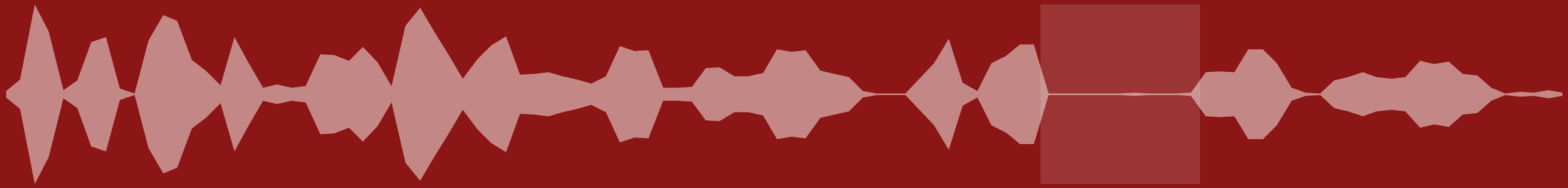
VLSP 2026 · SHARED TASK SPEECH QUALITY ASSESSMENT

Simulation-Guided DSP–SSL Hybrids for Cross-Lingual Speech Quality Diagnosis

Tri-Nhan Do, Long-Hai Tran, Minh-Toan Tran

Team SpeechWorld · Independent Researchers, Vietnam

VLSP 2026 Workshop · 15/11/2026

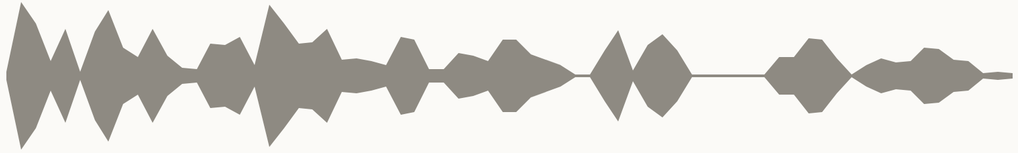


BÀI TOÁN

Không chỉ chấm chất lượng, mà phải chỉ ra lỗi gì, ở đâu

Input: một bản ghi bị lỗi, không có tín hiệu tham chiếu

waveform



Output: một quality score và các fault interval

noise

distortion

pause — trống: file không có lỗi này

quality — **3.4 / 5**

Ví dụ thật: file private pt_00925, nhãn đã công bố, kèm điểm do model của chúng tôi dự đoán. Fault interval được chấm trên lưới 20 ms.

Cách tính điểm

$$\text{Final} = 0.4\ \text{PCC} + 0.3\ \text{F1} + 0.3\ \text{IoU}$$

Quality PCC 40%	Noise 20%		Distortion 20%		Pause 20%	
mỗi lỗi = F1 + IoU → 10% + 10%, đúng tỉ lệ	F1	IoU	F1	IoU	F1	IoU

60% số điểm nằm ở phần chẩn đoán, không phải điểm chất lượng.

F1 tính theo file. **IoU** chỉ tính trên các file thật sự có lỗi đó.

Mỗi lỗi nặng 20%, dù hiếm đến đâu. Chi tiết này sẽ quan trọng về sau.

Thiếu quality score hay thiếu fault interval thì bài nộp đều không hợp lệ.

THÁCH THỨC

Lệch cả domain lẫn ngôn ngữ: train tiếng Anh, test tiếng Việt

QualiSpeech: 14,577 clip tiếng Anh, khoảng một nửa là TTS

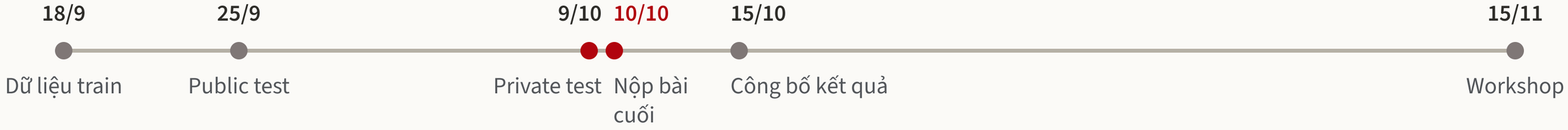
Private: 302 file tiếng Việt



Độ rộng đúng tỉ lệ. Tập private chỉ bằng khoảng 2% QualiSpeech, và nhãn chỉ được công bố sau khi kết thúc đánh giá.

ĐIỀU NÀY CÓ NGHĨA LÀ	CÁCH CHÚNG TÔI XỬ LÝ
Không có tiếng Việt nào được gán fault interval	→ **Mô phỏng lỗi trên tiếng Việt sạch**
The encoder understands speech only	→ **Thêm đặc trưng xử lý tín hiệu, độc lập ngôn ngữ**

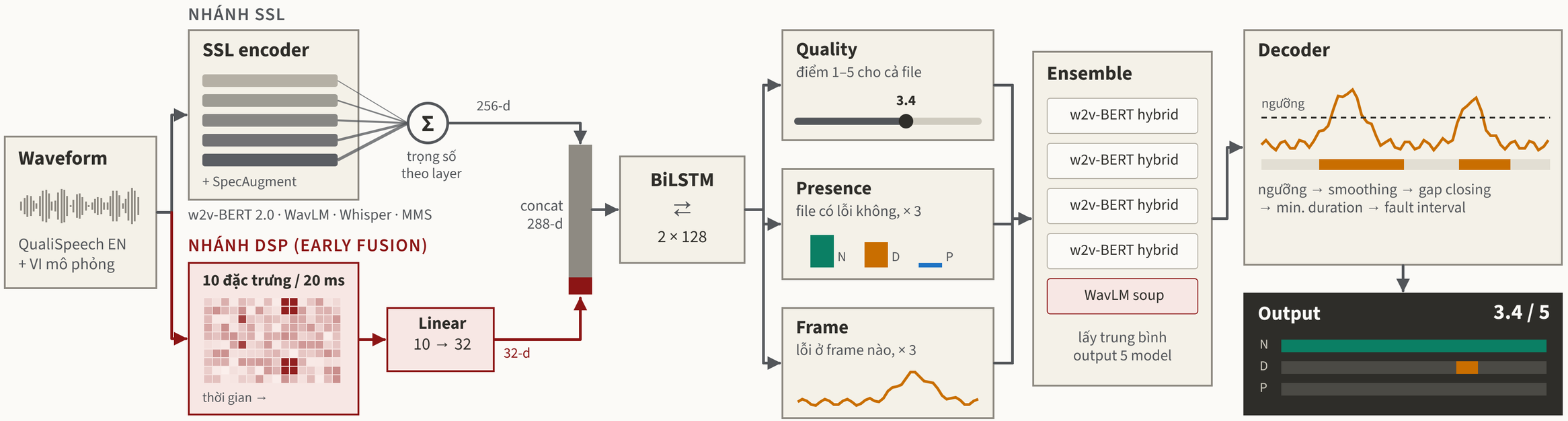
MỐC THỜI GIAN 2026 — Ba tuần từ lúc có dữ liệu train đến private test, thêm sáu ngày là có kết quả.



