Rakugo speech synthesis using segment-to-segment neural transduction and style tokens — toward speech synthesis for entertaining audiences

Shuhei KATO[1,2], Yusuke YASUDA[1,2], Xin WANG[2], Erica COOPER[2], Shinji TAKAKI[3], Junichi YAMAGISHI[1,2,4]

[1]SOKENDAI (The Graduate University for Advanced Sciences), Japan

[2]National Institute of Informatics, Japan

[3]Nagoya Institute of Technology, Japan

[4]The University of Edinburgh, UK

Can TTS entertain the audience?

Can TTS entertain the audience?

Speech as media

- Speech transfers information to listeners.
    - Contents, emotions, personality, intention, …
- Text-to-speech (TTS) research has mainly aimed to improve TTS to play this role well.
- **Some TTS systems can already produce speech as natural as human speech.**

Speech can stir listeners' emotions

- Verbal entertainment, including *rakugo*, can entertain audiences through the medium of speech.
- **Can the current TTS perform as well as a professional does?** **No**
- How about end-to-end (seq-to-seq) TTS?

Rakugo: A traditional Japanese form of verbal entertainment

- Like **one-person stand-up comedy + comic storytelling.**
- History: 300+ years.
- Performs **improvisationally or from memory** alone on a stage.
- **Plays multiple characters**, and their **conversations make the story progress**.

Shumputei Shotaro, who is a professional rakugo performer, is performing rakugo.

This photo is transformed from “DP3M2471” by akira kawamura licensed under CC BY 2.0.

Rakugo is popular even now

- About **600** professional performers (*hanashika*) are active in Tokyo.
- In Tokyo, four major *yoses* exist.
    - Yose is a theater that mainly performs rakugo **every day**.
- Some TV and radio programs are broadcasted every week.
- Thousands of CDs and DVDs.

Rakugo performance

- Performer sits on a *zabuton* (cushion) **alone** on a stage.
- Uses no properties other than a *sensu* (folding fan) and a *tenugui* (hand towel).
- Almost **no narrative sentences exist** in the main part of a rakugo story.

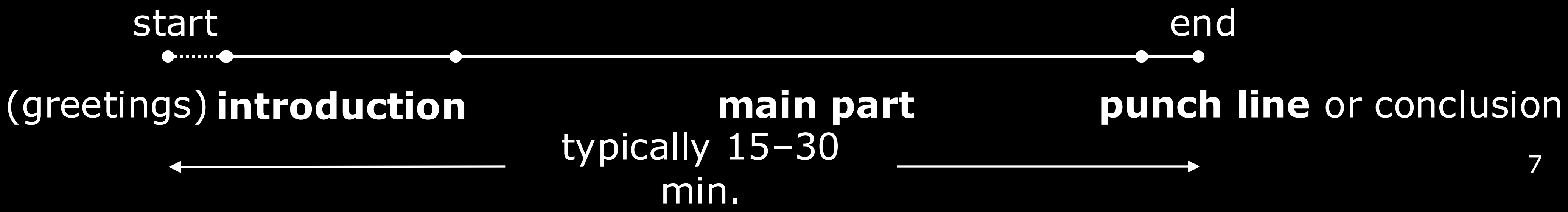
Zabuton

This photo is "Zabuton.jpg" by Lusheeta licenseced under CC BY 3.0.

Sensu and tenugui

Structure of a rakugo story

- A rakugo story has five parts: *maeoki* (greeting), ***makura*** (introduction), **main part**, ***ochi*** (punch line), and *musubi* (conclusion).
- Makura is often **improvised**, but performers basically don’t have conversations with audiences unlike stand-up comedy.
- Ochi (punch line) is most important part of rakugo.
    - The word “rakugo”（落語）is derived from “a story with ochi（落ち）. ”



start

end

(greetings) **introduction**

main part

punch line or conclusion

typically 15–30 min.

Dialects used in traditional rakugo stories

- Rakugo stories are generally divided into **standards** (established –1920s) and modern stories (created 1930s–).
- Japanese language used in standards are **slightly old-fashioned**.
  - Automatic analysis/tagging are practically **impossible**.
- Characters appearing in standards speak different Japanese dialects, sociolects, or idiolects according to their genders, ages, or social ranks.

