VLSP 2026 · VINV-TTS SHARED TASK · BÁO CÁO KỸ THUẬT

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

BÀI TOÁN

Tổng hợp tiếng nói kèm bốn loại âm thanh non-verbal

Input: văn bản tiếng Việt có tag chèn giữa câu, kèm một giọng tham chiếu

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

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

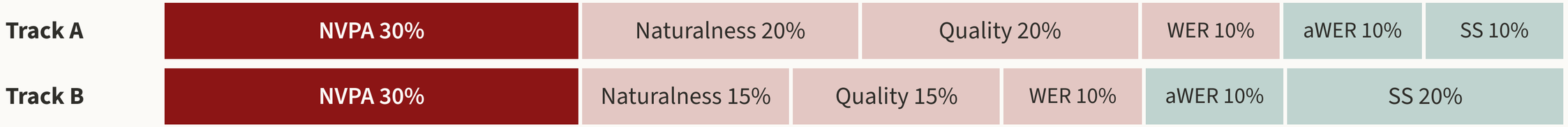
Track A

Người nói đã có trong dữ liệu train

Track B

Người nói chưa gặp, zero-shot voice cloning

Tag nào cũng phải xuất hiện, đúng loại, đúng vị trí. Điểm chính thức:



Hồng: người của BTC chấm · xanh: tự động. Track A nặng naturalness và quality; Track B nặng speaker similarity.

NVPA: tỉ lệ tag được sinh đúng loại, đúng vị trí. Lúc phát triển, team dùng **NVPA proxy**: MMS align vị trí tag, embedding AST, probe tuyến tính train trên audio thật BTC.

THÁCH THỨC

Ít mẫu sạch, lại không có model nào để bắt đầu

12 / 69

đoạn hít mũi còn lại sau khi nghe kỹ

19 / 117

đoạn hắng giọng còn lại; nhiều tag thực ra là tiếng ho

0 / 12

model TTS open-weight vừa nói được tiếng Việt, vừa có đủ bốn âm thanh

Khi chỉ có vài chục mẫu sạch, model được thấy mẫu **nào** quan trọng hơn là thấy bao nhiêu mẫu. Vì vậy, team để người nghe chọn.

KHẢO SÁT

Model public nào hỗ trợ non-verbal?

Model open-weight	Tiếng Việt	Cười	Thở	Hít mũi	Hắng giọng	Phủ
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

✓ đúng loại âm · ≈ chỉ có âm gần (thở dài, ho) · ✗ không có · ? tag dạng instruction tự do, chưa thử. Nguồn: model card, paper, docs chính thức.

0 / 12

model vừa nói tiếng Việt, vừa có đủ bốn tag. Dia đủ 4/4 nhưng chỉ nói tiếng Anh.

FUN FACT · 13/10/2026

Cuộc thi vừa xong thì Fish Audio Drama 3 ra mắt

Open-weight, 35B MoE (3B active), 83 ngôn ngữ, có tiếng Việt. Tag là instruction tự do trong ngoặc vuông. Team thử cả bốn tag **cười, thở, hít mũi, hắng giọng**: đều đọc được.