Nguồn: trang bài toán VLSP 2026 SQA; QualiSpeech (Wang et al., ACL 2025)

HỆ THỐNG

Thêm một nhánh DSP bên cạnh SSL encoder



DSP bù cho chỗ SSL bỏ sót

SSL encoder học nội dung tiếng nói, nên một chuỗi zero, một đoạn clipping hay chỗ mất băng tần gần như không để lại dấu vết. Đặc trưng DSP thì bắt được ngay.

HỆ THỐNG

Bộ mô phỏng tạo lỗi tiếng Việt kèm fault interval chính xác

Tiếng Việt sạch
PhoAudioBook, Vivoice, LSVSC, ViSEC, Dolly

↓

Chèn lỗi, ghi lại chính xác fault interval
noise, distortion, pause

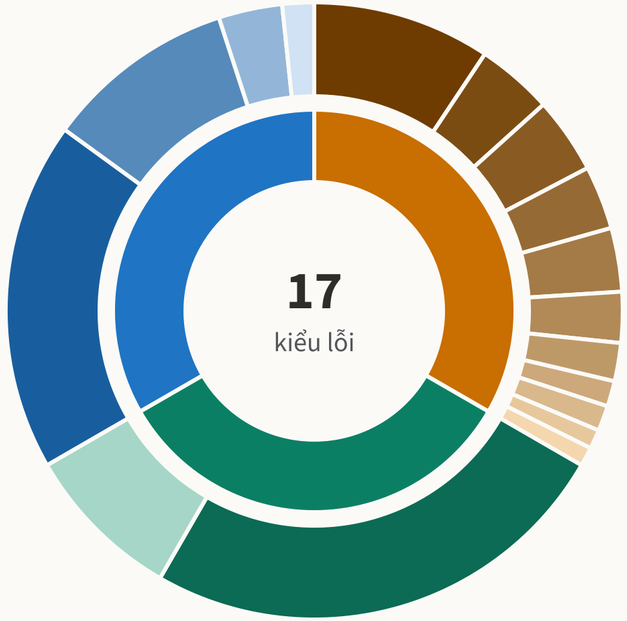
↓

12,000 file, 24.8 giờ
1,000 file để làm dev, không gán nhãn tay

Bộ mô phỏng qua bốn lần cải tiến. Riêng bước từ v1 lên v3/v4 đã đưa distortion IoU trên private từ **.411 lên .448**, dù không dùng nhãn lỗi tiếng Việt thật nào.

Mỗi nhóm lỗi gồm những kiểu nào (bộ mô phỏng cuối)

Vòng trong: 3 nhóm lỗi, được chèn với xác suất gần như nhau. Vòng ngoài: tỉ lệ từng kiểu trong nhóm.



Distortion

Cụm dropout	28%
Clip + attenuate	12%
Musical noise	12%
Low-pass theo đoạn	10%
Lượng tử hoá codec	10%
AGC pumping	8%
Overdrive	6%
Bit-crushing	4%
Pitch shift	4%
Griffin-Lim	3%
Aliasing	3%

Noise

Bản ghi thật	75%
Static tổng hợp	25%

Unnatural pause

Một khoảng lặng	55%
Nhiều khoảng lặng	30%
Stutter	10%
Ngắt quãng	5%

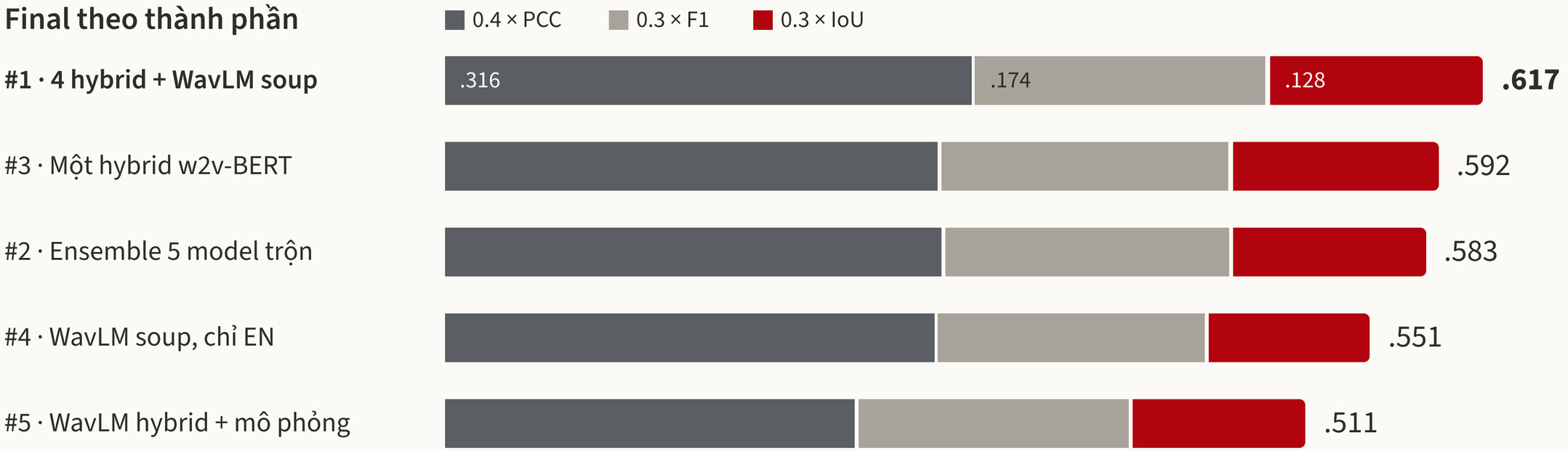
25% số file không có lỗi; ở các file còn lại, mỗi nhóm lỗi được chèn với xác suất 0.40–0.45. Noise có SNR 10–28 dB.

Một lựa chọn sẽ nhắc lại sau: khoảng lặng chèn vào được gán nhãn pause, giống QualiSpeech.

Cụm dropout: 3–9 khoảng zero số dài 40–100 ms. Pause: khoảng lặng 0.25–0.85 s được lấp bằng room tone.

KẾT QUẢ

Kết quả chính thức: bài tốt nhất trong năm bài nộp đạt 0.617



Mô phỏng giúp định vị lỗi

Từ soup chỉ học tiếng Anh (#4) sang một hybrid có dữ liệu mô phỏng (#3), distortion IoU tăng từ **.242 lên .462**.

Ensemble cải thiện quality

#1 giữ được khả năng định vị của các hybrid, đồng thời nâng PCC từ **.740 lên .789**, cao nhất của chúng tôi.

