

Gemini listens: Analyzing speaking tasks



JALTCALL 2025

Paul Daniels, Kochi University of Technology

Prosody



- Melody of speech
- Pitch
- Stress
- Rhythm

Research on AI assessment

Writing

Agreement of 84.72% (AI & human) <https://doi.org/10.1080/09588221.2024.2371395>

Agreement with humans is achievable, sufficient for low-stakes, formative assessment
<https://doi.org/10.1016/j.caeai.2024.100255>

ChatGPT excels in grammar, spelling, sentence structure, relevance, and supporting evidence, but with thematic consistency, human evaluators outperform AI.
<https://doi.org/10.1007/s44217-024-00320-6>

Grammar scores of human raters and the ChatGPT positively correlated, $r(152) = .55$, $p < .001$. DOI:10.5281/zenodo.10402530

Speaking

Research mainly focuses on how AI can help learners improve speaking skills.
<https://doi.org/10.1016/j.system.2024.103254>

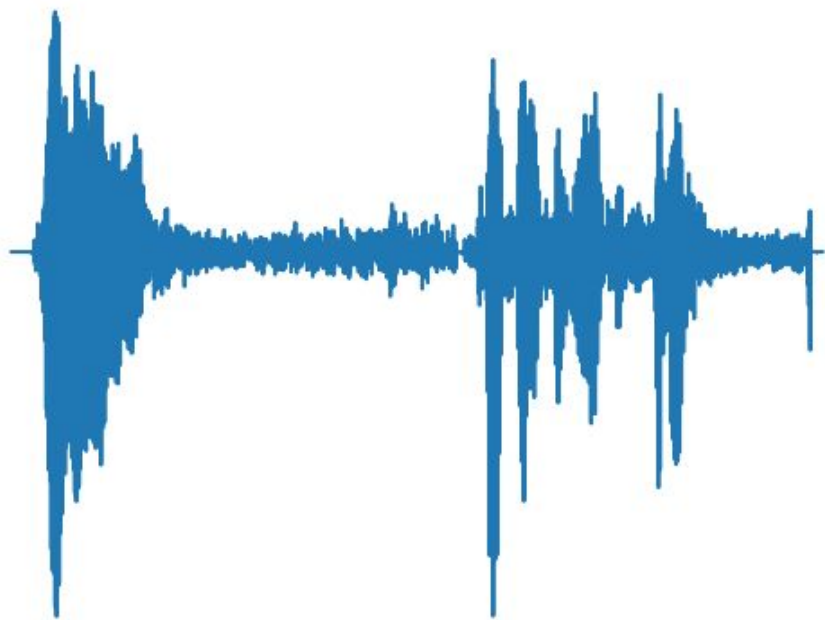
Speaking skill Assessment

Text

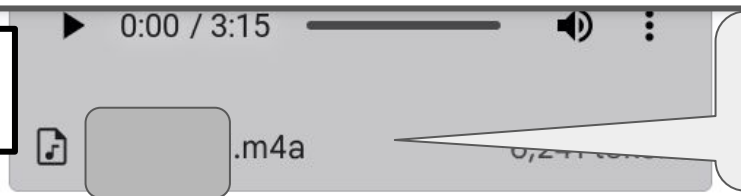
- Word & sentence count
- Vocabulary assessment
- Grammar assessment

Sound

- Intonation
- Stress, rhythm, loudness
- Fluency



aistudio.google.com



media

Please rate the fluency and prosody of the speaker in this audio file on a scale of 1 to 100.

Prompt

I'd rate the speaker's fluency around 60/100 and their prosody around 55/100.

Fluency:

- There are frequent pauses and hesitations, disrupting the flow of speech.
- Some words are mispronounced or repeated.
- The speaker sometimes struggles with grammatical structures, leading to slightly unnatural phrasing.