BACKBONE

Ưu tiên model có khả năng voice cloning và sinh audio chất lượng

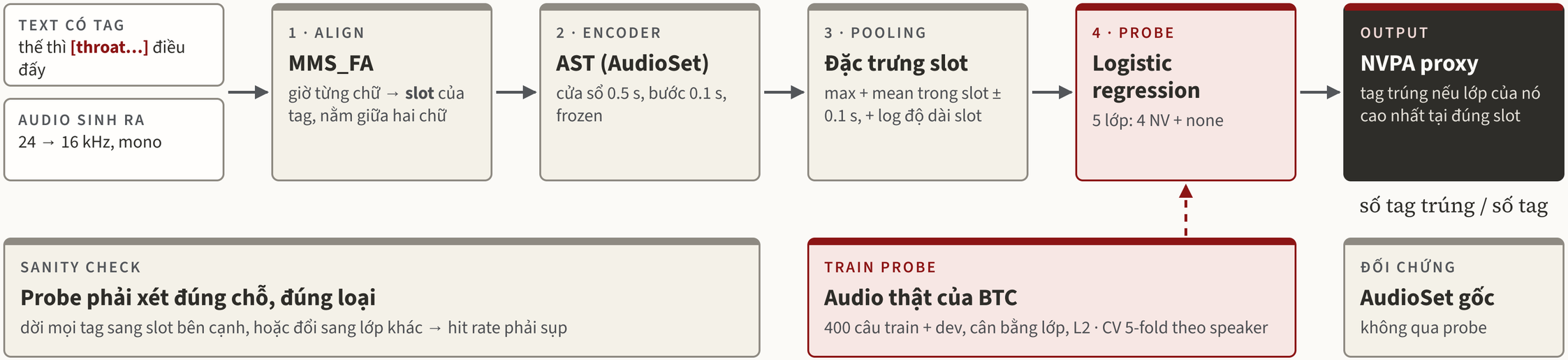
Nghe thấy gì	Cười	Thở	Hít mũi	Hắng giọng
Corpus (mục tiêu)	bật ra gọn	hít vào nhẹ	hít vào mạnh	hay lẫn tiếng ho
OmniVoice	nhẹ, ngắn	thành thở dài	không có	không có
Higgs	thêm “haha”	lúc có lúc không	không có	rõ
VieNeu-TTS	không có	thành thở dài	như đang khóc	bị hallucinate
Breeze-TTS-2	giống hít mũi	lúc có lúc không	quá yếu	lẫn tiếng ho

Không model nào tạo ổn định cả bốn âm thanh. Breeze-TTS-2 còn hay bị hallucinate.

OmniVoice cho tiếng nói sạch nhất và clone giọng tốt nhất, nên team xuất phát từ đây rồi dạy thêm các âm thanh còn thiếu.

