1. TITLE

INNER VOICE

Psychoacoustic Style-Delta Extraction for Voice Perception Modeling

Meriç Çetin Öztürk
IEM — University of Music and Performing Arts Graz

SUPERVISOR: PROF. ALJANAKI

June 2025

2. OUTLINE

Outline

1. **Research Question** — Can we model the internal voice?
2. **Background** — Bone conduction psychoacoustics
3. **Method** — Skull resonance filter & ΔV extraction
4. **Findings** — CAM++ invariance, $\|\Delta V\| = 1.191$
5. **Applications** — Clinical voice preservation
6. **Future Work** — Roadmap to December

3. RESEARCH QUESTION

Research Question

"Can the perceptual difference between
air-conducted (external) and **bone-conducted (internal)** voice
be captured as a style delta vector
in speaker embedding space?"

External (Recording)

What others hear.
Air conduction only.
Full frequency spectrum.

Internal (Self-Perception)

What the speaker hears.
Air + bone conduction.
Lowpass filtered + occlusion boost.

Research Gap: Speaker embedding space × bone conduction psychoacoustics intersection is unexplored. No prior work models the "voice in your head" as a structured delta in embedding space.

4. BACKGROUND — BONE CONDUCTION

Background: Bone Conduction

Békésy (1949)

- First systematic study of bone conduction
- Skull vibration → cochlea directly (compressional wave)
- Lowpass filter characteristic at ~4 kHz
- Nobel Prize in Medicine (1961)
- Established BC as an independent auditory pathway

Stenfelt (2006)

- BCTF (Bone Conduction Transfer Function)
- 4 mechanisms: osseotympanic, inertial, compressional, occlusion effect
- Below 500 Hz: ~10 dB less sensitive than AC
- Frequency-dependent: BC efficiency peaks at 1-4 kHz

Occlusion Effect (Carillo 2021)

- Ear canal blocked → LF boost (10-15 dB)
- Resonance peak at ~2.8 kHz
- Frequency-dependent: affects 1-4 kHz range
- Explains "hollow" self-voice perception

Filter Design Rationale

- 3-stage cascade digital filter
- LP 4kHz (6th order Butterworth) ← Békésy
- EQ peak 2.8kHz +8 dB (Q=1.8) ← Carillo occlusion
- Peak filter 800Hz +5 dB (Q=1.2) ← middle-ear resonance
- Torchaudio biquad implementation

5. CAM++ ARCHITECTURE

CAM++ Speaker Embedding

What is CAM++?

A speaker embedding model — takes audio and produces
a **192-number voice fingerprint** of who is speaking.

Built on DenseTDNN architecture.
Seed-VC's default style encoder.

Why CAM++?

- **Native integration** — Seed-VC's built-in encoder,
  no pipeline modifications needed
- **Proven performance** — 200k+ speakers trained,
  verification state-of-the-art
- **Reproducible** — open source on HuggingFace,
  anyone can verify our results
- **192-dim L2-normalized** — ideal for delta vector computation and manipulation

What We Discovered: CAM++ is trained with spectral augmentation
— SpecAugment, random EQ, channel simulation — to *ignore* spectral differences.
Our skull resonance filter is interpreted as a "channel effect."
This is Finding 1 — not a bug, but a discovery.

Reference: 3D-Speaker — CAM++, 2023 (Alibaba DAMO Academy, HuggingFace)

6. METHOD — SKULL RESONANCE FILTER

Method: Skull Resonance Filter

Stage	Type	Parameters	Models
1	Lowpass (6th order Butterworth)	4 kHz, Q=0.707	Bone conduction attenuation
2	EQ Boost (Peaking)	2.8 kHz, +8 dB, Q=1.8	Occlusion effect resonance
3	Peak Filter (Peaking)	800 Hz, +5 dB, Q=1.2	Ear canal / middle-ear resonance

Pipeline: Recording → CAM++ → **V_raw** Recording → **Skull Filter** → CAM++ → **V_perceived**
→ **ΔV = V_perceived − V_raw**

CAM++: Alibaba DAMO Academy, 192-dim L2-normalized embedding, DenseTDNN architecture

7. SYSTEM IMPLEMENTATION — override_style

System Implementation

override_style

3-line change in Seed-VC's vc_wrapper.py.
Injects arbitrary style vector into CFM conditioning.

```
if override_style is not None:
    target_style = override_style.to(device, dtype)
else:
    target_style = self.compute_style(...)
```

ΔV Injection

style_modified = style_raw + ΔV × intensity
Standard: normal Seed-VC
Perceived: V_ref + ΔV

Gradio 3-tab UI (VC, VIS, INFO) | Lazy loading (552M model) | MPS AR cache | Streaming fix

8. SEED-VC ARCHITECTURE

Seed-VC V2 Architecture

Pipeline (552M params)

- HuBERT + ASTRAL → content encoder
- AR (autoregressive) transformer
- CFM / DiT diffusion decoder
- BigVGAN neural vocoder

Two Conditioning Paths

- **Style embedding** (192-dim) → CFM conditioning
- ← Our override hook
- **target_mel** → AR transformer prompt
- ← Carries stronger speaker identity

Finding 3 Preview: target_mel conditioning dominates style embedding.
Overriding style (our ΔV) changes speaker identity partially,
but target_mel carries the primary speaker information.

