

*Prompted but **misheard**...*

How to provide **personal** and **accurate** speaking feedback in current spoken dialogue systems

Dr Sha Liu

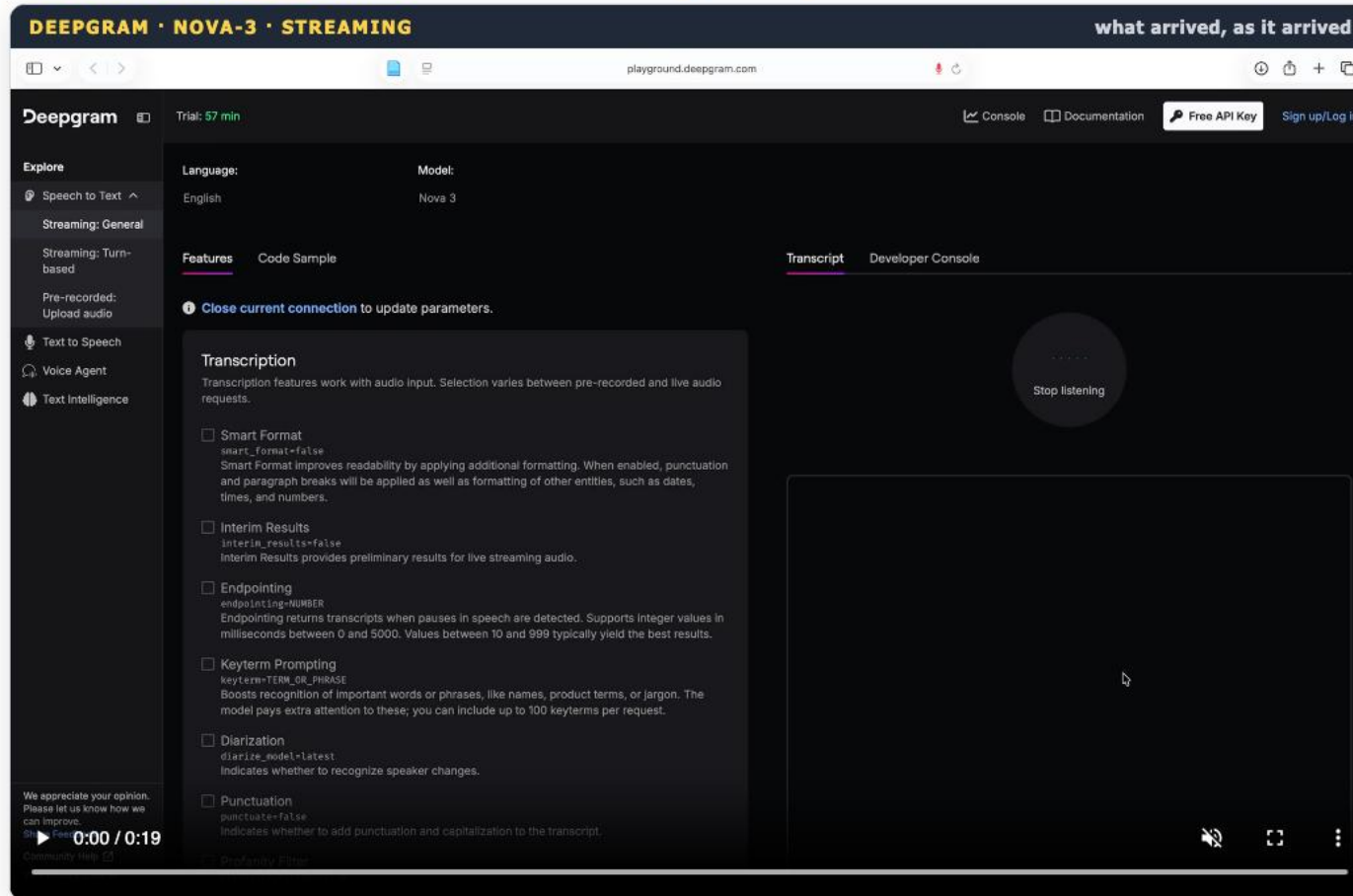


Test Development Researcher

British Council, UK

How well does current ASR (automatic speech recognition) hear English across first languages?