ĐÁNH GIÁ

NVPA proxy: probe AST tại vị trí của tag

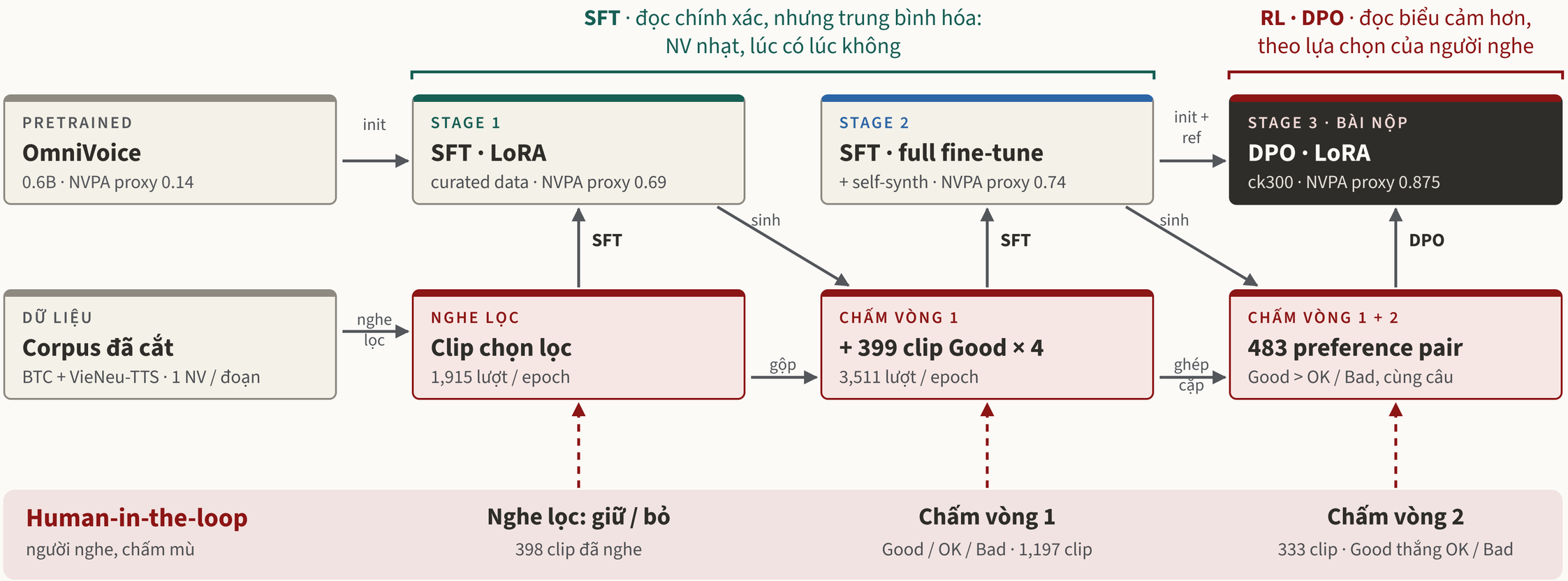


⚠ NVPA proxy chỉ mang tính tương đối, dùng để so sánh checkpoint. Khi nghe kiểm tra lại, tỉ lệ đúng của NVPA proxy chỉ khoảng 60%.

Note: metric này là **Agent Vibe**, do AI agent vibe-code, chưa được kiểm định kỹ.

HƯỚNG TIẾP CẬN

Human-in-the-loop xuyên suốt ba stage



Metric tự động chỉ dùng để so sánh checkpoint, không bao giờ dùng để chọn dữ liệu train hay preference pair.

STAGE 1 → 3 · DỮ LIỆU

Luồng chọn dữ liệu: người nghe lọc ở mọi bước

	Cười	Thở	Hít mũi	Hắng giọng	Không tag	Tổng
Nguồn BTC đã align + cắt	**744**	**4,267**	**69**	**117** + 4,556 VieNeu-TTS	**4,592** lời sạch	
Nghe lọc giữ hay bỏ từng clip	**giữ hết** 647 train, không nghe	**89 / 212** giữ / đã nghe	**12 / 69** giữ / đã nghe	**19 / 117** + 60 / 409 VieNeu-TTS	**500** chọn ngẫu nhiên	
Data Stage 1 lượt / epoch, lớp hiếm lặp lại	**647** × 1	**356** 89 × 4	**96** 12 × 8	**316** 79 × 4	**500** × 1	**1,915** lượt / epoch
Chấm vòng 1 chấm mù output Stage 1	**92 / 300** Good	**92 / 300** Good	**119 / 300** Good	**96 / 297** Good	**—** không sinh	**399** Good / 1,197
Data Stage 2 thêm clip Good × 4	**1,015** 647 + 92 × 4	**724** 356 + 92 × 4	**572** 96 + 119 × 4	**700** 316 + 96 × 4	**500** như Stage 1	**3,511** lượt / epoch
Chấm vòng 2 chấm mù output Stage 2	**62 / 85** Good	**48 / 84** Good	**64 / 84** Good	**40 / 80** Good	**—** không sinh	**214** Good / 333
Data Stage 3 · DPO vòng 1 + vòng 2 · Good > OK / Bad	**119** 106 + 13	**108** 94 + 14	**132** 126 + 6	**124** 110 + 14	**—**	**483** cặp · 436 + 47

Hàng đỏ do người nghe quyết định. Chấm vòng 2: model Stage 2 sinh mỗi câu hai bản, bám chữ và biểu cảm. Thêm clip OK vào SFT cho kết quả kém hơn, nên chỉ dùng clip Good.

