How similar or Different Is Rakugo Speech Synthesizer to Professional Performers?

Shuhei Kato[*,†], Yusuke Yasuda[*,†], Xin Wang[†], Erica Cooper[†], Junichi Yamagishi[†]

[*]The Graduate University for Advanced Studies, SOKENDAI, Japan
[†]National Institute of Informatics, Japan

How well does TTS entertain audiences?

Towards TTS that entertains audiences

Professional rakugo performers*

✓ Information

Natural speech

✓ Stirring emotion

Audiences

As natural as human speech in the case of narrative style

Text-to-speech (TTS)

✓ Information

Synthesized speech

? Stirring emotion

Audiences

*This photo is transformed from "DP3M2471" by akira kawamura licensed under CC BY 2.0.

Rakugo: A traditional Japanese form of verbal entertainment

Rakugo…

- Is like **one-person stand-up comedy and comic storytelling**.
- Has over 300 years of history.

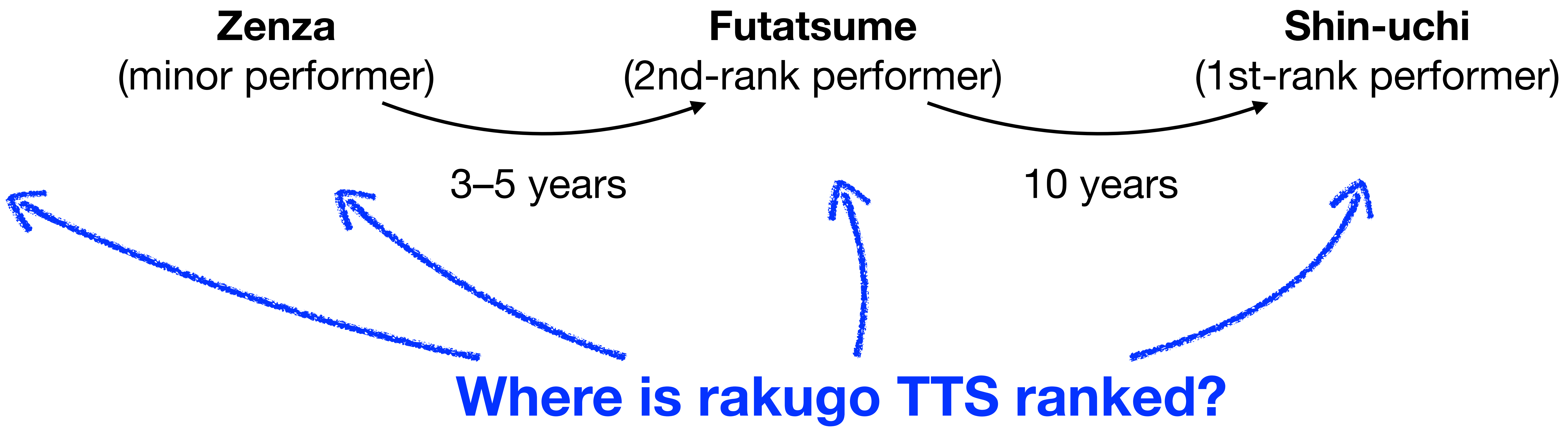
A rakugo performer…

- Performs **improvisationally** or **from memory alone on a stage**.
- **Plays multiple characters**, and **conversations between characters make the story progress**.

Shumputei Shotaro performing rakugo on a stage.

Motivation

- Professional rakugo performers are ranked at **three levels:**



Zenza
(minor performer)

3–5 years

Futatsume
(2nd-rank performer)

10 years

Shin-uchi
(1st-rank performer)

Where is rakugo TTS ranked?

- To investigate this, **we compared synthesized rakugo speech with ones by professional performers through a listening test.**

Listening test

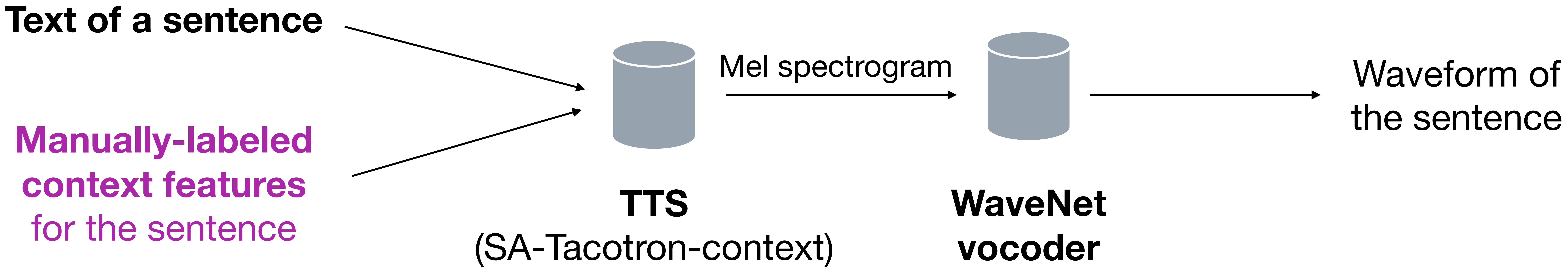
Speech samples (professional performers)

- We recorded performances of a story called “Misomame” by three ranks of professional performers.
  - “Misomame” is a short rakugo story (duration: 2–4 minutes)
  - **Wording and expression is different from performers** because rakugo stories have no explicit scripts.