Results

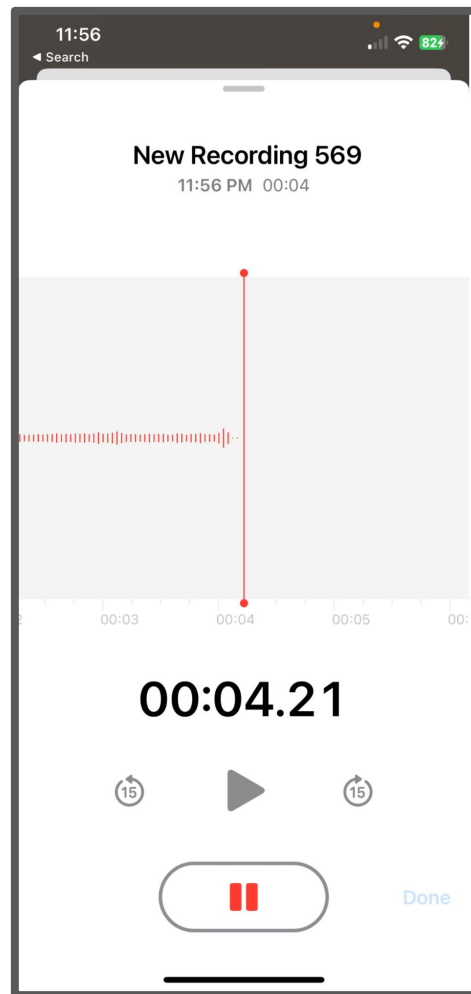
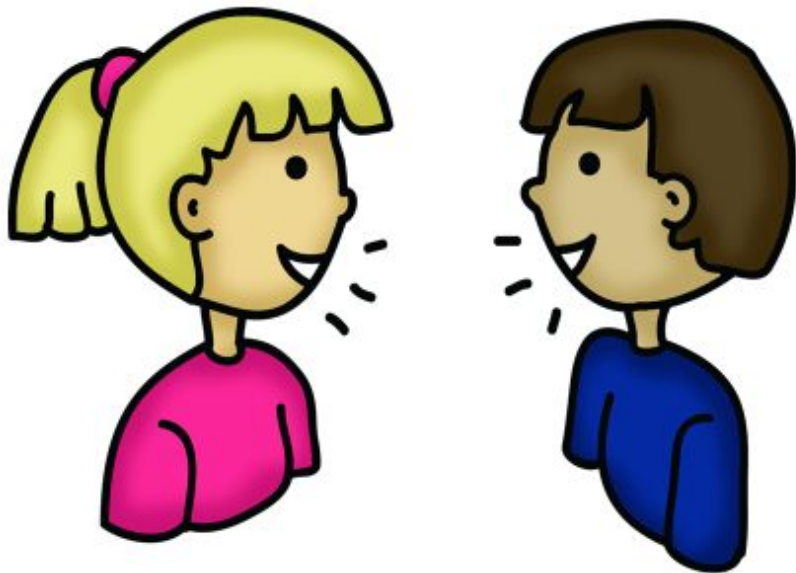
Gemini Prompt

Rate the fluency & prosody of the speaker in this audio file on a scale of 1 to 100. (speak smoothly) (rhythm / intonation)

I would rate the fluency & prosody 60/100.

- There are frequent **pauses** and hesitations, disrupting the flow of speech.
- Some words are **mispronounced** or repeated.
- The speaker sometimes struggles with grammatical structures, leading to slightly unnatural phrasing.
- The overall **pace** is quite **slow**, which contributes to the impression of disfluency.
- The **intonation** is relatively **flat** and monotonous. There isn't much variation in pitch or rhythm, making the speech sound less engaging.
- **Stress** placement is sometimes **inaccurate**, which can affect the clarity and meaning of the sentences.
- There's a lack of expressive phrasing; it sounds as if the speaker is **reading** rather than conveying a message with enthusiasm or naturalness.