Example of a rakugo paragraph

Tomi	Whoa! Oh no! Oh no! Oh no! Oh no!
Friend	Wait Tomi. What are you doing?
Tomi	Oh, I’m chasing after a thief.
Friend	Seriously? Aren’t you the fastest man in this town? He is unlucky.
Friend	Which direction did he escape?
Tomi	He’s catching up with me.

Rakugo TTS

Rakugo TTS vs. audiobook TTS

Differences between audiobook and rakugo:

- Main part of a rakugo story **consists of conversations by characters**.
- Rakugo speech is more casually pronounced because it is produced improvisationally or from memory.
- **Rakugo is inherently an entertainment**.
  - Rakugo TTS **has to** entertain the audience.

NII rakugo speech database

- Commercial rakugo recordings are not suitable for TTS modeling.
  - Most of them are live recordings including noise and reverberation.
- We recorded rakugo speech ourselves.
- **Performer**: Yanagiya Sanza (20+ years of professional career).
- **Content**: 25 standards (13.2 hours).

Recording conditions

- Recording was conducted in a recording booth.
- No audiences or reactions from ones.
- Didn’t retook on account of mispronunciation or restatements except in cases where the performer asked us to do so.

Transcription

- The first author transcribed pronunciation of the recorded speech.
- **No special symbols** for mispronunciation, fillers, or laughs.
- **Didn’t use accent symbols**.
  - Automatic estimation: Impossible
  - Manual labeling: Very time consuming!

Phonemes

- a, b, by, ch, cl, d, dy, e, f, fy, g, gw, gy, h, hy, i, j, k, kw, ky, m, my, n, N, ny, o, p, py, r, ry, s, sh, t, ts, ty, u, v, w, y, z

Pauses

- pau (,), sil (start/end of the sentence), qsil (interrogative ending)