1 OF 2 Me, a Chinese-English speaker, via Nova-3



Watch the right-hand panel as the lines arrive. **Does the text match what I said?**

How well does current ASR (automatic speech recognition) hear English **across first languages?**

1 OF 2 Me, a Chinese-English speaker, via Nova-3

NOVA-3 HEARD

(dropped) [Um ...] **i joined** [join] *the sports club
because because i want to stay healthy*

NOVA-3 HEARD

we meet at four on monday

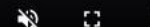
NOVA-3 HEARD

the students paint and make things there

DEEGRAM · NOVA-3 · STREAMING

what arrived, as it arrived

▶ 0:00 / 0:19



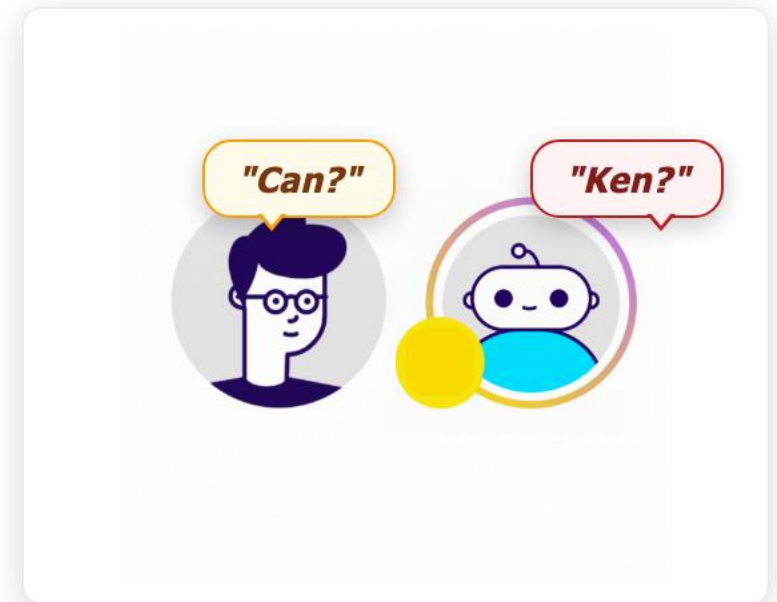
How well does current ASR hear English **across first languages?**

2 OF 2 Singaporean learners via Whisper

WHISPER HEARD *I don't know, you give me suggestion, **Ken** [can?].*

WHISPER HEARD *You explain to me, **Ken** [can?]*

WHISPER HEARD *Run and **swear** [elsewhere] no fun ah.*



The learner's original speech is shown in square brackets. Source: [Aryadoust & Zakaria \(2026\)](#)

Research on ASR errors

ERROR TYPE	EXAMPLE	EVIDENCE
Mishears a word	can → Ken; Sport Club → sparkle	Three turns by Singaporean learners through Whisper (Aryadoust & Zakaria, 2026); Grade 8 learners in Vietnam through Nova-2 and Nova-3 (<i>this study</i>).
Silently repairs the learner's mistake	I join → I joined	In L2 read speech, "82% of utterances contained an overcorrected mispronunciation", and the ASR engine was "most confident when it overcorrected" (Park, 2026).
Adds or drops word endings	join that → joined our	Error types are predictable from the first language's phonology (Choe et al., 2022); proficient Vietnamese speakers "omitted word-final /s/ and /z/ in 28% of expected instances" (Disney & Le, 2024).
► Inserts words never spoken	Anna like → I don't like	ASR engines "can also insert words that were never spoken" (Koenecke et al., 2024).
Drops fillers and hesitations	Um ... → (nothing)	Filled pauses are omitted at widely varying rates (Husain et al., 2026 ; McGuire, 2025).
► Attributes the bot's words to the learner	the chatbot's turn transcribed as the learner's	In dialogue, "the interlocutor's words may also be attributed to the learner" (McKnight et al., 2023).

Where do ASR errors enter current spoken dialogue systems?

Typical pipeline: turn-based, prompt-wrapped LLM cascade



Audio is transcribed to text, then to LLM, then back to audio. **The LLM never sees the audio itself.**

↳ **Feedback** is written by the LLM after the task, **from the ASR transcript of the learner's turns** and the chatbot's turns, not from the audio.

