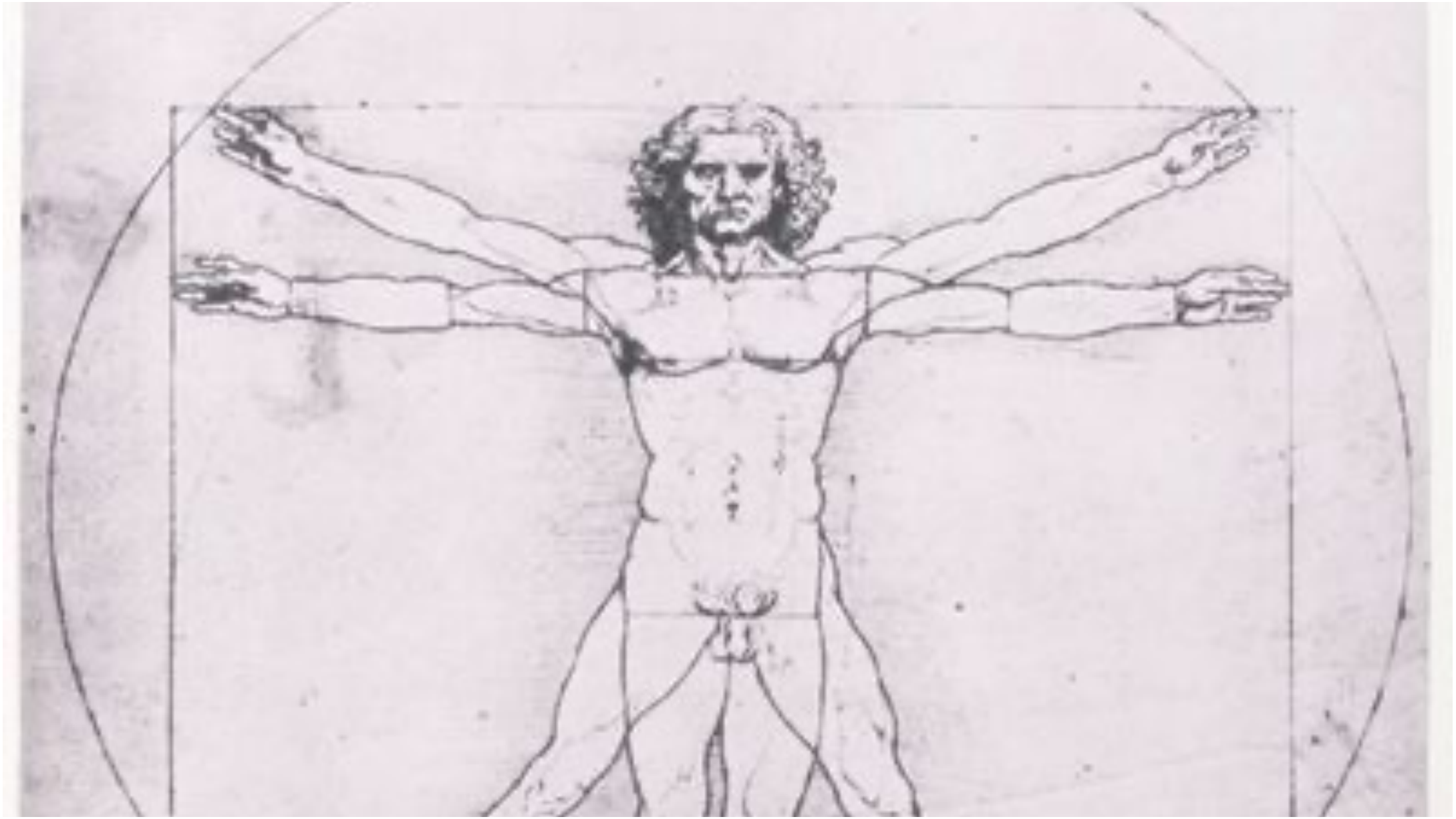
A sound ("pop") makes a brief visual cue ("pip") easier to find.

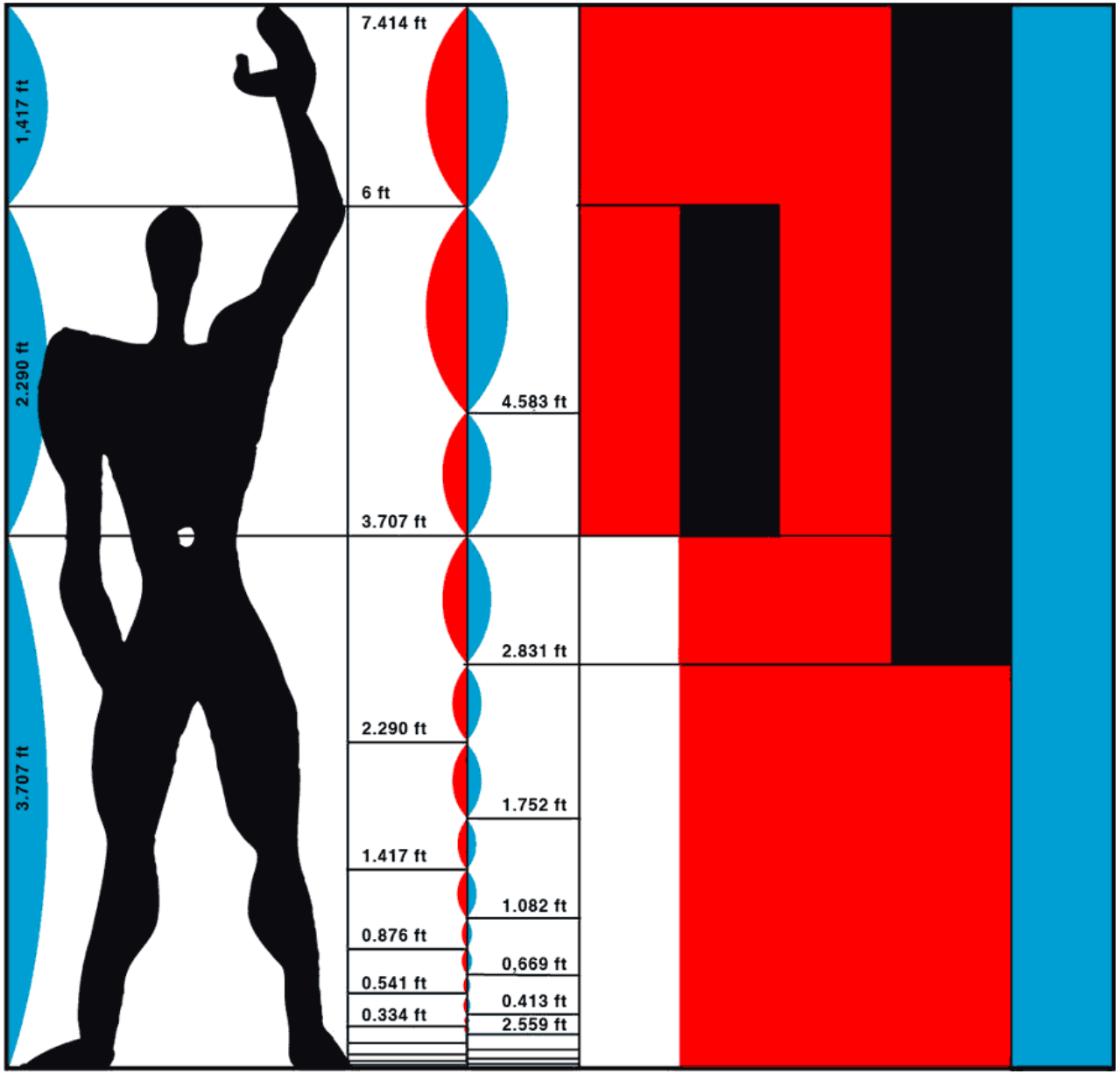
Cross-modal Plasticity

In deaf individuals, visual information can recruit auditory cortex.



Ernst, Marc O. "A Bayesian view on multimodal cue integration." *Human body perception from the inside out* 131 (2006): 105-131.



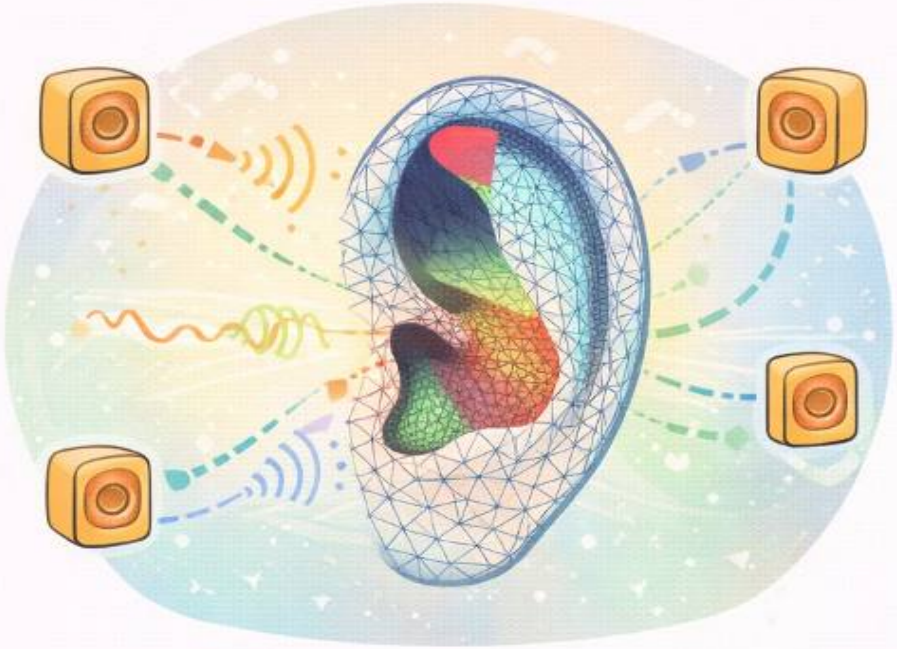


Modulor man Lecorbusier


HATS I

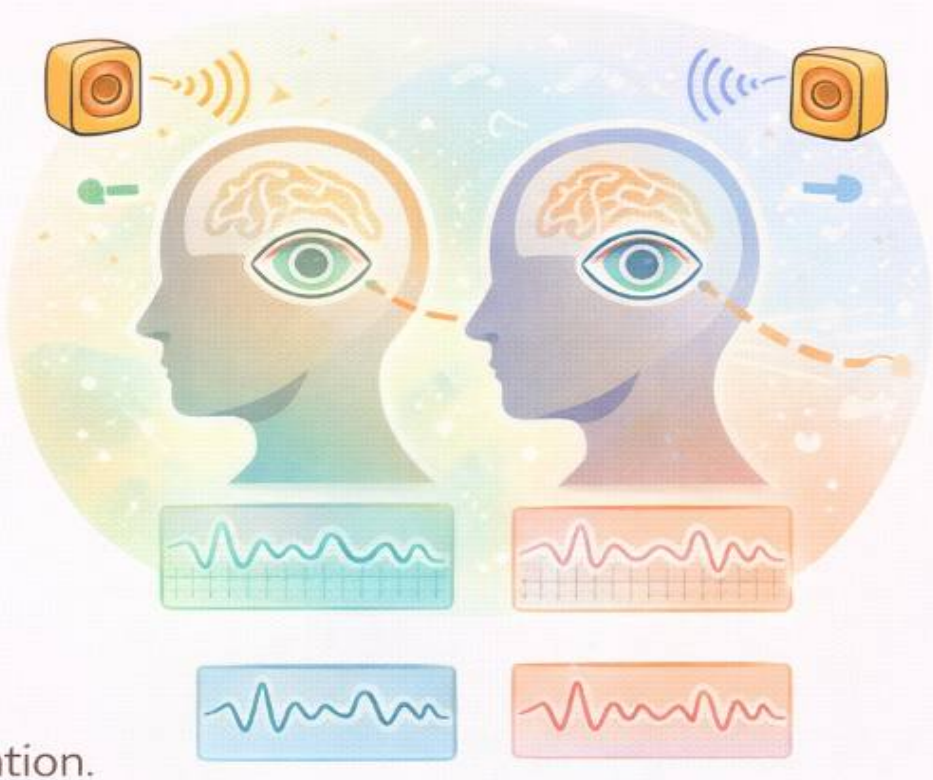


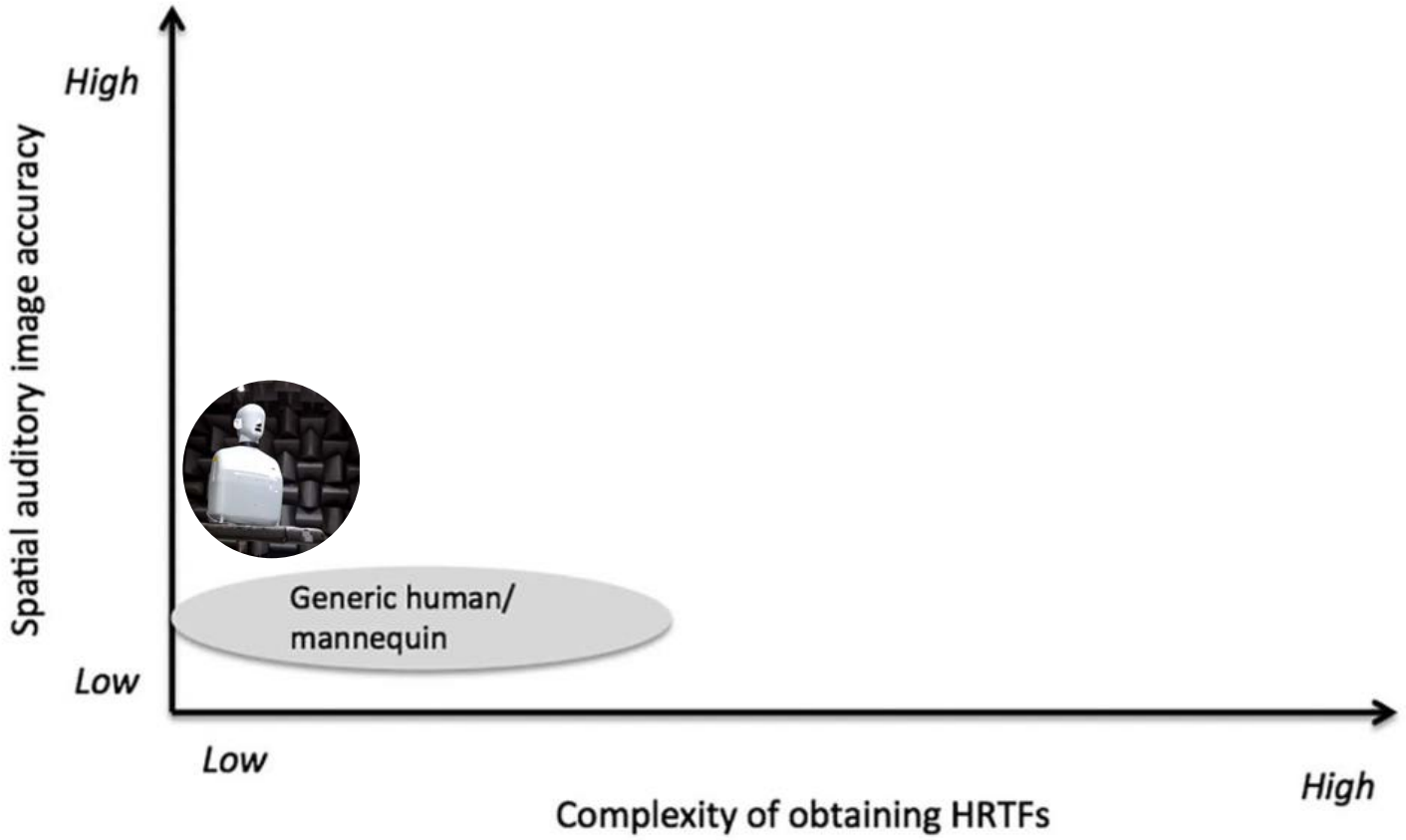
Your Ear is an Acoustic Fingerprint



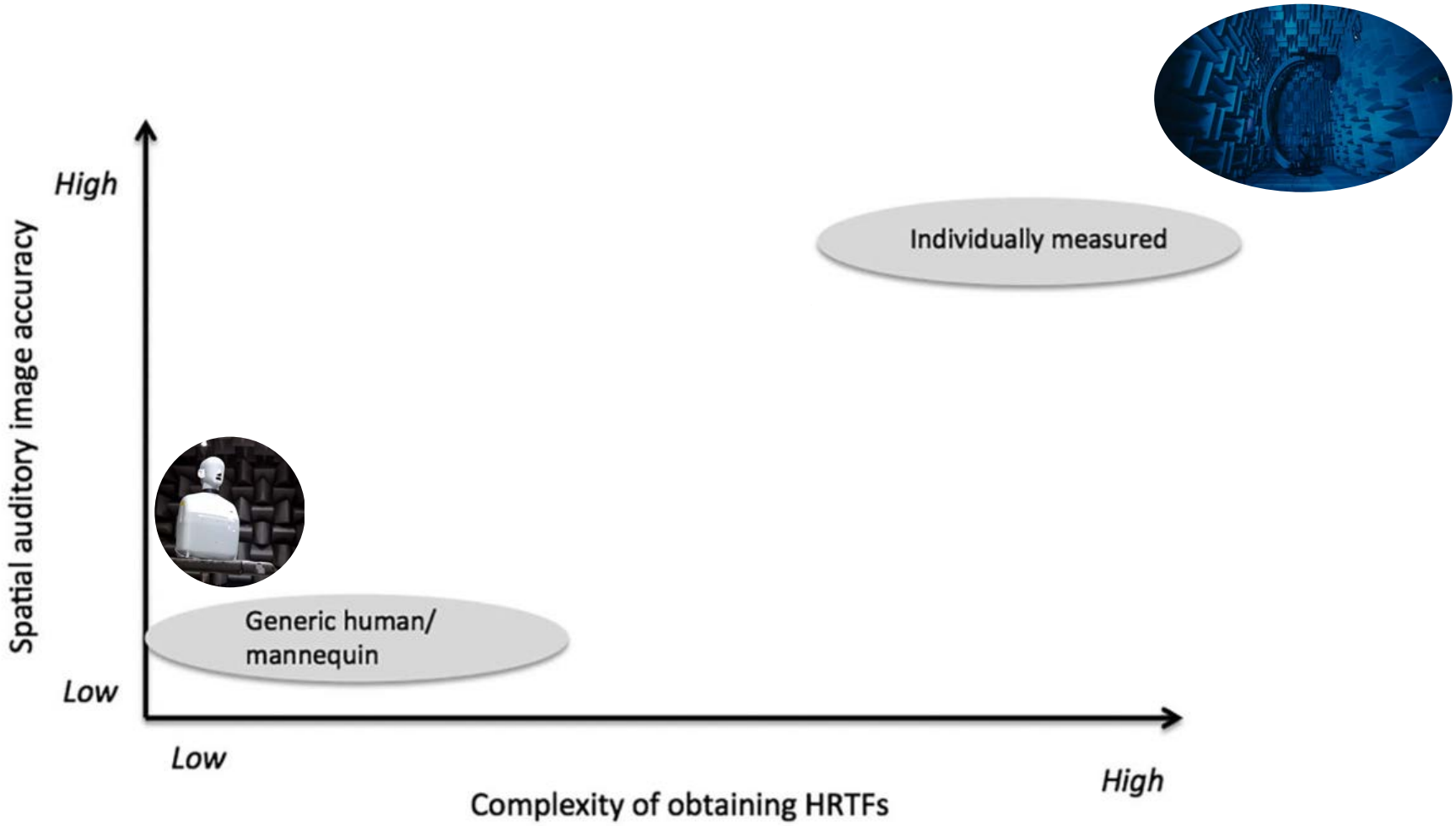
The ear acts as a directional acoustic filter that encodes spatial information.

