

LISTEN, DIAGNOSE, AND INTERRUPT: COUPLING MISPRONUNCIATION DIAGNOSTICS INTO A FULL-DUPLEX SPEECH LLM

Linkai Peng¹ Baorian Nuchged² Kaiqi Fu³ Yuyang Yao⁴ Jinsong Zhang⁵

¹University of Connecticut

²The University of Texas at Austin

³X Square Robot

⁴University of Oxford

⁵Beijing Language and Culture University

linkai.peng@uconn.edu

ABSTRACT

Corrective feedback can be particularly useful near the moment of the error, which often requires the diagnosis to reach the dialogue model while the learner is still speaking. We couple a time-aligned mispronunciation detection signal into a full-duplex speech LLM through a residual channel that carries the signal frame by frame, so the model can begin responding shortly after a word is flagged and speak the correction incrementally. With the offline detector, sentence-level detection F1 is nearly unchanged. The channel relays 83.7% of detector diagnoses, and 43.5% of its feedback matches human annotation, compared with 48.4% for an adapted turn-end text handoff. Unlike text handoff, it neither waits for the turn end nor adds sequence positions. Median first-audio latency is 516 ms after the flagged word end. The model’s PASS verdict also reduces the false interruption rate (FIR) from 55.0% to 39.8% on a native-speech stress test, although no reduction is observed on the smaller in-domain clean L2 subset.

Index Terms— mispronunciation detection and diagnosis, computer-assisted pronunciation training, full-duplex spoken dialogue, speech large language models

1. INTRODUCTION

In 1994, Project LISTEN could already follow a child reading aloud, notice a mispronunciation, and intervene [1]. In the decades since, mispronunciation detection and diagnosis (MDD) has advanced from goodness of pronunciation (GOP) scoring [2] through end-to-end CTC models [3] to detectors built on self-supervised speech representations [4, 5, 6], including training-free retrieval over them [7]. More recently, audio LLMs [8, 9] can be prompted to comment on pronunciation directly, and full-duplex speech LLMs [10, 11, 12, 13] listen while they speak, opening the prospect of a humanlike tutor. Timing, however, has not kept pace. LLM-based tutors typically wait for the learner to finish before responding, even when the underlying model is full-duplex and could speak while the learner is still talking.

Much progress in the field has been measured by how accurately the detector finds errors [14, 15]. Yet detection is only part of tutoring. The LLM can add natural-language explanation and the judgment of whether to intervene. We ask how the two can form a single interactive system. One straightforward design is turn-based, with the diagnosis produced after the sentence ends and written into the prompt as text. What this design gives up is timing. Corrective feedback supports L2 pronunciation development [16, 17], and teachers routinely give it orally within the interaction, in response to the error [18, 19]. A turn-based tutor, by contrast, can only respond once the sentence has ended.

Rather than waiting for the sentence to end, in this paper we propose a tutoring system that keeps the diagnosis time-aligned and couples it into a full-duplex speech LLM while the learner is still speaking. An MDD detector, a WavLM encoder [6] with phone-level classification heads, follows the learner’s speech and flags mispronounced words as they occur. Its diagnosis enters the LLM, adapted with low-rank adaptation (LoRA) [20] on a frozen backbone, through a zero-initialized residual channel rather than through the text prompt.

Our contributions are twofold. First, we propose a *time-aligned diagnostic channel* that adds the detector’s per-frame state to the residual stream of the host and, at the end of a flagged word, names the word and its phone-level error. Because the channel is added to the residual stream rather than to the token sequence, a diagnosis reaches the model shortly after the word ends without waiting for the sentence to finish, rebuilding the text prompt, occupying sequence positions, or displacing learner audio. Second, we build and evaluate the complete tutoring loop, in which the LLM decides from its first token whether to intervene or stay silent and speaks through incremental text-to-speech (TTS). Beyond detection F1, we report the false interruption rate (FIR), the content correctness of the feedback against human annotation. To our knowledge, this is the first system to couple time-aligned MDD diagnostics into a full-duplex speech LLM through a residual channel that occupies no sequence positions.