Recording (shin-uchi)

Speech samples (synthesis)

- Samples were synthesized through a Tacotron-based TTS system extended with self-attention (*SA-Tacotron-context* model from our previous study*) because this model was evaluated as the best one.



Text of a sentence

Manually-labeled context features for the sentence

Mel spectrogram

TTS
(SA-Tacotron-context)

WaveNet vocoder

Waveform of the sentence

- TTS model and WaveNet vocoder were trained with a rakugo speech database we built for the previous study.
  - Performer: the shin-uchi (1st-rank) above.

*Kato *et al*., “Modeling of Rakugo Speech and Its Limitations: Toward Speech Synthesis That Entertains Audiences,” *IEEE Access*, 8, 138149–138161, Jul 2020.

Training conditions

Data	**16 rakugo stories** (7,341 sentences, 4.31 hours). We didn’t used speech which duration < 0.5 seconds or > 20 seconds.
Sampling rate / bit depth / channel	48kHz / 16bit / mono
Training set	6,362 sentences (3.67 hours)
Validation set	706 sentences (0.42 hours)
Test set	273 sentences (0.22 hours)
Acoustic features	80-th mel spectrogram
Vocoder	WaveNet vocoder Input: mel spectrogram Output: **24kHz** / 16bit mono waveform

Test conditions

- Speech samples of “Misomame” performed by the three professional performers or TTS were used for the listening test.
  - Speech samples of TTS were synthesized **sentence by sentence**. Durations of pauses between sentences are the same as those of natural recording.
  - Non-speech sounds like mastication sounds were not modeled. Natural samples were used for such sounds.
- 292 listeners evaluated one of the speech samples of the whole story.

Test conditions

- Listeners answered five 5-scale MOS-based questions:

  1. Naturalness
  2. How accurately did you think you could distinguish each character?
  3. How well did you think you could understand the content?
  4. How well were you entertained?
  5. How high was the rakugo skill level of the performer?

TTS

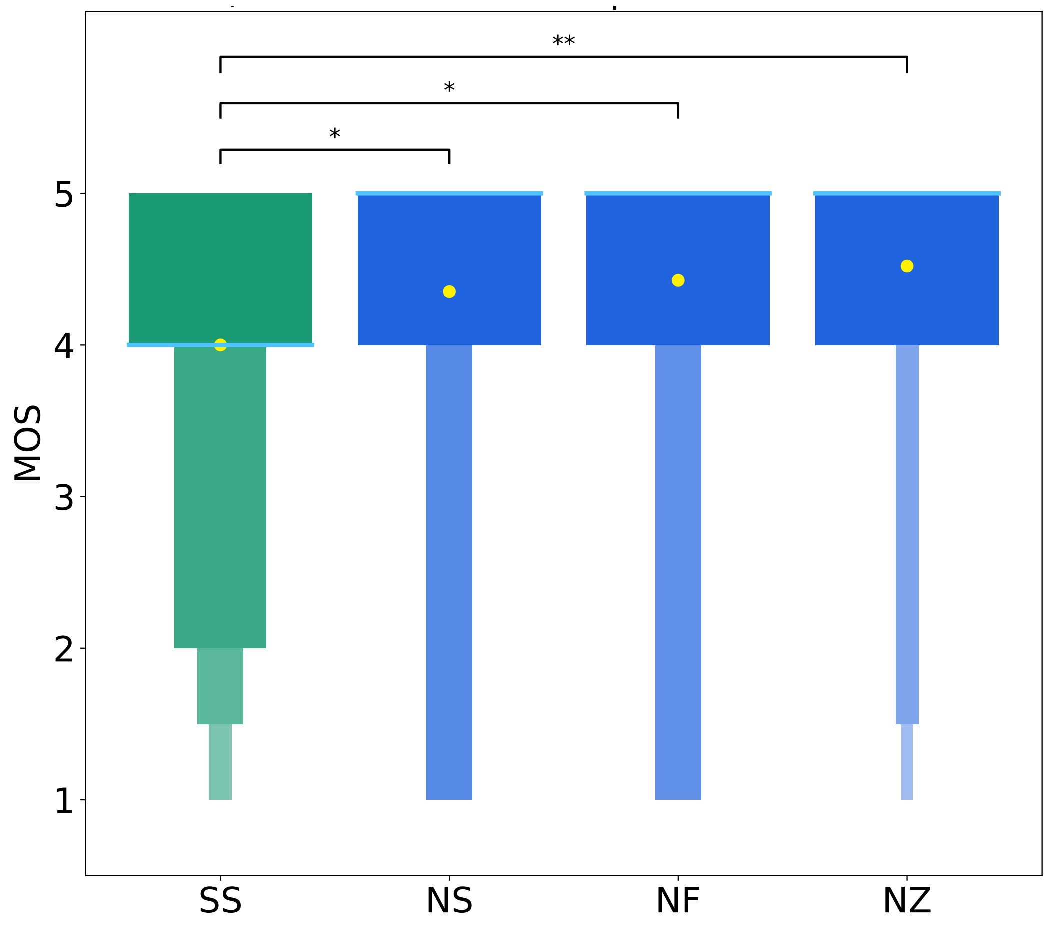
Shin-uchi
(1st-rank)