Generic HRTF vs Personalized HRTF



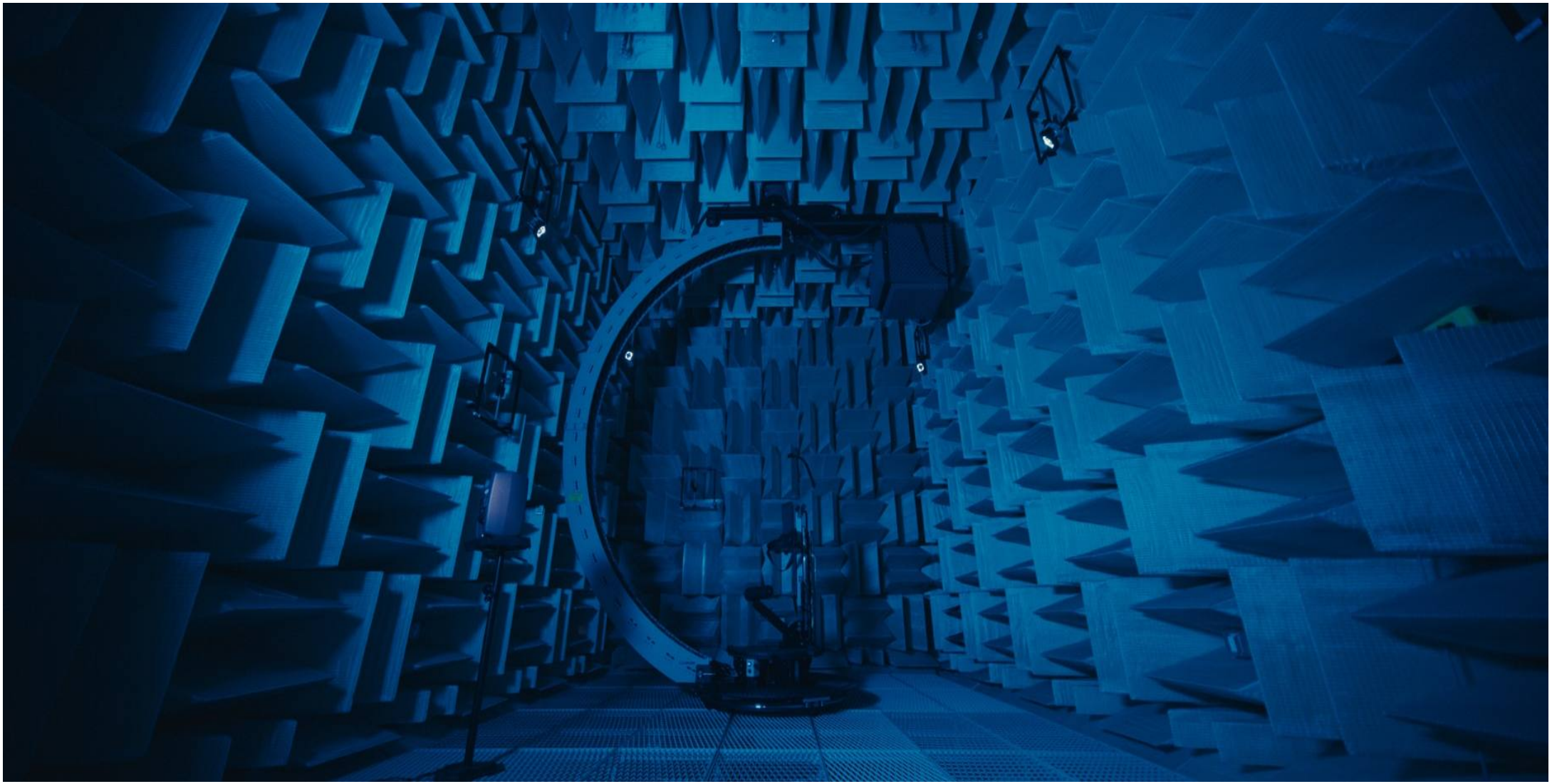


Adapted from: Roginska, A., & Geluso, P. (2017). *Immersive Sound*. Focal Press.



Adapted from: Roginska, A., & Geluso, P. (2017). *Immersive Sound*. Focal Press.

- https://facebookresearch.github.io/SS2_HRTF/

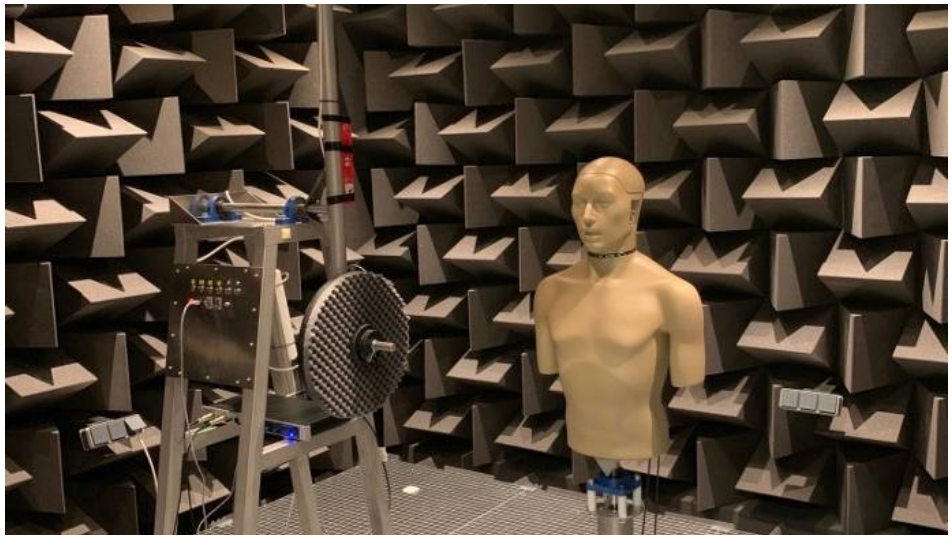


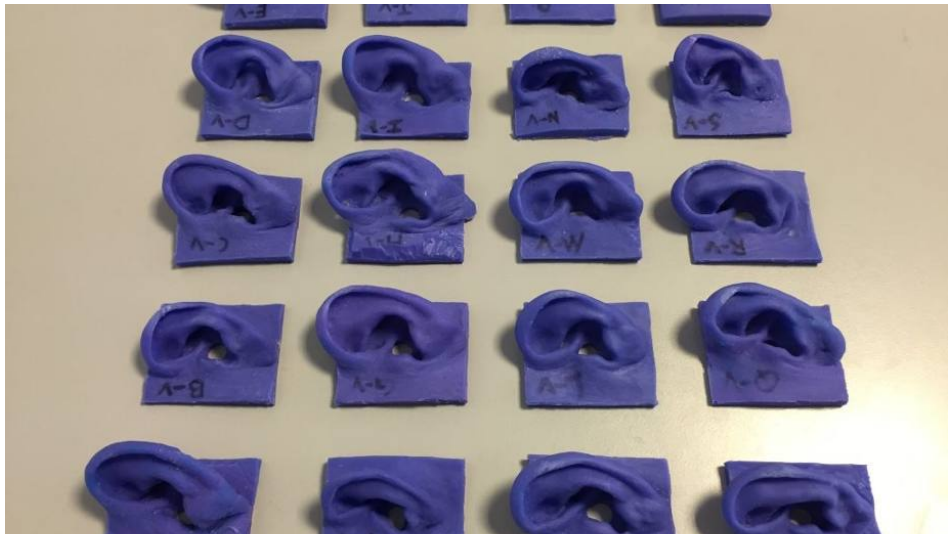
AALBORG UNIVERSITET

IT'S A DIVE

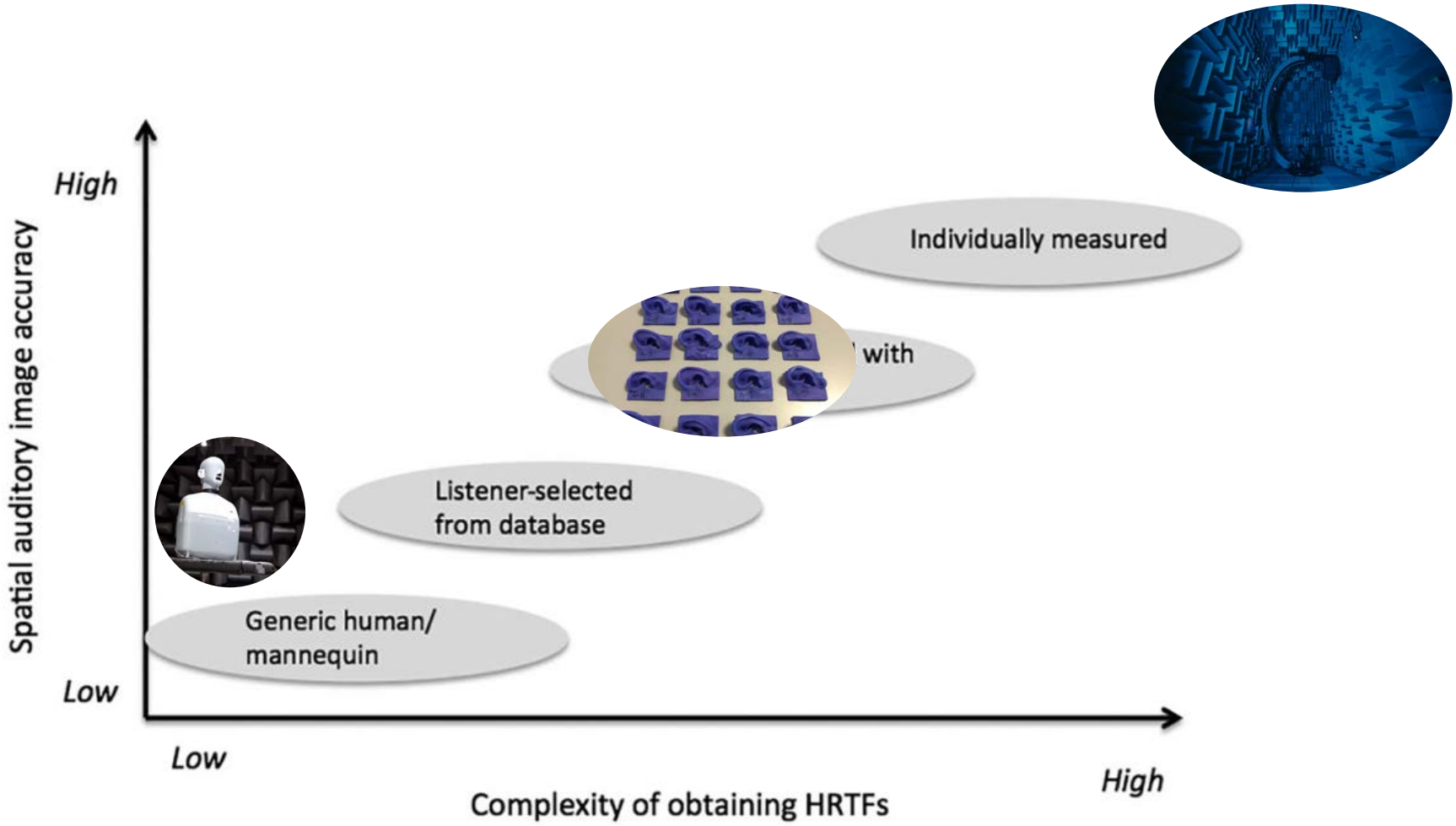
INDIVIDUAL THREE-DIMENSIONAL SPATIAL AUDITORY DISPLAYS FOR IMMERSIVE VIRTUAL ENVIRONMENTS

This project is fundend by the European Union



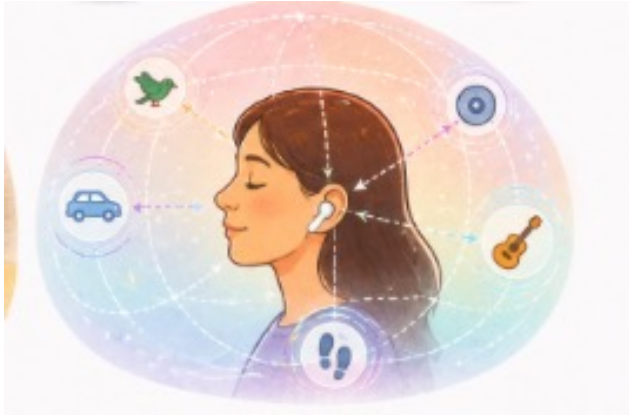


Simone Spagnol, 2021



Adapted from: Roginska, A., & Geluso, P. (2017). *Immersive Sound*. Focal Press.

Which HRTF FOR WHOM?



Volume 12 - 2018 |
https://doi.org/10.3389/fnins.2018.00021

Computational Audio

View all 12 articles >

Generic HRTFs May be Good Enough in Virtual Reality. Improving Source Localization through Cross-Modal Plasticity

Christopher C. Berger[1,2†] Mar Gonzalez-Franco[1*†]

Ana Tajadura-Jiménez[3,4] Dinei Florencio[1] Zhengyou Zhang[1,5]

[1] Microsoft Research, Redmond, WA, United States
[2] Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, United States
[3] UCL Interaction Centre, University College London, London, United Kingdom
[4] Interactive Systems DEI-Lab, Universidad Carlos III de Madrid, Madrid, Spain
[5] Department Electrical Engineering, University of Washington, Seattle, WA, United States

SCIENTIFIC REPORTS

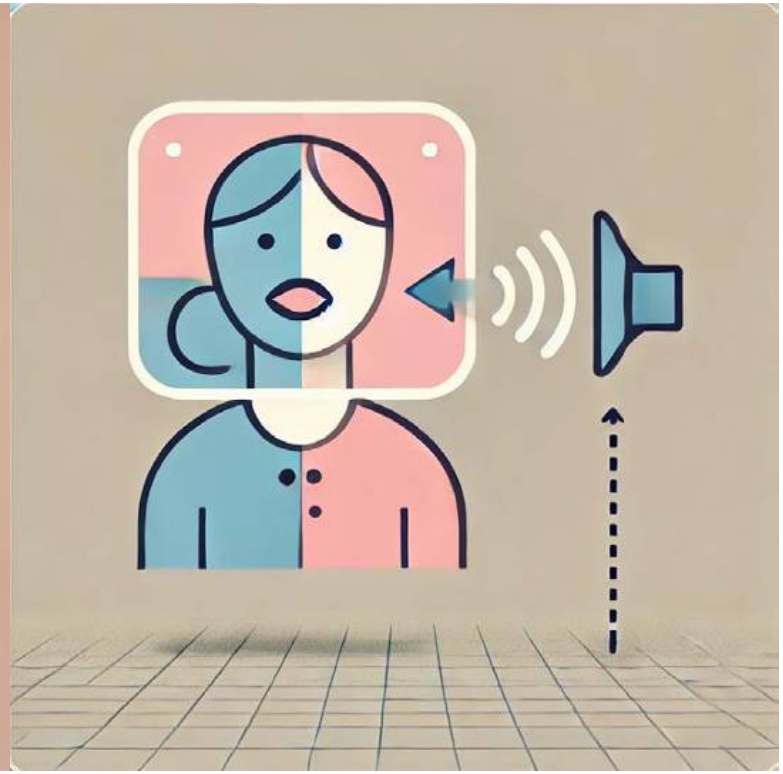
OPEN

Auditory Accommodation to Poorly Matched Non-Individual Spectral Localization Cues Through Active Learning

received: 8 May 2018
accepted: 17 December 2018
published online: 31 January 2019

Peter Stitt[1], Lorenzo Picinali[2] & Brian F. G. Katz[3]

Brain plasticity



Ventriloquism effect

A new step to optimize sound localization adaptation through the use of vision

Tristan-Gael Bara[1, 2], Alma Guilbert[2], and Tifanie Bouchara[1]

[1] *CEDRIC (EA4626), CNAM, HeSam Université, 75003 Paris, France*
[2] *VAC Laboratory (EA 7326), Université de Paris, 92774 Boulogne-Billancourt, France*

It depends

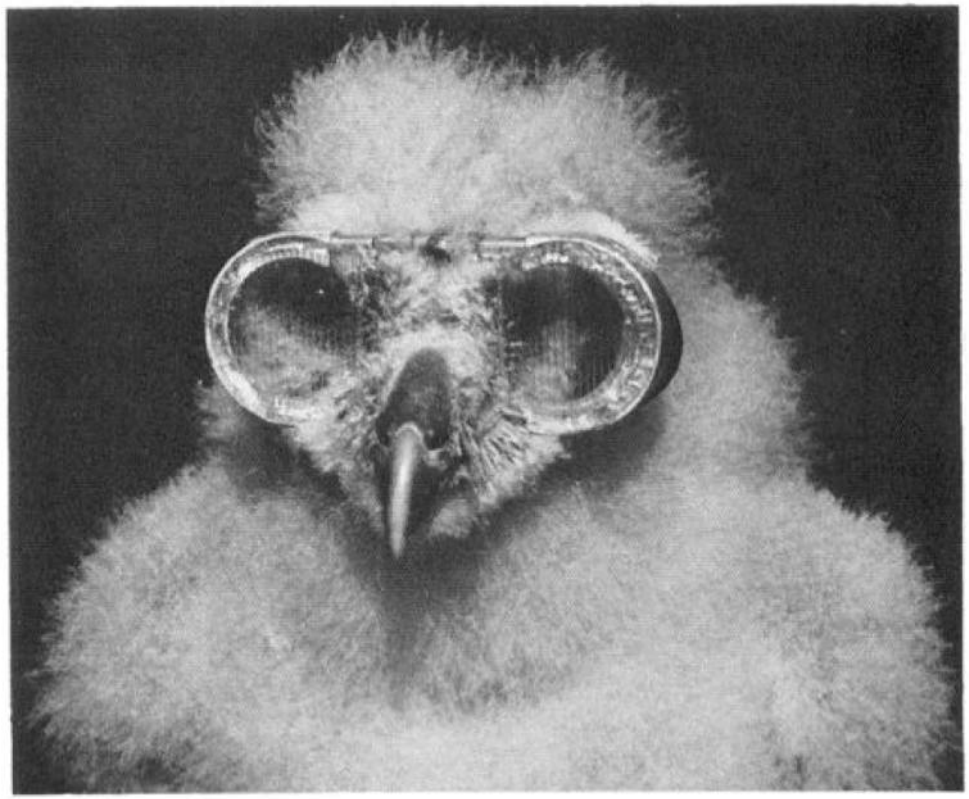
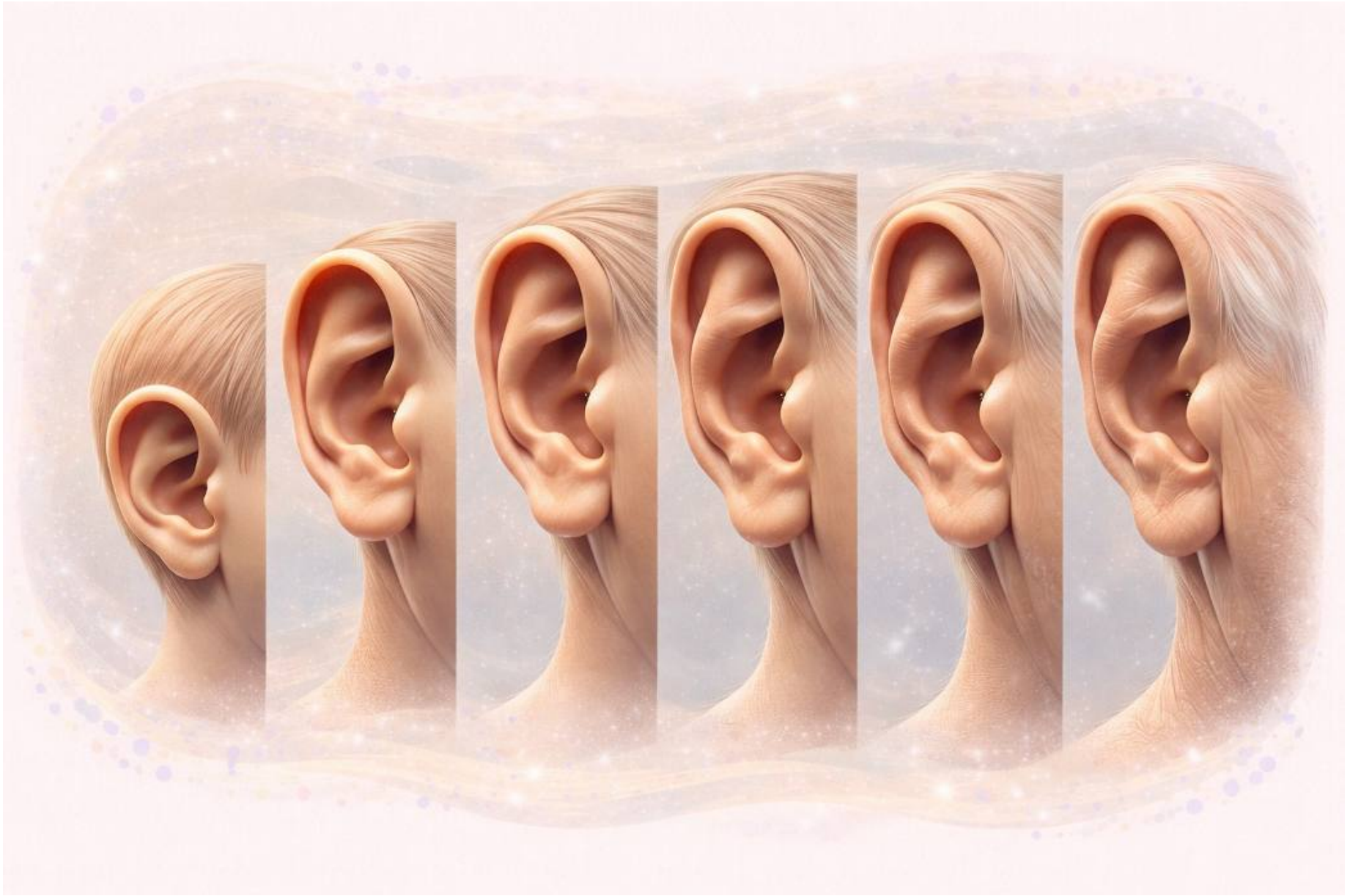