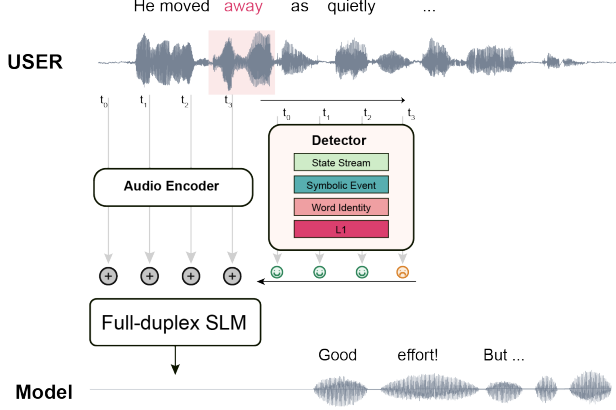


Fig. 1. Our residual coupling system. The detector emits a streaming state for every candidate word and one flag, at the end of the flagged word. Four inputs share three zero-initialized projections and are summed into the host’s residual stream before the decoder stack.

2. PROPOSED SYSTEM

Fig. 1 illustrates the architecture: an MDD detector, a residual channel that carries its per-frame diagnostic state, and a full-duplex speech LLM that hosts the dialogue. VoiceChat’s built-in incremental TTS (not shown) voices the response.

Detector. The detector consists of a WavLM-large encoder [6] followed by two classification heads over the encoder’s final hidden states. The **slot head** treats each reference phone as a slot and classifies its force-aligned frames as correct, deleted, or substituted by one of 39 phonemes. Its labels come from aligning the human phoneme annotation to the reference sequence. In contrast, the **CTC head** is a phone recognizer trained with CTC against the annotated realized phoneme sequence. At inference it is decoded freely and compared with the reference to obtain substitution and deletion rates.

A word-level logistic regression fuses the outputs of the two heads. For each word, its ten-dimensional input comprises summary statistics of phone-level GOP scores [2], error posteriors from the slot head, phone substitution and deletion rates, and the number of reference phones. The regression produces a *badness* score b , with a higher value indicating a greater likelihood of mispronunciation. A word is marked as a *flagged word* when its score exceeds a threshold selected on the development set. Its end time is denoted t_0 .

Speech-LLM backbone. VoiceChat-11B [13] (the *host*) serves as the speech-dialogue backbone and receives the same learner-audio stream as the MDD detector. Its coaching prompt is fixed and contains the reference sentence. The host processes audio in 80 ms frames indexed by t . Let $\mathbf{u}_t$ be its own fused input at frame t , that is, what would reach its decoder with no channel attached.

Residual coupling. The residual channel carries four

inputs to the host, three derived from the detector and one session-level constant. Each is written frame by frame, projected, and added to the residual stream before the decoder (Fig. 1).

Symbolic event. The channel looks up, in the host’s frozen embedding table, pronunciation error, the word index, the intended phoneme, and the phoneme judged as actually produced, and sums their embeddings with position encodings. Let the sum be $\mathbf{c}_t \in \mathbb{R}^H$, where H is the host’s embedding dimension.

Word identity. $\mathbf{e}_t \in \mathbb{R}^H$ is the mean embedding, taken from the same table, of the subword tokens into which the host tokenizer splits the flagged word.

State stream. For each candidate word, the detector submits a state once shortly after the word end. It writes four values $\mathbf{s}_t^{\text{det}} \in \mathbb{R}^4$, namely its badness and three GOP z -scores, which remain held until the next submission. The three GOP z -scores preserve the detector’s confidence in the underlying judgment, information that badness compresses away.

L1 lanes. The learner’s native language is encoded as a six-way one-hot vector $\mathbf{s}^{\text{L1}} \in \{0, 1\}^6$, a session-level constant held on every frame and all zeros for native speakers (TIMIT-400).