Reference: Seed-VC — Liu et al., arXiv:2411.09943, 2024

9. DEMO

Demo

[VC TAB]

Source Audio • Reference Audio
ΔV Toggle (ACTIVE / INACTIVE)
Intensity Slider (0.0 – 2.0)
Metrics: V_RAW, V_PER, ΔV, COS
→ STANDARD & PERCEIVED outputs

[VIS + INFO]

Filter frequency response (log scale)
ΔV heatmap (192-dim, green/+ magenta/-)
Audio spectrogram
Documentation & references

→ Filtered audio playback: *"This is what bone-conducted voice sounds like."*

10. FINDING 1 — CAM++ INVARIANCE

★ Finding 1: CAM++ Spectral Invariance

‖ΔV‖ = 1.191

ΔV norm is only ~6% of V_raw norm (19.81).
The skull resonance filter produces a minimal shift in CAM++ embedding space.

cos(V_raw, V_perceived) = 0.9967

Nearly identical vectors. CAM++ is highly invariant to spectral filtering.

Why? CAM++ is trained with spectral augmentation (SpecAugment, random EQ)
→ explicitly designed to ignore channel effects and spectral coloration.
This is the expected behavior of a well-trained verification model.

11. FINDING 2 — CONTROLLED ANALYSIS

★ Finding 2: Controlled Invariance Analysis

Filter Condition	‖ΔV‖	cos sim	ΔV / V_raw
Chosen (LP 4kHz + EQ 2.8kHz + Peak 800Hz)	1.191	0.9967	6.0%
Aggressive (LP 2kHz + EQ +12dB)	1.347	0.9958	6.8%
Single LP 4kHz (2-pole)	0.872	0.9979	4.4%
Random EQ (10-band ±6dB)	0.954	0.9972	4.8%

Even removing half the speech spectrum (LP 2kHz) → ΔV increases by only 13%.
Consistent with Wang et al. (2022) and Kwon et al. (2023) on embedding robustness.

12. FINDING 3 + CLINICAL APPLICATION

★ Finding 3 + Clinical Application

target_mel Dominance

CFM model uses target_mel as conditioning prompt.

The mel-spectrogram carries stronger speaker identity
than the 192-dim style embedding.

Two independent sources of speaker conditioning
→ override_style alone is insufficient.

Clinical Application

Laryngeal Cancer Rehabilitation

Pre-surgery: record voice, compute ΔV
Post-surgery: TEP speech → apply ΔV
→ preserves vocal identity + self-perception

Finding 3 published in Seed-VC technical report but not previously analyzed.

13. FUTURE WORK

Future Work

#	Task	Timeline	Impact
1	**target_mel override** — inject filtered reference mel into CFM	July	High
2	Replace CAM++ with **WavLM / HuBERT** (more filter-sensitive embeddings)	Aug–Sep	High
3	Subject study (n > 10) + listening tests	Sep–Oct	High
4	Aggressive filter parameters (LP 2kHz, EQ +12dB)	July	Medium
5	Clinical pilot — laryngeal cancer patients	Nov–Dec	Very High
6	Publication target: ICASSP / ISMIR	Dec	—

Immediate Next Step: target_mel override — filtered reference mel-spectrogram replaces original target_mel in CFM conditioning. 3-line change in `vc_wrapper.py` (parallel to override_style).

Project timeline: June → December 2025 (Master's thesis)

14. THANK YOU + Q&A

THANK YOU

Key References:
1. Seed-VC — Liu et al., arXiv:2411.09943, 2024
2. Bone conduction — Békésy, JASA 21(3), 1949
3. CAM++ — 3D-Speaker / Alibaba DAMO, HuggingFace
4. BCTF — Stenfelt, Audiology Research, 2006
5. CFM — Lipman et al., arXiv:2302.00482, 2023

Questions?
ozturkmericcetin@gmail.com
GitHub: github.com/mericnumeric/innervoice

IEM — INSTITUTE OF ELECTRONIC MUSIC AND ACOUSTICS