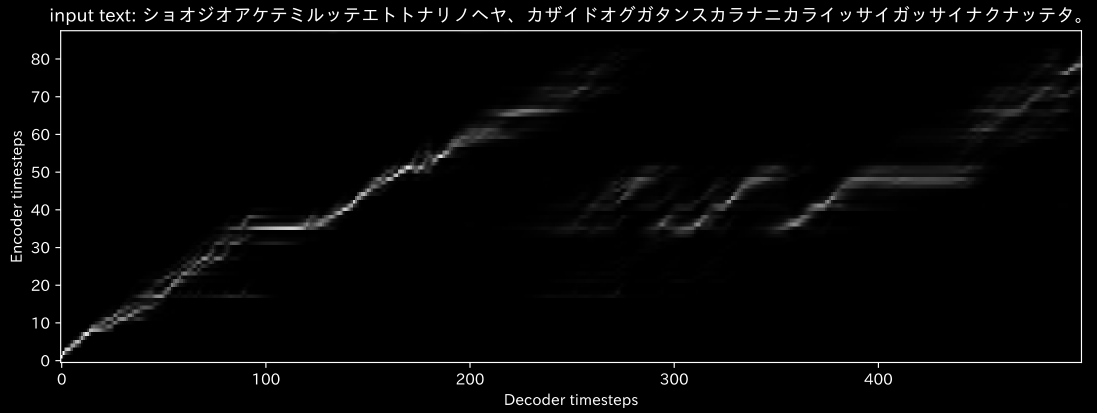
Context labels

Group	Name	Description
ATTRibution	**Role** of character	**Gender**: Hanashika*, male, female **Age**: Hanashika, child, young, middle-aged, old **Rank**: Hanashika, *samurai* (soldier), artisan, merchant, other townsperson, countryperson, with other dialect, modern, other
	Individuality of character	Hanashika, fool
CONDition	**Condition** of character	neutral, admiring, admonishing, affected, angry, begging, buttering up, cheerful, complaining, confident, confused, convinced, crying, depressed, drinking, drunk, eating, encouraging, excited, fearing, feeling sketchy, feeling sick, feeling sleep, feeling sorry, feeling suspicious, find it easier than expected, freezing, frustrated, ghostly, happy, hesitating, interested, justifying, *kakegoe*, loud voice, laughing, leaning on, lecturing, looking down, panicked, pet directed speech, playing dumb, putting up with, rebellious, refusing, sad, seducing, shocked, shouting, small voice, soothing, straining, surprised, swaggering, teasing, telling off, tired, trying to remember, underestimating, unpleasant
SITuation	**Relationship** of the companions to talk with	Hanashika, narrative, soliloquy, superior, inferior
	N_companion (number of companions to talk with)	Hanashika, narrative, soliloquy, one, two or more
	Distance to the companions to talk with	Hanashika, narrative, near, middle, far
STRucture	**Part** of the story	Makura (including maeoki), main part, ochi (including musubi)

* “Hanashika” refers to improvised or narrative speech in makura.

Using end-to-end (seq-to-seq) TTS

- We have used end-to-end (seq-to-seq) TTS for modeling rakugo speech because **we can use only phonemes as input features**.
- Tacotron didn't learn alignments well for rakugo speech in (Kato *et al.*, Mar 2019).
- We thought that over-flexibility of soft attention mechanism cannot deal with diversity of rakugo speech.
  - Soft attention can assign any encoder time steps to any decoder ones, though alignments of speech must be left-to-right.
  - Soft attention is used in all the encoder-decoder TTS other than our new one.