Pause là ngoại lệ

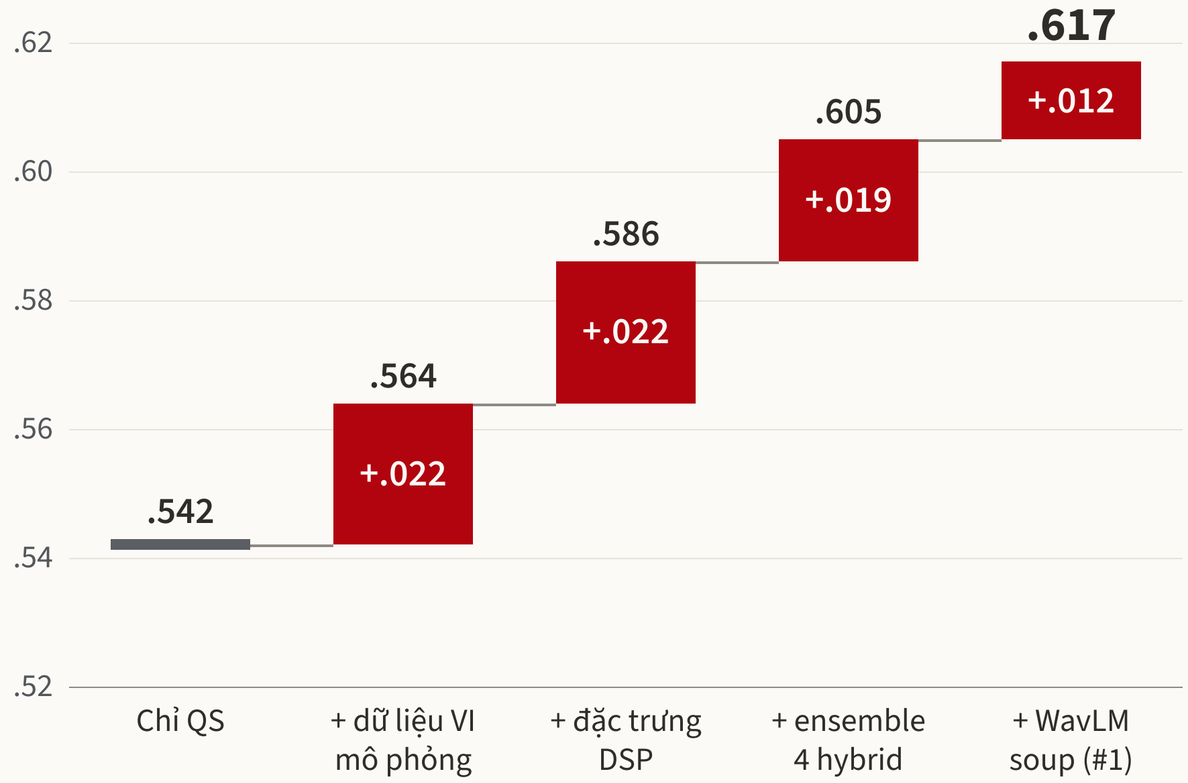
Pause F1 luôn **dưới .10** ở cả năm bài nộp, dù đổi backbone hay đổi dữ liệu train.

Điểm private test chính thức trên 302 file tiếng Việt. Soup: trung bình trọng số của WavLM từ epoch 3 đến 6.

ABLATION

Mỗi thành phần góp thêm khoảng .02 điểm

Private Final qua từng bước (w2v-BERT 2.0)



DSP cần dữ liệu mô phỏng

.542
Chỉ QS, không DSP

.534
Chỉ QS, có DSP

Không có dữ liệu mô phỏng thì DSP không giúp gì.

Encoder quan trọng không kém dữ liệu

.592
w2v-BERT 2.0

.494
WavLM-large

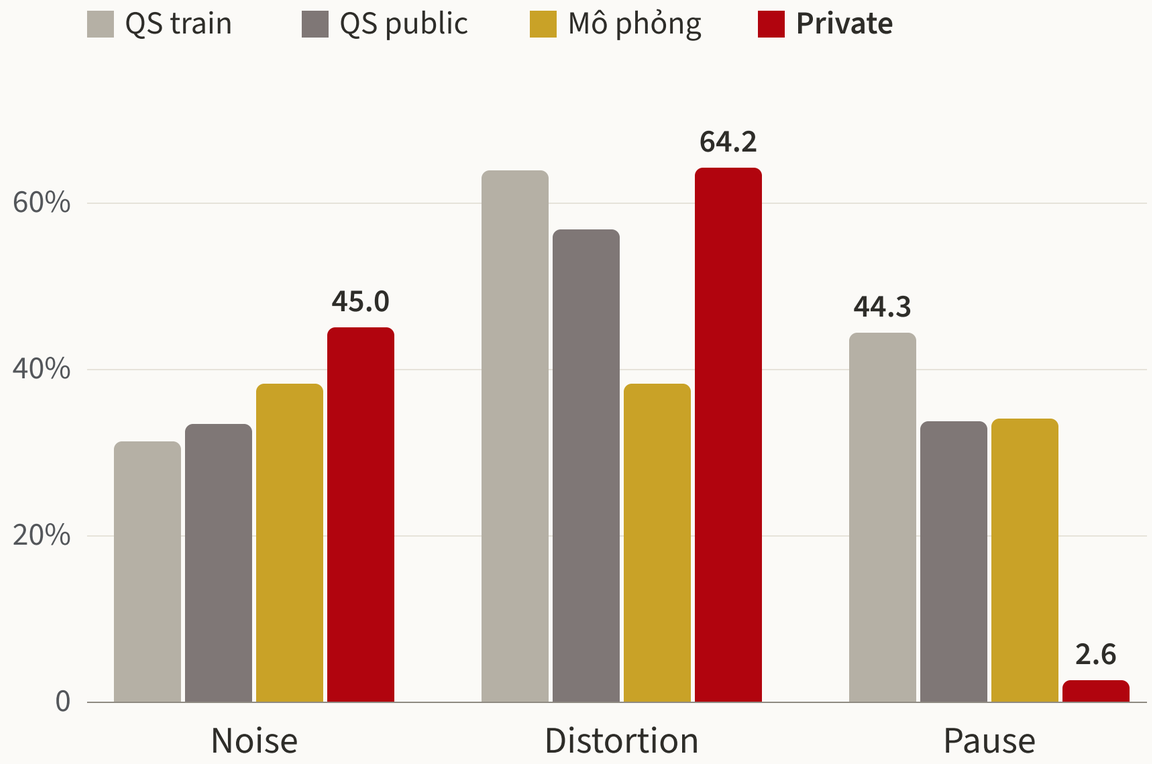
Cùng dữ liệu và cấu hình, chỉ khác encoder.

Mỗi hàng cộng dồn lên hàng trên. Hàng chỉ QS và hàng DSP lấy trung bình hai seed. Điểm private tính bằng nhãn đã công bố.

