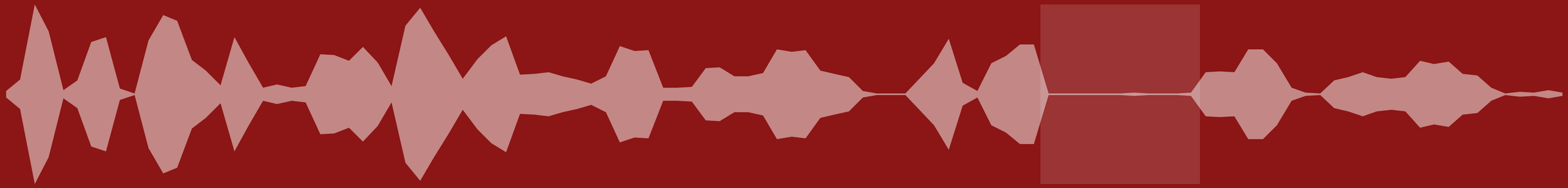
VLSP 2026 · SHARED TASK SPEECH QUALITY ASSESSMENT

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 November 2026

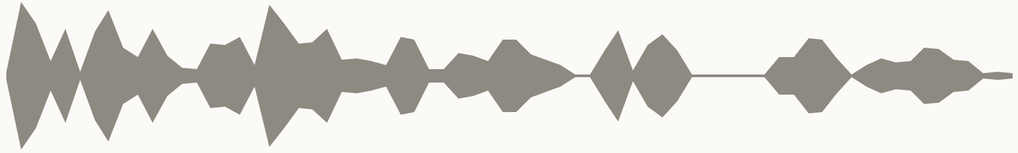


THE TASK

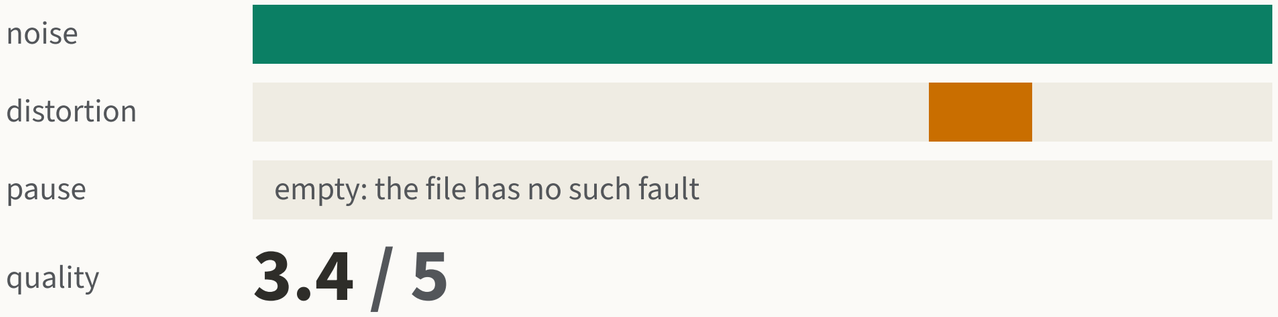
Not just rating quality, but saying what is wrong, and where

Input: a degraded recording, with no reference signal

waveform



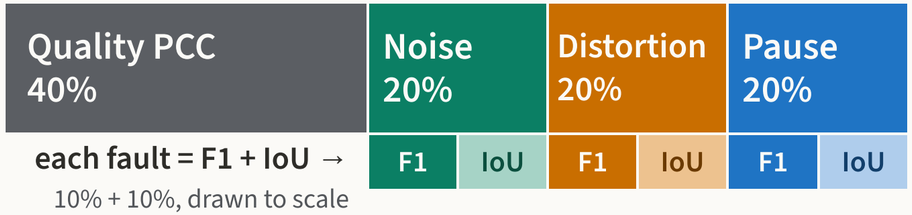
Output: a quality score and fault intervals



Real example: private file pt_00925 with its released labels, plus the scores predicted by our model. Fault intervals are scored on a 20 ms grid.

Scoring

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



60% of the score is diagnosis, not quality.

F1 is per file. **IoU** counts only files that actually have that fault.

Each fault weighs 20%, however rare it is. This detail matters later.

A submission missing either the quality score or the fault intervals is invalid.

Source: VLSP 2026 SQA task description

THE CHALLENGE

Two shifts at once: English train, Vietnamese test

QualiSpeech: 14,577 English clips, about half of them TTS

Private: 302 Vietnamese files

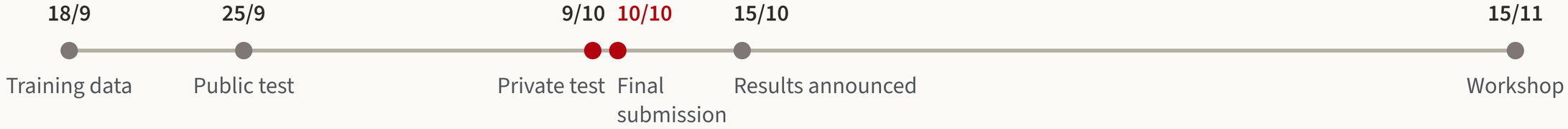


Widths to scale. The private set is only about 2% the size of QualiSpeech, and its labels were released only after the evaluation ended.

WHAT THIS MEANS	OUR RESPONSE
No Vietnamese speech has fault-interval labels	→ **Simulate faults on clean Vietnamese speech**
The encoder understands speech only	→ **Add language-independent signal-processing features**