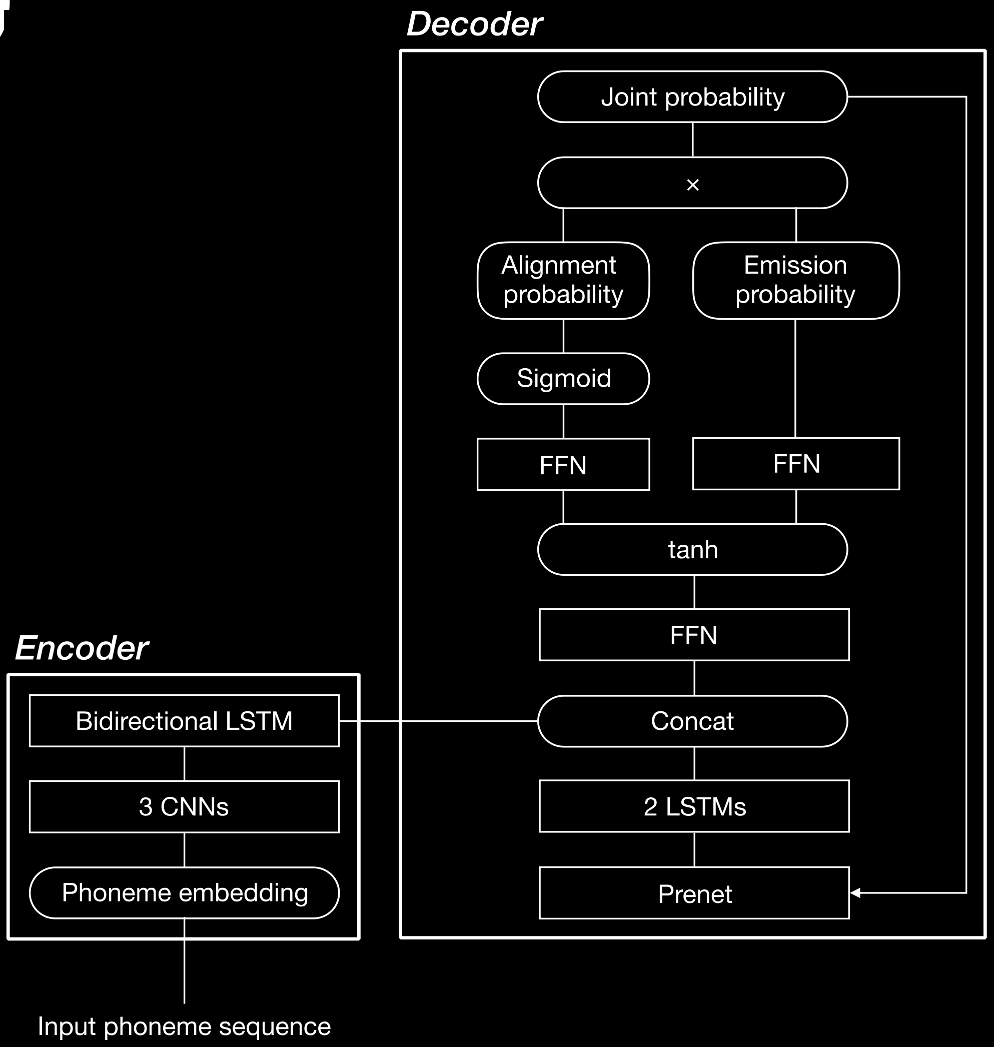


SSNT-based TTS (Oral session 6, 3rd day by Yasuda *et al*.)

- An encoder-decoder TTS that has **no attention network**.

Restrictions for alignment:

1. Alignment increases monotonically
2. One encoder step is assigned to one decoder step (hard alignment)



Transition of SSNT alignment

- Only two transition is allowed: **emit** or **shift**.