Futatsume
(2nd-rank)

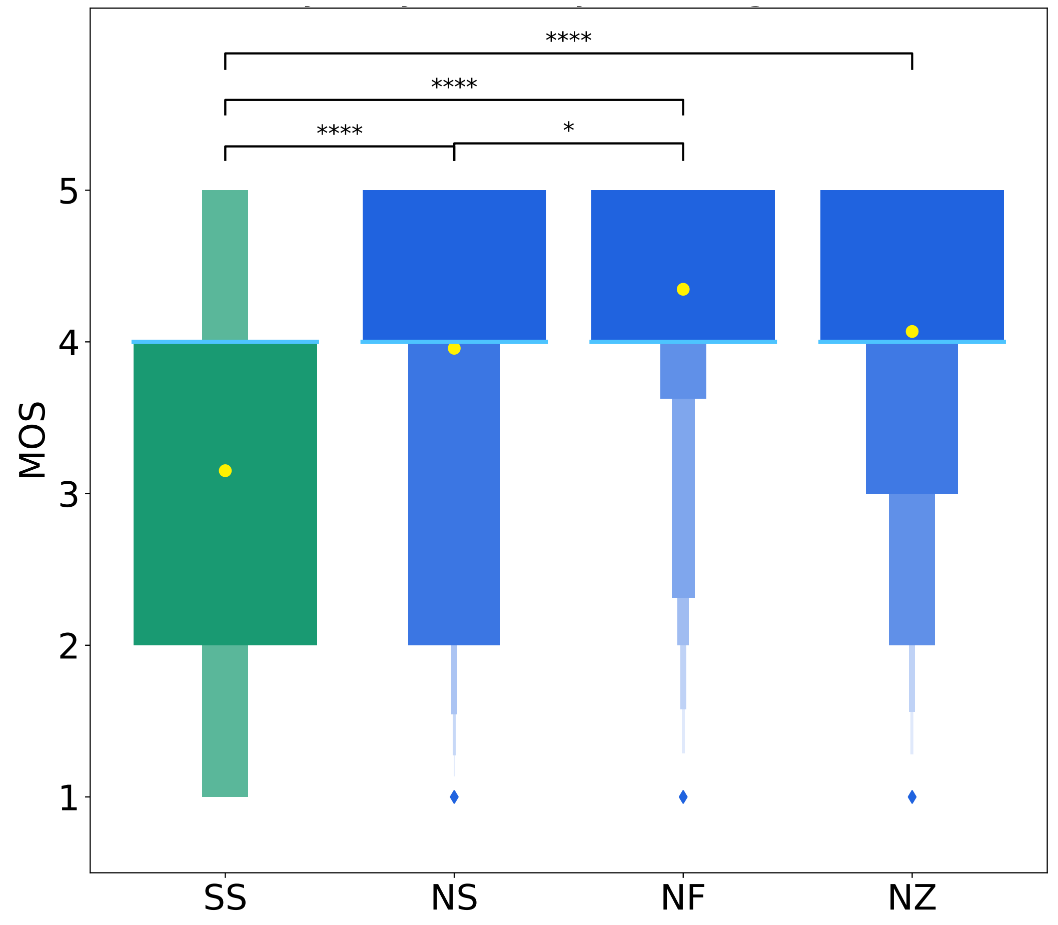
Zenza
(minor)