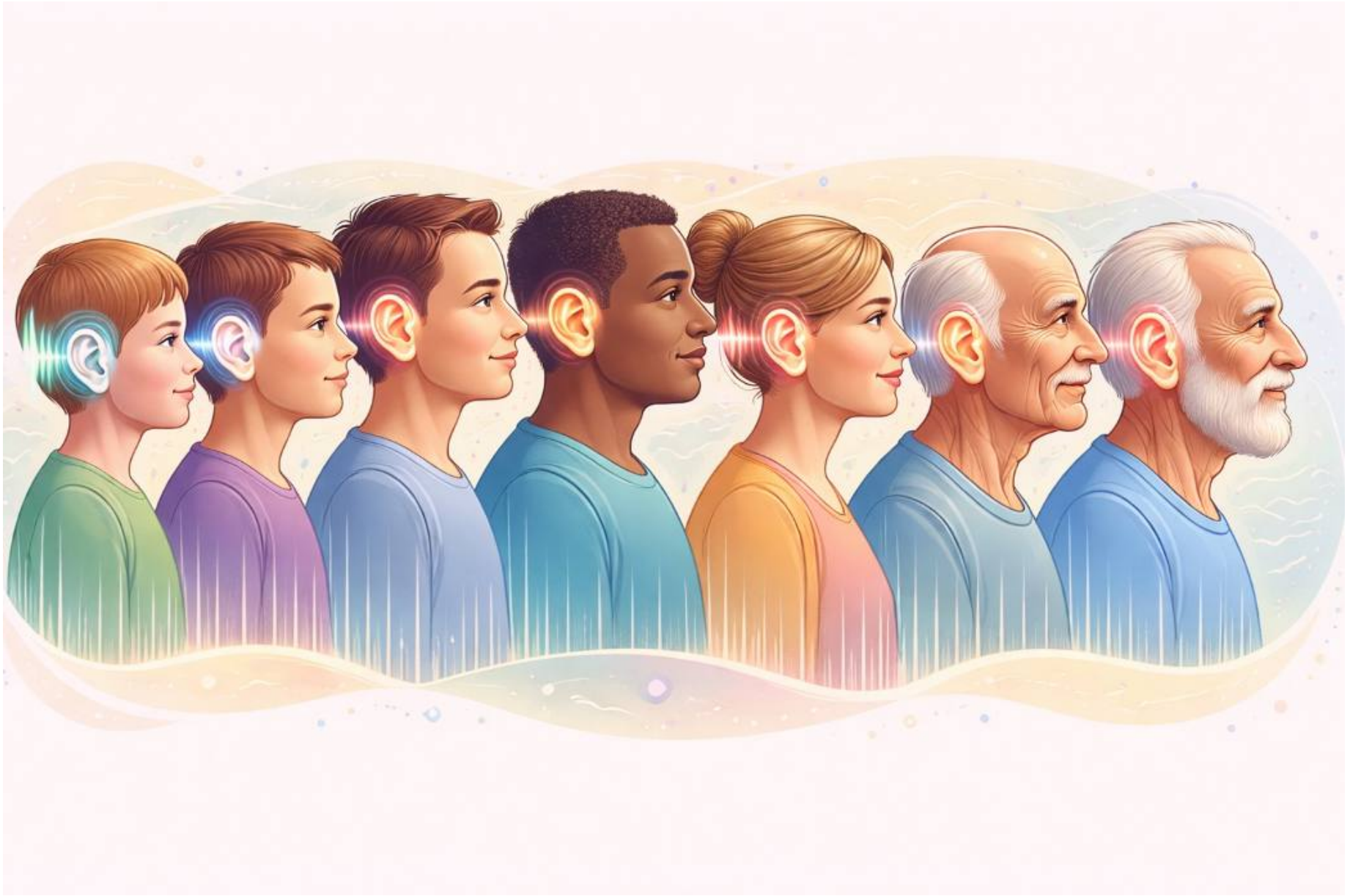
Figure 1. A baby barn owl, 28 d old, wearing binocular Fresnel prisms that displaced the visual field 34° to the right. Photograph indicates the optical quality of the prisms and the large visual field afforded by the spectacles. The vertical lines on the prisms are the edges of the individual prism elements that allow large optical displacements to be achieved with thin prisms.

Eric I. Knudsen & Phyllis F. Knudsen (1985), *Vision guides the adjustment of auditory localization in young barn owls*.

DANMARKS FRIE FORSKNINGSFOND
INDEPENDENT RESEARCH FUND DENMARK

What about blind or low vision individuals?





AURAL DIVERSITY

Andrew Hugill, 2022

Version 1.7

HEARING DIFFERENCES

"Disabilities", "disorders", "impairments", "loss" etc. affect 1 in 6 people and are typically classified on an audiometric scale from profound deafness to mild hearing loss. But differences are more prevalent than that. Less than 17% of people in Europe meet the international standard for "normal" hearing. Deaf (big D) = primarily communicates through sign language. deaf (small d) = does not sign. Hearing may be through interpretation. Hearing loss may be present from birth or acquired, in one or both ears, but n.b: not all forms of hearing difference involve loss.

SENSORINEURAL
Inner ear or nerve problems. c. 90% of all hearing loss.

- **Genetic** Inherited but not necessarily present from birth. Caused by mutation of dominant, recessive or mitochondrial genes.
- **Congenital** Present from birth. May be hereditary or caused by drugs, viruses or syndromes.
- **Acoustic Shock** Caused by unexpected loud sounds.
- **Auditory Neuropathy** Ear detects sound but has trouble passing it to the brain.
- **Disease** Viral, bacterial, fungal infections, diabetes, meningitis, tumour, inflammatory, Ménière's etc.
- **Ototoxic** Caused by drugs. Includes cochleotoxicity and vestibulotoxicity (one or two-sided).
- **Noise-induced** Caused by one brief intense sound or continuous exposure to loud sounds.
- **Other Causes** e.g. stroke, head trauma, perinatal conditions, hypothyroidism, cancers etc.
- **Recruitment** Reduced tolerance of loudness.

MIXED SENSORINEURAL & CONDUCTIVE
Genetic, age-related, noise-related, drugs, disease, trauma, ear wax, infections auditory system, etc.

CONDUCTIVE
Problems sending sound waves through the ear.

- **Blockages** Caused by wax, fluid, narrrow canal, abnormal bone growth, tumours, cholesteatoma, etc.
- **Barotrauma** Unequal air pressure between external and middle ear.
- **Inner ear** Enlarged Vestibular Aqueduct; Labyrinthine Fistula; Superior Canal Dehiscence.
- **Inflammations** Caused by viral, bacterial, fungal infections, allergies, otitis, Eustachian tube dysfunction, etc.

AUDITORY SYSTEM
Neurological differences that affect hearing

- **Tinnitus** Perception of sound when no sound is present. May be subjective, objective or pulsatile. Often involves clicking, buzzing, hissing or roaring sounds.
- **Amusia** Inability to perceive or produce musical sounds. Congenital or acquired (a.k.a. "tone deaf").
- **Palinacusis** Sounds repeat after inaudibility.
- **Diplacusis** Perception of pitch differences between the two ears.
- **Hyperacusis** Increased sensitivity to sound.
- **Misophonia** Negative emotional reactions to certain trigger sounds.
- **Phonophobia** Fear of specific sounds.
- **Hallucinations** Musical (hearing music when none is present); hearing voices etc.
- **Auditory Processing Disorder** "Normal" hearing, but sound cannot be processed by the brain.
- **Neurodivergence** Autistic people often have heightened sound perception and an ability to focus on detail.

UNIVERSAL VARIATIONS

- **Temporary** e.g. wax, viral, bacterial or fungal infections, common cold, tiredness etc.
- **Natural** e.g pinna shapes unique to every individual.
- **Non-cochlear** e.g. touch, vibration, textual, visual.
- **Historical** e.g. past differences in hearing, evidenced by hearing aids.
- **Ethnic** e.g. prevalence of hearing loss varies across ethnic groups.
- **Geographical** e.g. regional patterns of hearing loss.
- **Cultural** e.g. d/Deaf culture.
- **Economic** e.g. disparities as a determinant of hearing health
- **Environmental** e.g. noise exposure, lifestyle issues, etc.

TECHNOLOGY

- **Hearing Aids** Amplification devices. Behind-the-ear (BTE), In-the-ear (ITE), In-the-canal (ITC). Right-In-the-Ear (RITE) and others use acoustic waves. Bone-Conduction HA use mechanical waves to carry the signal to the cochlea. Contralateral Routing of Signals (CROS) aids transmit auditory information from one ear to the other.
- **Cochlear Implants** Surgically implanted neuroprostheses which bypass acoustic hearing by direct electromagnetic wave stimulation of the auditory nerve.
- **Other Devices** e.g. hearables, smart earbuds, assistive listening, apps, etc.
- **Machine Listening** a.k.a. Computer Audition, a branch of AI and computer science.
- **ACOUSTICS ENGINEERING** Includes many fields that affect aural diversity.

ANIMAL HEARING
Humans hear 20Hz-20kHz whereas bats hear 9 kHz-120 kHz and mice hear 1 kHz to 70–90 kHz. Pigeons hear infrasound (down to 0.5Hz). Dolphins, rats, horses, cats, dogs, elephants, owls, moths and many other species have specially adapted hearing which far exceeds human capabilities.

Trained Ears
Auditory (re-)training, and professional ear training for e.g. musicians, linguists, doctors, engineers.

"NORMAL" HEARING
defined by the International Standards Organisation as: "otologically normal persons in the age range from 18 years to 25 years inclusive" (ISO 226:2003) i.e. around 17% of the population.

Foetal hearing
A foetus's hearing is fully formed and functional from the last trimester of pregnancy.

→ **Infant hearing**
It typically takes six months for babies' hearing to develop to the point that they can fully hear and understand a range of sounds.

→ **Teens/young adults**
The normal hearing frequency range of a healthy young person is about 20 to 20,000Hz. A 'normal' audible range for loudness is 0-180dB SPL. Anything over 85dB SPL risks damaging hearing.

→ **Adults to old age**
Hearing deteriorates throughout life at varying rates from person to person. Male hearing typically declines at twice the rate of female hearing. Age related hearing loss (or presbycusis) may also be worsened by external factors such as environmental and medical conditions. High frequencies are lost first, and this usually becomes noticeable at c. 40 years, when our hearing range starts to dip below 14kHz. We may have difficulty understanding speech and distinguishing sounds in crowded noisy places. We may need to raise volume levels, ask people to repeat themselves, or experience tinnitus.

AURAL DIVERSITY

CoolHear Worskhop

live performance
multisensory workshop
interactive demos

19th April - 3.00 to 5.00 PM

ROYAL DANISH ACADEMY OF MUSIC
Rosenørns Alle 22, 1970 Frederiksberg

HoloBand
Augmented reality for perceptual music training
Balázs Iványi, Christian Tsalidis, Scott Naylor, Truls Bendik Tjemsland

Music trainining in VR
Vitual Reality application for training listening to music
Sine Marie Kromann Kristiansen, Emil Sønderskov Hansen

Name the instrument
Listening experiment where the system helps to identify what instruments are playing in a music piece
Doga Buse Cavdir, Francesco Ganis

Uncharted Chants
Mobile game for musical perception training
Erik Frej Knudsen, Jonas Slim Andersen, Helmer Nuijens

Lyt igen
Empowers young people with hearing loss to reach their full potential using VR
ME-Lab, CHBH and Decibel

VAM
High fidelity vibrotactile actuator
Razvan Paisa

Effects of spatialization
Listening experiment to examine how extreme spatialization influences musical appreciation and instruments segregation in a string quartet
Jesper Andersen

Multisensory live music
Bass, drums and voice duo can be felt in the entire body through custom designed furniture
Antonia Barišić, Peter Williams, Razvan Paisa, Francesco Ganis

Tickle Tuner
Haptic cellphone cover for musical training
Francesco Ganis

For more information and booking

melcph.create.aau.dk/coolhear-workshop

Contact us

Stefania Serafin
sts@create.aau.dk

Lone Marianne Percy-Smith
lone.percy-smith@regionh.dk

Event organized by

AALBORG UNIVERSITY

CHBC Center for Hørelse og Balance

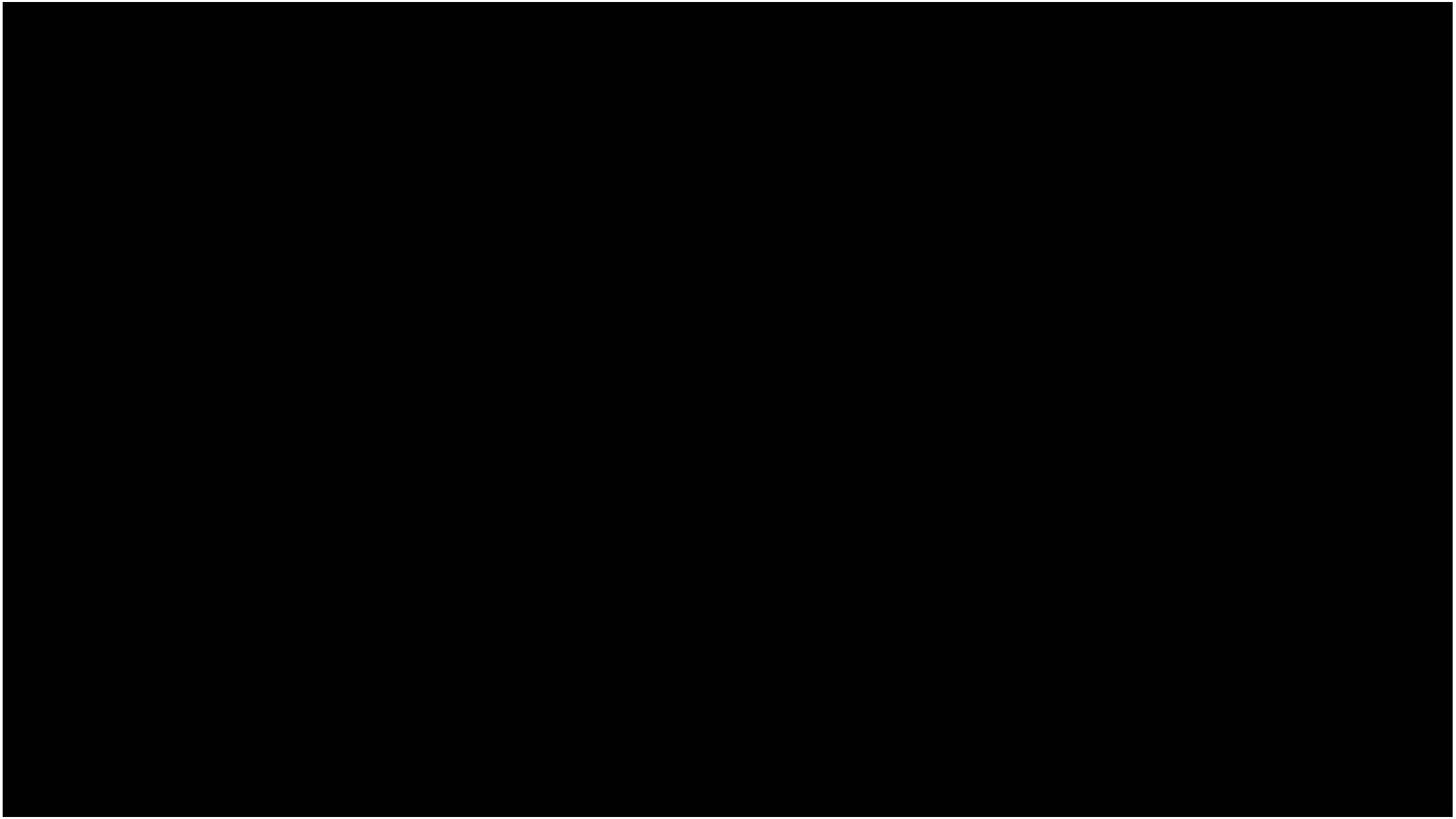
Rigshospitalet

THE ROYAL DANISH ACADEMY OF MUSIC

In collaboration with

Stanford University

oticon MEDICAL



MUSIC VISUALIZATION

Ivanyi, B., Tsalidis, C., Naylor, S., Tjemsland, T. B., Adjorlu, A., Kepp, N. E., & Serafin, S. (2022). HoloBand: An Augmented Reality Experience to Train Music Perception for the hard-of-hearing. In *2022 AES International Conference on Audio for Virtual and Augmented Reality, AVAR 2022* (pp. 1-10). Audio Engineering Society.