  - Alignment always monotonically increases.

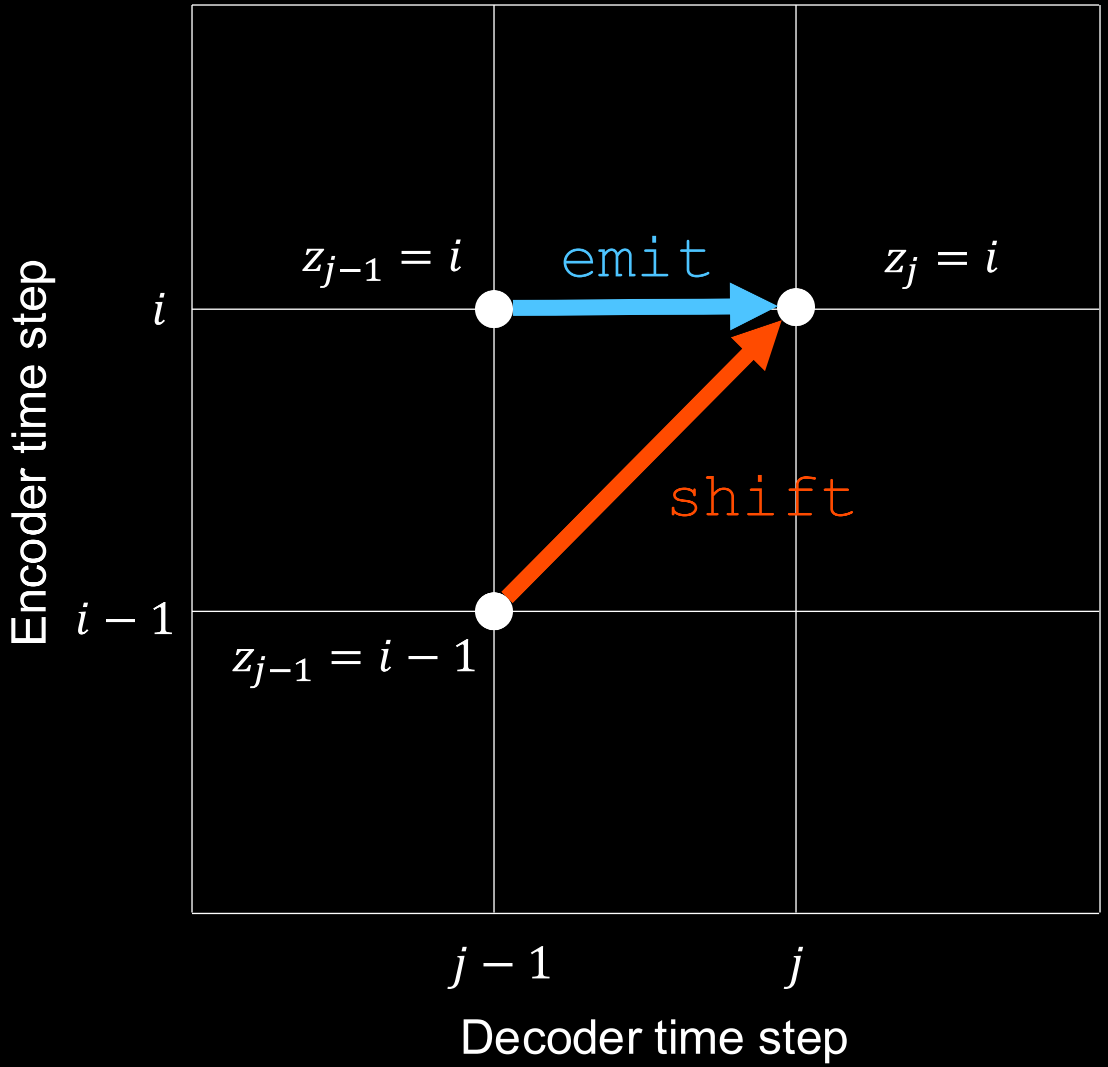
$$p\left(z_j = i \middle| z_{j-1}, \boldsymbol{y}_{1:j-1}, \boldsymbol{x}_{1:I}\right) = \begin{cases} 0 \text{ where } z_{j-1} > i \cup z_{j-1} < i-1 \\ p\left(a_{i,j} = \texttt{emit}\right) \text{ where } z_{j-1} = i \\ p\left(a_{i,j} = \texttt{shift}\right) \text{ where } z_{j-1} = i-1 \end{cases}$$