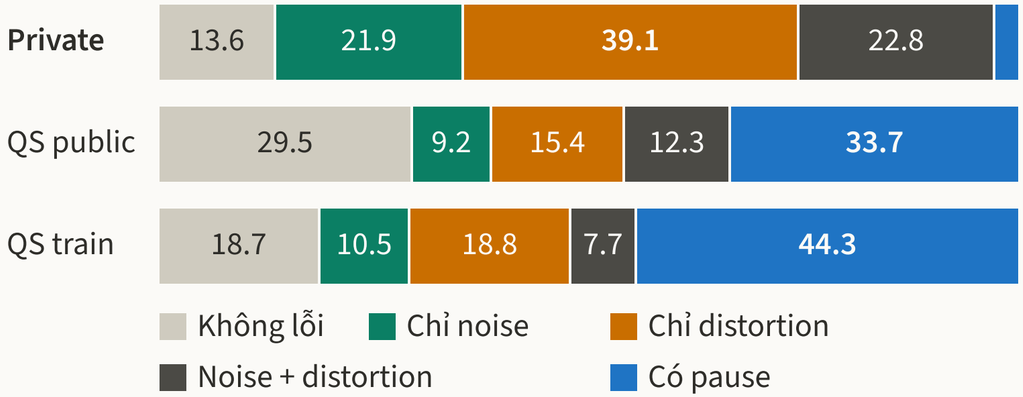
PRIVATE TEST

Shift thứ nhất từ nhãn private: pause gần như biến mất

Tỉ lệ file có từng loại lỗi



Tổ hợp lỗi, % số file



8 / 302
file có pause

0
file có đủ cả ba lỗi (QS public: 10.6%)

2.92
quality trung bình (QS public: 3.40)

Được tune trên dev có 34–47% pause, #1 dự đoán pause ở 98 file. Chỉ 8 file thật sự có.

Nhãn do ban tổ chức công bố sau đánh giá. Không phân tích nào ở đây được dùng cho các bài nộp.

PRIVATE TEST

Shift thứ hai: lỗi ở private ngắn hơn và nhẹ hơn QualiSpeech

1 Lỗi méo tiếng ngắn hơn hẳn

Trong một file bị méo (distortion), phần bị méo dài cỡ nào

QualiSpeech · đoạn méo thường dài **4 giây**



Private · đoạn méo thường chỉ **0.56 giây**



Cả thanh là cả file; phần tô màu là phần bị méo (trung vị).

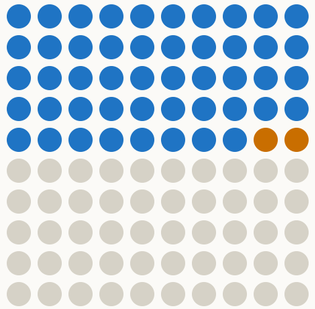
44% đoạn méo ngắn ở private thực ra là một quãng **gần như im bặt**, như bị mất tiếng

QualiSpeech: méo kéo dài cả file.
Private: méo là những "cú vấp" ngắn.

2 Khoảng lặng được gán nhãn khác

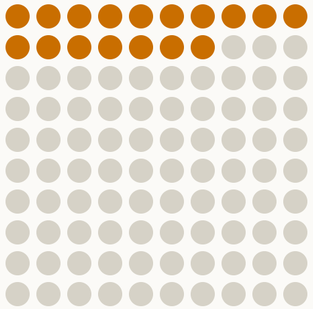
Cứ 100 khoảng lặng giữa câu, bao nhiêu cái bị gán là lỗi

QualiSpeech



48 bị gọi là pause

Private



0 bị gọi là pause
17 bị gọi là méo

pause · méo · không gán lỗi

Cùng một khoảng lặng: QualiSpeech hay gọi là "pause", private gần như không bao giờ.

3 Tiếng ồn nhẹ hơn, khó nhận ra

File có lỗi ồn (noise) ồn hơn file bình thường bao nhiêu

QualiSpeech



Private



Thanh càng dài, file có lỗi và file bình thường càng dễ phân biệt.

½ khoảng chênh chỉ còn khoảng một nửa: tiếng ồn dễ lẫn với tiếng nền bình thường

Mức "ồn là lỗi" học từ QualiSpeech sẽ bỏ sót tiếng ồn nhẹ ở private.

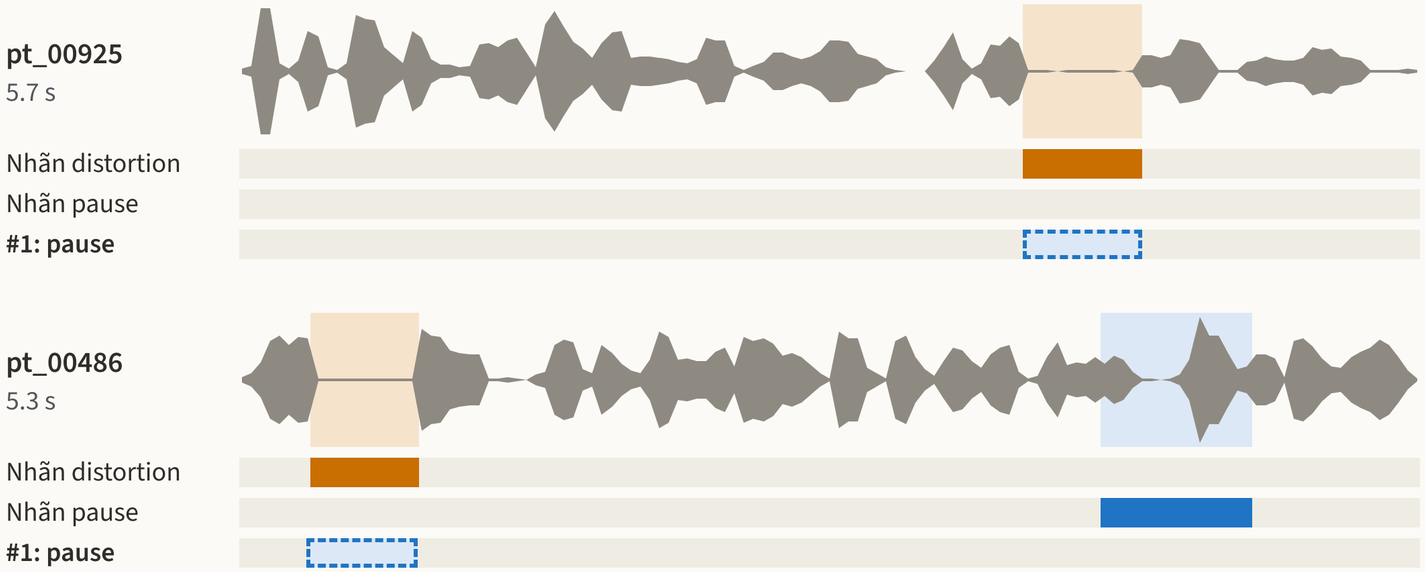
Lỗi ở private ngắn hơn, nhẹ hơn và được gán nhãn khác. Model chỉ học QualiSpeech sẽ khó nhận ra.