TIMELINE 2026

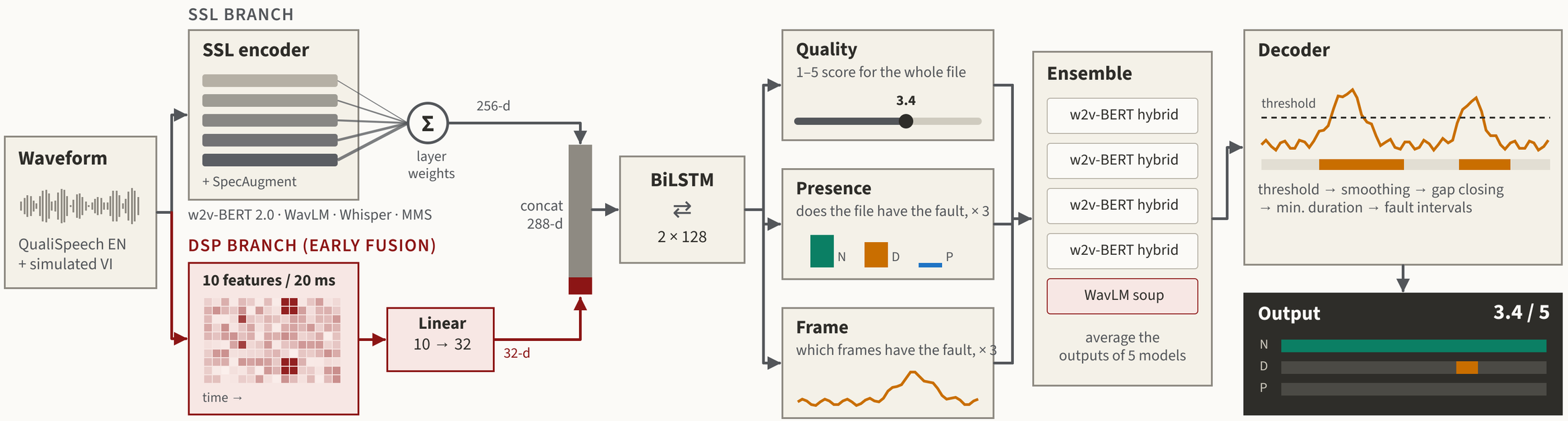
Three weeks from training data to private test, and six more days to results.



Source: VLSP 2026 SQA task page; QualiSpeech (Wang et al., ACL 2025)

OUR SYSTEM

A DSP branch next to the SSL encoder



DSP covers what SSL misses

An SSL encoder learns speech content, so a run of digital zeros, a clipped segment or a lost frequency band leaves almost no trace. DSP features catch them immediately.

Fine-tuned end-to-end on one A100-40GB: 2–7 hours per hybrid; under a minute per model for the 302 private files

OUR SYSTEM

A simulator creates Vietnamese faults with exact fault intervals

Clean Vietnamese speech
PhoAudioBook, Vivoice, LSVSC, ViSEC, Dolly

↓

Inject faults and record their exact intervals
noise, distortion, pause

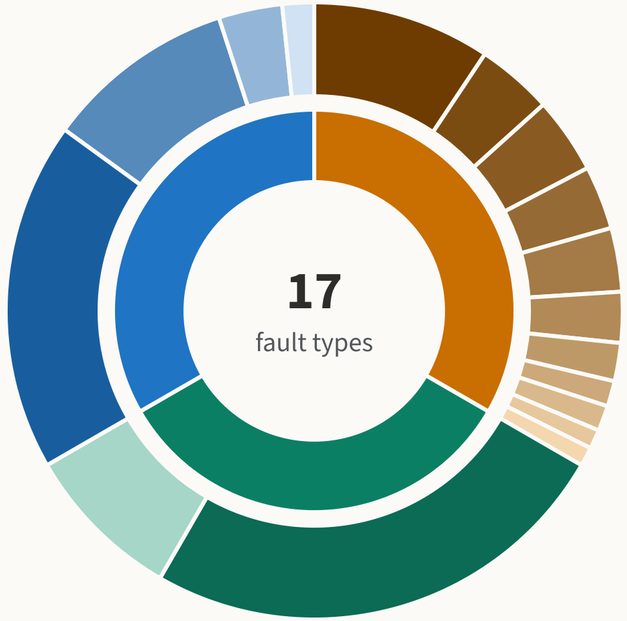
↓

12,000 files, 24.8 hours
1,000 files held out as dev, no manual labelling

The simulator went through four iterations. The step from v1 to v3/v4 alone raised private distortion IoU from **.411 to .448**, without a single real Vietnamese fault label.

Fault types in each group (final simulator)

Inner ring: 3 fault groups, injected with almost equal probability. Outer ring: share of each type within its group.



Distortion

Dropout clusters	28%
Clip + attenuate	12%
Musical noise	12%
Segment low-pass	10%
Codec quantisation	10%
AGC pumping	8%
Overdrive	6%
Bit-crushing	4%
Pitch shift	4%
Griffin-Lim	3%
Aliasing	3%

Noise

Real recordings	75%
Synthetic static	25%

Unnatural pause

Single gap	55%
Multiple gaps	30%
Stutter	10%
Choppy speech	5%

25% of files are fault-free; in the rest, each fault group is injected with probability 0.40–0.45. Noise at SNR 10–28 dB.

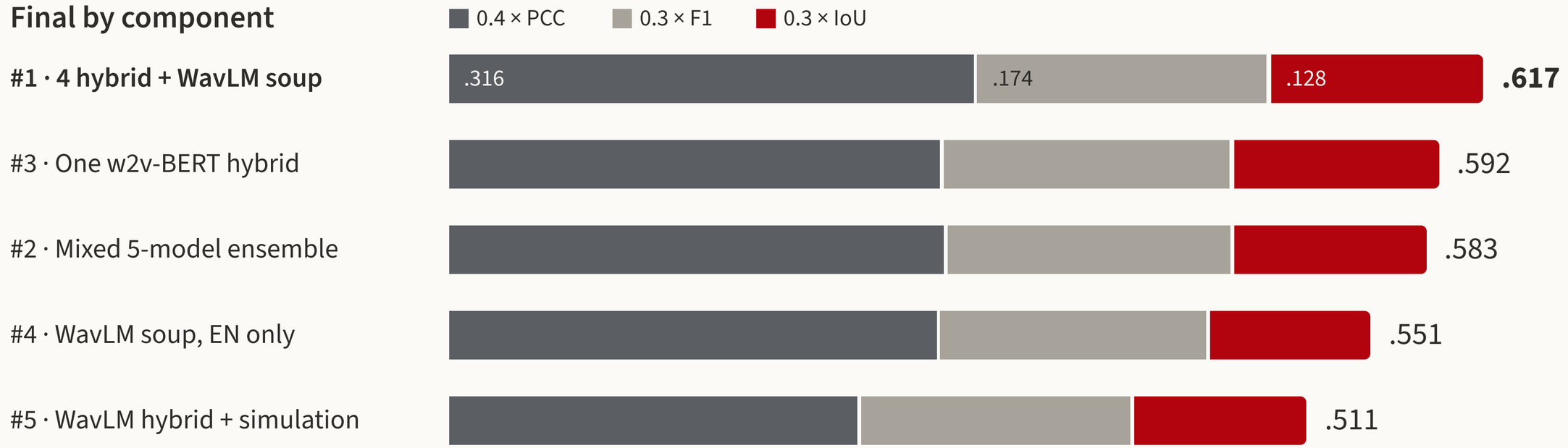
A choice we return to later: inserted silences are labelled as pauses, as in QualiSpeech.