$\boldsymbol{x}$: input phoneme sequence

$\boldsymbol{y}$: output acoustic feature sequence

$\boldsymbol{z}$: alignment

For more details, please check our presentation tomorrow!

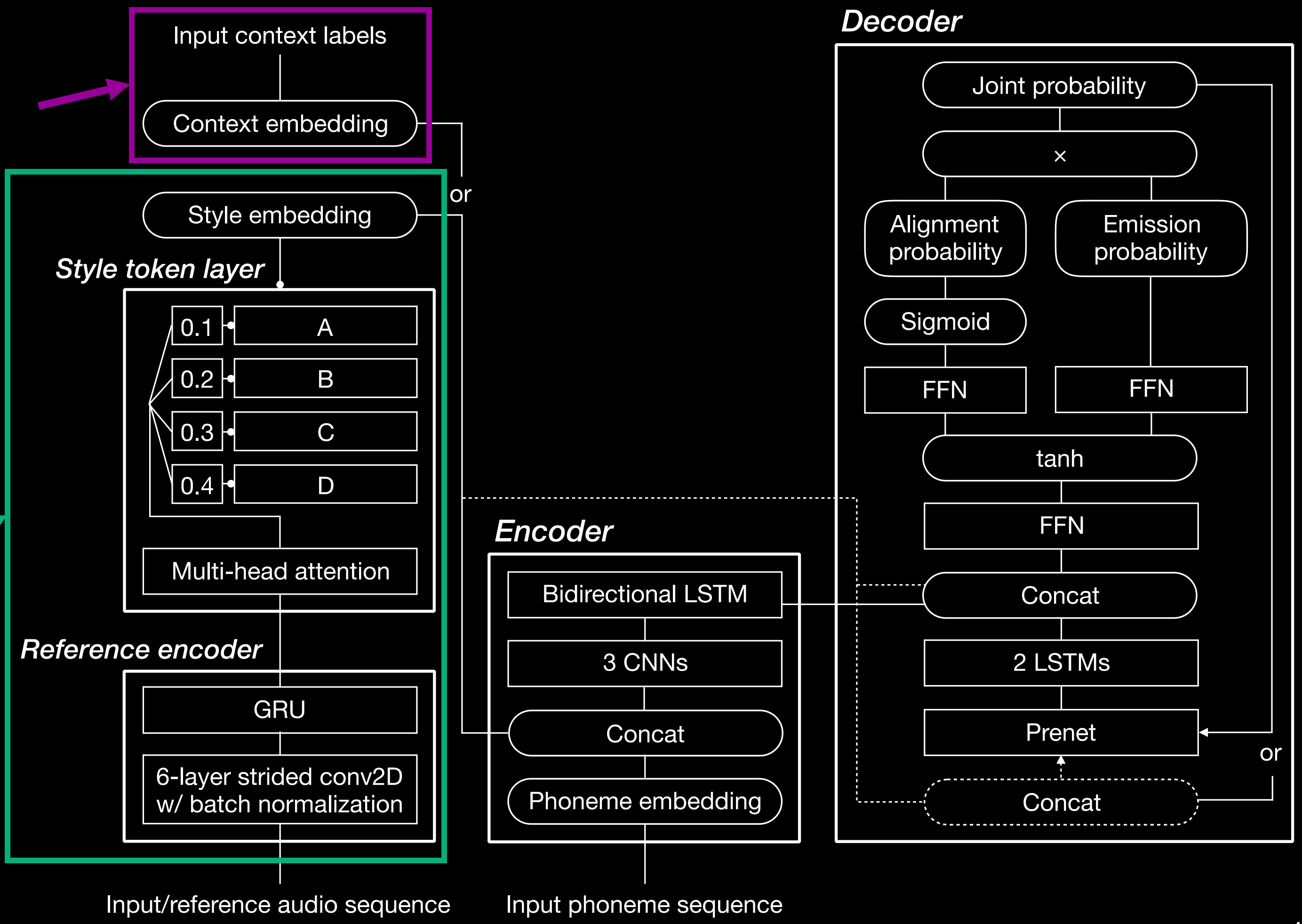


Modeling roles of characters and speaking styles of rakugo speech

Manually labeled context features
Context label for each sentence

or

Global style tokens (GST)
(Wang *et al*., 2018)
Estimates style embedding from reference audio



Experimental conditions

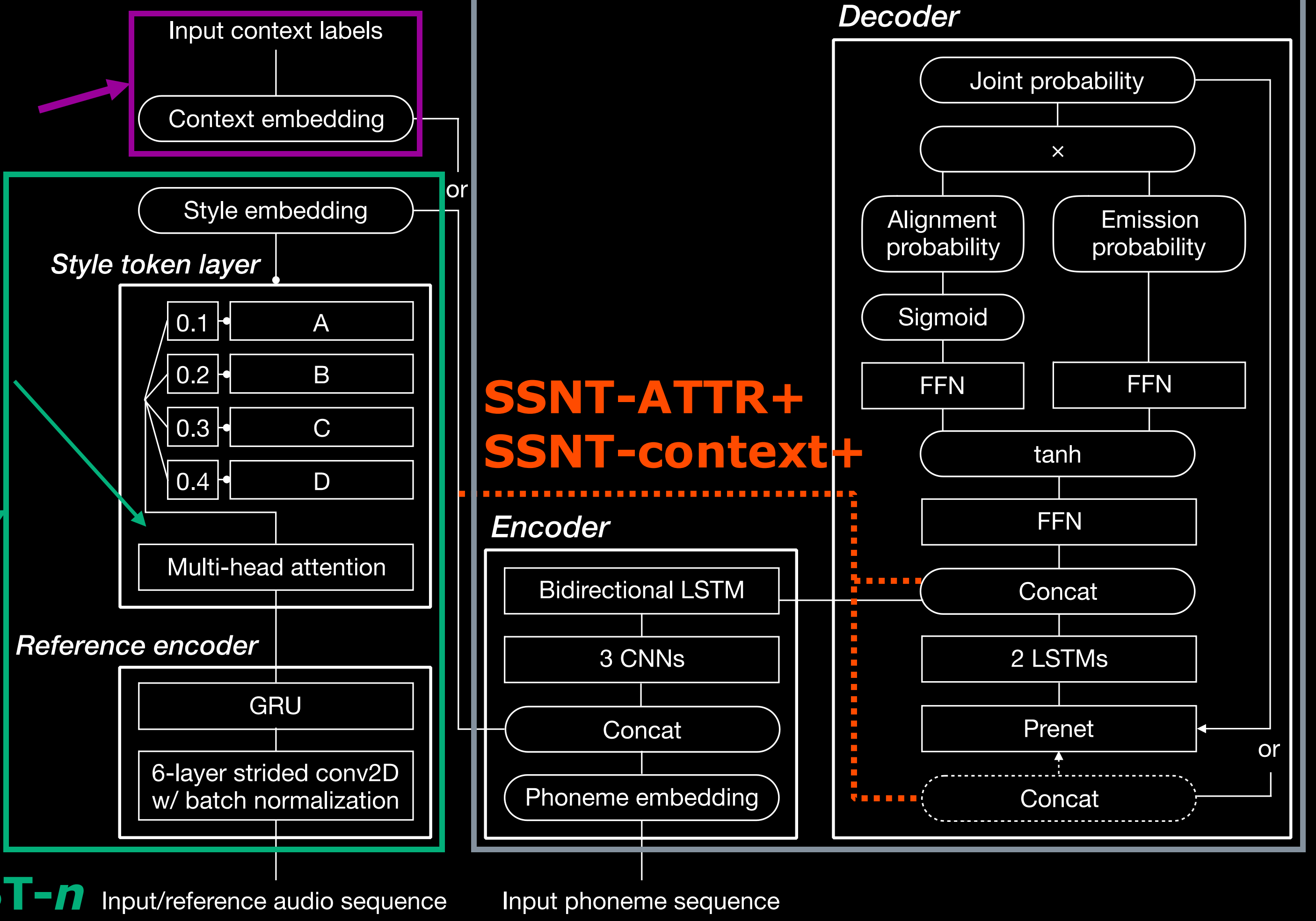
Data	**16 stories from NII rakugo speech DB (4.3 hours not including pauses between sentences, 7,337 sentences)**. Sentences which duration are < 0.5 s or ≥ 20 s were removed.
Sampling rate / bit / channels	48 kHz / 16 bit / mono
Training set	6,459 sentences
Validation set	717 sentences
Test set	161 sentences
Acoustic features	80-d mel spectrogram which were normalized to 0 mean and 1 stddev over all the test, validation, and test sets.
Reduction factor	2
Vocoder	**WaveNet vocoder** which was trained by all the test, validation, and test sets. **Input**: Mel spectrogram. **Output**: 16 kHz / 16 bit mono waveform
Number of style tokens	10