QualiSpeech: public test (tập train tương tự). Khoảng lặng: dài ≥ 0.26 s, nhỏ hơn tiếng nói ≥ 25 dB.

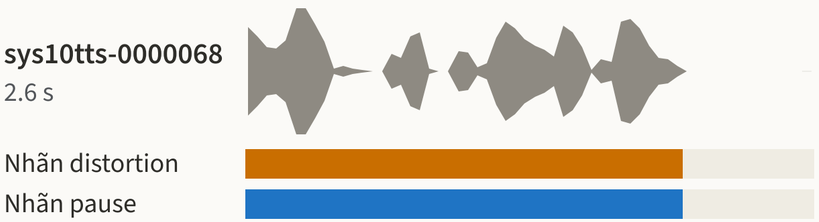
PRIVATE TEST

Model hay gọi nhầm distortion thành pause

Private: #1 gọi đoạn distortion là pause



Public: pause thường gán chồng lên distortion



“mechanical-sounding distortion accompanied by intermittent vocal sounds from 0s to 2s. Pause: 0-2s”

86%
file public có pause thì cũng có distortion

36%
khoảng pause trùng khít một khoảng distortion

67%
khoảng zero số được gán nhãn pause

77.9%
số frame im lặng mà #1 gọi là pause nằm trong nhãn distortion (ngẫu nhiên: 20.3%)

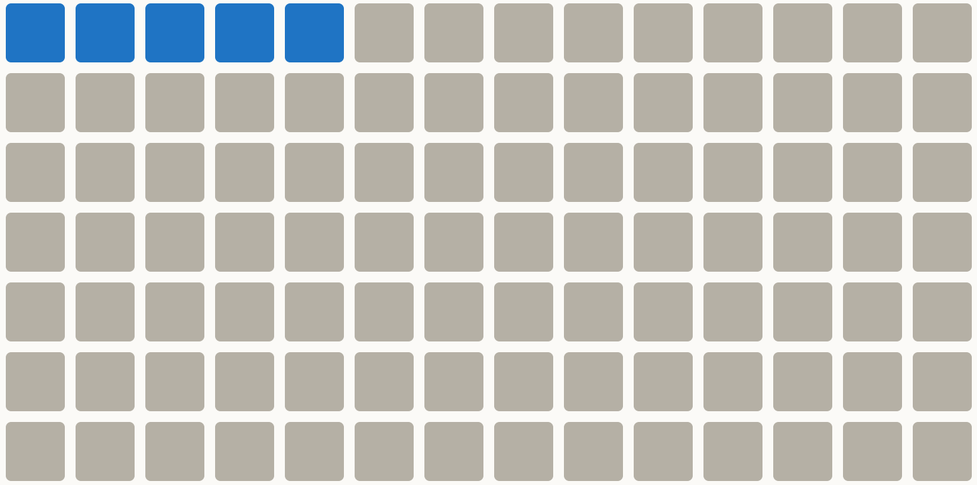
Model học theo QualiSpeech: ở đó pause hay được gán chồng lên distortion, nên gặp khoảng lặng bất thường là model gọi luôn là pause.

Public: 1,647 file có nhãn của QualiSpeech public test. Private: nhãn do ban tổ chức công bố; #1 là bài nộp tốt nhất.

PHÂN TÍCH LỖI

Lỗi pause chủ yếu là báo nhầm

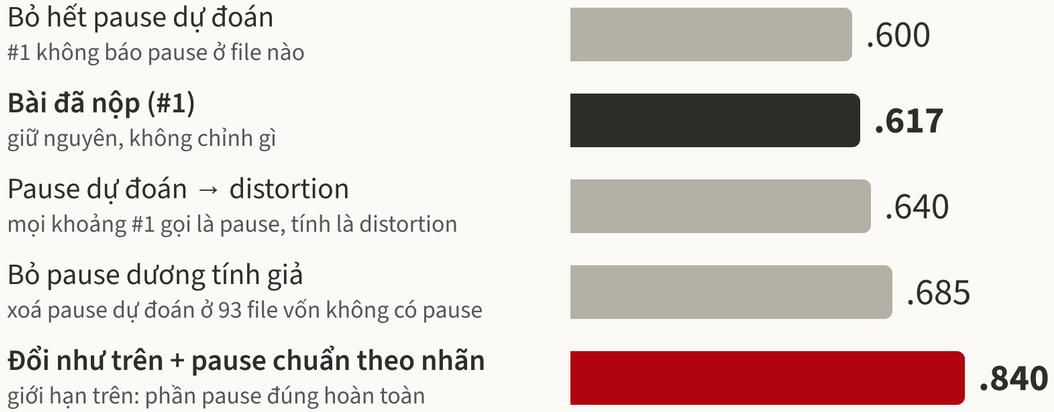
98 file bị đoán là pause, chỉ 8 file thật sự có



5 đoán đúng **93** báo nhầm **3** file pause thật bị bỏ sót

Pause rất nặng ký: mỗi lỗi nặng 20%, nên 8 file gánh một phần năm Final. Pause IoU của một file làm Final xê dịch **.0125**, bằng cả **24 file distortion**.

Thử giả định: Final của #1 tính lại sau cuộc thi



Chỉ riêng bước đổi pause → distortion:

.804 → .946
distortion F1

.499 → .760
distortion IoU

Phần lớn lỗi pause đến từ prior và quy ước gán nhãn, không phải do bỏ sót khoảng lặng.

Các phép thử giả định dùng nhãn đã công bố, không phải bài nộp. Trọng số IoU mỗi file trong Final: 0.1 / số file dương tính của lỗi đó.

Tổng kết

Strengths

Mô phỏng cùng ngôn ngữ: distortion IoU tăng từ .24 lên .45 mà không cần nhãn tay.

Encoder và DSP: w2v-BERT 2.0 hơn WavLM (.592 vs .494); DSP thêm .015 khi có dữ liệu mô phỏng.

Thành viên đa dạng: tương quan sai số .34, so với .68 giữa các model giống nhau.

Insights

Shift về nhãn: pause có ở 2.6% file private, so với 34–44% trong QualiSpeech.

Shift về lỗi: distortion chỉ dài 0.56 s, gần nửa số khoảng cục bộ là im lặng.

Một nhãn đắt giá: pause chiếm 48% phần lỗi còn lại.

