A Finetuned SpeechLLM for Joint Multi-Granular L2 Assessment and Natural-Language Rationales

Aditya Kamlesh Parikh (*aditya.parikh@ru.nl*) · **Cristian Tejedor-García** · **Catia Cucchiarini** · **Helmer Strik**

Centre for Language Studies, Radboud University, Nijmegen, The Netherlands

Radboud Universiteit

Motivation

Automated L2 assessment can assign proficiency labels but typically outputs **opaque numeric scores** with no actionable insight.

In practice learners need feedback that is interpretable across all granularities.

MODELS GIVE

`Fluency: 7/10`

LEARNERS NEED

What went wrong, and where.

Open gaps

G1 No single E2E model is jointly **multi-granular**: sentence *and* word/phoneme level and also generate interpretable rationales.

G2 Severe **label imbalance** (over 80% Good/Excellent) is problematic for ordinal training.

G3 Rationales may sound convincing yet **not match the labels the model actually assigned**.

RESEARCH QUESTION

To what extent can an E2E SpeechLLM **jointly predict** multi-aspect, multi-granular L2 labels and generate rationales **consistent with its predictions and human judgments**?

Model architecture

BACKBONE	**Qwen2-Audio-7B-Instruct**
WEIGHTS	**Frozen · 4-bit quantized**
ADAPTERS	**LoRA r=64, α=128, drop 0.05**

~115M trainable params

1.6% of the 7B model

Input

Audio + Rubric Instructions

Output

Sentence, word and phoneme level scores with rationales

Hybrid objective

$$L_{\text{total}} = L_{\text{BDPO}} + \lambda \cdot L_{\text{SFT}}$$

L_{SFT}: format learning via teacher forcing, prompt tokens masked.

L_{BDPO}: preference alignment on chosen vs. rejected labels.

Rejected ordinal labels are **near-misses** (Good vs. Excellent). BDPO bounds how far they fall, capped by δ.

Worst | Bad | Avg | **Good** | **Excel**

Data & rejected pairs

SpeechOcean762
English L2 Speech
L1 Mandarin

5,000 utterances

2500/2500 Train/Test Split

5 dims → 5 ordinal labels

SO762 label distribution, heavily skewed

One ground truth set per utterance → synthesize **rejected** labels by perturbing.

Asymmetric: **88% downgraded / 12% upgraded** to counter the "niceness bias".

Multi-granular results

TASK	PCC↑	RMSE↓	MCC↑
Sent. Acc	0.66	1.72	0.35
Fluency	**0.73**	1.33	**0.47**
Prosody	0.71	1.48	0.42
Word Acc	0.52	1.75	0.39
Phon. Acc	0.42	0.36	0.31

Best results for suprasegmental features (fluency, prosody). Word MCC exceeds Sentence MCC, so localized errors are categorized reliably.

Ours vs. SOTA systems (PCC)

TASK	BDPO-M	BDPO-S	GOPT	Azure	SimPO
Sent. Acc	**0.66**	**0.62**	0.71	0.70	0.68
Fluency	**0.73**	**0.72**	0.75	0.72	0.73
Prosody	**0.71**	**0.71**	0.76	0.84	0.73
Word Acc	**0.52**	**0.57**	0.53	0.62	0.51
Phon. Acc	**0.42**	**0.40**	0.61	n/a	0.38

BDPO-M (joint) and **BDPO-S** (single-granularity) are ours. Joint training improves sentence accuracy; **on phonemes it beats SimPO but trails GOP-based GOPT (0.42 vs 0.61).**

Rationale reliability

Plausible & self-consistent at the sentence level; **faithfulness degrades** at token level: references sparse, weakly aligned.

CONCLUSIONS

1. One model, one response: joint multi-granular labels and a rationale, matching single-granularity models.
2. BDPO improves rubric adherence and label–rationale consistency under data skewness.
3. Rationales are reliable for holistic diagnosis; but not yet for pinpointing errors on single words or phonemes.

Github Code Repository

Responsible AI for Voice Diagnostics
AiNed Fellowship Grants
NGF.1607.22.013

Interspeech 2026