Result

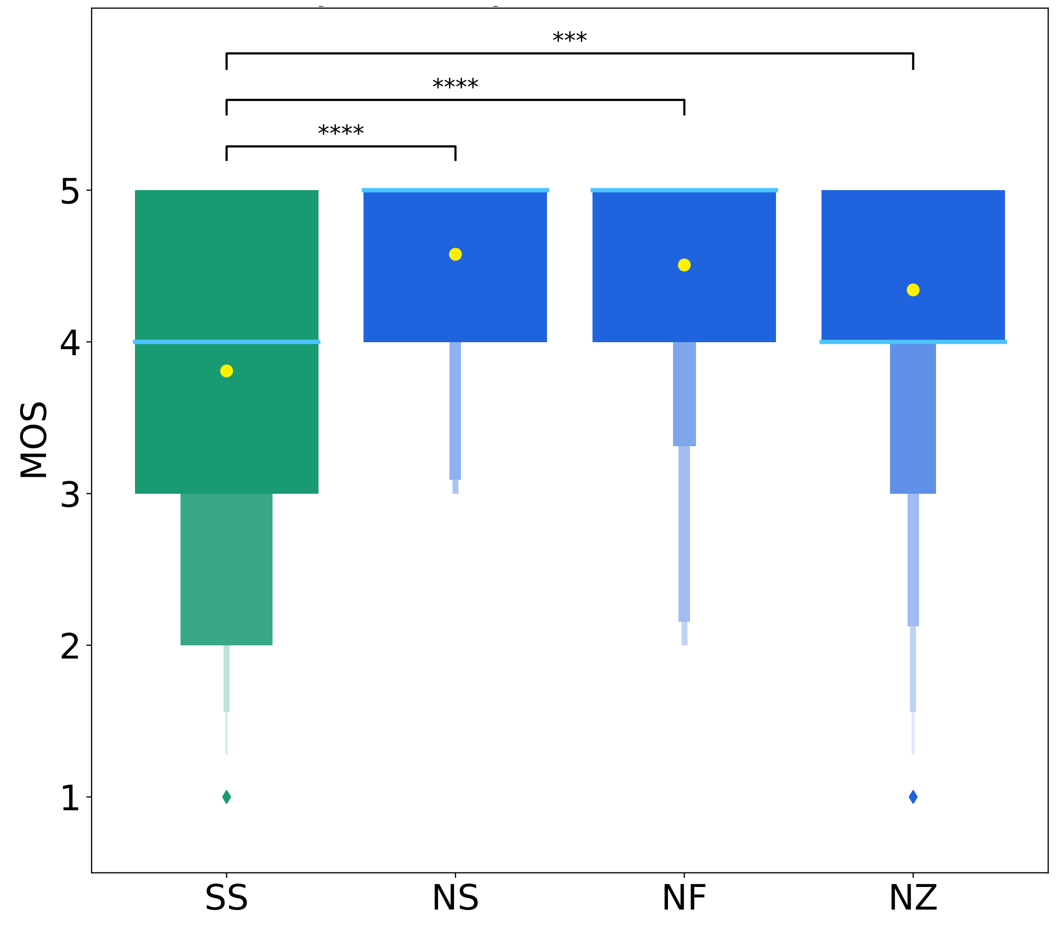
Q1: Naturalness



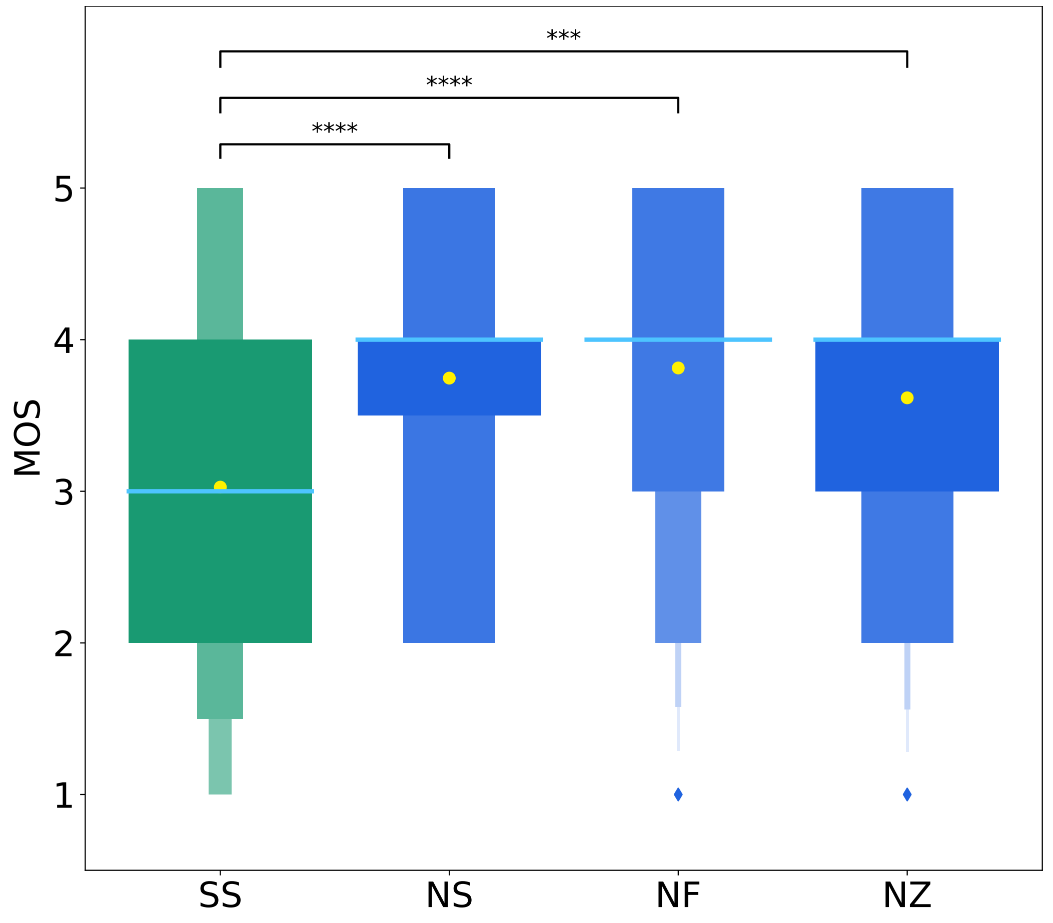
Q2: Distinguishability of characters



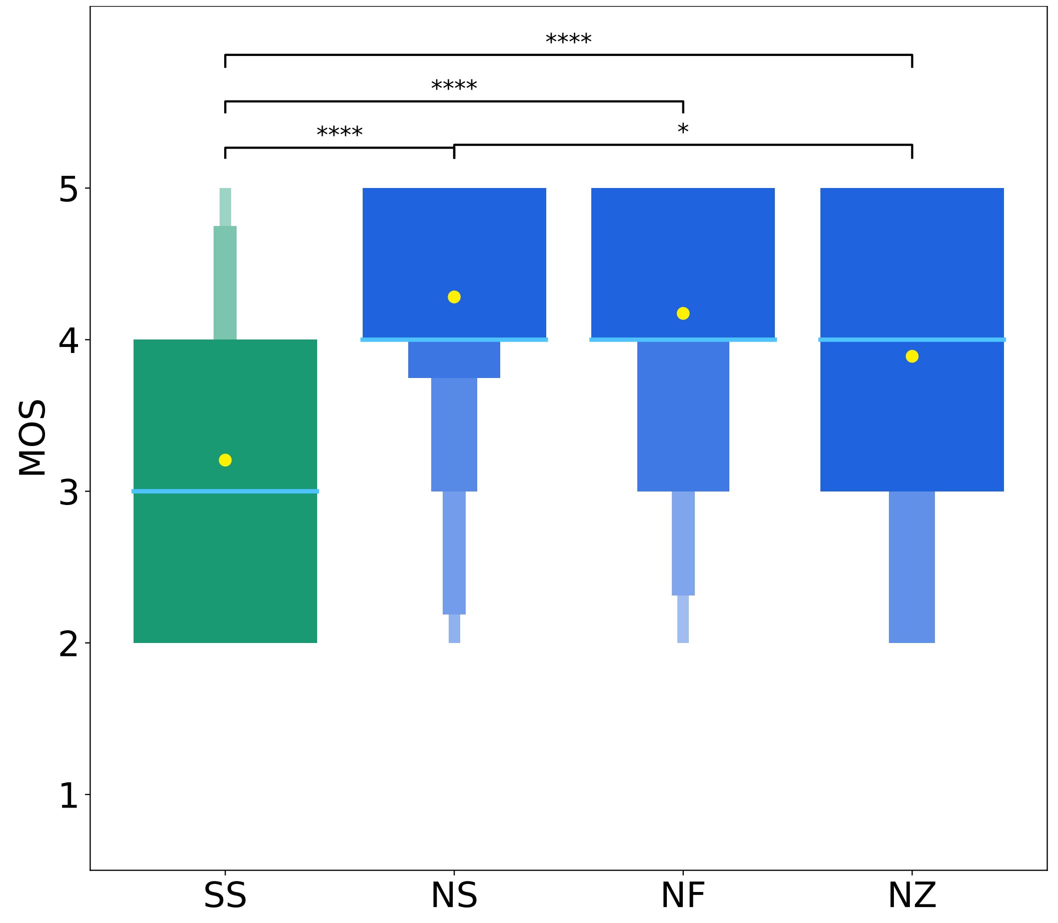
Q3: Understandability of contents



Q4: Degree of entertainment



Q5: Skill level



Legend

SS: Speech synthesis (TTS)

NS: Shin-uchi (1st-rank)

NF: Futatsume (2nd-rank)