ASR ENGINES MOST COMMONLY USED

Whisper (OpenAI) · Deepgram (Nova-2, Nova-3) · Google Cloud Speech-to-Text

What ASR errors do to speaking feedback

WHAT RESEARCH HAS FOUND

- Automatic grammatical error detection was less accurate on ASR transcripts than on manual ones: its F0.5 score, which weights precision above recall, fell from 42.5 to 29.4 ([Lu et al., 2019](#)).
- ASR errors produced spurious corrections in spoken grammatical error correction ([Qian et al., 2025](#)), and in a spoken practice chatbot they propagated into “nonsensical grammar feedback” ([Li et al., 2024](#)).
- The errors fall hardest on lower-level speaking.
 - A Whisper baseline's error rate at A2 “was roughly twice” that at C1 ([Sun et al., 2025](#)).
 - Vietnamese speakers had “the highest mean word error rate” of 22 first-language groups on one commercial ASR engine ([Chan et al., 2022](#)), and errors concentrate on “words that carry grammatical errors” ([Knill et al., 2018, 2019, as cited in Lu et al., 2019](#)).

For speaking feedback generated by LLMs, the risk of ASR errors propagating into feedback has been pointed out ([Ku, 2026](#); [Aryadoust et al., 2026](#)), *but not yet examined empirically.*

Spoken dialogue systems at scale inherit ASR errors.



**Can feedback that quotes the learner still be both
personal and accurate?**

This study

AIM

To test whether prompt design and model choice can make GenAI feedback on young, low-proficiency learners' speech both personalised and accurate.

QUESTIONS

- RQ1** How do the two quoting strategies (simple vs. governed) affect the feedback's quoting rate, quote faithfulness and message misquote rate?
- RQ2** How do the three models compare on these measures, and where do their misquotes originate?

DATA

94 dialogues between Grade 8 learners in Vietnam (aged 13–14, around A2) and a chatbot, all on one A2-level task, "Which school club do you like best?", with the ASR transcript and a human transcription of each.

ANALYSIS

- **Personalisation:** the quoting rate, the proportion of feedback messages that quote the learner.
- **Quoting accuracy:** quote faithfulness, the proportion of quoted spans that match the human transcription; and the message misquote rate, the proportion of messages with at least one misquote.

This study: two quoting strategies, written into the feedback prompt

	Simple quoting	Governed quoting
The request	One quotation of the learner, introduced by "You said"	The same request
The transcript	Ignore transcription issues	Treat it as fallible: it may contain words the learner did not say
What counts as the learner's own issue	—	Only issues that recur across the learner's turns
Conditions on a quotation	—	Copied exactly from a learner turn; plausible for the learner and the task; central to the point being made
When not to quote	—	Whenever no span qualifies, or the model is unsure what the learner said
Final check	—	Every quotation checked against the conditions before the message is sent

— : not in the prompt. All the prompts share a viability check: when there is too little usable speech, the model writes a generic message that quotes nothing.

This study: the three models

	<u>GPT-4o</u>	<u>GPT-6.1 Sol</u>	<u>Claude Sonnet 5.5</u>
Vendor	OpenAI	OpenAI	Anthropic
Identifier	gpt-4o-2024-08-06	gpt-6.1-sol	claude-sonnet-5-5
Interface	Responses API	Responses API	Message Batches API
Reasoning	No reasoning step	Reasoning effort medium (the default), with detailed reasoning summaries returned	Adaptive thinking at effort high (the default), with summarised thinking returned

Liu, Skidmore & Felice (in preparation).

Results: how often feedback quoted the learner, and how accurately

Quoting rate: the proportion of feedback messages that quote the learner's words.

Quote faithfulness: the proportion of quoted spans that match what the learner said, as established by human transcription.

Message misquote rate: the proportion of messages containing at least one misquote.