Three zero-initialized projections [21], $W_c, W_w \in \mathbb{R}^{H \times H}$ and $W_s \in \mathbb{R}^{H \times 10}$, map the four inputs into the host’s embedding space, with the state stream and the L1 lanes concatenated into a single $(4 + 6)$ -dimensional vector. The projected inputs are added to the residual stream as

$$\mathbf{h}_t = \mathbf{u}_t + W_c \mathbf{c}_t + W_s [\mathbf{s}_t^{\text{det}} \parallel \mathbf{s}^{\text{L1}}] + W_w \mathbf{e}_t, \quad (1)$$

where $\parallel$ denotes concatenation. Since the channel is frame-wise, the diagnosis need not be serialized into the dialogue prompt or delayed until the turn ends.

Response. Two stages decide whether the learner is interrupted. The first is the detector’s badness threshold described above. The host applies the second. Its response begins with a verdict token, INTERVENE when a correction is needed and PASS when it is not. The tutor stays silent only when $P(\text{PASS}) > 0.5$ and $b < \beta$. Thus the host can withdraw a flag below the badness ceiling β . An INTERVENE response continues into a correction naming the flagged word and both phones. This correction is passed incrementally to the TTS while the model is still receiving the learner’s audio.

3. EXPERIMENTS

3.1. Datasets

Our data split follows the commonly used partition of L2-ARCTIC [22]. We use the six annotated speakers NJS, TLV, TNI, TXHC, YKWK, and ZHAA as the test set. This set contains 900 manually annotated utterances. Across the remaining 18 speakers, we use a total of 2,159 utterances for training and 540 for development. We tune all model choices

and thresholds on the training and development sets and evaluate on the test set.

To measure the FIR, we need an evaluation pool that is clean by construction. L2-ARCTIC contains only 70 error-free test utterances, which is too small for this purpose. We therefore add TIMIT-400, a set of 400 utterances from the TIMIT test partition [23]. In this set, the manual phone transcriptions agree with the canonical pronunciations throughout, so it contains no pronunciation errors by construction.

3.2. Experimental Setup

Detector. The detector encoder and the two heads are trained in two stages. The first stage uses the 4,620 TIMIT sentences, and the second uses the L2-ARCTIC training split. We use learning rates of 10^{-5} for the encoder and 10^{-3} for the heads, accumulate gradients over 8 sentences, train for up to 25 epochs, and stop after 6 epochs without improvement. The word-level logistic regression is standardized, uses balanced class weights, and is fitted on the 20,333 words in the training split. The badness threshold, tuned on the development set, is 0.46.

Host. The host is fully frozen except for the three zero-initialized channel projections and the LoRA adapters [20]. Because the host is a hybrid Mamba-attention model whose last eight layers contain no attention, we place rank-16 adapters on 20 linear modules in layers 48–55 and on attention layers 30 and 39. The projections and adapters together contain 45.5M trainable parameters, 0.4% of the host. We train for 3,000 steps with gradient accumulation over 8 samples and a learning rate of 10^{-4} .

Each training sample pairs a sentence with the top-1 diagnostic event from the detector. Running the detector on the annotated sentences from the 18 training speakers yields 1,819 training samples. We also regenerate candidates through 4-fold cross-fitting to improve robustness to detector variation. We add 1,600 PASS samples from unannotated L2-ARCTIC sentences to cover cases where pronunciation is correct and no interruption is needed. These samples come from cross-validation with multiple ASR systems and have no human labels, so they may still contain residual pronunciation errors. At inference, we use a badness ceiling of $\beta = 0.8$, selected on the development set.

Metrics. The main metrics are Sent. F1, FIR, and Content. For Sent. F1, a detection counts as correct only if the top-1 flagged word is also marked as mispronounced in the human annotation. FIR is the share of TIMIT-400 native sentences that trigger an interruption. Content measures whether the feedback is correct, that is, the proportion of cases in which the mispronounced word, target pronunciation, and actual pronunciation in the response match the human annotation. Arrives and Seq. pos. describe timing and computational cost. Arrives is the earliest time when the diagnosis becomes available to the host, and Seq. pos. is the number of sequence

positions occupied by the diagnostic text. Relay is meaningful only for A–D and O, which share the same detector. It denotes the proportion of responses that fully relay the detector diagnosis and measures channel fidelity rather than diagnostic correctness.