Dropout clusters: 3–9 digital-zero gaps of 40–100 ms. Pauses: gaps of 0.25–0.85 s filled with room tone.

RESULTS

Official results: the best of our five submissions scored 0.617

Final by component



Simulation helps localisation

From an English-only soup (#4) to one hybrid with simulated data (#3), distortion IoU rises from **.242 to .462**.

The ensemble improves quality

#1 keeps the hybrids' localisation while raising PCC from **.740 to .789**, our highest.

Pause is the exception

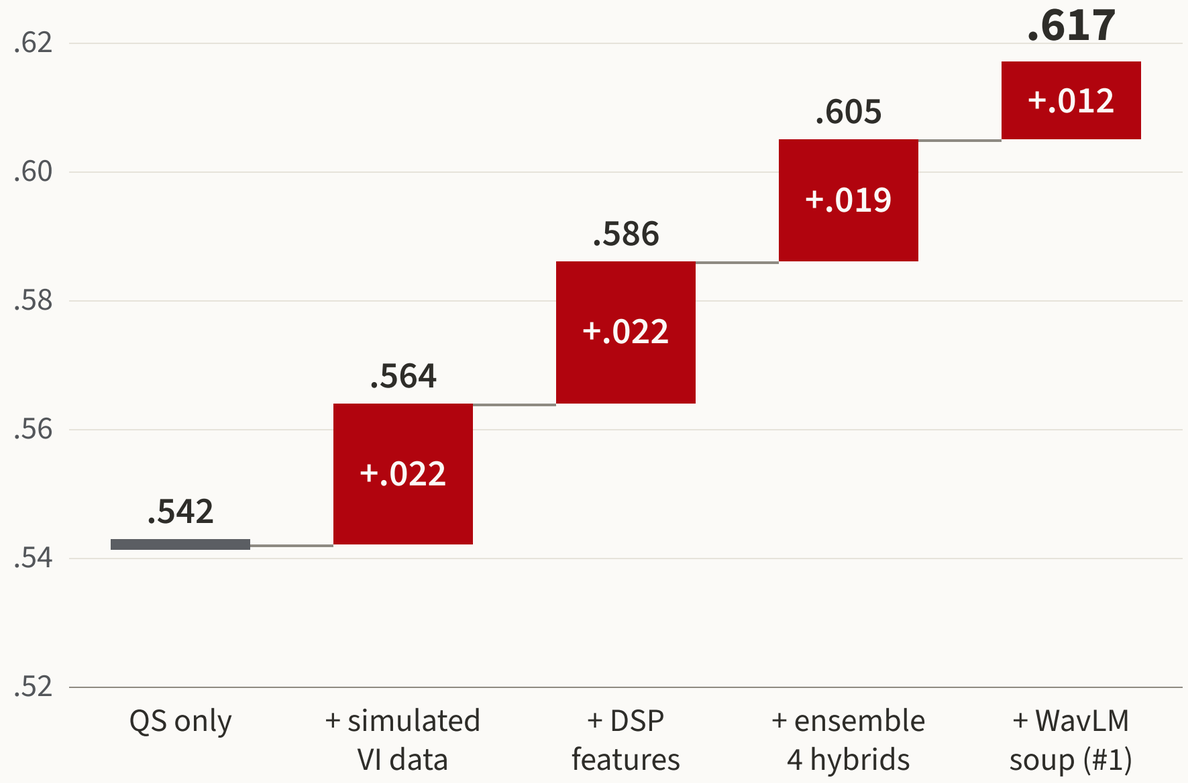
Pause F1 stays **below .10** in all five submissions, whatever the backbone or training data.

Official private-test scores on 302 Vietnamese files. Soup: weight average of WavLM checkpoints from epochs 3 to 6.

ABLATION

Each component adds about .02

Private Final step by step (w2v-BERT 2.0)



DSP needs simulated data

.542
QS only, no DSP

.534
QS only, with DSP

Without simulated data, DSP does not help.

The encoder matters as much as the data

.592
w2v-BERT 2.0

.494
WavLM-large

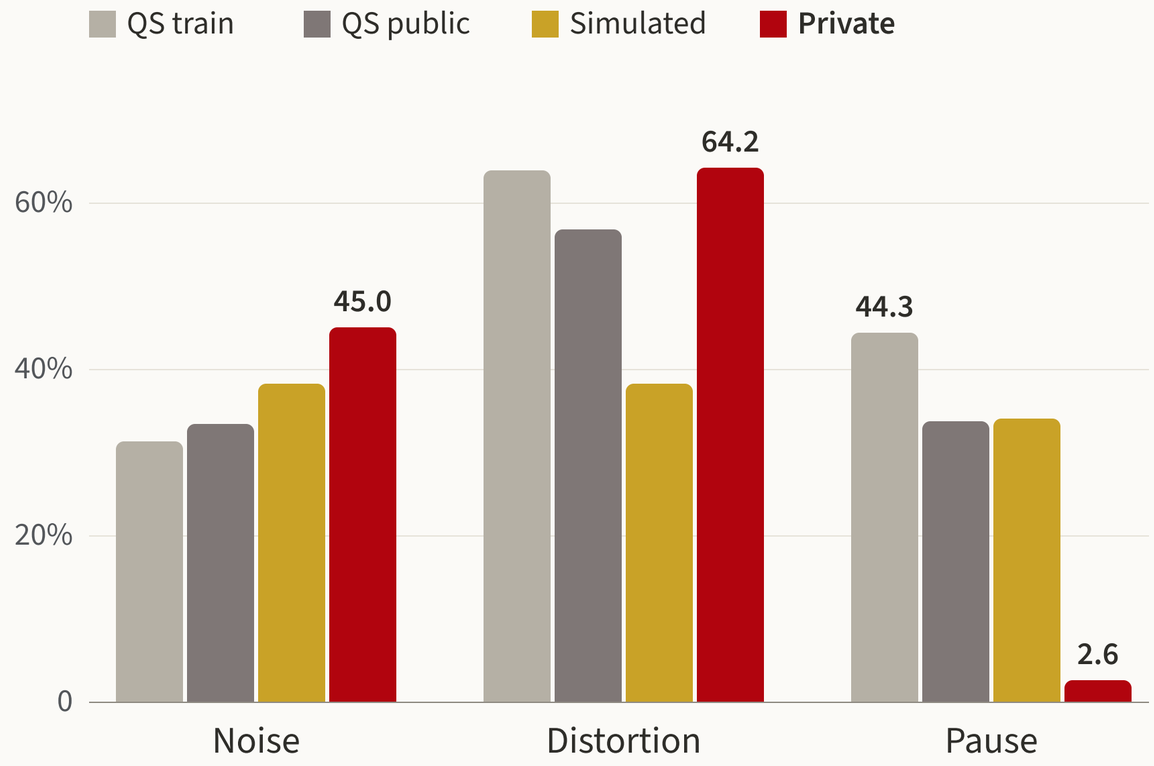
Same data and configuration, only the encoder differs.

