VLSP 2026 · VINV-TTS SHARED TASK · TECHNICAL REPORT

Human-in-the-Loop Data Curation and Preference Optimization for Vietnamese Non-Verbal Text-to-Speech

SpeechWorld

Tri-Nhan Do · Minh-Khang Le · Hoang-Minh Tran · Thanh-Hau Luong · Long-Hai Tran · Viet-Lam Phung

Independent Researchers, Vietnam

THE TASK

Synthesize speech with four non-verbal sounds

Input: Vietnamese text with inline tags, plus a reference voice

> Họ chơi cũng không phải là quá hay đâu **[breathing]** nhưng mà cái cú đó…

[laughter] [breathing] [sniff] [throatclearing]

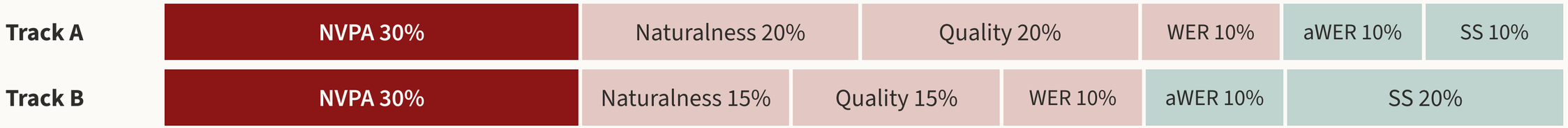
Track A

Speakers seen in training

Track B

Unseen speakers, zero-shot voice cloning

Every tag must be present, of the right type, at the right position. Official score:



Pink: rated by the organizers' listeners · green: automatic. Track A weighs naturalness and quality; Track B weighs speaker similarity.

NVPA: share of tags rendered with the right type at the right position. During development, the team used an **NVPA proxy**: MMS alignment of tag positions, AST embeddings, and a linear probe trained on the organizers' real audio.

THE CHALLENGE

Few clean examples, and no model to start from

12 / 69

sniff segments survive careful listening

19 / 117

throat-clearing segments survive; many tags are actually coughs

0 / 12

open-weight TTS models both speak Vietnamese and cover all four sounds

With only a few dozen clean examples, **which** examples the model sees matters more than how many. So the team let listeners choose.

SURVEY

Which public models support non-verbal sounds?

Open-weight model	Vietnamese	Laughter	Breathing	Sniff	Throat clearing	Coverage
OmniVoice	✓	✓	✗	✗	✗	1/4
Higgs TTS 3	✓	✓	≈	✓	≈	2/4
Fish Audio S2 Pro	✓	✓	✓	?	✓	3/4
VieNeu-TTS v3 Turbo	✓	≈	≈	✗	✓	1/4
VoxCPM2	✓	✗	✗	✗	✗	0/4
Dia	✗	✓	✓	✓	✓	4/4
Breeze-TTS-2	✗	✓	≈	✗	✓	2/4
Orpheus TTS	✗	✓	≈	✓	≈	2/4
OpenAudio S1-mini	✗	✓	≈	✗	✗	1/4
Chatterbox Turbo	?	✓	✗	✗	≈	1/4
Qwen3-TTS · Higgs Audio v2	✗	✗	✗	✗	✗	0/4

✓ the right sound · ≈ only a similar sound (sigh, cough) · ✗ not supported · ? free-form instruction tags, untested.
Sources: model cards, papers, official docs.

0 / 12

models both speak Vietnamese and support all four tags. Dia covers 4/4 but speaks English only.

FUN FACT · OCT 13, 2026

Fish Audio Drama 3 was released right after the competition ended

Open-weight, 35B MoE (3B active), 83 languages including Vietnamese. Tags are free-form instructions in square brackets. The team tried all four tags **laughter, breathing, sniff, throat clearing**: all of them are rendered.