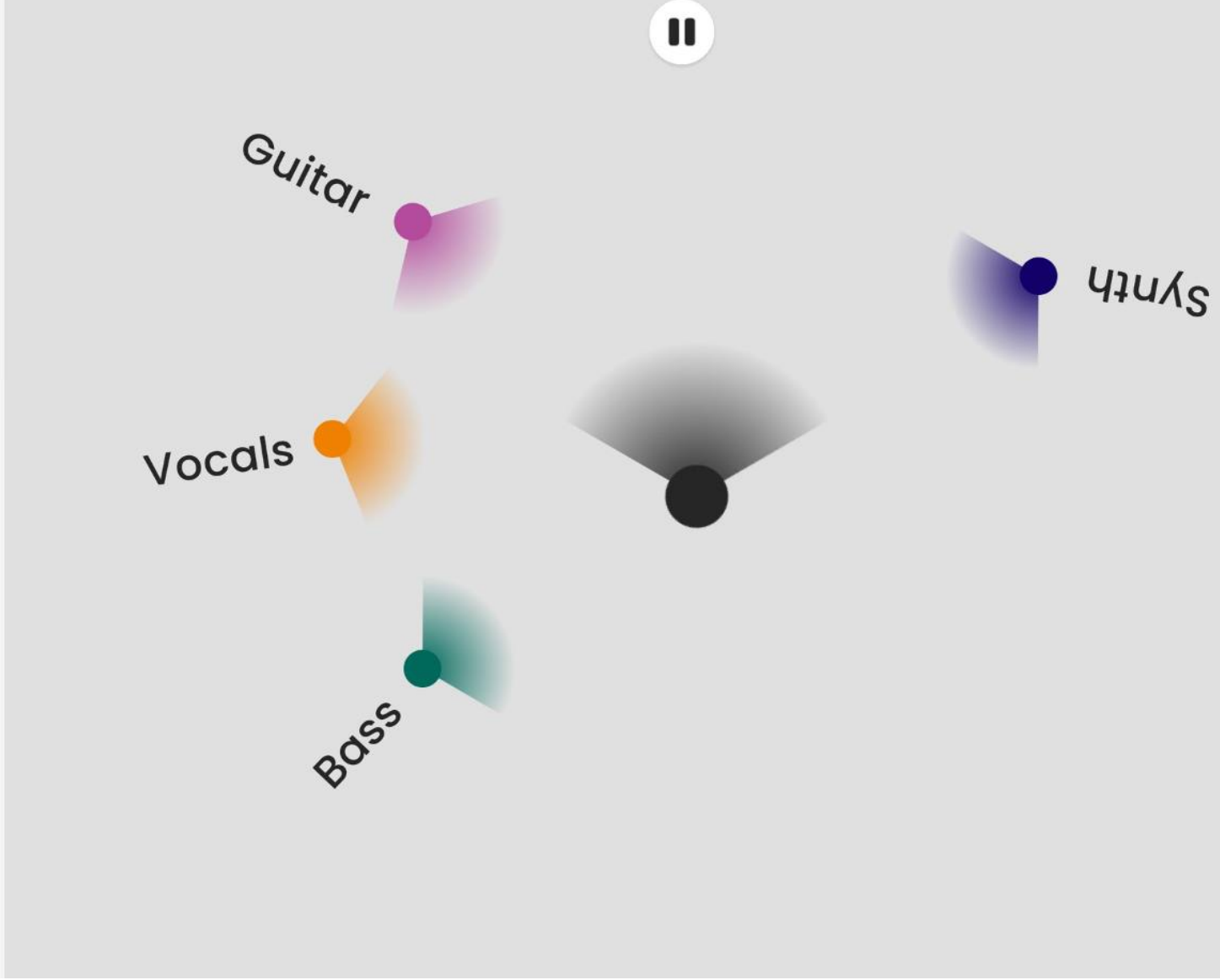
Guitar

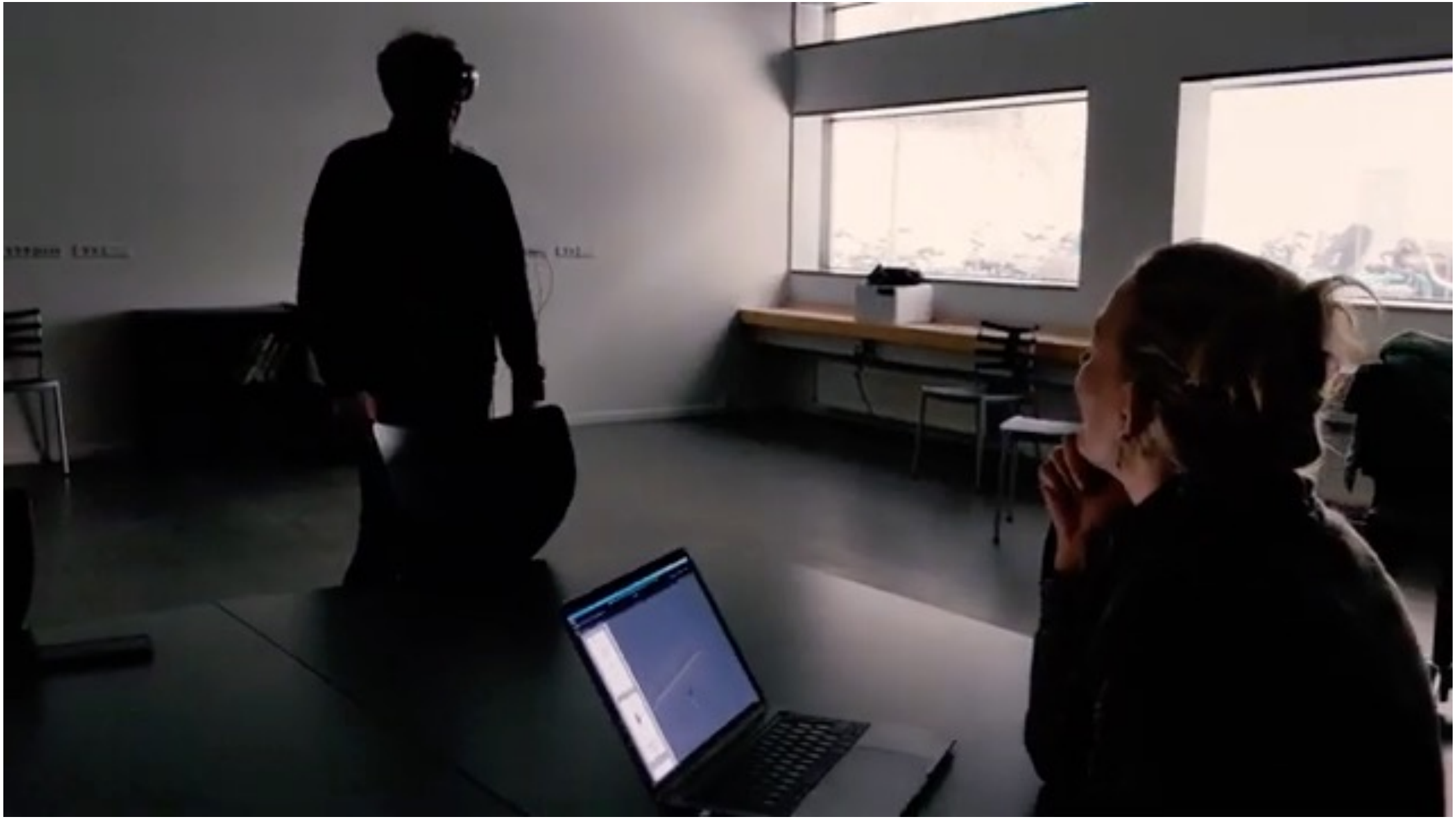
Visuals Solo

Bass

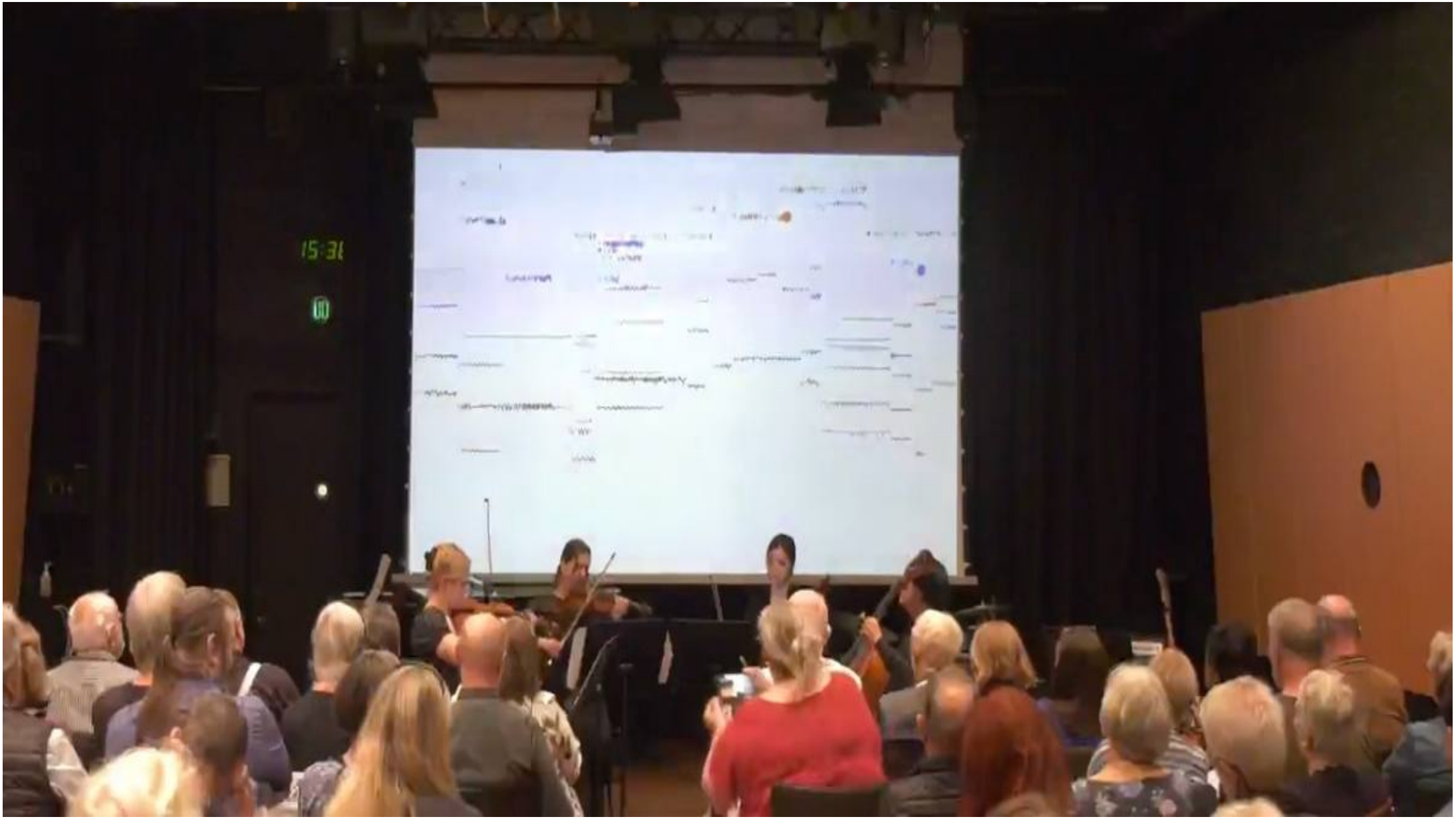
Visuals Solo

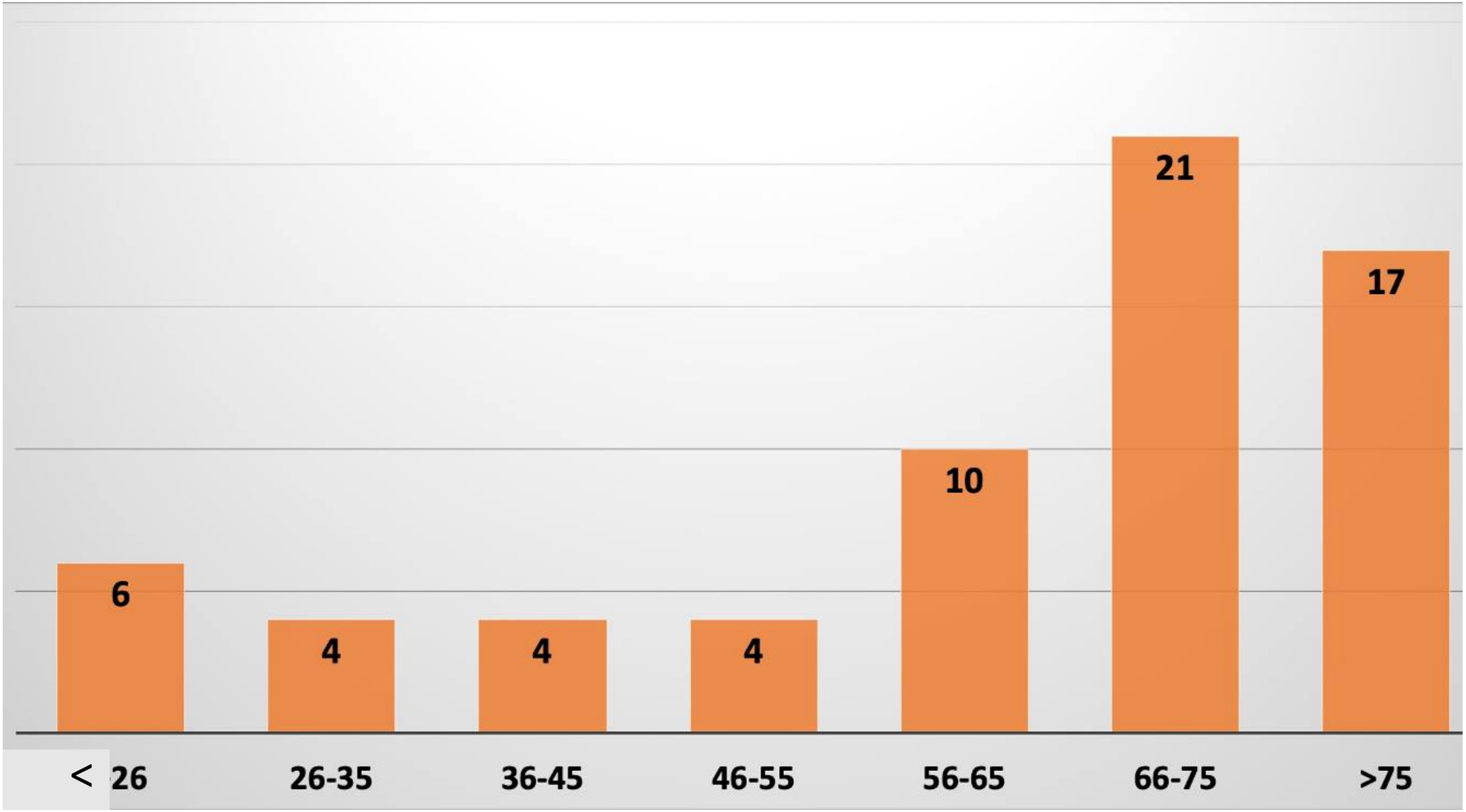
Vocals

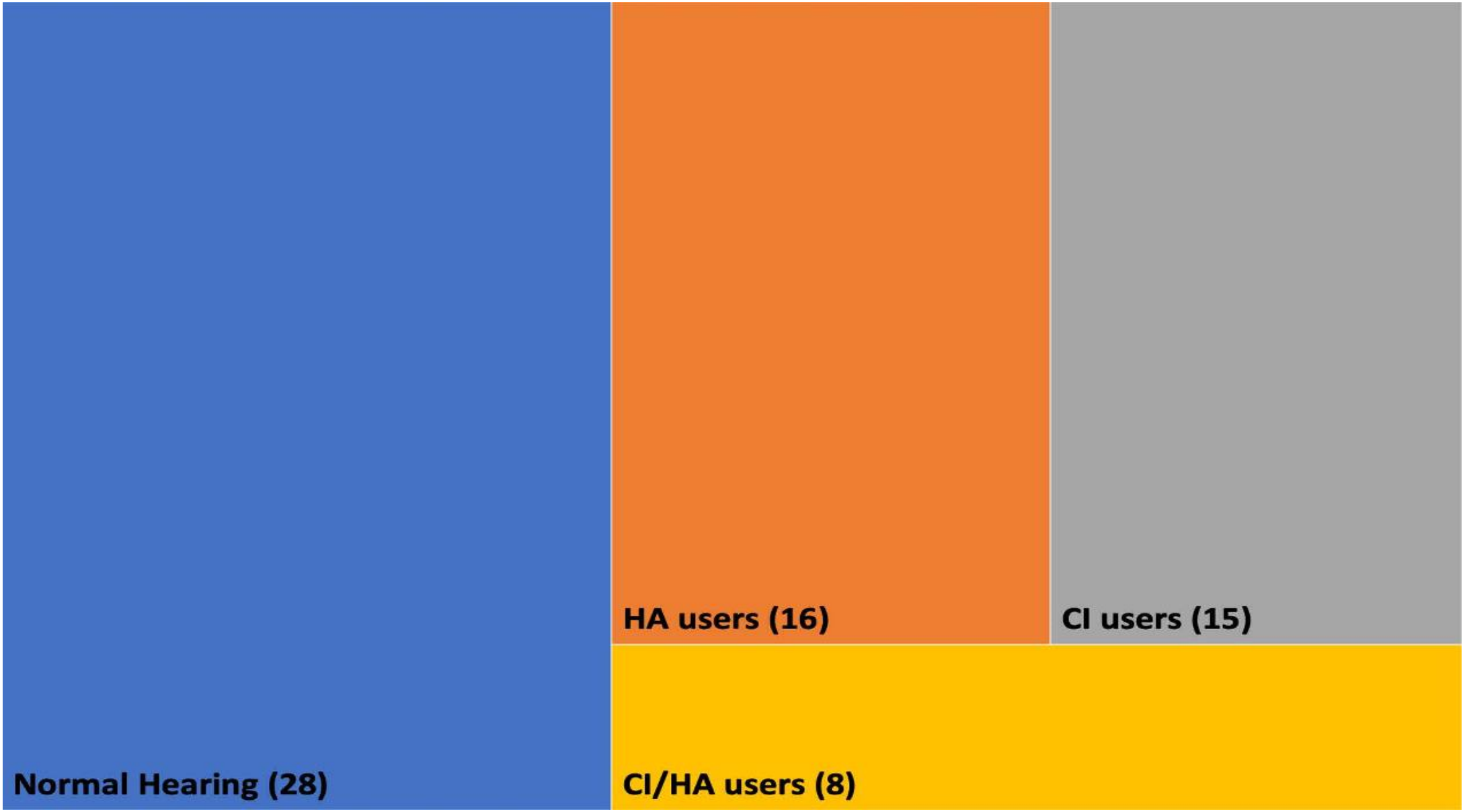


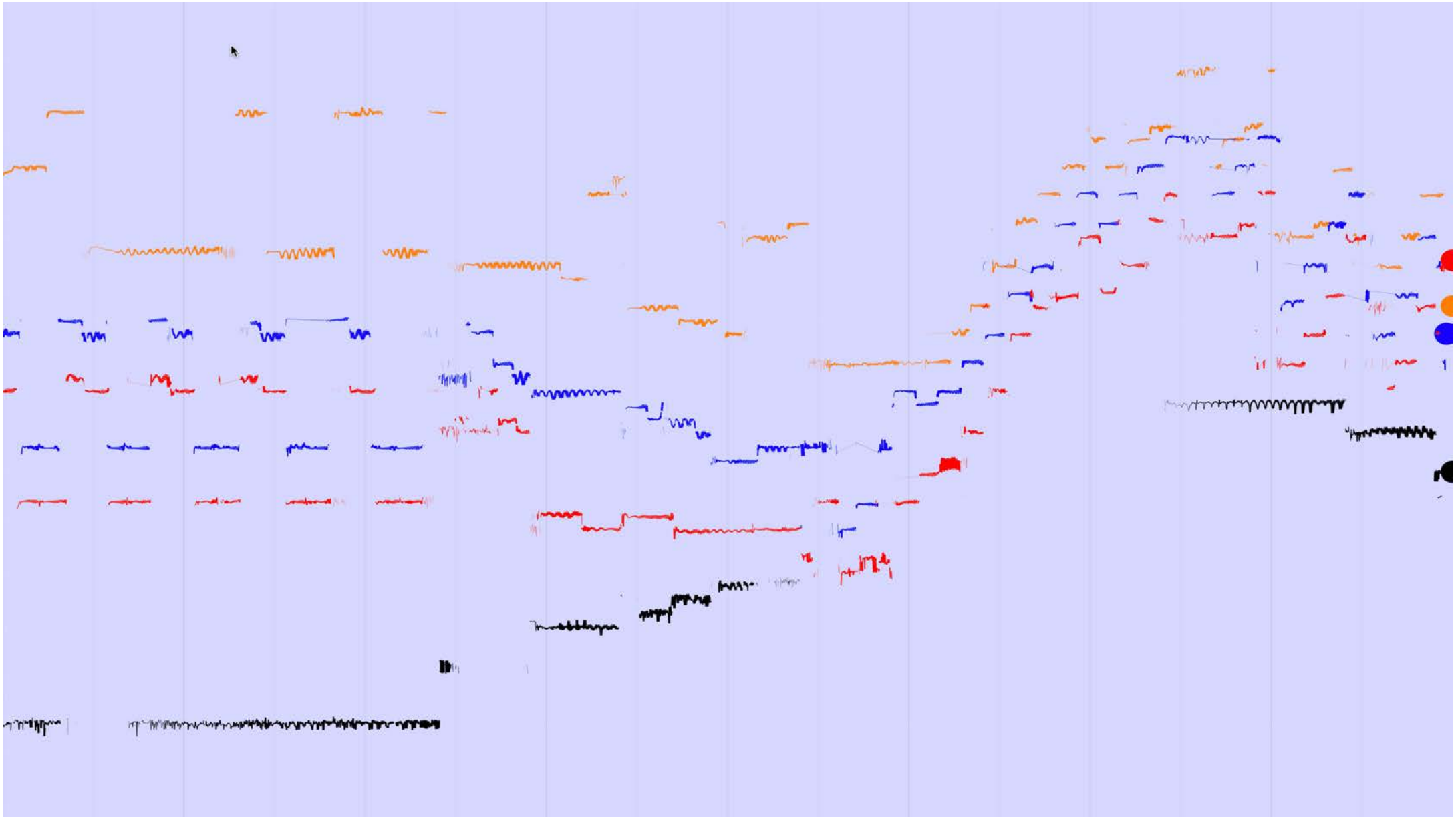


SOCIAL EXPERIENCES




6
4
4
4
10
21
17
< 26
26-35
36-45
46-55
56-65
66-75
>75


Normal Hearing (28)
HA users (16)
CI users (15)
CI/HA users (8)



RESULTS

Participants filled a questionnaire and 16 participated to interviews.
Here we focus on interviews

The visualization was useful

The visualization was distracting. I closed my eyes.

The visualization took focus away from the music.

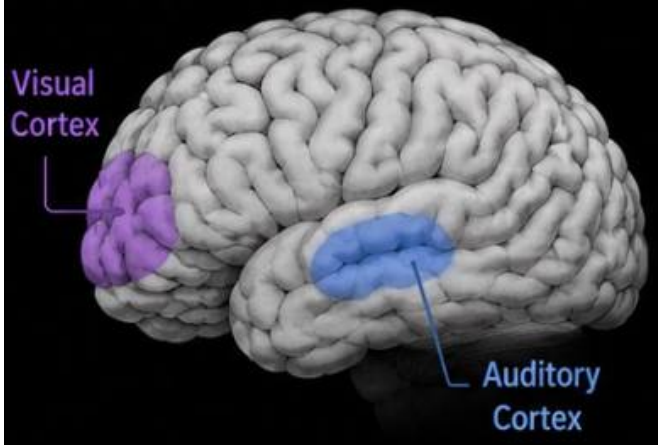
There should have been an introduction.