NZ: Zenza (minor)

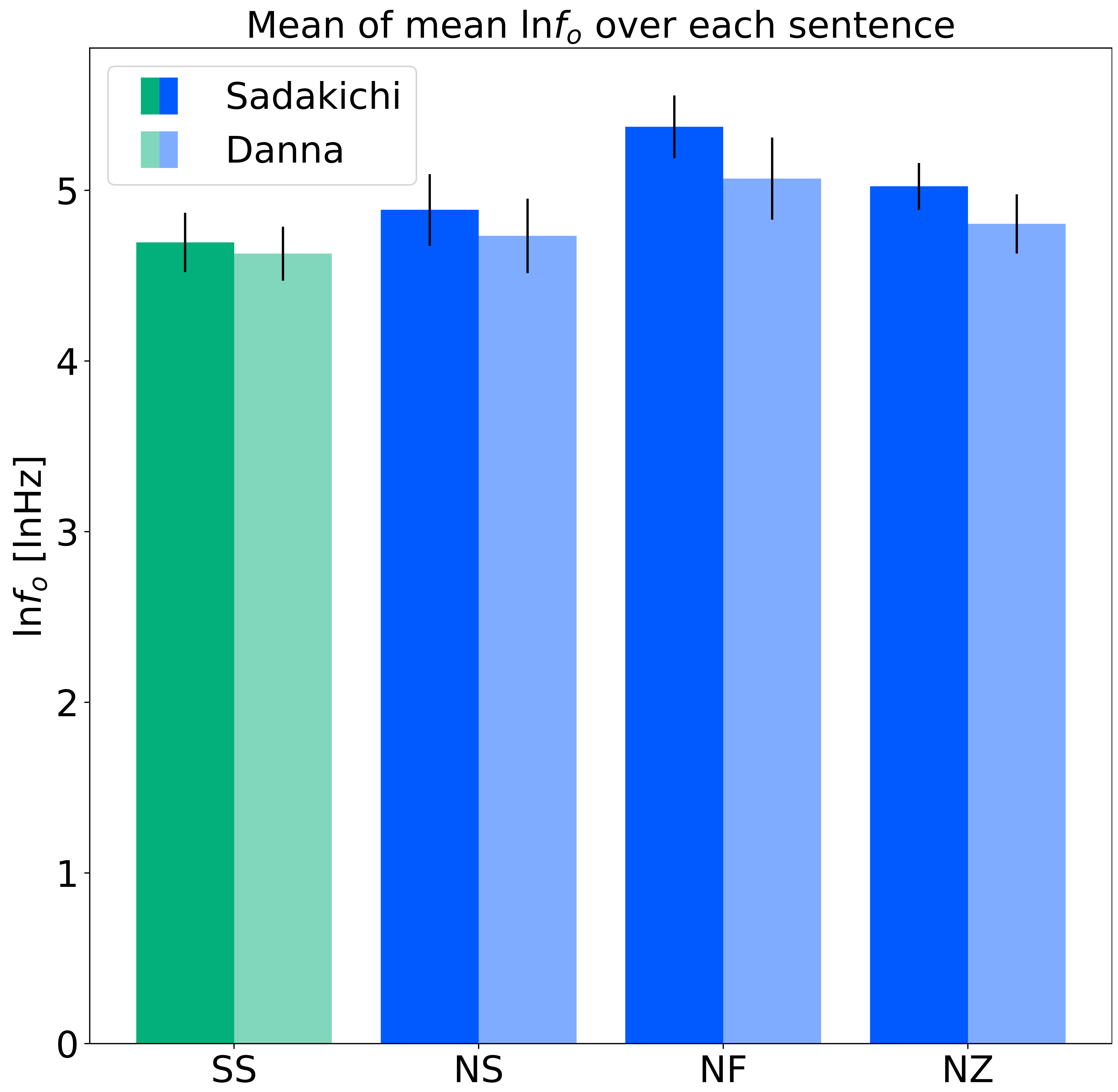
*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.005$, ****: $p < 0.001$

Correlations between scores of questions

	Q2 (character)	Q3 (content)	Q4 (entertaining)	Q5 (skill level)
Q1 (naturalness)	0.287	0.303	0.317	0.339
Q2 (distinguishability of characters)	-	0.538	**0.486**	0.580
Q3 (understandability of contents)	-	-	**0.597**	0.582
Q4 (degree of entertainment)	-	-	-	0.656

Degree of entertainment correlates stronger with distinguishability of characters and understandability of contents than naturalness.

