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

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

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.

BACKGROUND: BONE CONDUCTION

BÉKÉSY (1949)

- First systematic study of bone conduction
- Skull vibration → cochlea directly
- Lowpass filter at ~4 kHz
- Nobel Prize in Medicine (1961)

STENFELT (2006)

- BCTF measurements
- 4 mechanisms: osseotympanic, inertial, compressional, occlusion
- Below 500 Hz: 10dB less sensitive than AC

OCCLUSION EFFECT (CARILLO 2021)

- Ear canal blocked → LF boost
- Peaks at ~2.8 kHz
- Gain: 10–15 dB

OUR MODEL

- 3-stage digital filter
- LP 4kHz + EQ 2.8kHz + Peak 800Hz
- torchaudio biquad implementation

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

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

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."

★ 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.

★ 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.

★ 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.

FUTURE WORK

#	TASK	TIMELINE	IMPACT
1	**target_mel override** with filtered reference	July	High
2	Replace CAM++ with **WavLM / HuBERT**	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 workshop	Dec	—

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

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