Visual Stimuli Can Activate Auditory Cortex After Deafness

Cross-modal plasticity and outcomes with cochlear implantation

1. NORMAL HEARING

Vision activates visual cortex.
Sound activates auditory cortex.



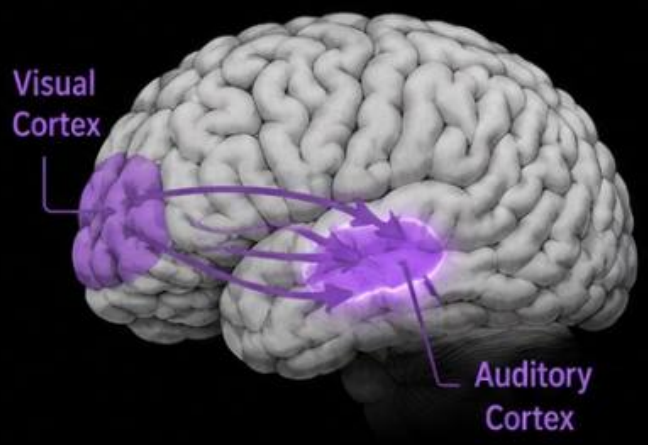
Visual input → Visual Cortex

Sound input → Auditory Cortex

Auditory cortex is dedicated

2. DEAFNESS

In the absence of sound, visual input recruits auditory cortex.

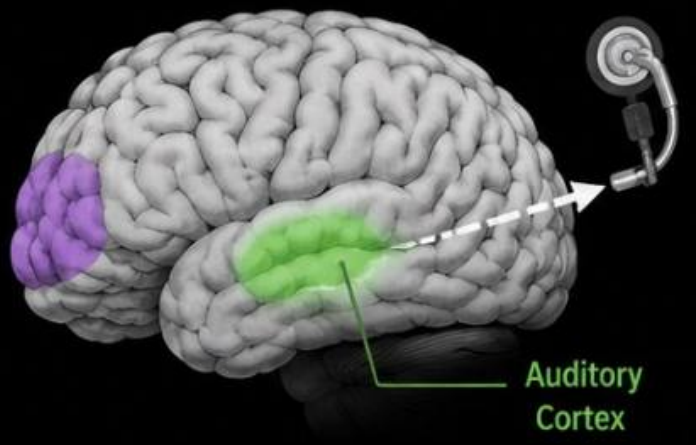


Visual input → Visual Cortex + Auditory Cortex

Auditory cortex is partially

3. COCHLEAR IMPLANT OUTCOMES

Restoring sound can re-engage auditory cortex.
Outcome depends on extent of reorganization.

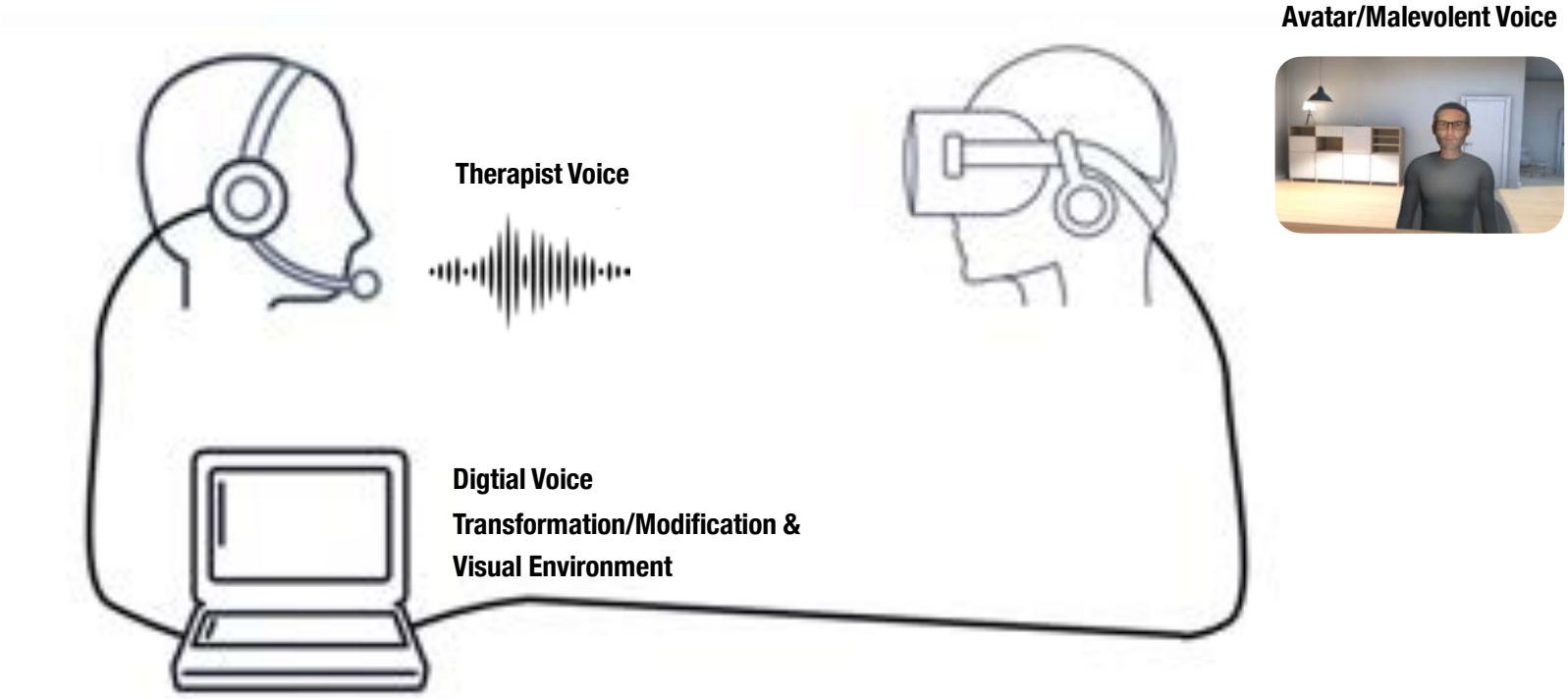


Less visual takeover → Better CI outcomes
(auditory cortex more available for sound)

More visual takeover → Poorer CI outcomes
(auditory cortex less available for sound)

HEARING VOICES

AVATAR THERAPY

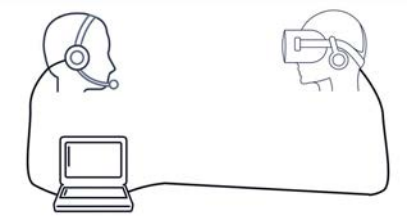


AVATAR THERAPY

0 1

Session setup

The patient is positioned in virtual reality (VR)



0 2

Avatar Creation

Avatar is created matching the perceived voice

0 3

Live Interaction

The therapist speaks and guides the patient

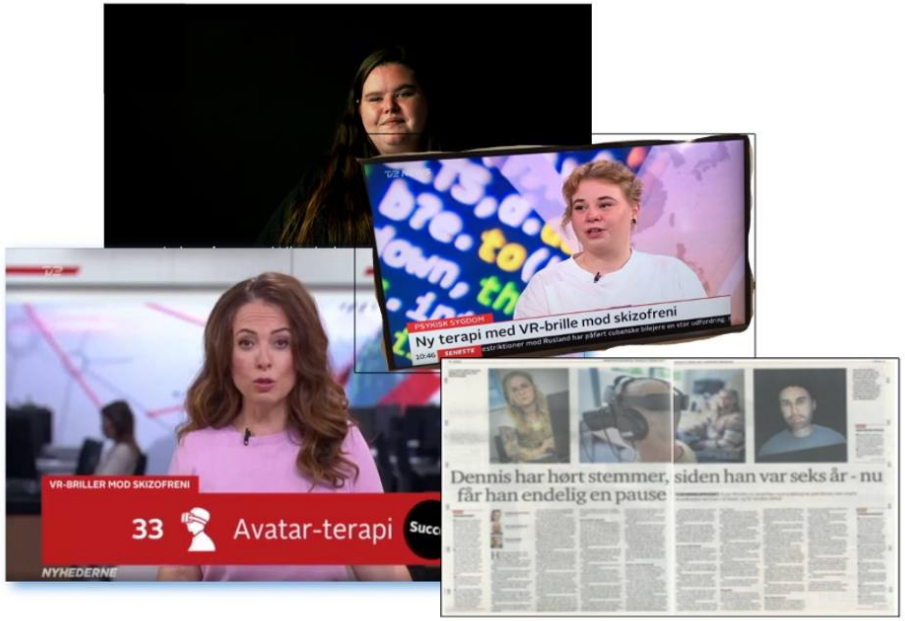
0 4

Observation

The therapist observes and iterates (domain knowledge)

AVATAR THERAPY

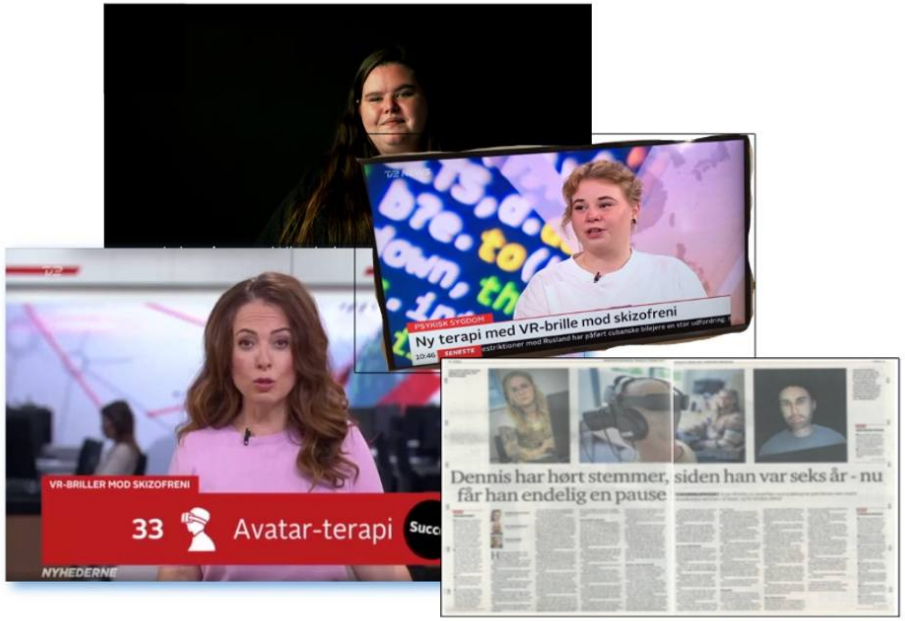
- Reduced severity and frequency of auditory hallucinations after 12 weeks.
- 11 out of 271 treatment resistant patients no longer heard voices by the 24 week follow up.
- Greater satisfaction with treatment.
- Not for everyone, with possibilities of improving general realism, immersion and user-experience.



[1Glenthøj, et al., “Impact of avatar features and presence on treatment outcomes in virtual reality-assisted therapy for auditory hallucinations”, Schizophrenia research, Elsevier, 2025

AVATAR THERAPY

- Reduced severity and frequency of auditory hallucinations after 12 weeks.
- 11 out of 271 treatment resistant patients no longer heard voices by the 24 week follow up.
- Greater satisfaction with treatment.
- Not for everyone, with possibilities of improving general realism, immersion and user-experience.



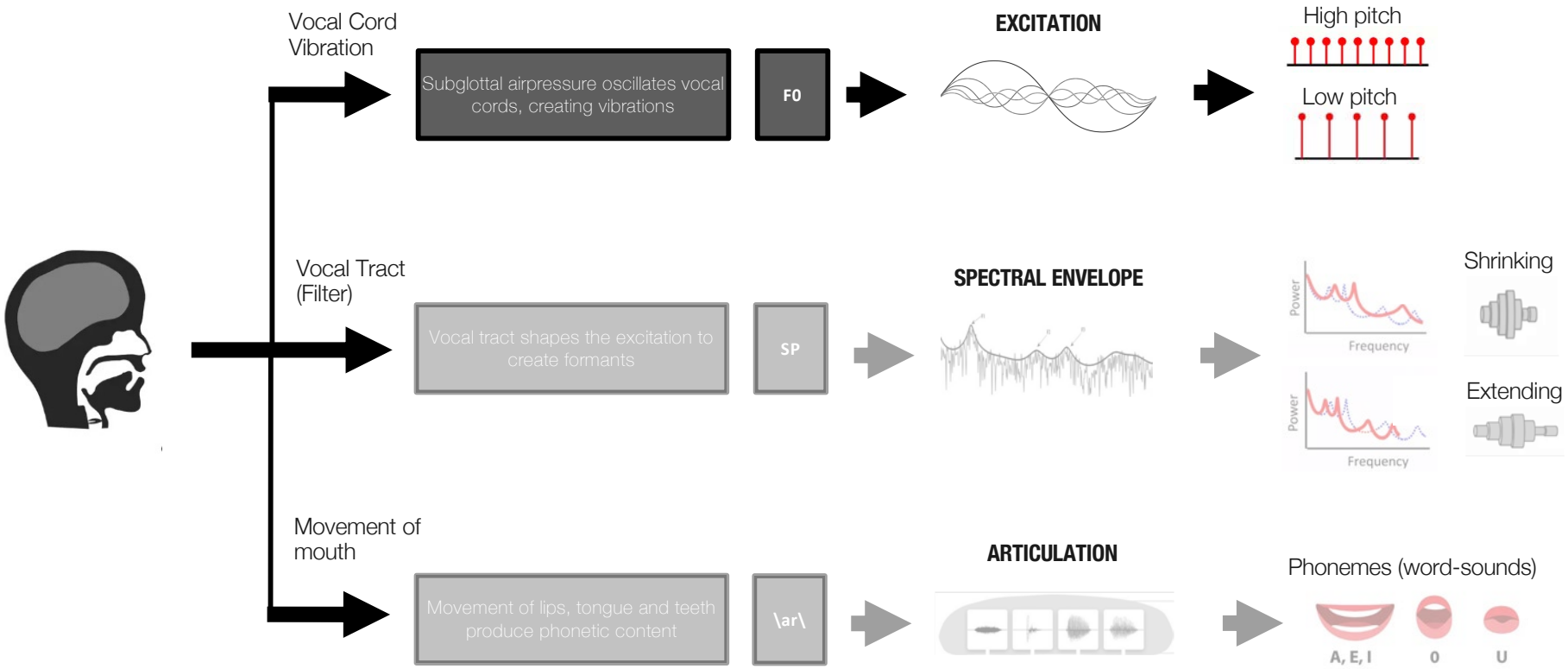
[L. Glenthøj, et al., “Impact of avatar features and presence on treatment outcomes in virtual reality-assisted therapy for auditory hallucinations”, Schizophrenia research, Elsevier, 2025

AVATAR THERAPY

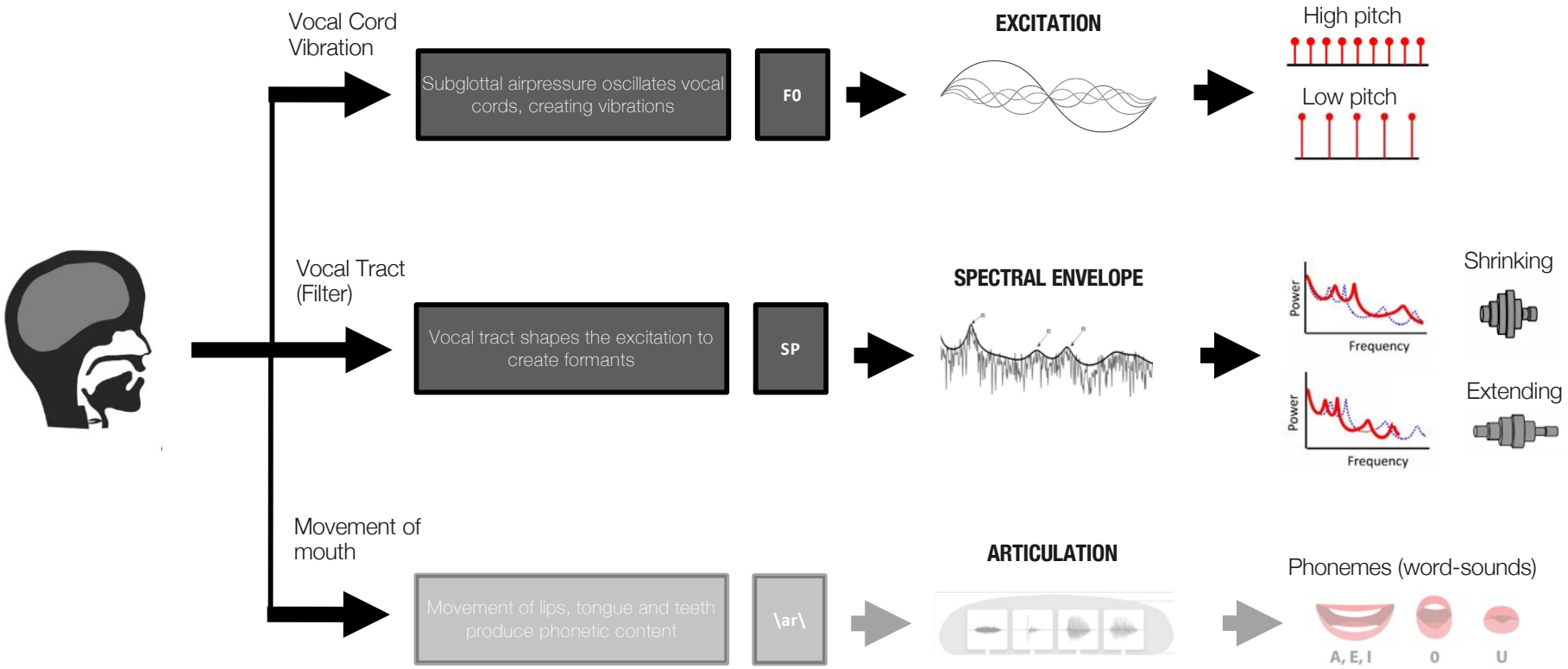
> *Realism of the avatar, particularly in its auditory dimension, may facilitate engagement and therapeutic processing.*

[1] L. Glenthøj, et al., “Impact of avatar features and presence on treatment outcomes in virtual reality-assisted therapy for auditory hallucinations”, Schizophrenia research, Elsevier, 2025