Differences in f_os between performers/characters



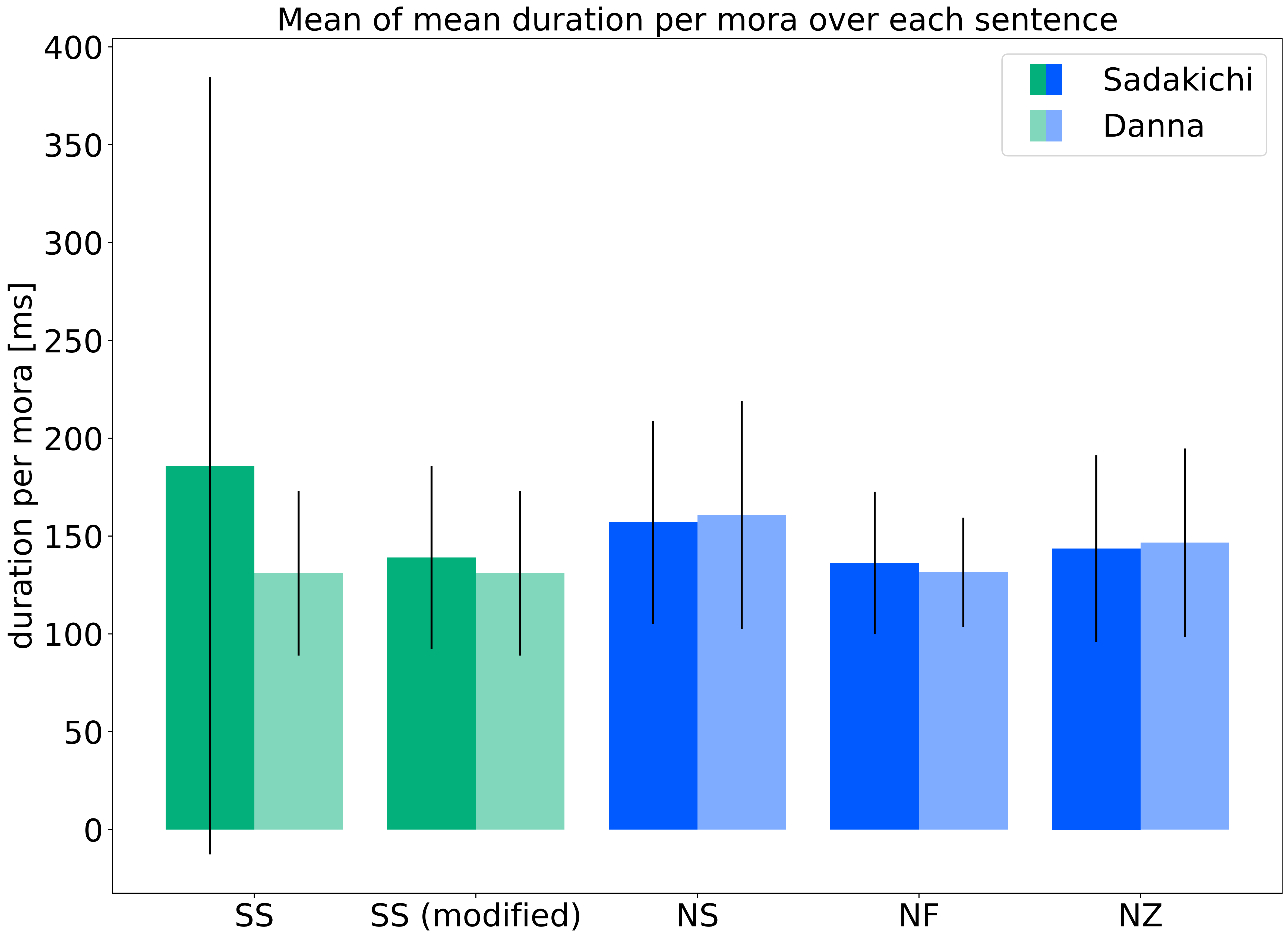
TTS | **Shin-uchi** (1st-rank) | **Futatsume** (2nd-rank) | **Zenza** (minor)

Sadakichi: boy
Danna: middle-aged man

TTS had a smaller difference of means of f_os between characters than professional performers**.**

TTS and zenza had smaller standard deviations of f_os for both characters than other performers.

Differences in speaking rates between performers/characters



TTS

Shin-uchi (1st-rank)

Futatsume (2nd-rank)