Each row builds on the row above. The QS-only and DSP rows average two seeds. Private scores computed with the released labels.

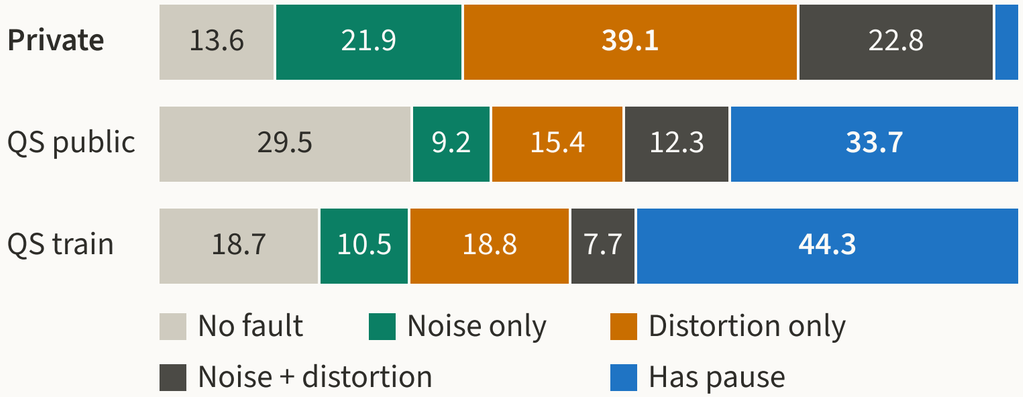
PRIVATE TEST

First shift from the private labels: pauses almost disappear

Share of files with each fault



Fault combinations, % of files



8 / 302
files with pauses

0
files with all three faults (QS public: 10.6%)

2.92
mean quality (QS public: 3.40)

Tuned on a dev set with 34–47% pauses, #1 predicted pauses in 98 files. Only 8 files actually have one.

Labels released by the organizers after the evaluation. None of this analysis was used for the submissions.

PRIVATE TEST

Second shift: private faults are shorter and milder

1 Distortion is much shorter

In a distorted file, how long is the distorted part

QualiSpeech · distortion usually lasts **4 seconds**



Private · distortion usually lasts only **0.56 seconds**



The whole bar is the whole file; the coloured part is the distorted part (median).

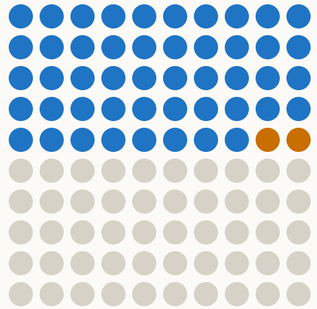
44% of short private distortions are actually a stretch of **near silence**, like a dropout

QualiSpeech: distortion spans the whole file.
Private: distortion is short “stumbles”.

2 Silences are labelled differently

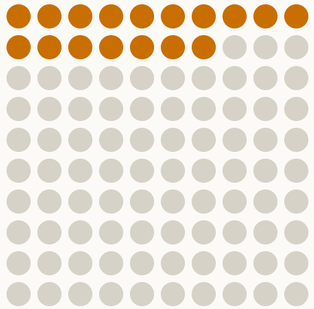
Out of every 100 mid-utterance silences, how many are labelled as a fault

QualiSpeech



48 called pause

Private



0 called pause
17 called distortion

pause · dist. · no fault

Same silence: QualiSpeech often calls it a “pause”, the private set almost never does.

3 Noise is milder and harder to notice

How much louder the noise floor is in noisy files than in clean ones

QualiSpeech



Private



The longer the bar, the easier it is to tell noisy files from clean ones.

½ the gap is only about half: the noise blends into normal background sound

A “noise is a fault” threshold learned from QualiSpeech misses the mild noise in the private set.

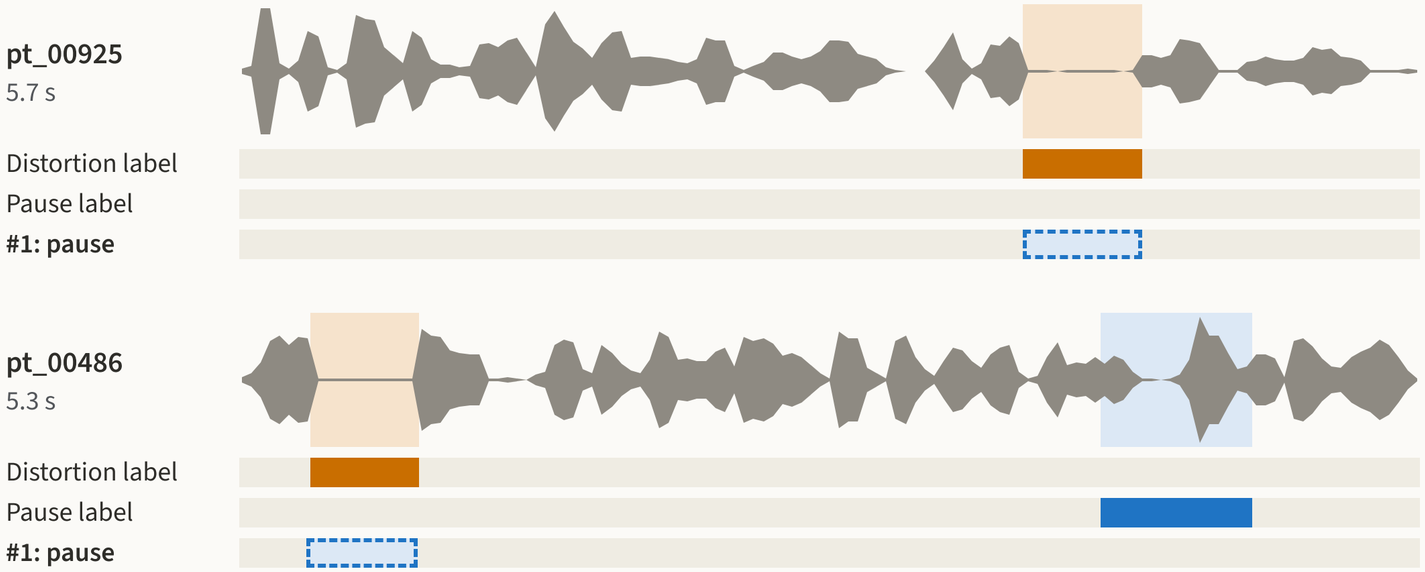
Private faults are shorter, milder and labelled differently. A model trained only on QualiSpeech struggles to recognise them.