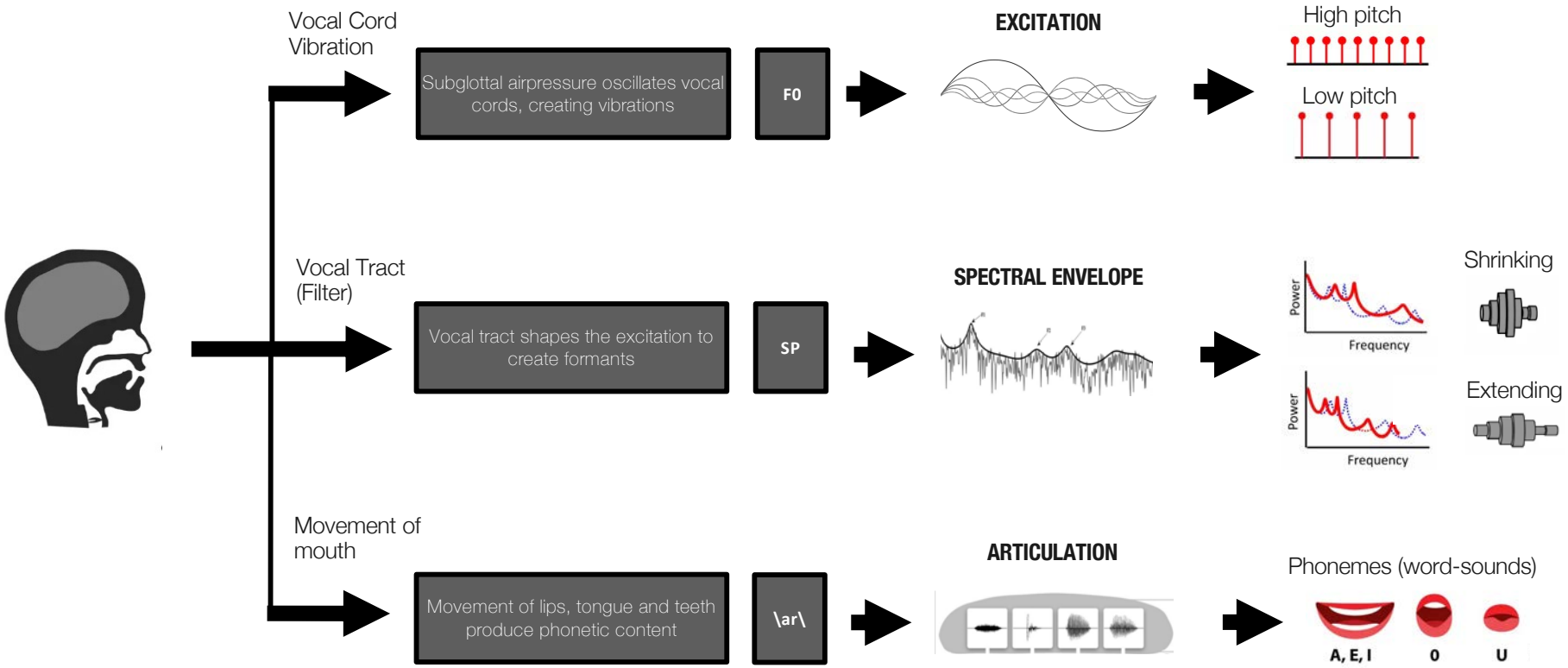
SPEECH PRODUCTION



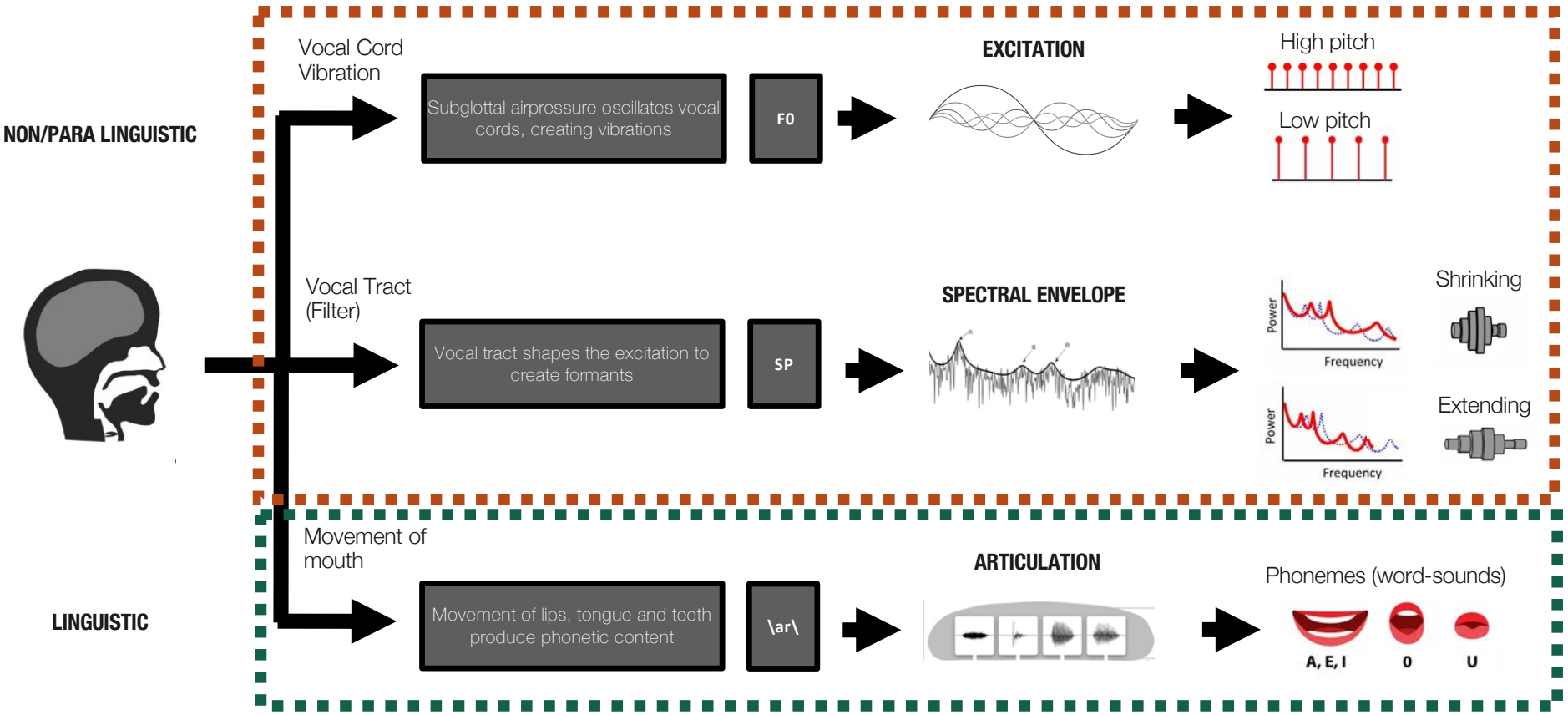
SPEECH PRODUCTION



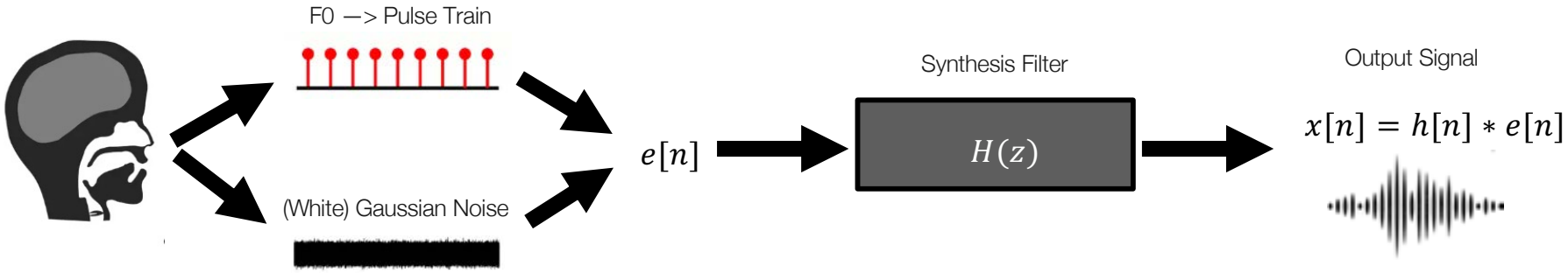
SPEECH PRODUCTION



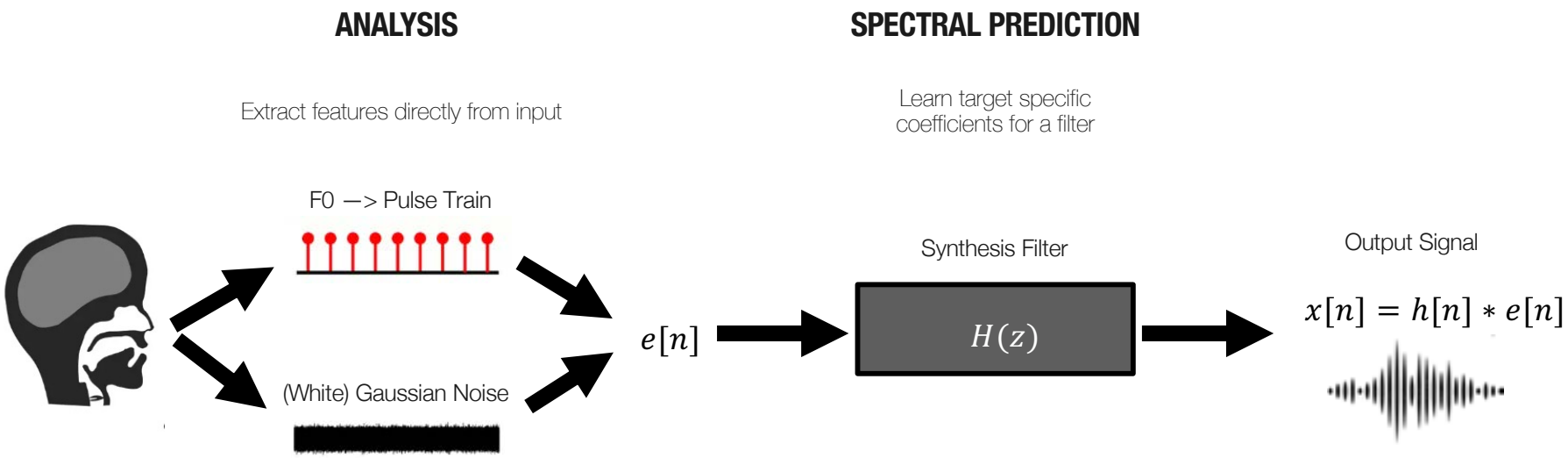
SPEECH PRODUCTION



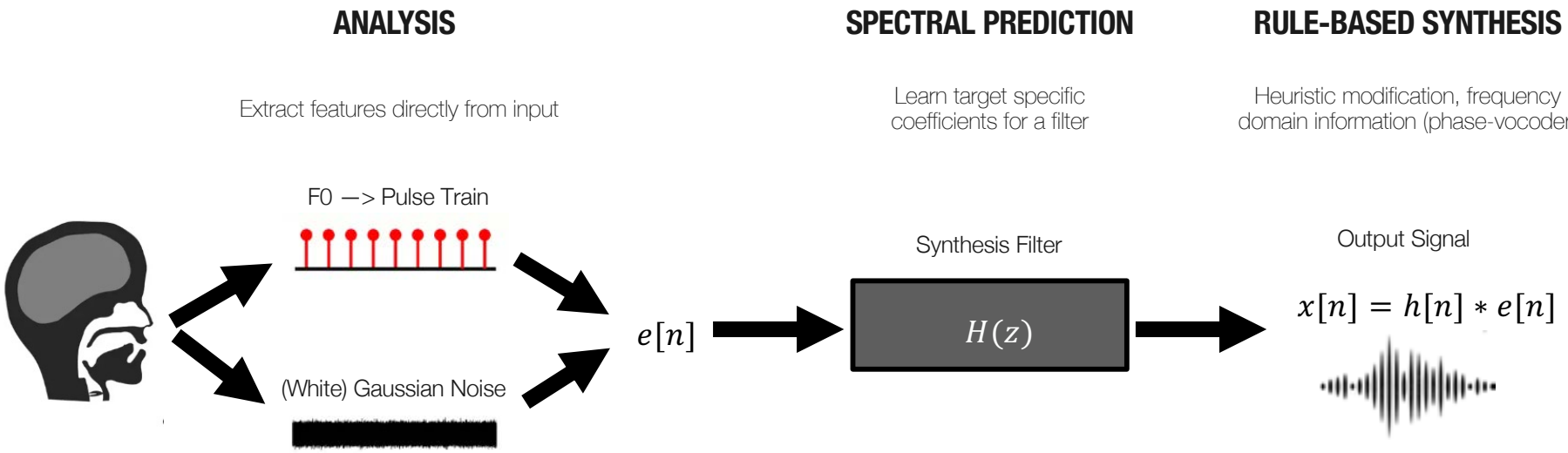
VOICE CONVERSION (NAIVE APPROACH)



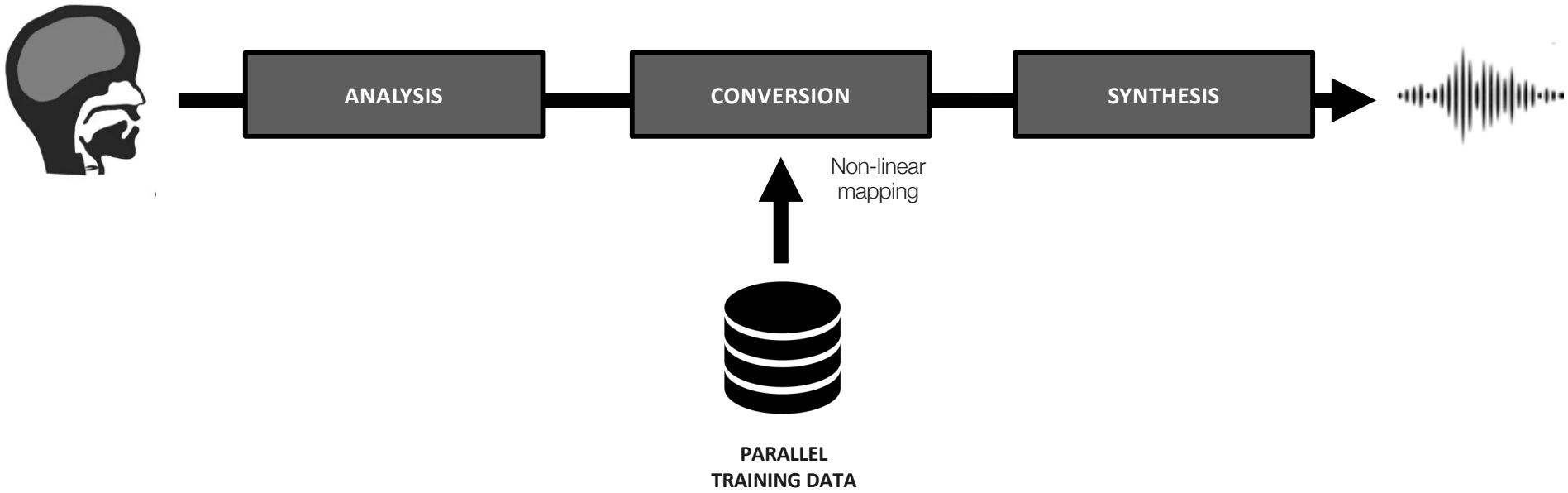
VOICE CONVERSION (NAIVE APPROACH)



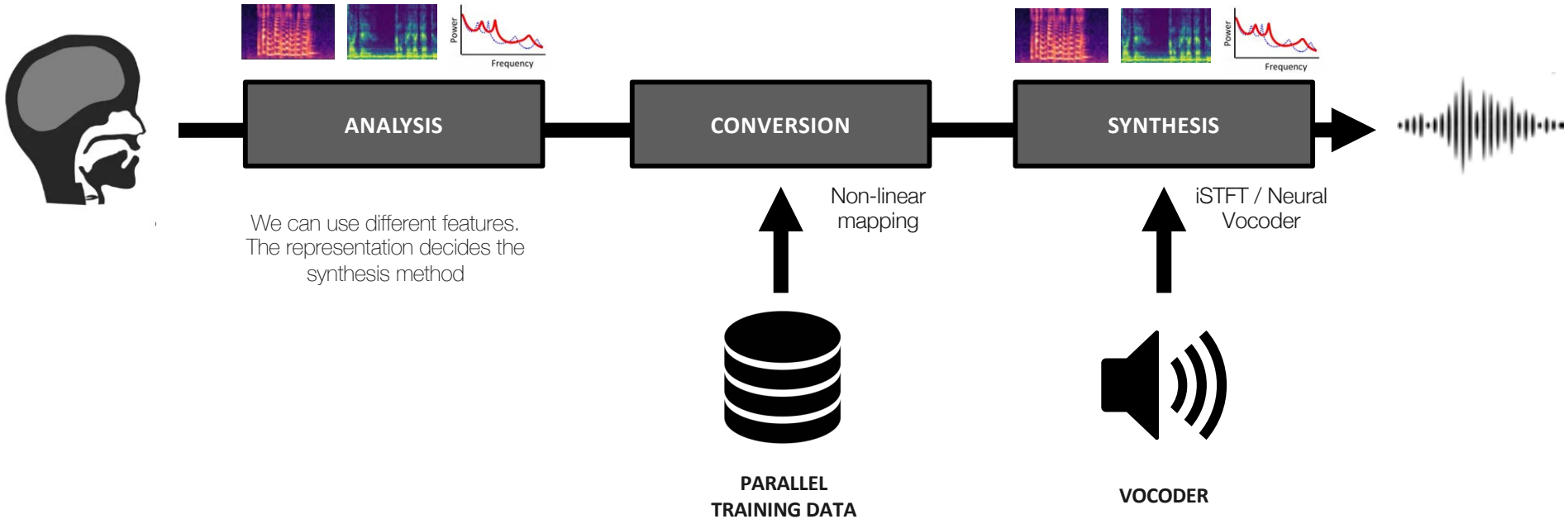
VOICE CONVERSION (NAIVE APPROACH)



VOICE CONVERSION



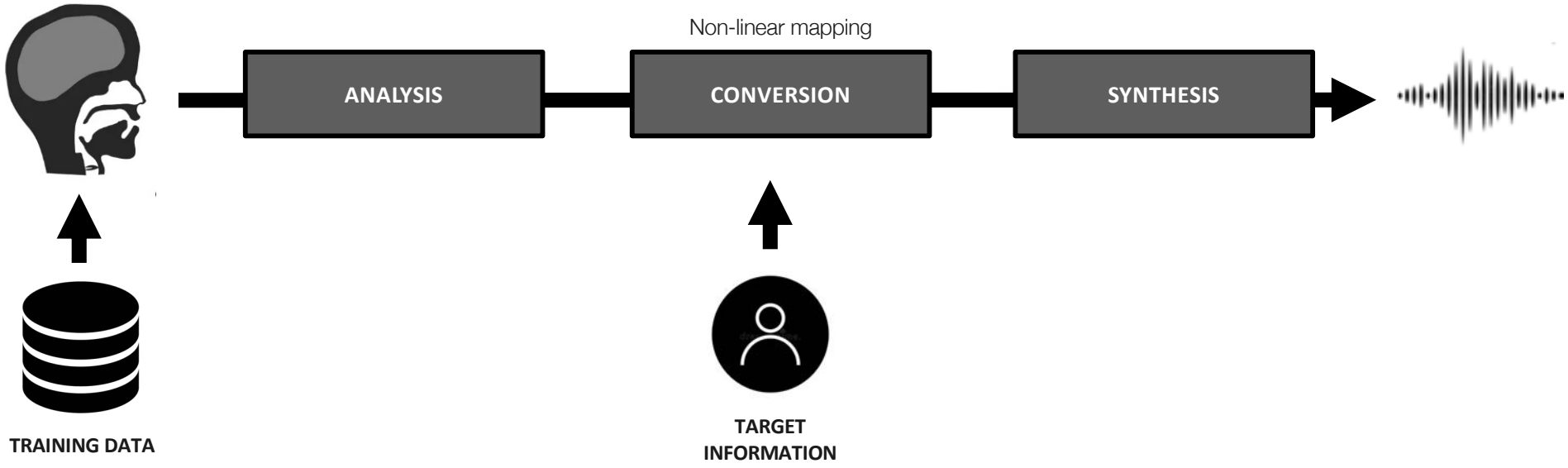
VOICE CONVERSION



VOICE CONVERSION (NON-PARALLEL)

Recognition: Extract desired information such as linguistics and eliminate unwanted source information.

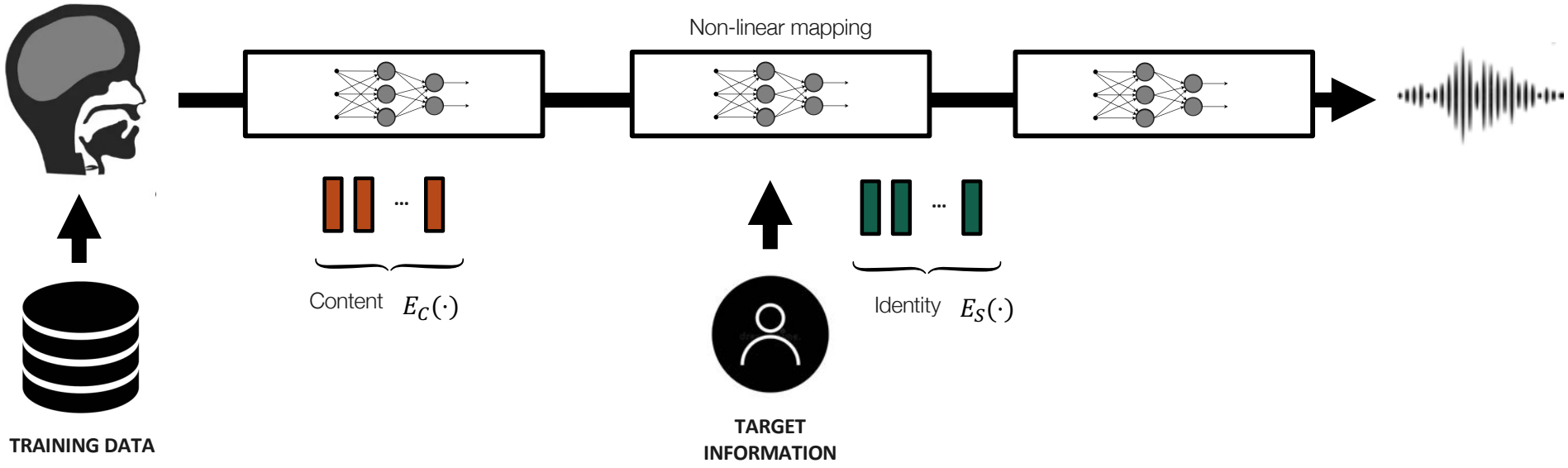
Synthesis: Inject and condition the generation on information from the target, such as identity.