Automate



Submit Audio



Search in Drive

Content > riko_survey

✓

☰









⋮

Type

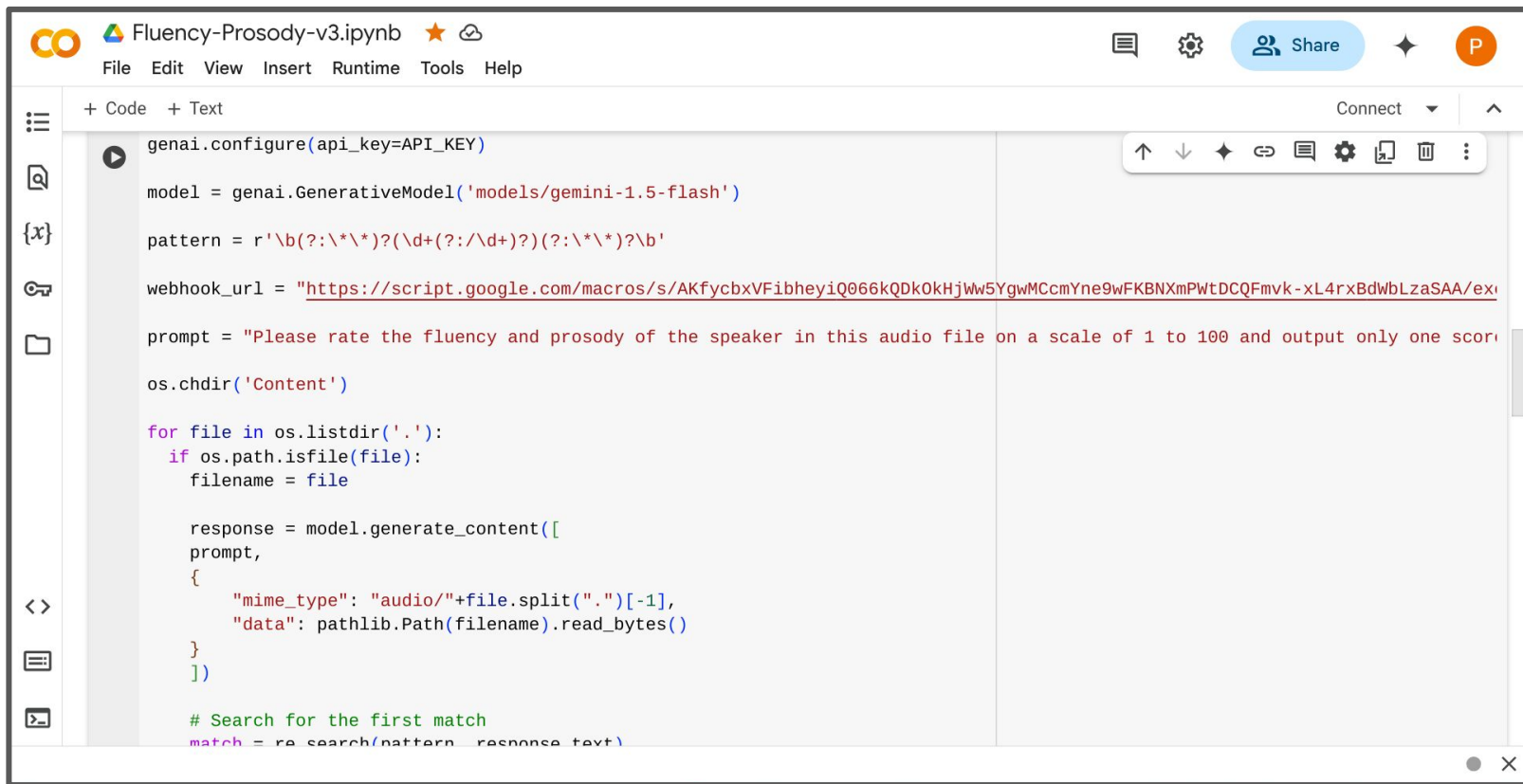
People

Modified

Source

Name	Owner	Last m...	File size
 submission_file_126037...	me	Feb 23, 2025	1.5 MB
 gnsignsubmission_file_1280...	me	Feb 23, 2025	615 KB
 gnsignsubmission_file_1280...	me	Feb 23, 2025	916 KB
 gnsignsubmission_file_1260...	me	Feb 23, 2025	475 KB
 gnsignsubmission_file_1280...	me	Feb 23, 2025	1.8 MB
 gnsignsubmission_file_1270...	me	Feb 23, 2025	1.4 MB
 gnsignsubmission_file_1260...	me	Feb 23, 2025	1.7 MB
 gnsignsubmission_file_1270...	me	Feb 23, 2025	868 KB

Google Colab & Gemini API



The screenshot shows a Google Colab notebook interface. The title bar at the top reads "Fluency-Prosody-v3.ipynb" with a star icon and a cloud icon. Below the title bar is a menu bar with "File", "Edit", "View", "Insert", "Runtime", "Tools", and "Help". On the right side of the title bar are icons for chat, settings, a "Share" button, a star icon, and a profile icon labeled "P".

The notebook has two tabs: "+ Code" (selected) and "+ Text". On the right side of the code editor, there is a "Connect" dropdown menu and a small toolbar with icons for undo, redo, insert, link, chat, settings, download, delete, and a menu icon.