BACKBONE

Voice cloning and audio quality come first

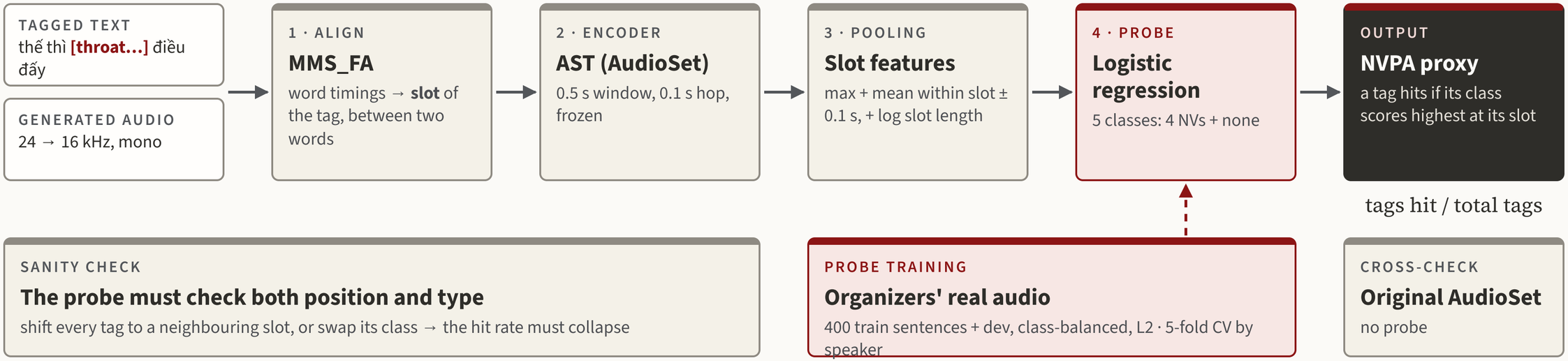
What we heard	Laughter	Breathing	Sniff	Throat clearing
Corpus (target)	crisp burst	quiet intake	sharp intake	often a cough
OmniVoice	light, short	becomes a sigh	missing	missing
Higgs	adds “haha”	intermittent	missing	clear
VieNeu-TTS	missing	becomes a sigh	like sobbing	hallucinates
Breeze-TTS-2	like a sniff	intermittent	too weak	mixed with coughs

No model renders all four sounds reliably. Breeze-TTS-2 also often hallucinates.

OmniVoice gave the cleanest speech and the best voice cloning, so the team started from it and taught it the missing sounds.