QualiSpeech: public test (the training set is similar). Silences: ≥ 0.26 s long, ≥ 25 dB below speech.

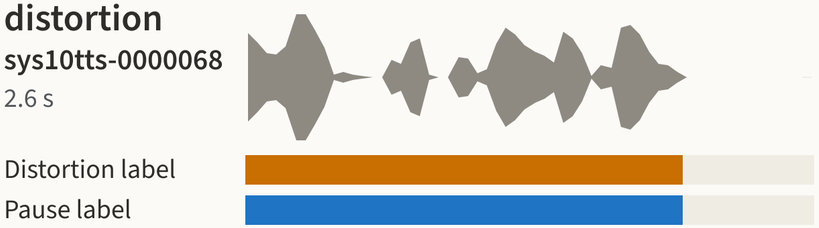
PRIVATE TEST

The model often calls distortion a pause

Private: #1 calls distortion segments pauses



Public: pauses are often labelled on top of distortion



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

86%
of public files with a pause also have distortion

36%
of pause intervals exactly match a distortion interval

67%
of digital-zero gaps are labelled as pauses

77.9%
of the silent frames #1 calls pause lie inside distortion labels (chance: 20.3%)

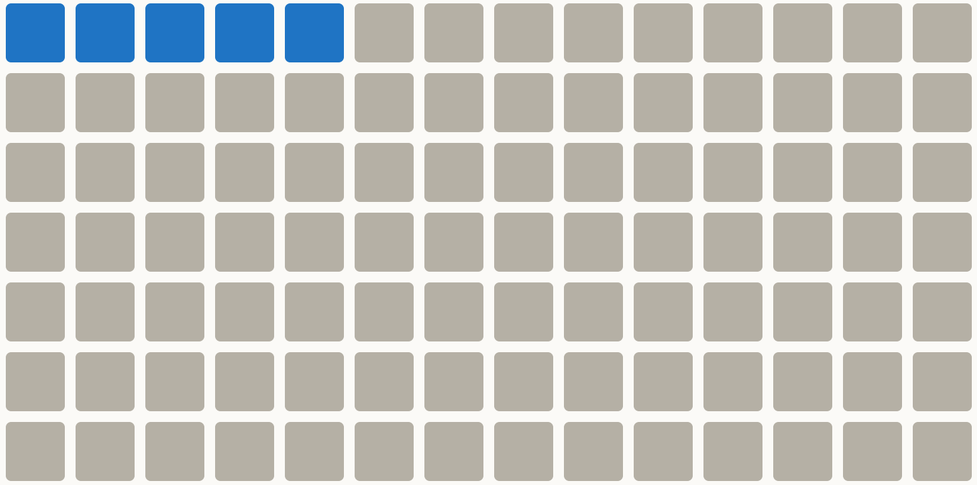
The model learned from QualiSpeech, where pauses are often labelled on top of distortion, so it calls any unusual silence a pause.

Public: 1,647 labelled files of the QualiSpeech public test. Private: labels released by the organizers; #1 is our best submission.

ERROR ANALYSIS

Pause errors are mostly false alarms

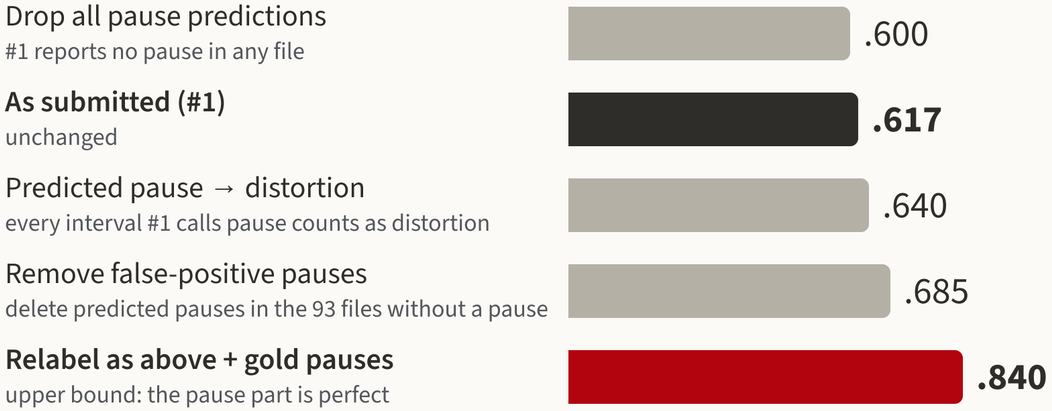
98 files predicted as pause, only 8 truly positive



5 correct **93** false alarm **3** true pause files missed

Why so costly: each fault weighs 20%, so 8 files carry a fifth of the Final. One file's pause IoU moves the Final by **.0125**, as much as **24 distortion files**.

What if? #1's Final recomputed after the competition



Relabelling pause → distortion alone:

.804 → .946
distortion F1

.499 → .760
distortion IoU

Most of the pause error comes from a prior and a labelling convention, not from missing silences.