Xin cảm ơn thầy cô trong BTC

Team SpeechWorld · dotrinhan99@gmail.com

Code: github.com/v-nhandt21/SQA-VLSP-2026-SpeechWorld-

SLIDE DỰ PHÒNG

Ablation từng bước, bảng đầy đủ

Hệ thống (mặc định w2v-BERT 2.0)	Public	Private	PCC	F1	IoU	Dist. IoU
Chỉ QS, không DSP (2 seed)	.676	.542 ± .015	.688	.545	.343	.240
Chỉ QS, có DSP (2 seed)	.679	.534 ± .013	.673	.539	.343	.242
+ dữ liệu VI mô phỏng, không DSP (seed 42)	.688	.564	.677	.570	.409	.448
+ đặc trưng DSP (2 seed)	.678	.586 ± .009	.732	.569	.408	.453
+ ensemble 4 hybrid	.684	.605	.757	.581	.427	.494
+ WavLM soup (#1, chính thức)	.696	.617	.789	.581	.425	.499
Ensemble chỉ QS, 12 model	.688	.571	.764	.519	.367	.242
Hybrid WavLM-large, cùng dữ liệu và cấu hình	.672	.494	.587	.527	.338	.347

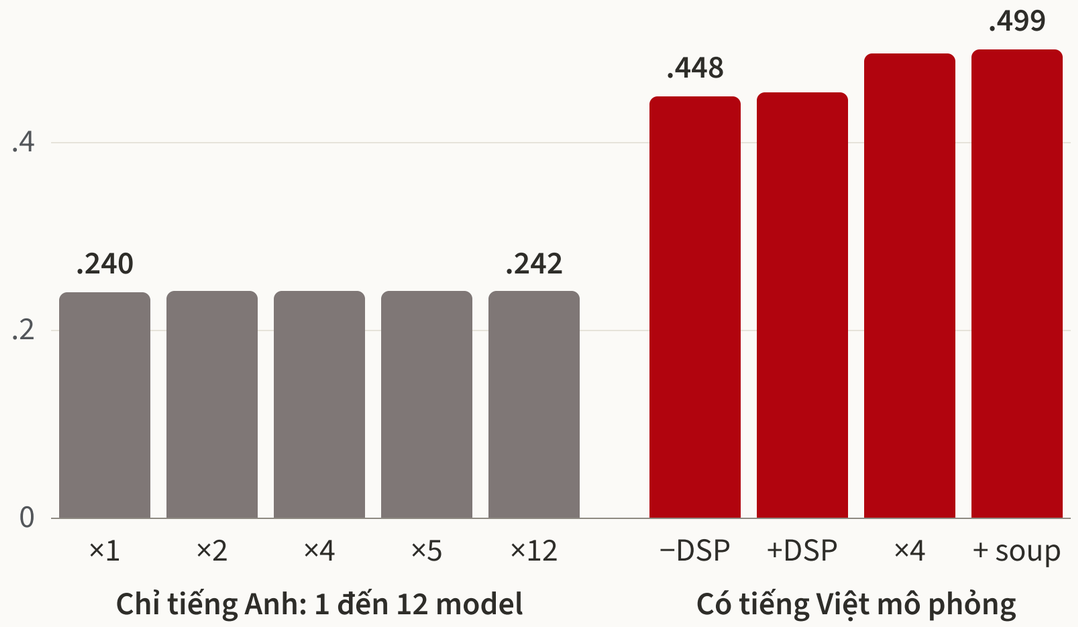
Cặp DSP cùng seed (seed 42, cùng dữ liệu và hyper-parameter): .579 khi có DSP, .564 khi không. Tăng +.015, paired bootstrap 95% CI [+.001, +.030] trên 302 file. Toàn bộ phần tăng nằm ở PCC (.723 vs .677); khả năng định vị không đổi (IoU .404 vs .409).

± là độ lệch chuẩn mẫu trên hai seed. Mọi decoder đều không dùng DSP late fusion ($\alpha = 0$).

ABLATION

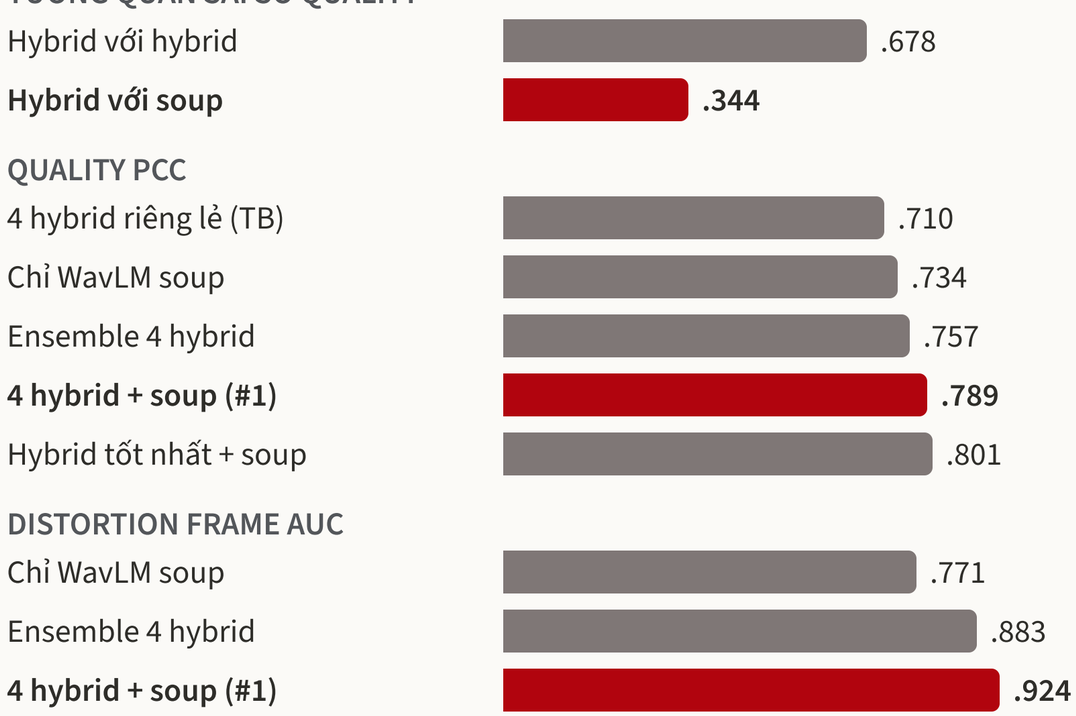
Thêm model chỉ có ích khi chúng mang lại điều mới

Distortion IoU trên private test



Mười hai model chỉ học tiếng Anh đẩy Final lên **.571**, nhưng distortion IoU vẫn đứng ở **.24**. Chỉ một model có dữ liệu mô phỏng đã đạt **.45**.