The code in the notebook is as follows:

```
genai.configure(api_key=API_KEY)

model = genai.GenerativeModel('models/gemini-1.5-flash')

pattern = r'\b(?:\s*\s*)?(?:\d+(?:/\d+)?)?(?:\s*\s*)?\b'

webhook_url = "https://script.google.com/macros/s/AKfycbxvFibheyiQ066kQDkOkHjWw5YgwmCcMvYne9wFKBNxmPwtDCQFmVnk-xL4rxBdWbLzaSAA/ex..."

prompt = "Please rate the fluency and prosody of the speaker in this audio file on a scale of 1 to 100 and output only one score"

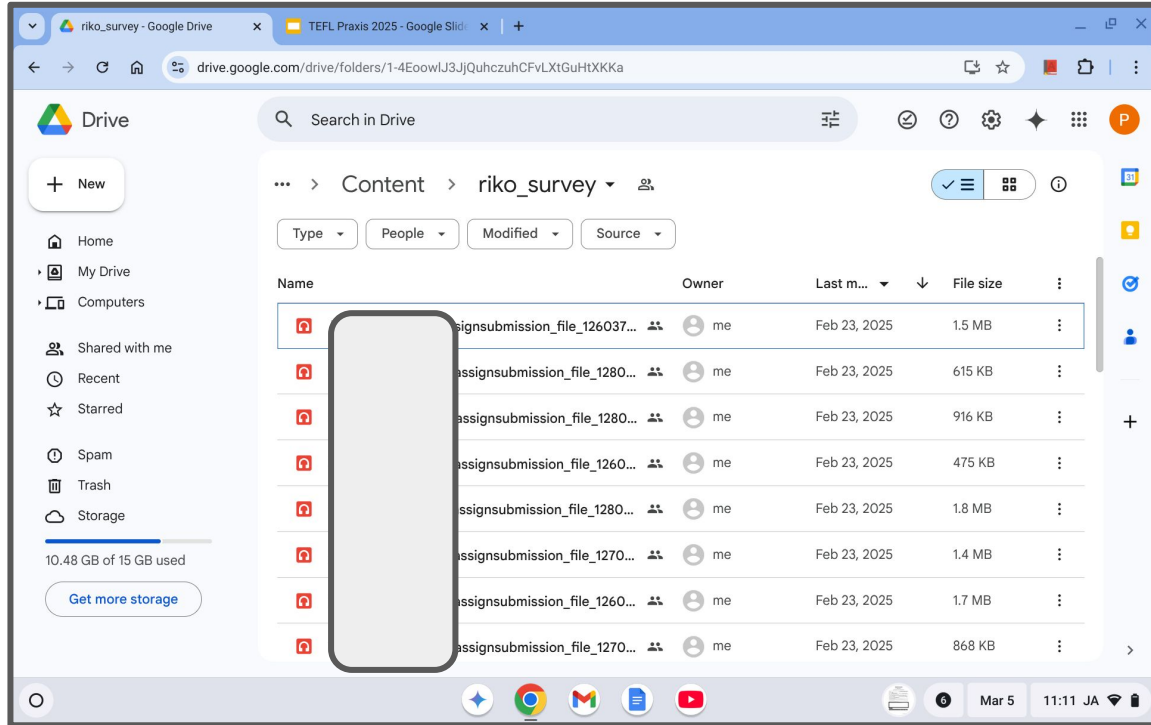
os.chdir('Content')

for file in os.listdir('.'):
    if os.path.isfile(file):
        filename = file

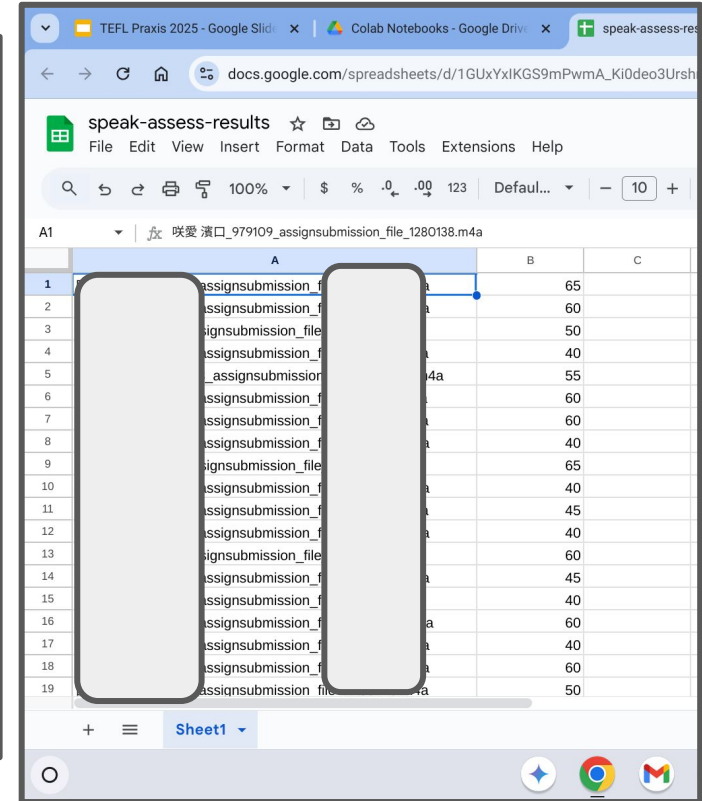
        response = model.generate_content([
            prompt,
            {
                "mime_type": "audio/"+file.split(".")[1],
                "data": pathlib.Path(filename).read_bytes()
            }
        ])

        # Search for the first match
        match = re.search(pattern, response.text)
```