4. RESULTS

4.1. Diagnosis Path

Table 1 compares systems of several paradigms. Z1 and Z2 prompt general-purpose speech LLMs, the frozen host and Qwen3-Omni-30B [9], for pronunciation assessment directly. Their Content is only 0.5 and 3.8, and their Sent. F1 is 0.30 and 0.32. Their FIR scores also differ substantially, at 23.5 and 58.5, suggesting that their interruption decisions may depend on factors beyond phoneme-level evidence. Z3 applies instruction tuning to Z1 and improves performance, with F1 0.59 and Content 27.2, but remains 21 percentage points below the detector-based template baseline (System A). One possible reason is that the host acoustic encoder is oriented toward semantics and lacks sufficient resolution for subtle pronunciation deviations. Z4 uses a simpler approach. It gives the learner’s ASR transcript from Parakeet-TDT [24] to Qwen2-7B [25] and asks it to compare the transcript with the reference text. It can identify some word-level errors but cannot report phoneme-level errors, and its Content is 0. Overall, these results support the use of a dedicated detector in this setting, even though faithfully relayed feedback is bounded by detector accuracy.

With the detector serving as the source, the remaining question is how its diagnosis should enter the host. B–D and O share the same detector and the same host. Text handoff is a natural option. B writes the detector result into the prompt and lets the frozen host generate feedback, but its Content score is only 9.8. C matches the template baseline’s Content score of 48.4 after adaptation. The adapted LLM reproduces the detector result, achieving a Relay score of 100. However, B and C receive the diagnosis only at the end of the turn, as 58 prompt tokens, so the prompt must be rebuilt before feedback can start. D uses VoiceChat’s function-call interface to insert the same diagnosis as a tool result at the end of the flagged word. Without adaptation, D reaches 14.2, higher than B, suggesting that the function channel can convey diagnostic information. Adaptation could improve diagnostic fidelity, but would not remove the sequence-position cost.

In D, each inserted token occupies an 80 ms step during which learner audio is replaced by silence. Passing the same diagnosis as C takes 70 steps (5.6 s). O adds the diagnosis to existing frames without consuming sequence positions, inserting silence, or rebuilding the prompt. It arrives shortly after the end of a word, with Relay 83.7 and Content 43.5, approximately 5 percentage points below C. Reading Relay and Content together suggests that this gap may partly reflect

Table 1. Comparison with systems of different paradigms. SELF means the model diagnoses the audio itself. Sent. F1 on the L2-ARCTIC test set. FIR on TIMIT-400. Relay and Content in %. Shaded: B–D forward every detector flag, so these values equal the shared detector’s (A).

	System	Backbone	Diagnosis	Adapted	Arrives	Seq. pos.	Sent. F1	FIR (%)	Relay (det.)	Content (human)
Z1	Prompted host	VoiceChat-11B	SELF	✗	–	–	0.30	23.5	–	0.5
Z2	Prompted audio LLM	Qwen3-Omni-30B	SELF	✗	–	–	0.32	58.5	–	3.8
Z3	Self-diagnosing host	VoiceChat-11B	SELF	✓	–	–	0.59	35.8	–	27.2
Z4	ASR cascade	Qwen2-7B	ASR mismatch	✗	turn end	–	0.30	18.8	–	0.0
A	Template read-out	–	MDD	–	word end	–	0.699	55.0	100.0	48.4
B	Text handoff	VoiceChat-11B	MDD	✗	turn end	58	0.699	55.0	18.5	9.8
C	Text handoff	VoiceChat-11B	MDD	✓	turn end	58			100.0	48.4
D	Function-channel text	VoiceChat-11B	MDD	✗	word end	70			27.2	14.2
O	Residual channel (ours)	VoiceChat-11B	MDD	✓	word end	0	0.694	39.8	83.7	43.5

Table 2. Channel ablation for system O on flagged test utterances. % of responses matching the detector’s word, phone pair, or both.

