09

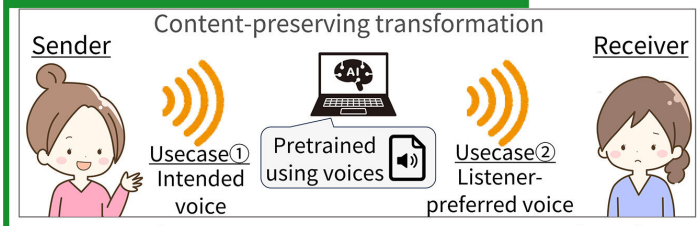
An idol voice dataset for research on speech generation AI

Providing a diverse range of appealing voices

Abstract

Research data for speech generation AI are often biased toward acted speech produced by professional speakers, such as voice actors. As a first step toward enabling AI systems to generate speech that is personalized to individual users' preferences and perceived as appealing, we construct the Japanese Idol Speech (JIS) corpus, a multi-speaker speech dataset featuring live idols as speakers with a wide range of vocal characteristics. JIS is the first large-scale idol speech dataset with properly established contractual arrangements for research use, comprising over 200 speakers and approximately 30 hours of audio. In addition to reading speech and everyday conversational utterances commonly found in existing speech datasets, JIS includes distinctive idol-specific speech styles, such as utterances simulating Instax photo session events. By learning from the voices of diverse individuals, we aim to develop an AI system that enables anyone to flexibly and effectively refine their own voice.

Communication augmentation through speech generation AI



Future goal：Generating attractive voices tailored to fit everyday life (Voice makeup)

What is a fascinating voice ?
①Differ from listener to listener
②Jointly determined by voice quality and style
} Diverse

As a first step, we collect diverse fascinating voices

	JVS	CSJ	JIS (Prop.)
Speaker	Voice actors / Narrators	Non-professional	Live Idols
Scale	~30 h 100 people	~660 h 1417 people	~30 h 204 people
Natural-ness	△	○	○
Attractive-ness	○	△	○

※JVS: Japanese Versatile Speech corpus, CSJ: Corpus of Spontaneous Japanese

Selected speaker genre

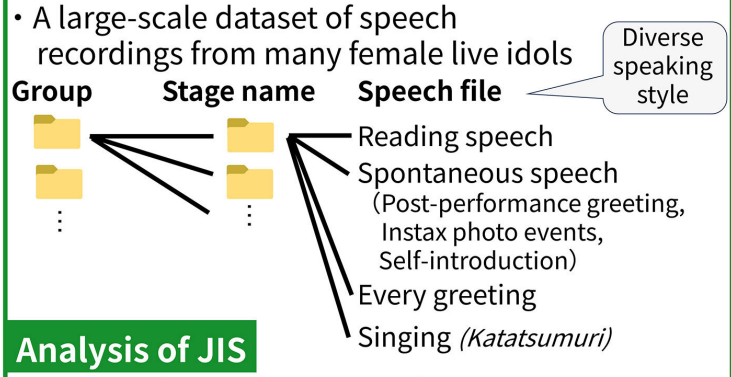
Live idol: Idols focused on live performances

Benefits for Research

- Balancing privacy protection and fan accessibility (via stage names)
- Detailed subjective evaluations by fans who understand speakers' appeal
- Feasible voice collection from many speakers

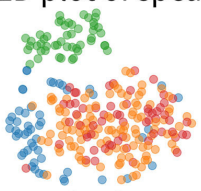
Japanese Idol Speech Corpus (JIS)

- A large-scale dataset of speech recordings from many female live idols



Analysis of JIS

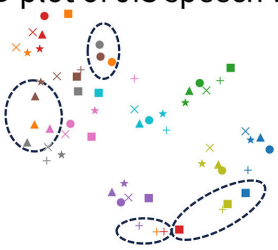
- 2D plot of speaker acoustic features



●: JVS male speakers
●: JVS female speakers
●: JIS female speakers recorded at studio
●: The other JIS female speaker
※Visualizing only reading speech

Separation between JVS female and JIS speakers
➡JIS covers speaker types insufficiently covered by JVS

- 2D plot of JIS speech with various speaking styles



■: Reading speech
▲: Post-performance greeting
●: Instax photo events
×: Self-introduction
★: Everyday greeting
+: Singing
※Same color indicates the same speaker

Not only the speaker differences
but also style-specific features (◌) are confirmed
➡ Suggesting that differences in speaking style are related to listener impressions (e.g., ▲⇒ resonant)
➡ Potential for scenario-aware voice generation

References

[1] Y. Kondo, H. Kameoka, K. Tanaka, T. Kaneko, "JIS: A Speech Corpus of Japanese Idol Speakers with Various Speaking Styles," in *Proc. INTERSPEECH*, pp. 4783-4787, 2025.

Contact

Yuto Kondo, Computational Modeling Research Group, Media Information Laboratory