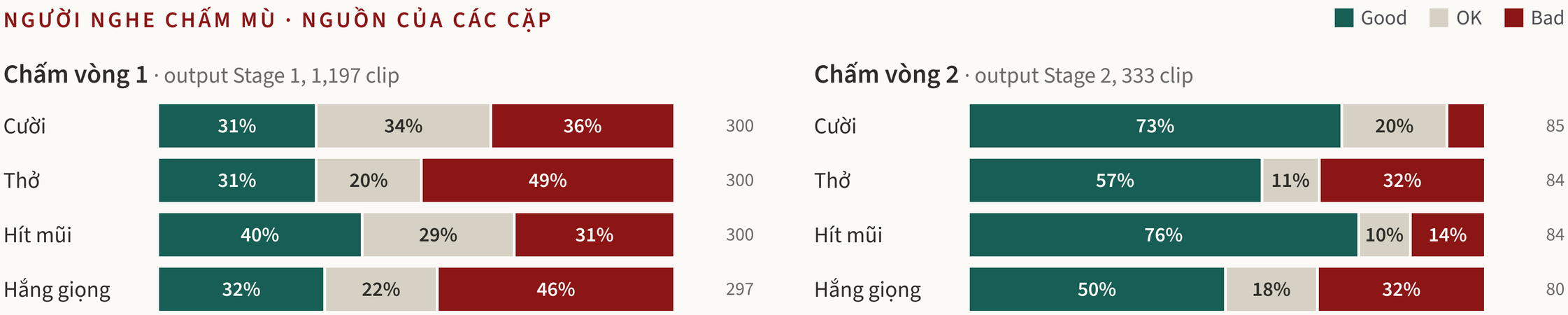
STAGE 3

DPO cho masked TTS model

VÌ SAO

SFT giúp đọc chính xác nhưng trung bình hóa: NV lúc có, lúc bỏ, lúc thành tiếng ho. RL với preference của người nghe giúp đọc biểu cảm hơn.

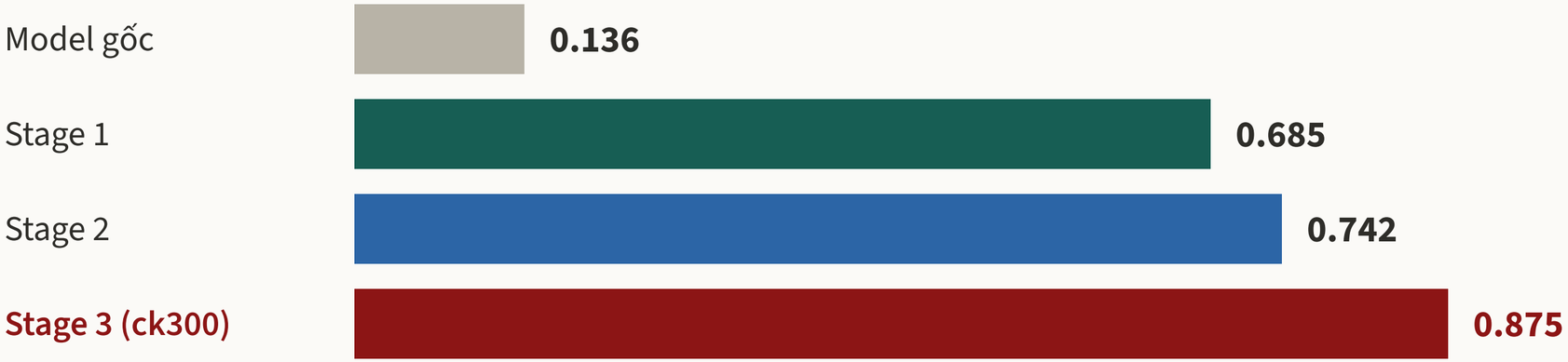
NGƯỜI NGHE CHẤM MÙ · NGUỒN CỦA CÁC CẶP



KẾT QUẢ · PUBLIC TEST

Stage nào cũng nâng độ chính xác vị trí NV

Độ chính xác vị trí NV (NVPA proxy) · public test BTC: 30 câu × 11 người nói (7 Track A, 4 Track B) = 330 câu



Phát hiện bằng AudioSet, không probe

13.9% → 83.8%

Khoảng lặng dài do hallucinate

12.1% → 4.2%

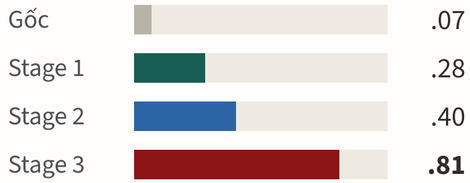
ASR WER

5.87% → 2.71%

INSIGHT · THEO TỪNG ÂM THANH

Thở khó nhất; nghe lọc giúp hít mũi và hắng giọng rõ

Cười



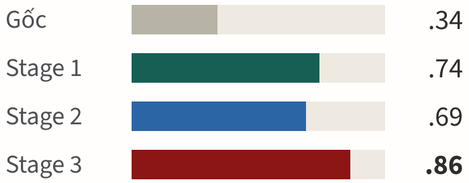
Lớp duy nhất không nghe lọc

Sau SFT vẫn thấp nhất: .28 → .40.