WavLM soup giúp ensemble đa dạng hơn

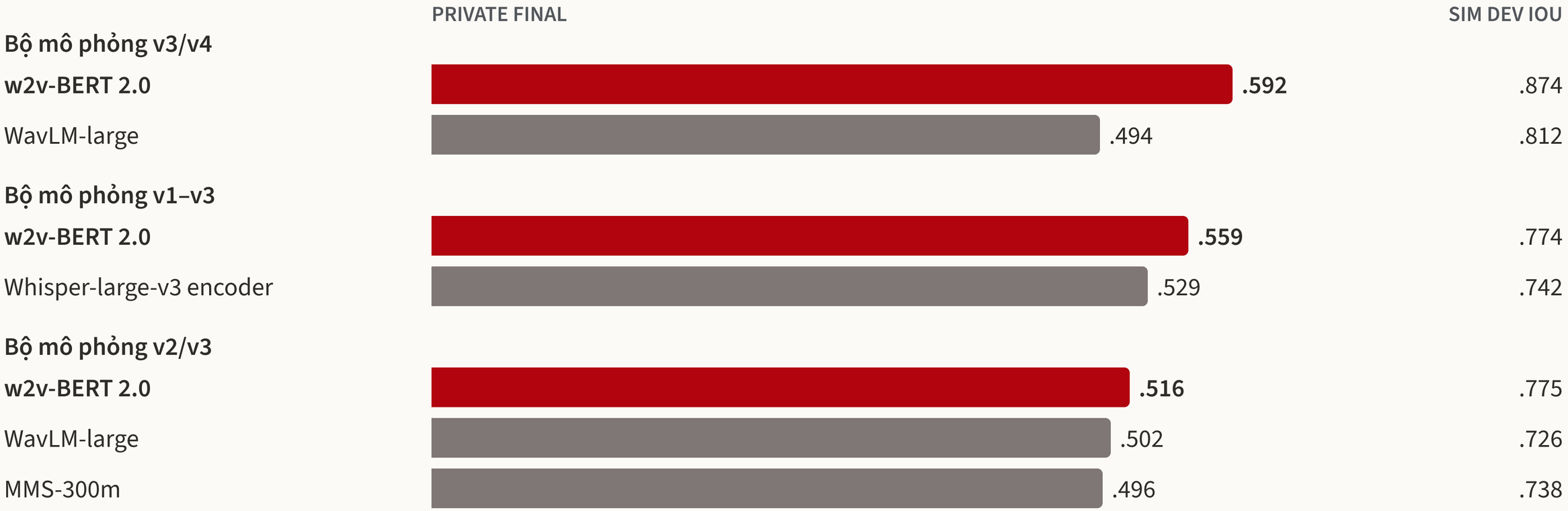


Chỉ hai model khác biệt đã thắng cả năm model về PCC, **.801 so với .789**. Thành viên yếu nhất về distortion lại kéo ensemble từ **.883 lên .924**.

Tính sau cuộc thi bằng nhãn đã công bố. Sai số: phần dư khi hồi quy output của từng model theo nhãn private.

SLIDE DỰ PHÒNG

Với dữ liệu tương đương, w2v-BERT 2.0 là encoder mạnh nhất



Chuyển từ bộ mô phỏng v2/v3 sang v3/v4, w2v-BERT 2.0 tăng từ .516 lên .592.

Các lần chạy đơn theo cùng một protocol, decoder không dùng late fusion. Cột bắt đầu từ 0.

SLIDE DỰ PHÒNG

10 đặc trưng DSP đo gì, và model được train thế nào

Đặc trưng	Nó đo điều gì	Giúp bắt lỗi
Zero fraction	Bao nhiêu phần của frame là im lặng tuyệt đối (tín hiệu = 0) tỉ lệ sample có $\|x\| \le 2^{-14}$	Distortion
Zero run	Frame nằm trong đoạn mất tín hiệu hoàn toàn, dài từ 40 ms trở lên chuỗi zero số ≥ 40 ms	Distortion
Clipping	Tiếng bị "cắt đỉnh": thu quá to nên sóng âm bị phẳng ở trên ≥ 3 sample liền nhau ở mức ≥ 98% đỉnh của file	Distortion
Energy	Frame này to hay nhỏ, so với đoạn to nhất của file log-energy của frame − p95	Tất cả
Noise floor	Ồn nền: chỗ yên tĩnh nhất xung quanh (±0.3 s) vẫn còn ồn đến đâu p10 của energy trong ±0.3 s − p95	Noise
Spectral tilt	Âm thanh nghiêng về tiếng cao (rè, xì) hay tiếng trầm (đục) energy của sai phân bậc một − energy của frame	Noise Distortion
Energy change	Âm lượng thay đổi đột ngột so với frame ngay trước chênh lệch energy giữa hai frame liên tiếp	Pause Distortion
Silent	Frame có đủ nhỏ để tính là khoảng lặng hay không energy < p5 + 0.35 (p95 − p5)	Pause
Silence run	Khoảng lặng chứa frame này kéo dài bao lâu, tính tới 1 s độ dài đoạn silent liên tục, tối đa 1 s	Pause
Band jump	Đột nhiên mất tiếng cao, như khi codec cắt bớt băng tần tilt − trung vị tilt của các frame lớn tiếng	Distortion

Cấu hình huấn luyện

Encoder: model nền đã học sẵn tiếng nói: w2v-BERT 2.0, WavLM-large, Whisper-large-v3 (cửa sổ 16 s), MMS-300m

Phần đầu ra: gộp các lớp của encoder thành vector 256 chiều, ghép thêm 32 chiều từ DSP, rồi qua BiLSTM 2 × 128

Tốc độ học: encoder đã học sẵn nên chỉnh rất nhẹ (1e-5 đến 2e-5); các head mới học nhanh hơn (1e-3). AdamW, weight decay 0.01

Lịch train: khởi động 6% số bước đầu, sau đó giảm dần (cosine); batch 8, bf16, 5–7 epoch

Dữ liệu mỗi epoch: 10.6k file QualiSpeech + 40–100% dữ liệu tiếng Việt mô phỏng

Augmentation: đổi âm lượng ±6 dB, cắt ngẫu nhiên đoạn 12 s, LayerDrop 0.05, SpecAugment