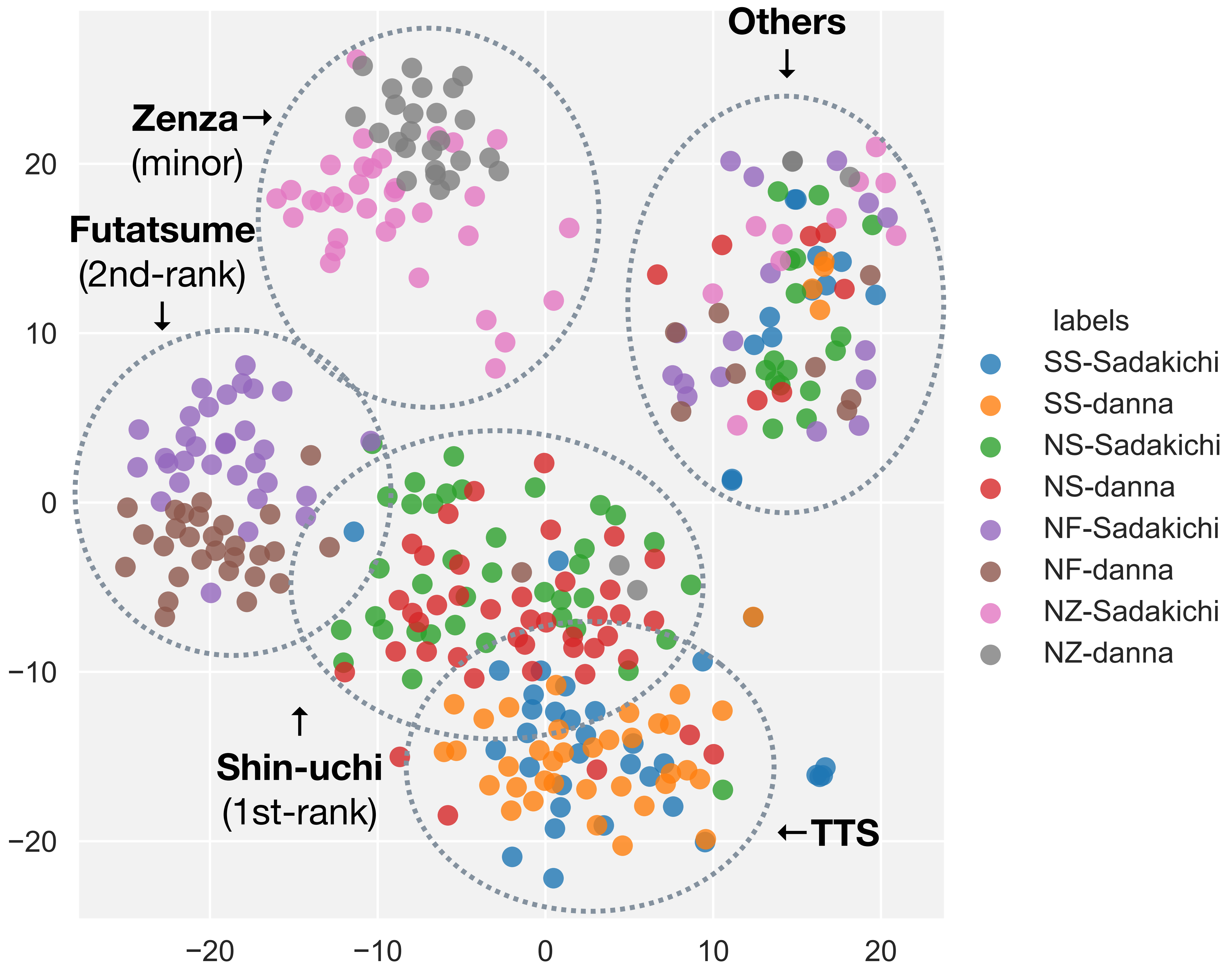
Zenza (minor)

TTS (excluding sentences with duration estimation errors)

Sadakichi: boy
Danna: middle-aged man

There were no better conditions in terms of speaking rates.

Visualization of x-vector for each sentence



Although we could find two sub-clusters corresponding to characters for zenza and futatsume, **we could not for TTS and shin-uchi.**

Shin-uchi performer may differentiate characters **using features that cannot be captured by x-vectors** (such as local features in time dimension).

(cf. The TTS model used in the listening test receives **global** features: context features labeled manually sentence by sentence))

Conclusions

Conclusions

- To investigate how high the level of rakugo TTS, we compared synthesized speech with natural speech performed by professional performers of three different ranks.

- There were significant differences between the evaluation for the current rakugo TTS and those for the professional performers.

- However, we obtained **valuable suggestions for further improvement of TTS**.

1. To more entertain audiences, we should not only improve naturalness **but also focus on the distinguishability of characters and the understandability of contents and improve them**.

2. Current rakugo TTS can be improved in terms of the **distinguishability of characters using f_os**.

3. To more differentiate characters, **we may need to model features that cannot be captured by x-vectors** (such as local features in time dimension).

Future work

1. **Designing an TTS architecture to better distinguish characters.**

- However, the frequency of the properties of the characters (gender, age, social rank, etc.) in common rakugo stories is **very unbalanced**.

2. **Working on other issues to be solved, such as estimating the durations of pauses between sentences and visual synthesis.**

- Rakugo is essentially a form of audio-visual entertainment.