Setup	Symbolic event	State stream	Word	Phone	Both
LoRA only	✗	✗	76.2	29.5	28.0
+ symbolic event	pulse	✗	80.9	71.6	63.0
+ state stream	pulse	✓	82.9	80.2	70.7
pulse → hold	hold	✓	82.5	85.9	75.0
+ L1 lanes	hold	✓	82.3	88.0	75.5
+ word identity	hold	✓	91.6	88.4	82.9

imperfect transmission of the detector diagnosis. Among the LLM paths, O alone combines word-end delivery and feedback correctness close to the template baseline with zero additional sequence positions.

O keeps Sent. F1 nearly unchanged (0.699 for the detector alone vs. 0.694) and, through the host’s PASS verdict, reduces FIR on TIMIT-400 from 55.0% to 39.8%. This reduction is observed only on the out-of-domain native-speech set. On the 70 clean sentences in the L2 test set, the FIR of O remains unchanged at 62.9%.

4.2. Channel Composition

Table 2 reports ablations of the channel components for system O. Evaluation uses flagged test sentences, a different subset from that used in Table 1, scored against the detector diagnosis. The LoRA-only baseline receives audio and the reference sentence and is forced to act as its own detector. It names the correct word in 76.2% of responses, but only 29.5% correctly state what is wrong. Adding the symbolic event raises Phone from 29.5% to 71.6%, and the detector state stream adds another 8.6 points.

The symbolic event can be written as a *pulse* lasting four frames from its submission, after which it returns to zero, or as a *hold*, kept from its submission until the detector’s next submission so that it stays present while the response

is generated. Switching from pulse to hold raises Phone by another 5.7 percentage points, and the L1 lanes by 2.1. Finally, adding the word embedding on the same frames raises the word identification rate to 91.6%. The channel therefore mainly improves attribution, that is, the ability to point out what is wrong.

4.3. Streaming Latency

The latency from the flagged word end to the first audio frame equals detector latency plus speaking latency. The streaming detector keeps listening for a right-context window (R-CTX) after each word end before submitting. On the development set, increasing R-CTX from 0 to 320 ms shifts the median submission from 120 to 400 ms and raises agreement, the fraction of submitted words whose streaming flag and phone pair match the offline full-utterance reading, from 0.789 to 0.857. The gain is mainly in the first 80 ms, adding 4.4 points, while the next 240 ms adds only 2.4 points. We therefore set the working point to 80 ms. On a single A100 (80 GB) GPU, it gives a median submission of 180 ms and a chunkwise real-time factor (RTF) of 0.24. Table 1 uses the offline detector while in this streaming setting, O reaches Sent. F1 0.605 and Content 35.8%. After the detector submits, the median latency to the first audio frame is 336 ms and the p90 is 379 ms. That is, feedback onset is at about 516 ms after the flagged word end.

5. CONCLUSIONS

We presented a residual channel that injects time-aligned pronunciation diagnostics into a full-duplex speech LLM frame by frame, so corrective feedback can begin shortly after the flagged word ends without occupying sequence positions. Against human annotation, its feedback correctness stays within 5 percentage points of an adapted text baseline that copies the diagnosis verbatim, with detection F1 nearly unchanged. False interruptions decrease on out-of-domain native speech but not on in-domain clean L2 utterances.