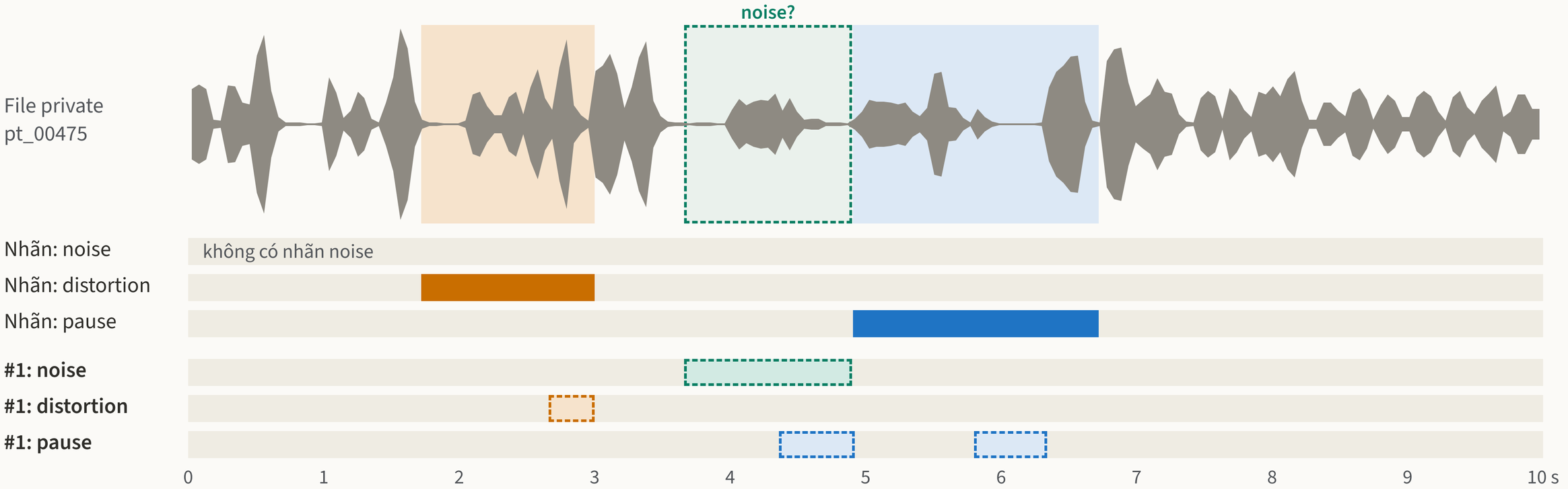
Decoder: biến xác suất từng frame thành fault interval: dò ngưỡng, làm mượt 1–25 frame, nối khoảng hở ≤ 0.6 s, bỏ đoạn ngắn hơn 0.3 s

Chọn model: theo điểm Final trên QS dev, cộng 0.3 F1 + 0.3 IoU trên dev mô phỏng

Mọi đặc trưng tính trên từng frame 20 ms và so với chính file đó. p5, p95: mức năng lượng thấp và cao điển hình của file.

PRIVATE TEST · THÊM MỘT VÍ DỤ

Model bắt được pause, và cả đoạn noise mà nhãn bỏ sót



Pause: phát hiện được

#1 đoán đúng file này có pause. Một đoạn nằm gọn trong vùng nhãn (5.80–6.34 s), một đoạn chạm mép đầu. Đoạn dự đoán còn ngắn nên IoU chỉ khoảng 0.23.

Noise: model nghe thấy, nhãn thì không

#1 báo noise ở 3.66–4.90 s. Nghe lại thì đoạn này có tiếng ồn nền thật, nhưng nhãn không ghi noise nào, nên khi chấm điểm bị tính là false positive.

Quality gần như khớp

MOS của nhãn là 3.175, #1 dự đoán 3.18. Distortion cũng bắt được phần cuối của vùng nhãn (2.66–3.00 s).

Một phần “false positive” có thể đến từ nhãn bỏ sót, không hẳn do model sai.

Nhãn do ban tổ chức công bố. Khối màu đặc: nhãn; khung nét đứt: bài nộp #1. IoU ước lượng từ các mốc thời gian.

SLIDE DỰ PHÒNG

17 kiểu lỗi giả lập nghe như thế nào

Kiểu lỗi	Nghe như thế nào	Tỉ lệ
Distortion		
Cụm dropout	Tiếng bị “nấc”, mất từng nhịp rất ngắn, như mạng chập chờn 3–9 đoạn tín hiệu bằng 0, mỗi đoạn 40–100 ms, trong vùng 0.6–2 s	28%
Clip + attenuate	Thu quá to nên vỡ tiếng, rồi cả file bị hạ nhỏ: vừa rè vừa bé cắt đỉnh ở 10–40% biên độ, rồi hạ đỉnh cả file xuống 0.04–0.3	12%
Musical noise	Tiếng “lóc chóc”, giọng mỏng đi: khử ồn làm quá tay spectral subtraction trừ gấp 2–5 lần mức ồn ước lượng	12%
Low-pass theo đoạn	Một đoạn bỗng đục, nghe như nói sau bức tường lọc thông thấp 0.8–2 kHz trên một đoạn 1–4 s	10%
Lượng tử hoá codec	Giọng nghe “kim loại”, lỗ chỗ, như gọi điện khi mạng yếu lượng tử phổ theo bước 3–8 dB, cắt trên 3–6 kHz, đục lỗ 10–35% ô phổ	10%
AGC pumping	Âm lượng nhảy lên xuống liên tục, như bộ tự chỉnh âm lượng bị loạn gain ngẫu nhiên −14 đến +6 dB, đổi mỗi 50–150 ms	8%
Overdrive	Tiếng bão hoà, rè đặc như loa bị mở quá to méo bằng hàm tanh, độ khuếch đại 8–30	6%
Bit-crushing	Tiếng sạn, lạo xạo như máy chơi game đời cũ giảm độ phân giải biên độ còn 3–5 bit	4%
Pitch shift	Giọng bị biến cao hẳn hoặc trầm hẳn, có khi rung rung dịch 3–7 bán cung; một nửa số lần thêm rung 6–20 Hz	4%
Griffin-Lim	Giọng “robot”, rè vang như TTS đời cũ bỏ pha, dựng lại bằng Griffin-Lim chỉ 2–7 vòng lặp	3%
Aliasing	Tiếng rít, méo ở âm cao do lấy mẫu quá thưa giữ 1 trong 2–5 sample rồi lặp lại	3%

Kiểu lỗi	Nghe như thế nào	Tỉ lệ
Noise		
Bản ghi thật	Ồn thật: nhạc, xe cộ, trong ô tô, người nói nền, TV, mưa MUSAN và RIRS point-source, SNR 10–28 dB	75%
Static tổng hợp	Tiếng lẹt xẹt, xì sóng radio, hoặc ù điện crackle 45%, radio hiss 35%, ù 50/100 Hz 20%	25%
Unnatural pause		
Một khoảng lặng	Đang nói thì ngừng hẳn một quãng bất thường chèn 0.25–0.85 s, lấp bằng tiếng nền của chính file	55%
Nhiều khoảng lặng	Như trên, nhưng lặp lại 2–3 lần trong file 2–3 khoảng 0.25–0.85 s ở vị trí ngẫu nhiên	30%
Stutter	Lặp một mẩu tiếng như bị vấp: “tô… tô… tôi” lặp mẩu 80–250 ms 2–4 lần, xen lặng 50–200 ms	10%
Ngắt quãng	Nói nhát gừng, từng mẩu ngắn xen khoảng lặng mẩu 0.15–0.5 s xen lặng 0.2–0.5 s, trong vùng 1.5–5 s	5%

Không tính là lỗi, không gán nhãn: vang phòng nhẹ (30% file), ồn nền rất nhỏ (thấp hơn tiếng nói 30–50 dB), lọc băng tần kiểu điện thoại.

Bộ mô phỏng v4 (src/simulate_vi.py). Tỉ lệ là tỉ lệ từng kiểu trong nhóm, giống vòng ngoài của biểu đồ ở slide 5.