DPO tăng mạnh nhất: .40 → .81

Preference chỉnh thành tiếng cười **bật ra mạnh, hào sảng hơn**.

Thở



Khó nhất · rơi vào đầu câu

Người nói trong data BTC hay hít vào trước khi bắt đầu, model học luôn pattern đó.

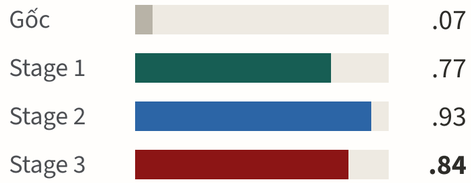
Khó nhất · thế nào là thở

Thở trong BTC nhẹ, gần như khoảng lặng; lúc chấm DPO team nghiêng về kiểu đó.

Hướng xử lý

Gắn tag cho tiếng thở có sẵn, đệm khoảng lặng vào reference; nêu rõ kiểu thở mục tiêu khi chấm.

Hít mũi



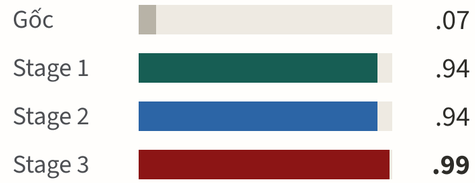
Ít mẫu nhưng rõ nhất

Chỉ 12 clip sau nghe lọc, nhưng phát âm **rõ hơn cả tiếng thở**: chỉ 3/88 bị thành tiếng thở.

Không phải lúc nào cũng tăng

Giảm .93 → .84 sau DPO.

Hắng giọng



Nghe lọc sửa được

Clip VieNeu-TTS hay nhầm tiếng ho với hắng giọng. Nghe 409, giữ 60.

Âm đã rõ

Đạt .99 sau DPO, cao nhất trong bốn lớp.

NVPA proxy, 88 tag mỗi âm thanh. Stage 1 = LoRA, Stage 2 = full fine-tuning, Stage 3 = DPO ck300.

TRADE OFF

Âm thanh tốt hơn, đổi lại mất chút speaker similarity

	Trước DPO	Sau DPO · nộp	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 giảm dần theo số bước DPO, giảm nhiều nhất với người nói đã gặp.

Vì sao vẫn nộp bản DPO

NVPA nặng gấp 3 lần speaker similarity ở Track A và 1.5 lần ở Track B, nên phần tăng NV lớn hơn phần mất.

Trong các bước DPO, bản nộp có NVPA proxy cao nhất, WER thấp nhất và proxy Track B tốt nhất.

DPO cũng sinh thêm khoảng một NV mỗi câu ở những vị trí không có tag.

TỔNG KẾT

Insight

- Chọn lọc quan trọng hơn số lượng: 12 clip sạch là đủ dạy hít mũi
- Output tốt nhất của chính model là dữ liệu rẻ mà hữu ích
- Cùng các đánh giá đó, dùng làm preference, cho mức tăng lớn nhất
- Nhãn từ người nghe giúp metric không bị "ăn gian"

Future Work

- Thêm thành phần speaker similarity vào hàm mục tiêu preference
- Khử nhiễu audio tham chiếu lúc inference
- Chọn clip tham chiếu riêng cho từng âm thanh lúc inference
- Tự động lọc bỏ các tiếng cười yếu
- Detect lại silence dài giữa các word và thêm tag breath phù hợp, theo threshold volume và signal pattern
- Trộn thêm dữ liệu NV tiếng Anh, hoặc xuất phát từ model như Dia hay Chatterbox-Turbo

Hạn chế: kết quả đo trên public test, cũng là tập dùng để chọn checkpoint; mỗi câu ở Stage 2 chỉ có một người chấm; kết quả private test chưa được công bố.

SPEECHWORLD · VLSP 2026 VINV-TTS

Xin cảm ơn thầy cô trong BTC

Hỏi đáp và thảo luận

dotrinhan99@gmail.com

Code, nhãn của người nghe và script đánh giá đã được chia sẻ riêng với ban tổ chức.