These what-ifs use the released labels and are not submissions. Per-file IoU weight in the Final: 0.1 / number of positive files for that fault.

Takeaways

Strengths

Same-language simulation: distortion IoU from .24 to .45 with no manual labels.

Encoder and DSP: w2v-BERT 2.0 beats WavLM (.592 vs .494); DSP adds .015 when simulated data is present.

Diverse members: error correlation .34, versus .68 between similar models.

Insights

Label shift: pauses in 2.6% of private files, versus 34–44% in QualiSpeech.

Fault shift: distortions last only 0.56 s; nearly half of the local intervals are silence.

One costly label: pause accounts for 48% of the remaining error.

Thank you to the organizers

Team SpeechWorld · dotrinhan99@gmail.com

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

Step-by-step ablation, full table

System (w2v-BERT 2.0 unless noted)	Public	Private	PCC	F1	IoU	Dist. IoU
QS only, no DSP (2 seeds)	.676	.542 ± .015	.688	.545	.343	.240
QS only, with DSP (2 seeds)	.679	.534 ± .013	.673	.539	.343	.242
+ simulated VI data, no DSP (seed 42)	.688	.564	.677	.570	.409	.448
+ DSP features (2 seeds)	.678	.586 ± .009	.732	.569	.408	.453
+ 4-hybrid ensemble	.684	.605	.757	.581	.427	.494
+ WavLM soup (#1, official)	.696	.617	.789	.581	.425	.499
QS-only ensemble, 12 models	.688	.571	.764	.519	.367	.242
WavLM-large hybrid, same data and configuration	.672	.494	.587	.527	.338	.347

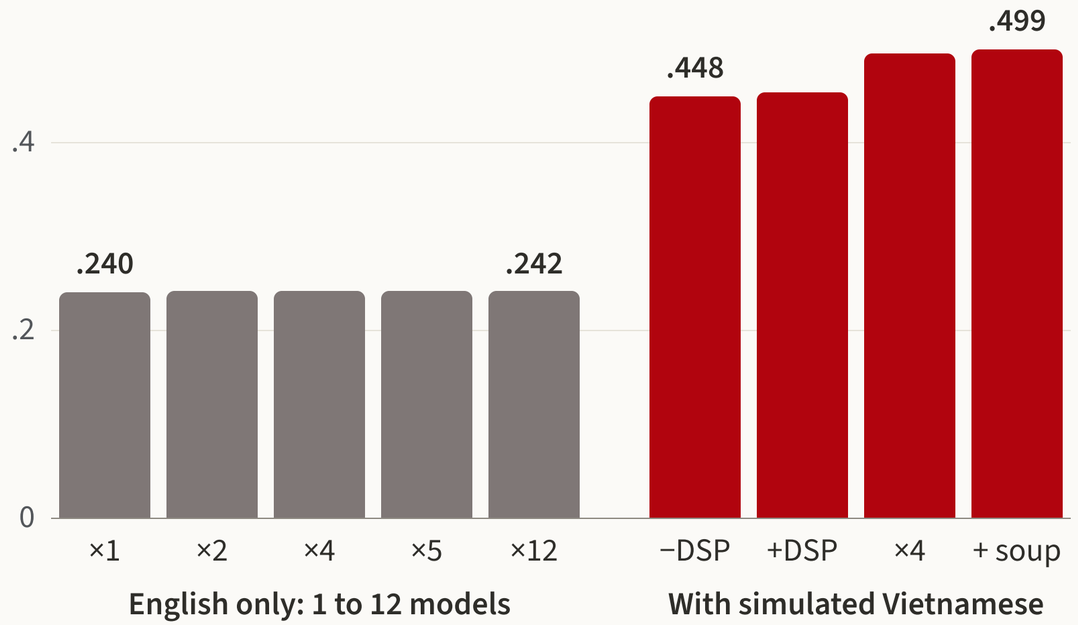
Same-seed DSP pair (seed 42, same data and hyper-parameters): .579 with DSP, .564 without. A gain of +.015, paired bootstrap 95% CI [+.001, +.030] over 302 files. The whole gain is in PCC (.723 vs .677); localisation is unchanged (IoU .404 vs .409).

± is the sample standard deviation over two seeds. No decoder uses DSP late fusion ($\alpha = 0$).

ABLATION

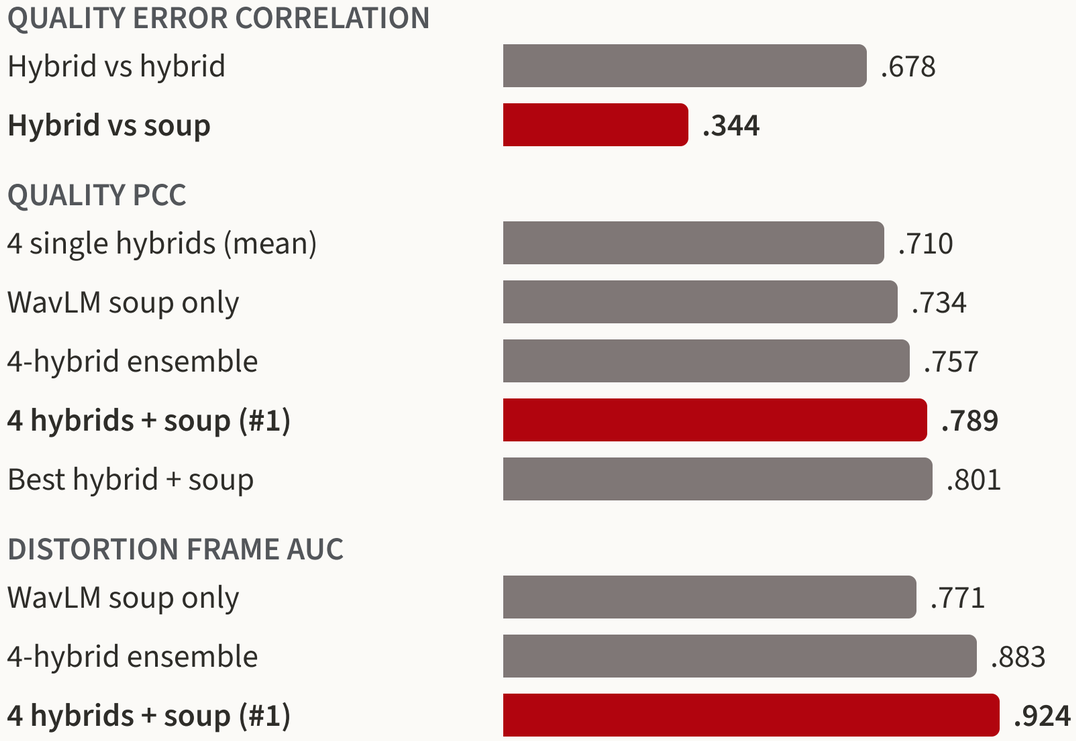
More models help only when they bring something new

Distortion IoU on the private test



Twelve English-only models push the Final to **.571**, but distortion IoU stays at **.24**. A single model with simulated data already reaches **.45**.