EVALUATION

NVPA proxy: an AST probe at the tag position

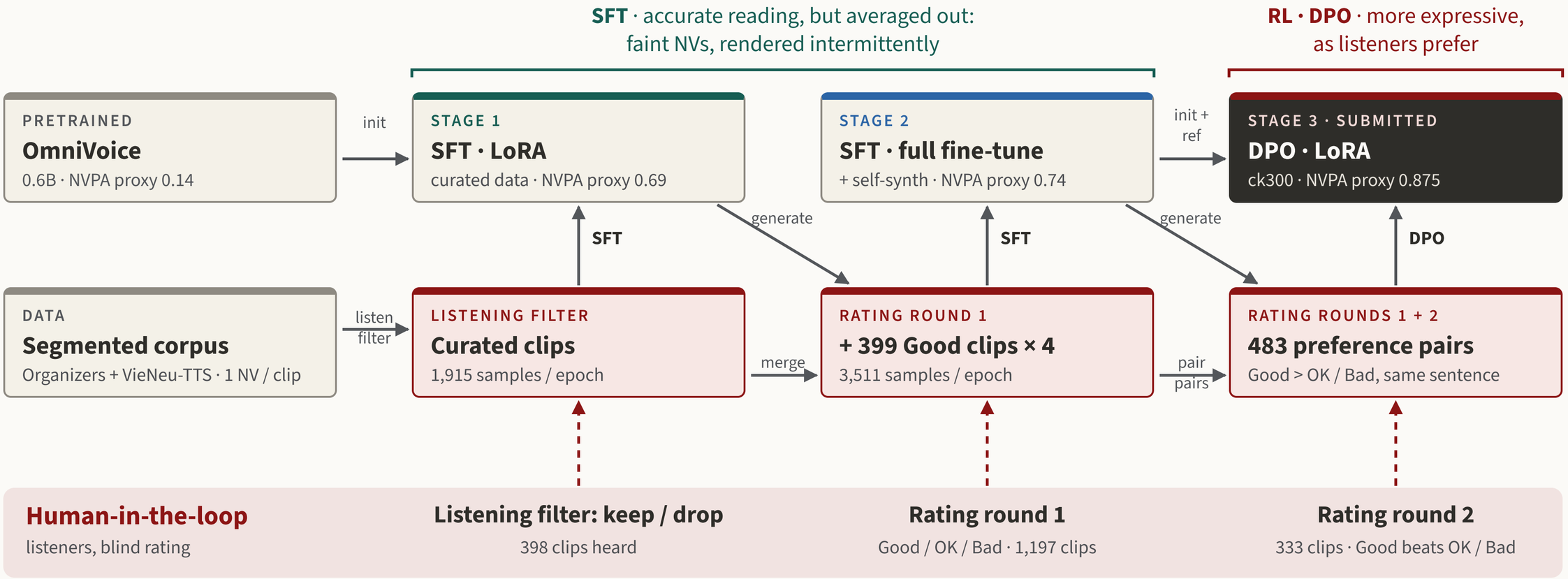


⚠ The NVPA proxy is only relative and is used to compare checkpoints. When checked by listening, the proxy was right only about 60%.

Note: this metric is **Agent Vibe**, vibe-coded by an AI agent and not yet thoroughly validated.

APPROACH

Human-in-the-loop across all three stages



Automatic metrics are used only to compare checkpoints, never to select training data or preference pairs.

STAGE 1 → 3 · DATA

Data selection flow: listeners filter at every step

	Laughter	Breathing	Sniff	Throat clearing	No tag	Total
Source Organizer data, aligned + cut	**744**	**4,267**	**69**	**117** + 4,556 VieNeu-TTS	**4,592** clean speech	
Listening filter keep or drop each clip	**keep all** 647 train, not heard	**89 / 212** kept / heard	**12 / 69** kept / heard	**19 / 117** + 60 / 409 VieNeu-TTS	**500** random sample	
Data Stage 1 samples / epoch, rare classes repeated	**647** × 1	**356** 89 × 4	**96** 12 × 8	**316** 79 × 4	**500** × 1	**1,915** samples / epoch
Rating round 1 blind rating of Stage 1 output	**92 / 300** Good	**92 / 300** Good	**119 / 300** Good	**96 / 297** Good	**—** not generated	**399** Good / 1,197
Data Stage 2 add Good clips × 4	**1,015** 647 + 92 × 4	**724** 356 + 92 × 4	**572** 96 + 119 × 4	**700** 316 + 96 × 4	**500** as in Stage 1	**3,511** samples / epoch
Rating round 2 blind rating of Stage 2 output	**62 / 85** Good	**48 / 84** Good	**64 / 84** Good	**40 / 80** Good	**—** not generated	**214** Good / 333
Data Stage 3 · DPO rounds 1 + 2 · Good > OK / Bad	**119** 106 + 13	**108** 94 + 14	**132** 126 + 6	**124** 110 + 14	**—**	**483** pairs · 436 + 47

Red rows are decided by listeners. Rating round 2: the Stage 2 model generates two versions per sentence, faithful and expressive. Adding OK clips to SFT performed worse, so only Good clips are used.