6. REFERENCES

- [1] Jack Mostow, Steven F. Roth, Alexander G. Hauptmann, and Matthew Kane, “A prototype reading coach that listens,” in *Proc. AAAI*, 1994, pp. 785–792.
- [2] Silke M. Witt and Steve J. Young, “Phone-level pronunciation scoring and assessment for interactive language learning,” *Speech Communication*, vol. 30, no. 2–3, pp. 95–108, 2000.
- [3] Wai-Kim Leung, Xunying Liu, and Helen Meng, “CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis,” in *Proc. IEEE ICASSP*, 2019, pp. 8132–8136.
- [4] Linkai Peng, Kaiqi Fu, Binghuai Lin, et al., “A study on fine-tuning wav2vec2.0 model for the task of mispronunciation detection and diagnosis,” in *Proc. Interspeech*, 2021, pp. 4448–4452.
- [5] Eesung Kim, Jae-Jin Jeon, Hyeji Seo, and Hoon Kim, “Automatic pronunciation assessment using self-supervised speech representation learning,” in *Proc. Interspeech*, 2022, pp. 1411–1415.
- [6] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, et al., “WavLM: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [7] Huu Tuong Tu, Ha Viet Khanh, Tran Tien Dat, et al., “Mispronunciation detection and diagnosis without model training: A retrieval-based approach,” in *Proc. IEEE ICASSP*, 2026, pp. 16312–16316.
- [8] Yunfei Chu, Jin Xu, Xiaohuan Zhou, et al., “Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models,” arXiv:2311.07919, 2023.
- [9] Jin Xu, Zhifang Guo, Hangrui Hu, et al., “Qwen3-Omni technical report,” arXiv:2509.17765, 2025.
- [10] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour, “Moshi: A speech-text foundation model for real-time dialogue,” arXiv:2410.00037, 2024.
- [11] Ziyang Ma, Yakun Song, Chenpeng Du, et al., “Language model can listen while speaking,” in *Proc. AAAI*, 2025, pp. 24831–24839.
- [12] Xiong Wang, Yangze Li, Chaoyou Fu, et al., “Freeze-Omni: A smart and low latency speech-to-speech dialogue model with frozen LLM,” in *Proc. ICML*, 2025, pp. 63345–63354.
- [13] Jagadeesh Balam, Travis Bartley, Edresson Casanova, et al., “NemotronLabs VoiceChat: An open full-duplex speech-to-speech model with tool calling capabilities,” arXiv:2609.21967, 2026.
- [14] Minglin Wu, Kun Li, Wai-Kim Leung, and Helen Meng, “Transformer based end-to-end mispronunciation detection and diagnosis,” in *Proc. Interspeech*, 2021, pp. 3954–3958.
- [15] Wenxuan Ye, Shaoguang Mao, Frank Soong, et al., “An approach to mispronunciation detection and diagnosis with acoustic, phonetic and linguistic (APL) embeddings,” in *Proc. IEEE ICASSP*, 2022, pp. 6827–6831.
- [16] Kazuya Saito and Roy Lyster, “Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /r/ by Japanese learners of English,” *Language Learning*, vol. 62, no. 2, pp. 595–633, 2012.
- [17] Junkyu Lee, Juhyun Jang, and Luke Plonsky, “The effectiveness of second language pronunciation instruction: A meta-analysis,” *Applied Linguistics*, vol. 36, no. 3, pp. 345–366, 2015.
- [18] Roy Lyster and Leila Ranta, “Corrective feedback and learner uptake: Negotiation of form in communicative classrooms,” *Studies in Second Language Acquisition*, vol. 19, no. 1, pp. 37–66, 1997.
- [19] Roy Lyster, Kazuya Saito, and Masatoshi Sato, “Oral corrective feedback in second language classrooms,” *Language Teaching*, vol. 46, no. 1, pp. 1–40, 2013.
- [20] Edward J. Hu, Yelong Shen, Phillip Wallis, et al., “LoRA: Low-rank adaptation of large language models,” in *Proc. ICLR*, 2022.
- [21] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala, “Adding conditional control to text-to-image diffusion models,” in *Proc. IEEE/CVF ICCV*, 2023, pp. 3836–3847.
- [22] Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, et al., “L2-ARCTIC: A non-native English speech corpus,” in *Proc. Interspeech*, 2018, pp. 2783–2787.
- [23] John S. Garofolo, Lori F. Lamel, William M. Fisher, et al., “TIMIT acoustic-phonetic continuous speech corpus,” Linguistic Data Consortium, LDC93S1, 1993.
- [24] NVIDIA, “Parakeet-TDT-0.6B-v2,” Hugging Face model card, 2025.
- [25] An Yang, Baosong Yang, Binyuan Hui, et al., “Qwen2 technical report,” arXiv:2407.10671, 2024.