The WavLM soup makes the ensemble more diverse

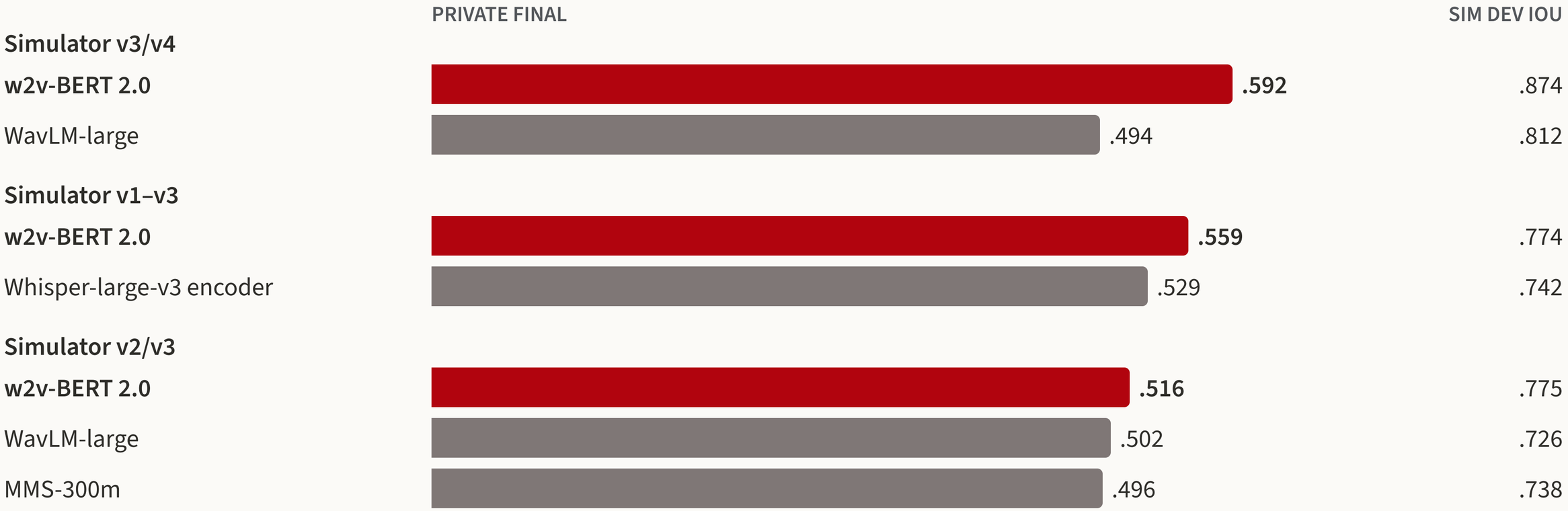


Just two different models beat all five on PCC, **.801 vs .789**. The weakest member on distortion lifts the ensemble from **.883 to .924**.

Computed after the competition with the released labels. Error: residual of regressing each model's output on the private labels.

BACKUP SLIDES

With comparable data, w2v-BERT 2.0 is the strongest encoder



Moving from simulator v2/v3 to v3/v4, w2v-BERT 2.0 rises from .516 to .592.

Single runs under the same protocol, decoder without late fusion. Bars start at 0.

BACKUP SLIDES

The 10 DSP features, and how the model is trained

Feature	What it measures	Helps catch
Zero fraction	How much of the frame is absolute silence (signal = 0) share of samples with $\|x\| \le 2^{-14}$	Distortion
Zero run	Frame lies in a complete signal dropout lasting 40 ms or more digital-zero run ≥ 40 ms	Distortion
Clipping	"Clipped" sound: recorded too loud, so the waveform is flattened at the top ≥ 3 consecutive samples at ≥ 98% of the file's peak	Distortion
Energy	How loud this frame is, relative to the loudest part of the file frame log-energy − p95	All
Noise floor	Background noise: how noisy the quietest nearby part (±0.3 s) still is p10 of energy within ±0.3 s − p95	Noise
Spectral tilt	Whether the sound leans toward highs (harsh, hissy) or lows (muffled) energy of the first difference − frame energy	Noise Distortion
Energy change	Sudden loudness change from the previous frame energy difference between consecutive frames	Pause Distortion
Silent	Whether the frame is quiet enough to count as silence energy < p5 + 0.35 (p95 − p5)	Pause
Silence run	How long the silence containing this frame lasts, capped at 1 s length of the continuous silent run, up to 1 s	Pause
Band jump	Sudden loss of high frequencies, as when a codec cuts the band tilt − median tilt of loud frames	Distortion

Training setup

Encoder: pretrained speech models: w2v-BERT 2.0, WavLM-large, Whisper-large-v3 (16 s window), MMS-300m

Heads: encoder layers pooled into a 256-d vector, concatenated with 32 DSP dimensions, then a 2 × 128 BiLSTM

Learning rate: the pretrained encoder is tuned very gently (1e-5 to 2e-5); the new heads learn faster (1e-3). AdamW, weight decay 0.01

Schedule: 6% warm-up, then cosine decay; batch 8, bf16, 5–7 epochs

Data per epoch: 10.6k QualiSpeech files + 40–100% of the simulated Vietnamese data