VOICE CONVERSION (NON-PARALLEL)

Recognition: Extract desired information such as linguistics and eliminate unwanted source information.

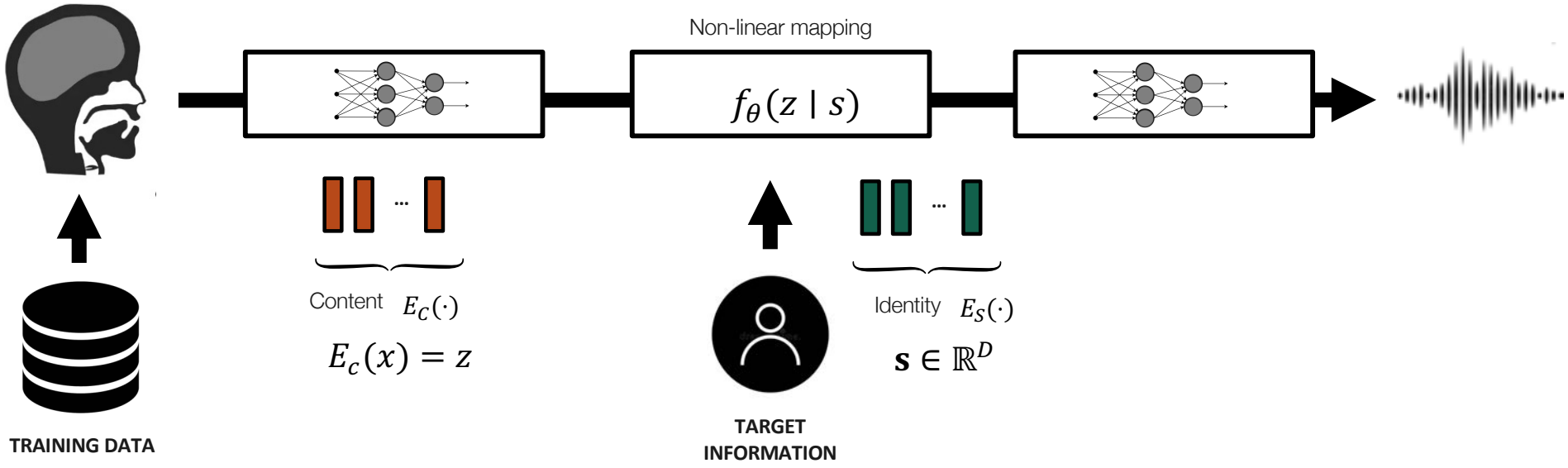
Synthesis: Inject and condition the generation on information from the target, such as identity.



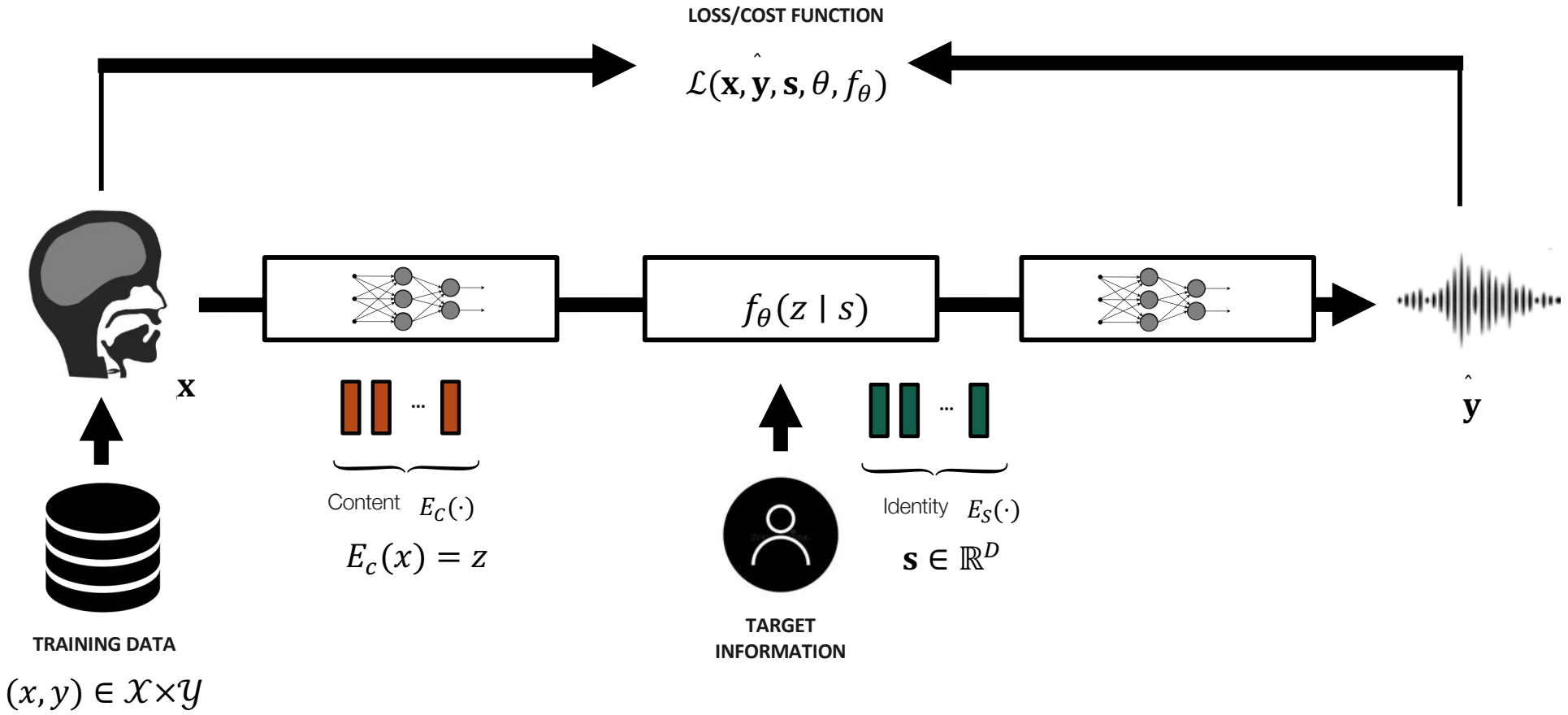
VOICE CONVERSION (NON-PARALLEL)

Recognition: Extract desired information such as linguistics and eliminate unwanted source information.

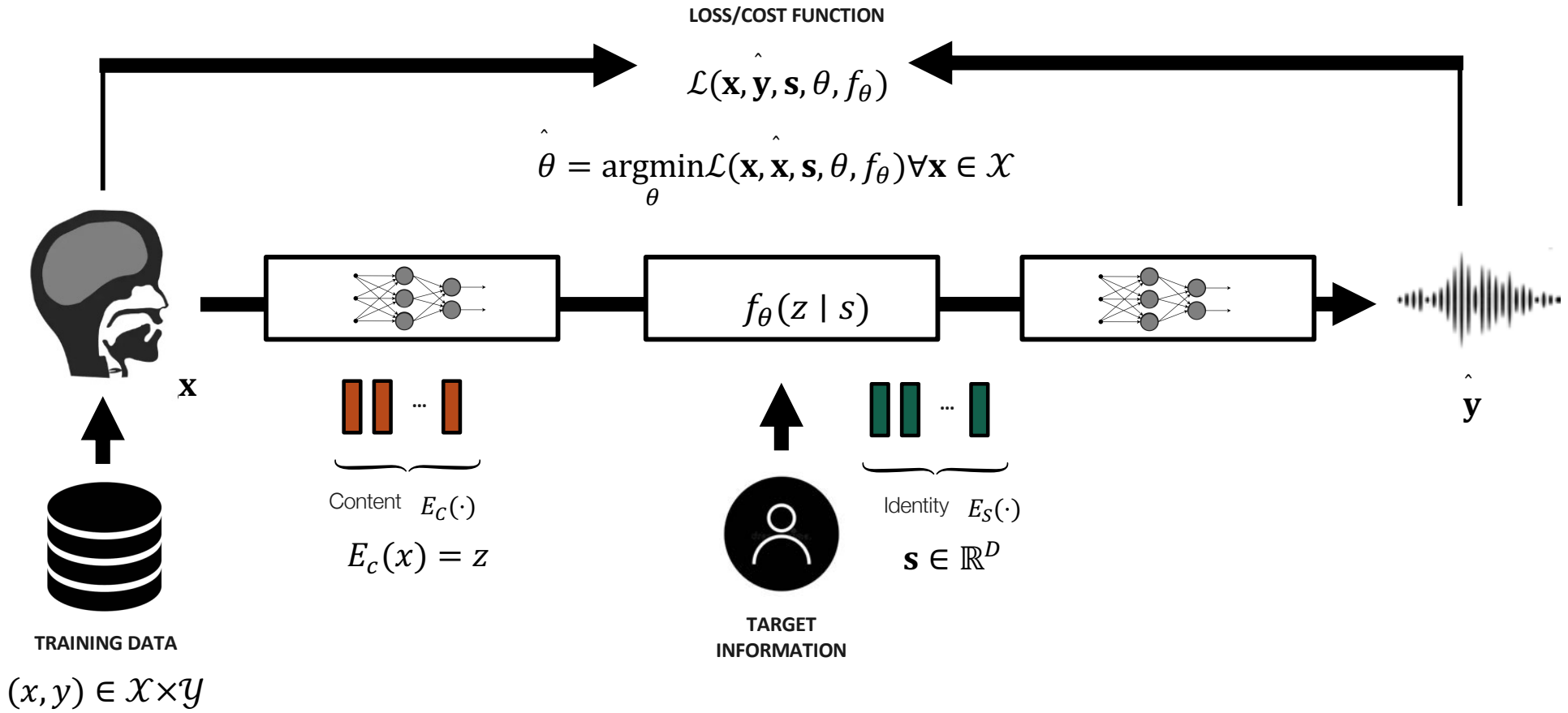
Synthesis: Inject and condition the generation on information from the target, such as identity.



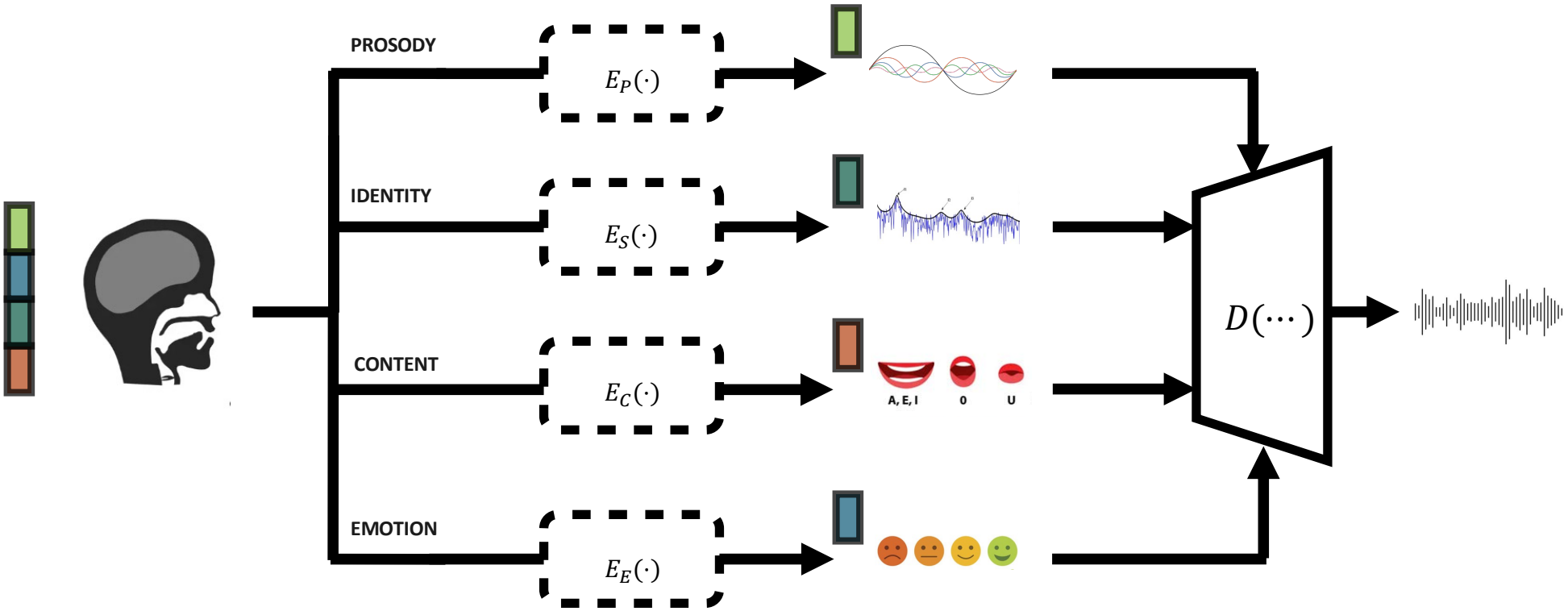
VOICE CONVERSION (NON-PARALLEL)



VOICE CONVERSION (NON-PARALLEL)



VOICE CONVERSION (NON-PARALLEL)

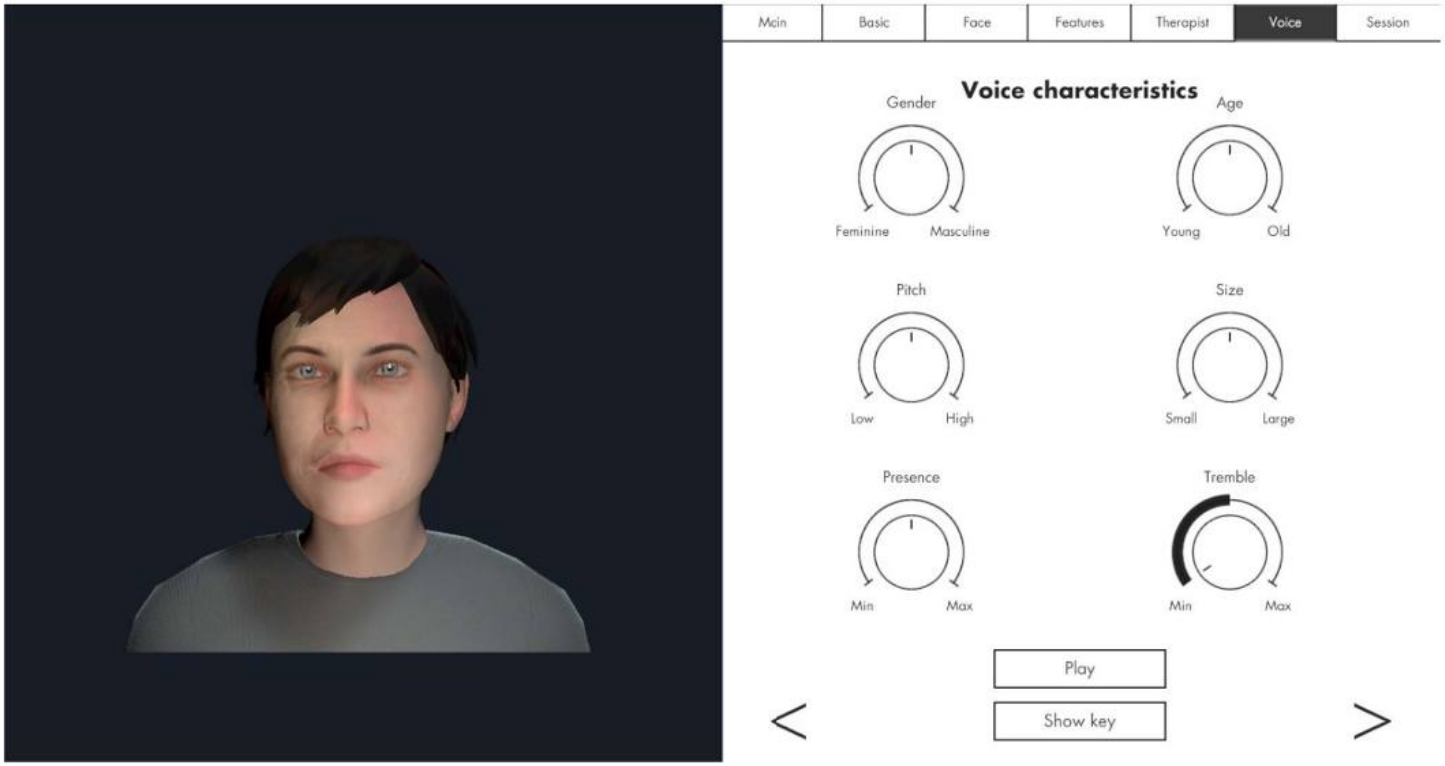


The CHALLENGE Project: Fighting Auditory Hallucinations by using Virtual Reality

Luis Viera, Peter Fisher and Simon Lajboschitz *
Khora Virtual Reality

Stefania Serafin †
Multisensory Experience Lab, Aalborg University Copenhagen

Merete Nordentoft ‡
Psykiatrisk Center København



AVATAR THERAPY

0 1

Voice Manipulation

AI-driven voice synthesis to make the avatar voice sound more natural

TTS / Voice Cloning

Naturalness & Prosody

0 2

Controllability

The voice should be controllable, natural and match human descriptors.

Controllability

Personification

0 3

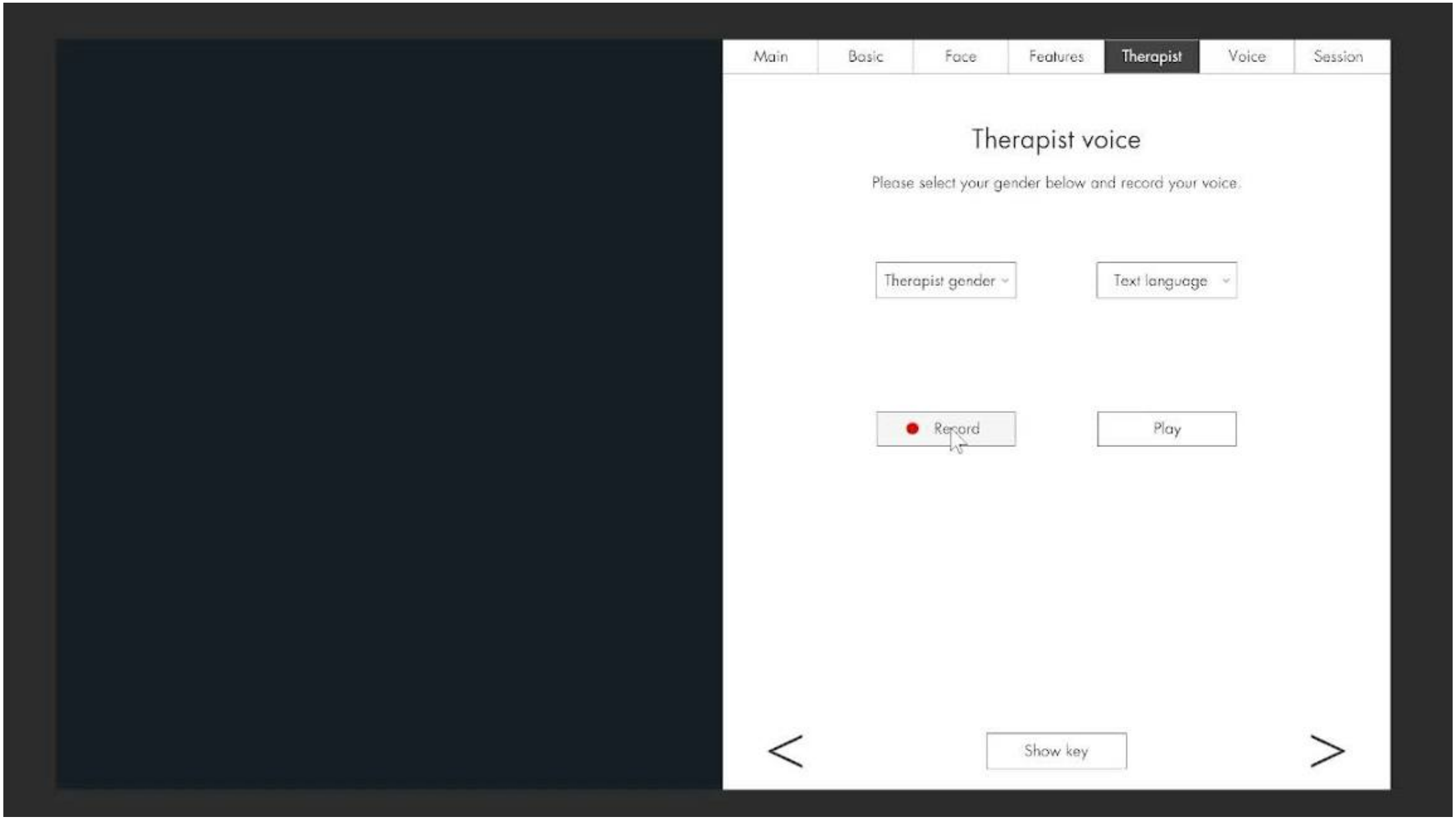
Technical Specs

Latency < 100 ms, with the ability to run the system on a consumer CPU (Unity).