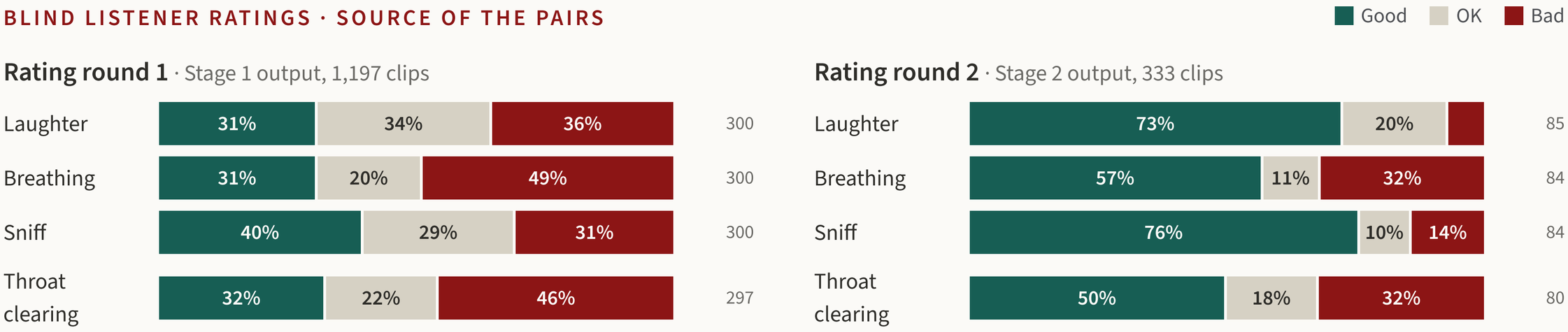
STAGE 3

DPO for a masked TTS model

WHY

SFT reads accurately but averages out: an NV is sometimes rendered, sometimes skipped, sometimes turned into a cough. RL on listener preferences makes the delivery more expressive.

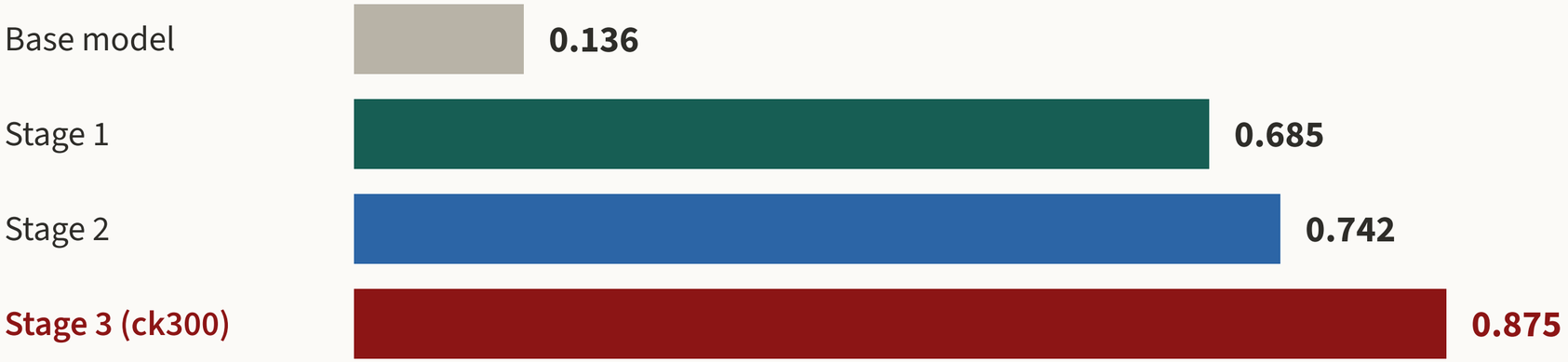
BLIND LISTENER RATINGS · SOURCE OF THE PAIRS



RESULTS · PUBLIC TEST

Every stage improves NV placement

NV placement accuracy (NVPA proxy) · organizers' public test: 30 sentences × 11 speakers (7 Track A, 4 Track B) = 330 utterances