Augmentation: gain ±6 dB, random 12 s crops, LayerDrop 0.05, SpecAugment

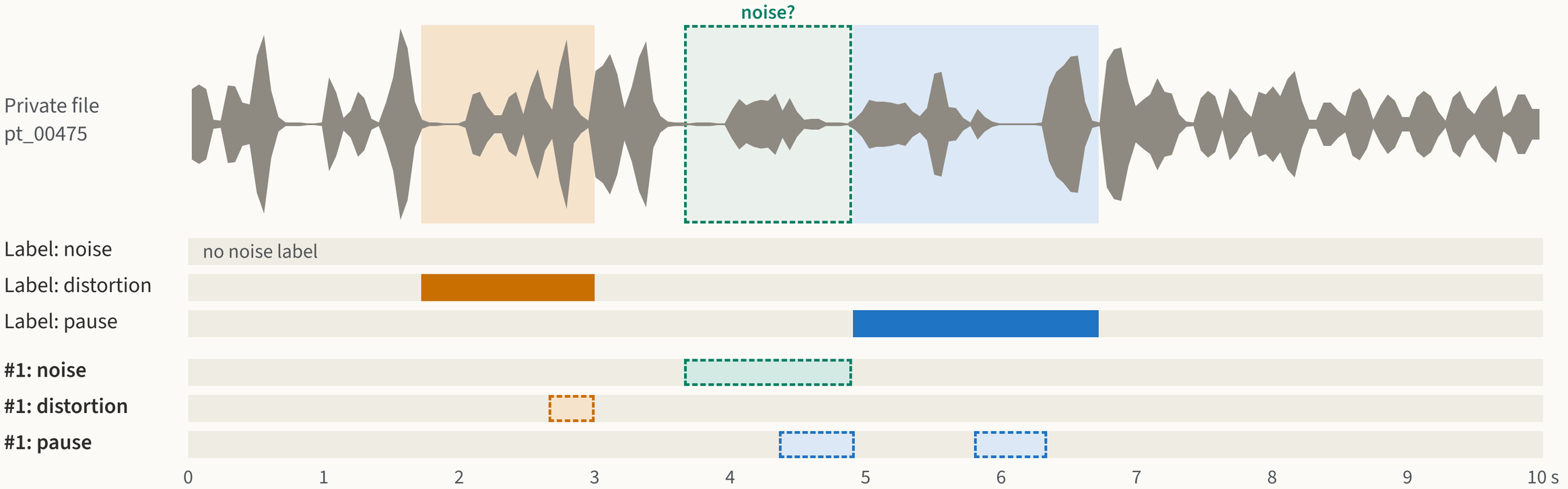
Decoder: turns frame probabilities into fault intervals: threshold search, 1–25 frame smoothing, closing gaps ≤ 0.6 s, dropping segments shorter than 0.3 s

Model selection: by Final on QS dev, plus 0.3 F1 + 0.3 IoU on the simulated dev

All features are computed per 20 ms frame and relative to the file itself. p5, p95: the file's typical low and high energy levels.

PRIVATE TEST · ONE MORE EXAMPLE

The model catches the pause, and noise the labels missed



Pause: detected

#1 correctly finds that this file has a pause. One segment lies inside the labelled region (5.80–6.34 s), another touches its start. The predicted segments are short, so IoU is only about 0.23.

Noise: the model hears it, the labels do not

#1 reports noise at 3.66–4.90 s. On listening, this segment does have real background noise, but the labels contain no noise, so it is scored as a false positive.

Quality almost matches

The labelled MOS is 3.175 and #1 predicts 3.18. Distortion also catches the end of the labelled region (2.66–3.00 s).

Some “false positives” may come from missing labels, not necessarily model errors.

Labels released by the organizers. Solid blocks: labels; dashed boxes: submission #1. IoU estimated from the timestamps.

BACKUP SLIDES

What the 17 simulated fault types sound like

Fault type	What it sounds like	Share
Distortion		
Dropout clusters	Speech “hiccups”, losing very short beats, like a flaky network 3–9 zero-signal gaps of 40–100 ms each, within a 0.6–2 s region	28%
Clip + attenuate	Recorded too loud so it breaks up, then the whole file is turned down: both harsh and quiet clipped at 10–40% of amplitude, then the file peak lowered to 0.04–0.3	12%
Musical noise	“Watery” chirping, thinned voice: over-aggressive denoising spectral subtraction at 2–5× the estimated noise level	12%
Segment low-pass	A segment suddenly turns muffled, like speaking behind a wall 0.8–2 kHz low-pass on a 1–4 s segment	10%
Codec quantisation	“Metallic”, holey voice, like a call on a weak network spectral quantisation in 3–8 dB steps, cut above 3–6 kHz, 10–35% of spectral bins punched out	10%
AGC pumping	Volume keeps jumping up and down, like a broken auto-gain random gain −14 to +6 dB, changing every 50–150 ms	8%
Overdrive	Saturated, thick buzz, like a speaker turned up too loud tanh distortion, gain 8–30	6%
Bit-crushing	Gritty, crunchy sound like an old game console amplitude resolution reduced to 3–5 bits	4%
Pitch shift	Voice shifted much higher or lower, sometimes wobbling 3–7 semitone shift; half the time with 6–20 Hz vibrato	4%
Griffin-Lim	“Robotic”, buzzy voice like old TTS phase discarded, rebuilt with Griffin-Lim in only 2–7 iterations	3%
Aliasing	Hissing, harsh highs from sampling too sparsely keep 1 of every 2–5 samples and repeat it	3%

Fault type	What it sounds like	Share
Noise		
Real recordings	Real noise: music, traffic, in-car, background talkers, TV, rain MUSAN and RIRS point sources, SNR 10–28 dB	75%
Synthetic static	Crackle, radio hiss, or mains hum crackle 45%, radio hiss 35%, 50/100 Hz hum 20%	25%
Unnatural pause		
Single gap	Speech stops for an unusually long stretch insert 0.25–0.85 s, filled with the file's own background	55%
Multiple gaps	As above, but repeated 2–3 times in the file 2–3 gaps of 0.25–0.85 s at random positions	30%
Stutter	Repeats a fragment as if stumbling: “I… I… I” repeat an 80–250 ms fragment 2–4 times, with 50–200 ms gaps	10%
Choppy speech	Halting speech, short fragments separated by silences 0.15–0.5 s fragments with 0.2–0.5 s gaps, within a 1.5–5 s region	5%

Not counted as faults, not labelled: light room reverb (30% of files), very low background noise (30–50 dB below speech), telephone-style band-pass filtering.

Simulator v4 (src/simulate_vi.py). Shares are per type within its group, as in the outer ring of the chart on slide 5.