Latency

Deployment

[C. J. Edwards, et al., "Thevoicecharacterisation checklist: psychometric properties of a brief clinical assessment of voices as socialagents," Frontiers in Psychiatry, vol. Volume 14, 2023.

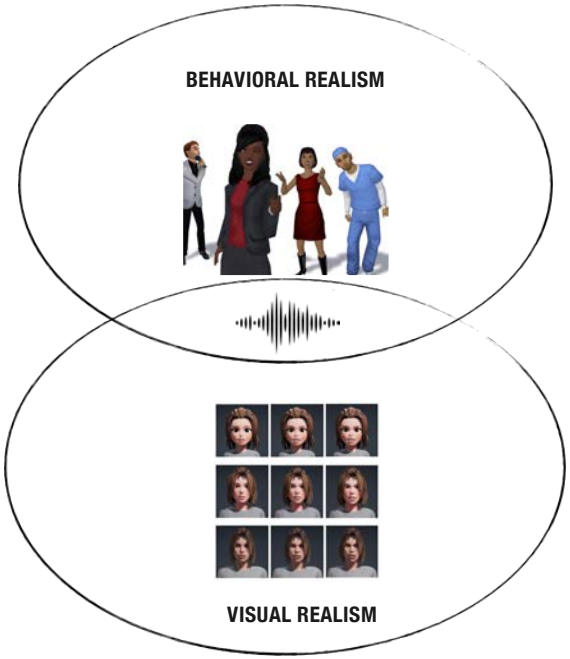

Main
Basic
Face
Features
Therapist
Voice
Session
Therapist voice
Please select your gender below and record your voice.
Therapist gender
Text language
Record
Play
Show key

MOTIVATION

- A cartoonish avatar can maintain a **strong social presence** if it exhibits human-like behaviour and voice.

- Synthetically created voices are perceived as less appealing when paired with photorealistic avatars.

- Virtual characters with human voices are generally considered more understandable and expressive than their synthetic counterparts.



Zegaran et. Al, “Does avatar design in educational games promote a positive emotional experience among learners?” *SAGE E-Learning and Digital Media*, 2021.
Higgins et. Al, “Sympathy for the digital: Influence of synthetic voice on affinity, social presence and empathy for photorealistic virtual humans”, *Computers & Graphics*, 2022.
[Cabral et. Al, “The Influence of Synthetic Voice on the Evaluation of a Virtual Character”, *ISCA Interspeech,* 2017.

EXPERIMENT

1. Realism, **2**. Social Presence, **3**. Emotional, **4**. Likeability

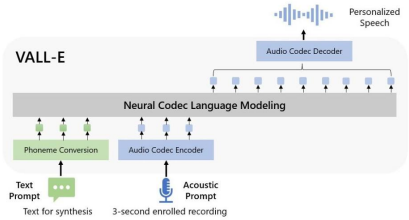
21 Participants (within subjects)

Bargum, A. R., Hansen, E. S., Erkut, C., & Serafin, S. (2025, March). Exploring the Impact of AI-Generated Speech on Avatar Perception and Realism in Virtual Reality Environments. In *2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (pp. 647-652). IEEE.

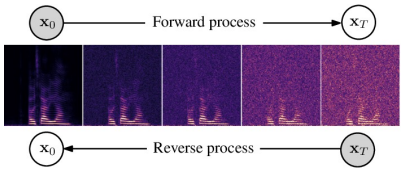
EXPERIMENT

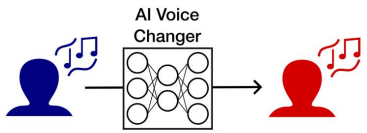
a) Coqui (TTS)

b) VALL-E (TTS)

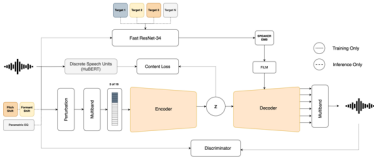


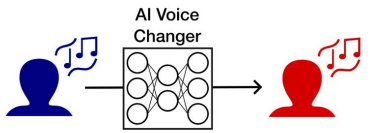
c) Diff-VC (STS)





d) S-RAVE (STS)





EXPERIMENT

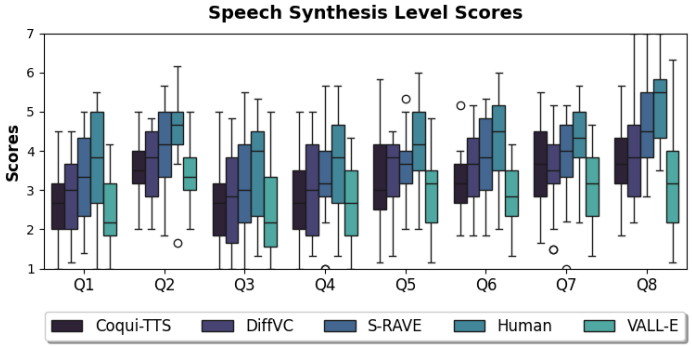
Category	Item	Question	ID
Realism	Overall realism	*"I found the person realistic overall"*	Q1
	Appearance	*"The appearance of the person matched the voice"*	Q2
Social Presence	General Presence	*"It feels like there was a presence of another person in the room with me"*	Q3
	Believeability	*"The person did not seem to be alive"*	Q4
Empathy	Emotion	*"I did not feel that the person's emotions seemed genuine"*	Q5
	Alignment	*"The emotions of the person matched the voice"*	Q6
	Expressivity	*"I felt the person was expressive"*	Q7
Voice Trait	Likeability	*"I disliked the voice of the person"*	Q8

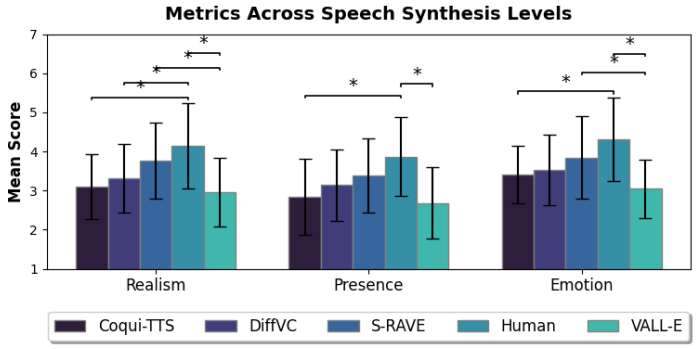
RESULTS

- None of the avatars were deemed realistic or present.

- Avatars using **STS-based** speech approached the human baseline on all metrics.

- STS methods outperform TTS systems in conveying **emotional subtleties**.
- **TTS systems have emotional limitations**: lack of intonation, pitch and natural phonetic coding within text-based-tokens.



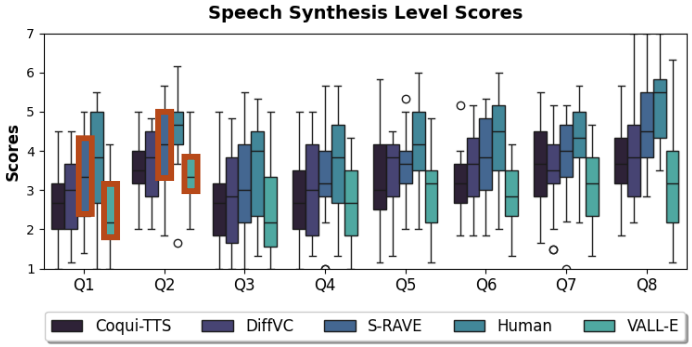


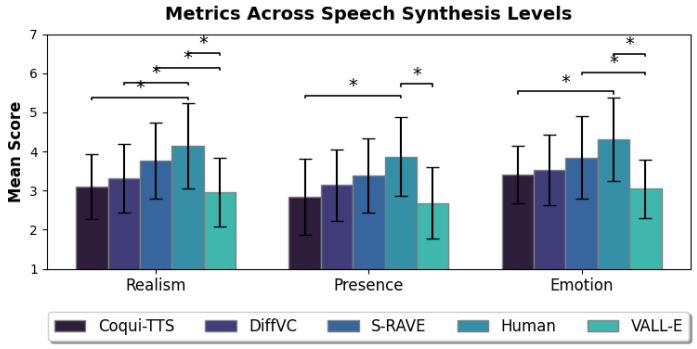
RESULTS

- None of the avatars were deemed realistic or present.

- Avatars using STS-based speech approached the human baseline on all metrics.

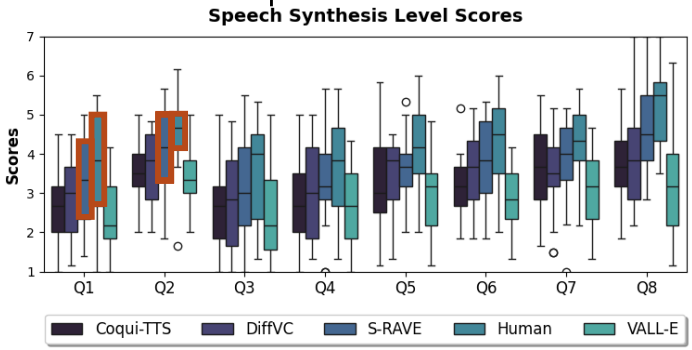
- STS methods outperform TTS systems in conveying emotional subtleties.
- **TTS systems have emotional limitations**: lack of intonation, pitch and natural phonetic coding within text-based-tokens.





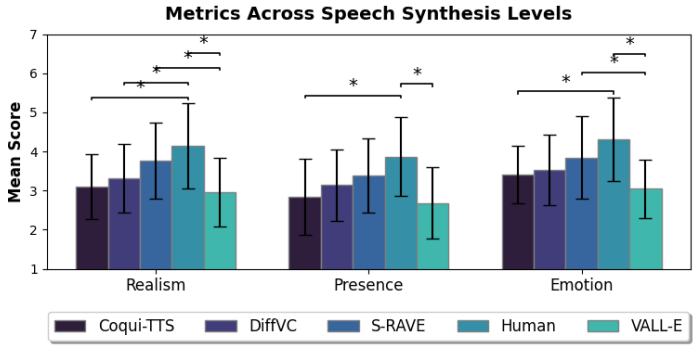
RESULTS

- None of the avatars were deemed realistic or present.



- Avatars using STS-based speech approached the human baseline on all metrics.

- STS methods outperform TTS systems in conveying emotional subtleties.
- **TTS systems have emotional limitations**: lack of intonation, pitch and natural phonetic coding within text-based-tokens.

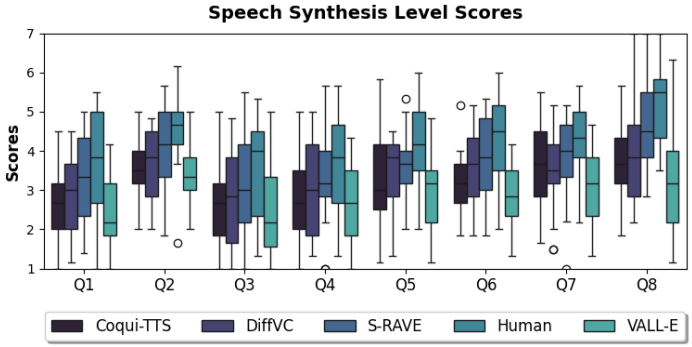


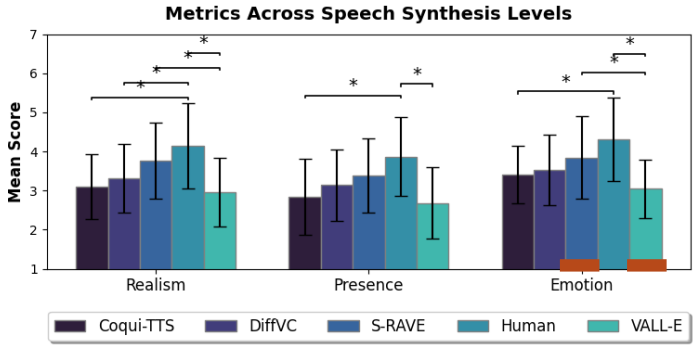
RESULTS

- None of the avatars were deemed realistic or present.

- Avatars using **STS-based** speech approached the human baseline on all metrics.

- STS methods outperform TTS systems in conveying **emotional subtleties**.
- **TTS systems have emotional limitations**: lack of intonation, pitch and natural phonetic coding within text-based-tokens.

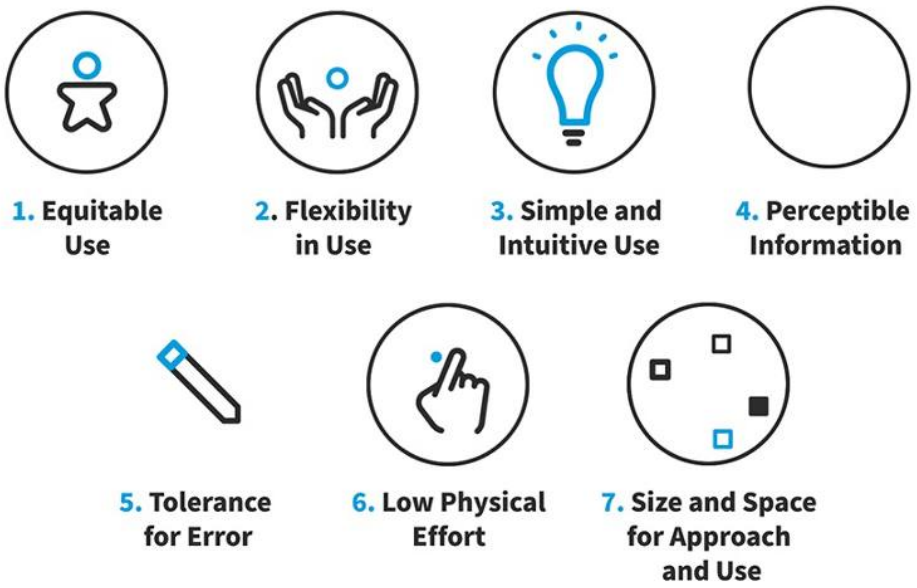


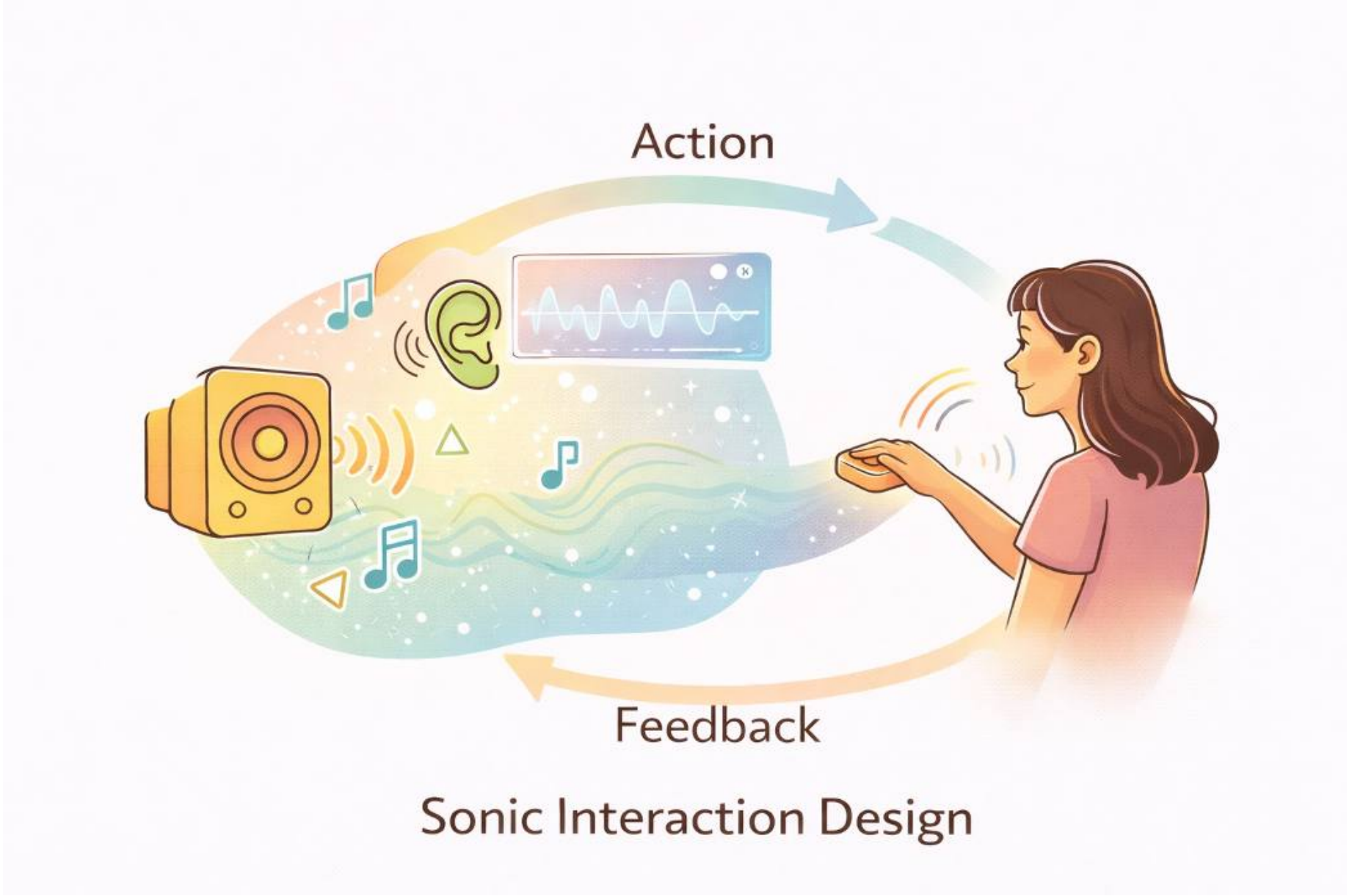


UNIVERSAL DESIGN

- Ronald Mace “The design of products and environments to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design.”

PRINCIPLES OF UNIVERSAL DESIGN




Action
Feedback
Sonic Interaction Design

from Sonic Interaction Design

to Sonic Universal Design

7 PRINCIPLES OF SONIC UNIVERSAL DESIGN

Designing sound experiences that are usable, comfortable, and meaningful for all, across diverse hearing abilities, contexts, and technologies.

1. Equitable Use

Sound experiences are usable and meaningful for people with a wide range of hearing abilities and technologies.

Key idea:
No one is excluded from the sonic experience.

2. Flexibility in Use

Sound can be experienced, customized, and interacted with in multiple ways to suit individual preferences and needs.

Key idea:
Provide choice, control, and adaptability.

3. Simple and Intuitive Use

Sonic interactions are easy to understand, learn, and use, regardless of experience, language, or hearing ability.

Key idea:
Make sound interactions clear and predictable.

4. Perceptible Information

Important information is conveyed through sound in ways that are detectable and understandable across diverse listening conditions.

Key idea:
Communicate clearly through multiple sonic cues.

5. Tolerance for Error

Sonic interactions are forgiving and support recovery from mistakes, misunderstandings, or missed cues.

Key idea:
Reduce stress and support confident interaction.

6. Low Physical Effort

Sound interactions require minimal physical, cognitive, and listening effort to use and perceive.

Key idea:
Make listening and interacting effortless and comfortable.

7. Size and Space for Approach and Use

Sonic interfaces scale across devices, environments, and situations, supporting orientation and access for all users.

Key idea:
Design for diverse bodies, spaces, and situations.

SID versus SUD

Past / Present	**Future**
Sonic Interaction Design	Sonic Universal Design
Sound as interaction modality	Sound as inclusive infrastructure
Designing interactions with sound	Designing for perceptual diversity
Adding sound to interfaces	Rethinking interfaces through sound
Generic users	Diverse users

THANKS!

Stefania Serafin

stefse@dtu.dk

https://www.linkedin.com/in/stefaniaserafin/

Copenhagen Hearing and Balance Centre

CHBC