Google Drive



Google Sheets



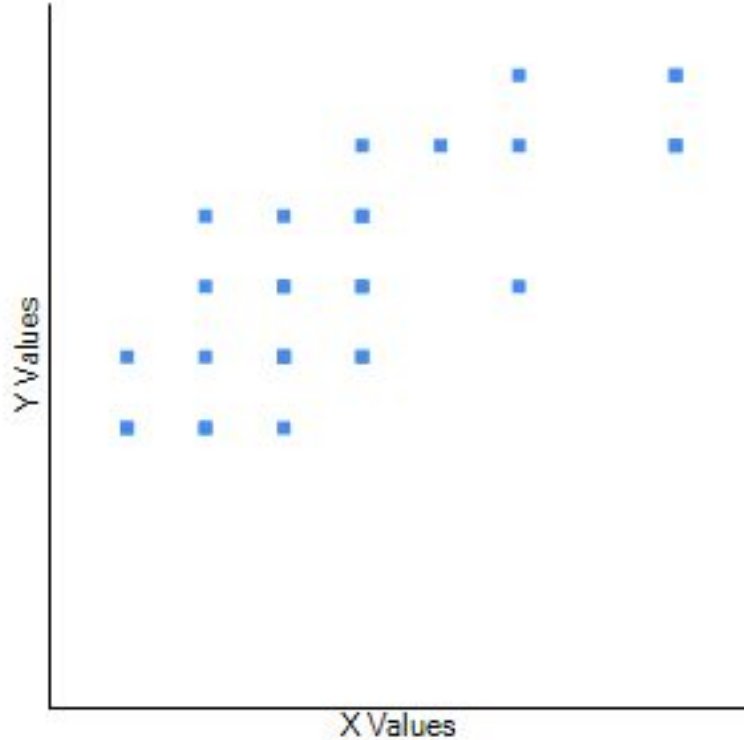
Score verification

Verify low scores

Verify high scores

30	35
30	25
25	30
20	20
15	20
50	45
30	25
25	20
30	30
15	25
50	40

Human & AI scores



$R = 0.7953$

$N = 31$

Positive
correlation

AI Fluency & prosody scores

4/5



2/5



Wrapping it up

- Frequent HUMAN score verification
- Student feedback
- Student check
- Prompt tuning
- Sound & text analysis

Thank you!

daniels@kochi-tech.ac.jp

Abstract

Generative AI is transforming language teaching and learning in areas such as translation, feedback, and evaluation. This presentation examines AI's ability to analyze speaking tasks in the language learning classroom. Most generative AI tools, such as ChatGPT, first convert speech to text and then analyze the transcript—an approach that overlooks important prosodic features. However, Google's Gemini can process raw audio directly, capturing intonation, stress, rhythm, and loudness without relying on text-based transcription. This study compared the accuracy and efficiency of human and AI ratings of pair-work speaking tasks, focusing on Gemini 2.0's multimodal ability to analyze natural prosody and intonation. The findings revealed a moderate positive correlation between human and AI ratings of speaking tasks, indicating that Gemini 2.0 aligns well with human judgments of intonation and rhythm in language learner speech.