AudioSet detection, no probe

13.9% → 83.8%

Long hallucinated pauses

12.1% → 4.2%

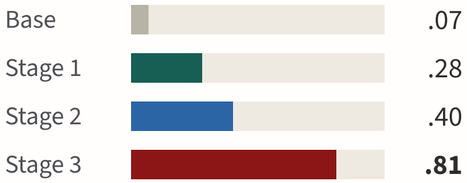
ASR WER

5.87% → 2.71%

INSIGHT · PER SOUND

Breathing is hardest; listening fixed the rare sounds

Laughter



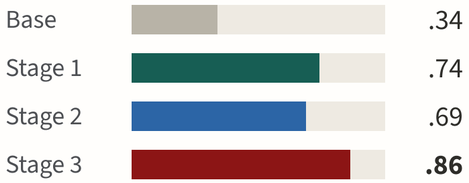
Only class not filtered by listening

Still the lowest after SFT: .28 → .40.

DPO gives the biggest jump: .40 → .81

Preferences shaped the laugh into a **stronger, heartier burst**.

Breathing



Hardest · lands at sentence start

Organizer speakers often inhale before speaking, and the model copies it.

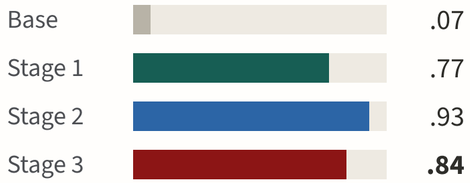
Hardest · what counts as a breath

Organizer breaths are light, almost a pause; our DPO ratings favoured that.

Next

Tag existing breaths, pad the reference with silence, and define the target breath when rating.

Sniff



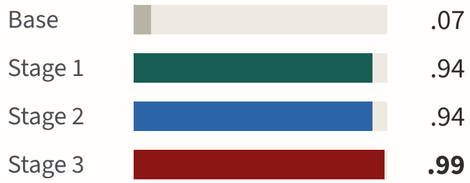
Few examples, yet the clearest

Only 12 clips after listening, yet rendered **more clearly than breathing**: only 3/88 turned into breathing.

Not always rising

Drops .93 → .84 after DPO.

Throat clearing



Listening fixed it

VieNeu-TTS clips often confuse a cough with throat clearing. We heard 409 and kept 60.

Now clear

Reaches .99 after DPO, the highest of the four classes.

NVPA proxy, 88 tags per sound. Stage 1 = LoRA, Stage 2 = full fine-tuning, Stage 3 = DPO ck300.

TRADE-OFF

Better sounds cost some speaker similarity

	Before DPO	After DPO · submitted	Δ
NVPA proxy	.742	**.875**	**+.133**
Speaker sim.	.748	**.678**	**−.070**
ASR WER	3.02%	**2.71%**	**−0.31**
Proxy score A	.804	**.853**	**+.049**
Proxy score B	.764	**.838**	**+.074**

Speaker similarity falls with each DPO step, most for seen speakers.

Why we still submitted the DPO model

NVPA weighs 3× speaker similarity in Track A and 1.5× in Track B, so the NV gain outweighs the loss.

Among the DPO checkpoints, the submitted one has the highest NVPA proxy, the lowest WER and the best Track B proxy.

DPO also adds about one extra NV per utterance at untagged positions.

TAKEAWAYS

Insights

- Curation beats quantity: 12 clean clips were enough to teach sniff
- The model's own best outputs are cheap, useful data
- The same ratings, used as preferences, gave the largest gain
- Listener labels keep the metrics from being “gamed”

Future Work

- Add a speaker-similarity term to the preference objective
- Denoise the reference audio at inference
- Pick a reference clip per sound at inference
- Filter out weak laughs automatically
- Re-detect long silences between words and add matching breath tags, using volume thresholds and signal patterns
- Mix in English NV data, or start from models such as Dia or Chatterbox-Turbo

Limitations: results are on the public test, which was also used to select checkpoints; each Stage 2 sentence had a single rater; private-test results have not been released.

SPEECHWORLD · VLSP 2026 VINV-TTS

Thank you to the organizers

Questions and discussion

dotrinhan99@gmail.com

Code, listener labels and evaluation scripts have been shared privately with the organizers.