MODEL	Quoting rate (%)		Quote faithfulness (%)		Message misquote rate (%)	
	SIMPLE	GOVERNED	SIMPLE	GOVERNED	SIMPLE	GOVERNED
<u>GPT-4o</u>	81.4	73.9	63.9	80.9	53.7	32.4
<u>GPT-6.1 Sol</u>	89.4	75.0	82.2	85.2	31.4	16.5
<u>Claude Sonnet 5.5</u>	90.4	74.5	71.0	86.0	45.2	20.2
All three models	87.1	74.5	72.0	83.7	43.4	23.0

Governed quoting raised quote faithfulness from 72.0% to 83.7% and nearly halved the messages containing a misquote (43.4% → 23.0%), *at the cost of quoting less often (87.1% → 74.5%).*

Liu, Skidmore & Felice (in preparation). First run, with the human transcription being verified against the recordings. Governed minus simple, all models pooled, with 95% bootstrap confidence intervals: quoting rate -12.6 points [-17.0, -8.5]; quote faithfulness +11.6 [8.7, 14.6]; message misquote rate -20.4 [-25.7, -15.2].

Results: where the misquotes came from

ASR-CARRIED

The model copied its input accurately, but the transcript misrepresented the learner.

MISATTRIBUTED

The chatbot's words were quoted as the learner's.

FABRICATED

Words that no one in the dialogue produced in that form.

MODEL	MISQUOTED SPANS	ASR-CARRIED	MISATTRIBUTED	FABRICATED
<u>GPT-4o</u>	244	238	0	6
<u>GPT-6.1 Sol</u>	109	109	0	0
<u>Claude Sonnet 5.5</u>	190	190	0	0
All three models	543	537	0	6

Governed quoting cut misquotation, and the models' own errors were rare: no model quoted the chatbot's words as the learner's, and fabricated quotations were six in all, confined to GPT-4o. *However, ASR-carried misquotation persisted: 537 of the 543 misquoted spans were copied exactly from the ASR transcript.*

Sample: ASR-carried misquotation in feedback

THE LEARNER SAID

*"They **join** the Sports Club to try to stay healthy."*



ASR HEARD

*"They **joined** the sports club to try to stay healthy"*



THE FEEDBACK THEN SAYS

*"You said 'They **joined** the sports club'. To talk about what students usually do, try saying 'They join the sports club'."*

Potential impact: the learner had it right and is told to change it. *Feedback like this can leave a young learner unsure whether the fault is theirs or the machine's.*

Implications

- 1. Prompting.** Governed quoting made feedback more accurate (faithfulness up, misquotes nearly halved) while keeping most feedback specific, but it could not remove misquotation: the error is already in the transcript.
- 2. Model choice.** The newer models (GPT-6.1 Sol, Claude Sonnet 5.5) misquoted less than GPT-4o under both strategies, yet still carried ASR errors into their quotations.

The quoting accuracy of feedback should be monitored. *Learners, especially young and less proficient ones, should be reminded that ASR errors exist.*

Looking ahead: what changes with a voice agent?

IN USE TODAY

TYPICAL PIPELINE: TURN-BASED, PROMPT-WRAPPED LLM CASCADE



Audio is transcribed to text, then to LLM, then back to audio. **Feedback is written from the ASR transcript.**

EMERGING PIPELINE: SPEECH-TO-SPEECH VOICE AGENT, CASCADE COLLAPSED INTO ONE MODEL



One model hears and speaks; the reply needs no transcript. **Feedback can be written from the model's own transcript, or from a separate ASR running alongside.**

TO REFLECT ON

- 1 The model hears the speaker: pronunciation, pace and hesitation now reach it. What should feedback address, and what should it leave alone?
- 2 If feedback still reads an ASR transcript, which of today's errors go away, and which come with it?

Two questions for the discussion

- 1 How can teachers help their learners make the best of AI-mediated speaking feedback **in the face of ASR errors?**
- 2 Would it be better to have a human verify the transcript before it is used for feedback? **What would that mean for speaking assessment at scale?**

Thank you

— and the conversation continues with the LAQ SI 2027 →

★ **Forthcoming**

LANGUAGE ASSESSMENT QUARTERLY

Special Issue 2027 · Construct Transformation and Validity Challenges in AI-Mediated L2 Speaking Assessment

MANUSCRIPTS DUE **30 January 2027**

PUBLICATION **2027 · Issue 4/5**

GET IN TOUCH sha.liu@britishcouncil.org · osf.io/5adyv/overview