Systems

SSNT-ATTR (role only)
SSNT-context (all the contexts)

Manually labeled context features
Context label for each sentence

n = num of heads
4, 8, 16, 32, 64



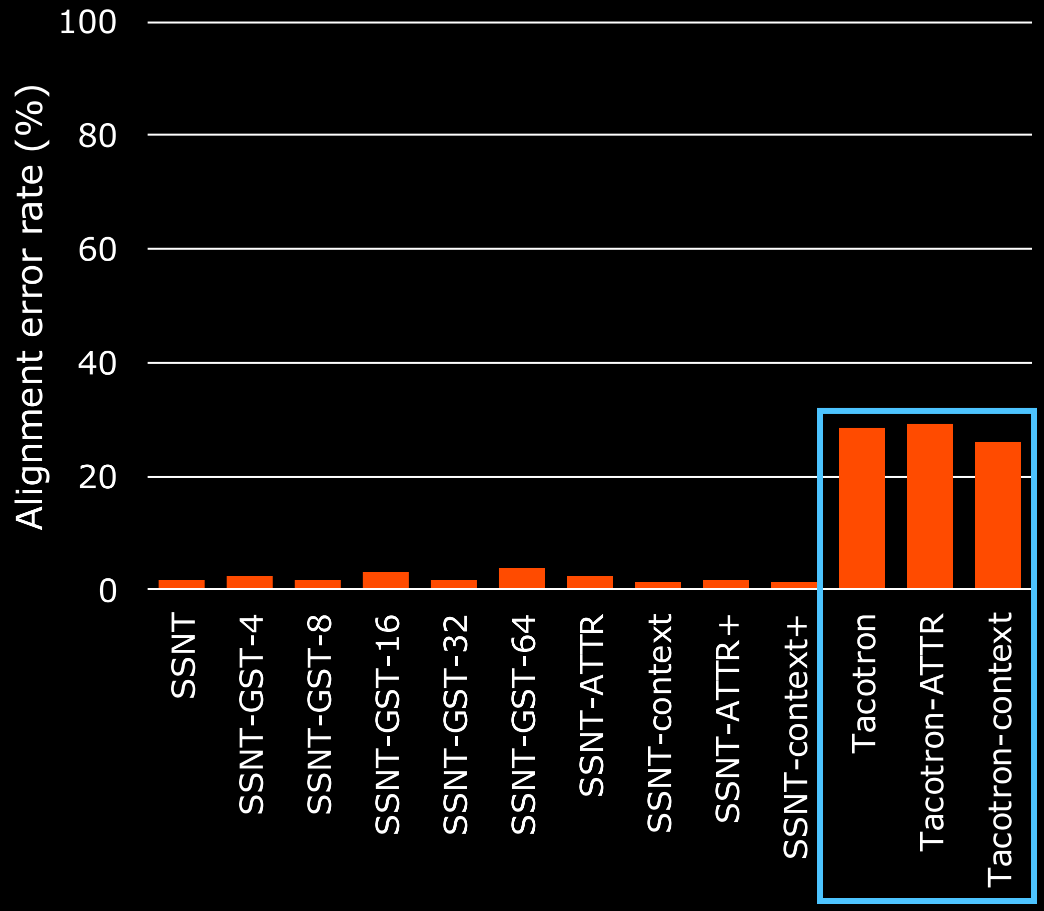
SSNT-ATTR+
SSNT-context+

SSNT

Global style tokens (GST)
(Wang *et al*., 2018)
Estimates style embedding from reference audio

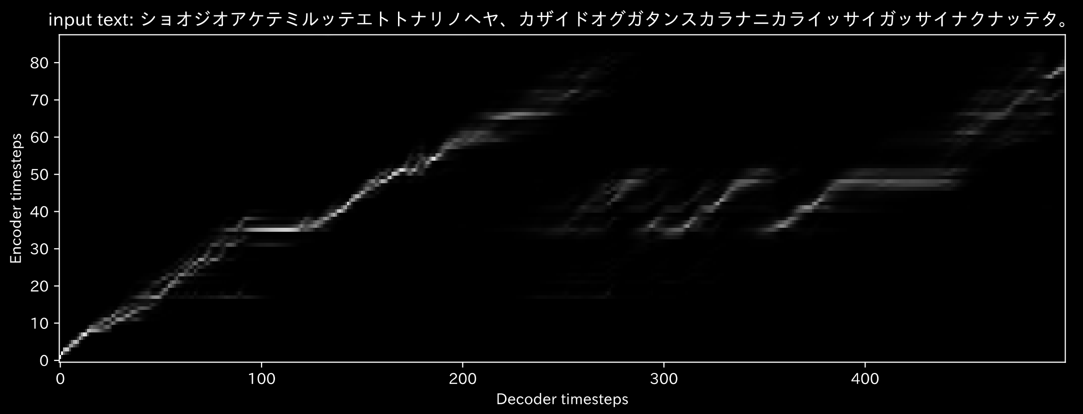
SSNT-GST-*n*
Reference audio = ground truth one

Result: Alignment error rates



Ratio of sentences containing obvious alignment errors:

- Skipping
- Incompleteness (did not consume all the inputs)



Example of alignment errors

Listening tests

- We prepared 12 short paragraphs (made of 161 test sentences) for the listening tests.
    - We combined speech synthesized sentence by sentence. The lengths of pauses were equal to ones of ground truth.
- Listeners evaluated speech **paragraph by paragraph**, **not** sentence by sentence.
- We didn’t use ground truth speech.
- 5-point scale MOS test was conducted.
    - Questions: 1) Naturalness, 2) how accurately you think you could distinguish each character, 3) how properly you think you could understand the content
    - In one evaluation round, listeners listened to the same short paragraph generated by different 13 systems.
    - 135 paid listeners evaluated 453 rounds.

Audio sample

🔊 Synthesized 🔊 Natural

Young man	Oh, look, look, look!
Young man	This crab … this crab looks strange. Crabs walk sideways, do they? It walks straight. What happened?
Performer	Then the crab raises its face and says:
Crab	Excuse me. I’m drunk now.

Audio sample

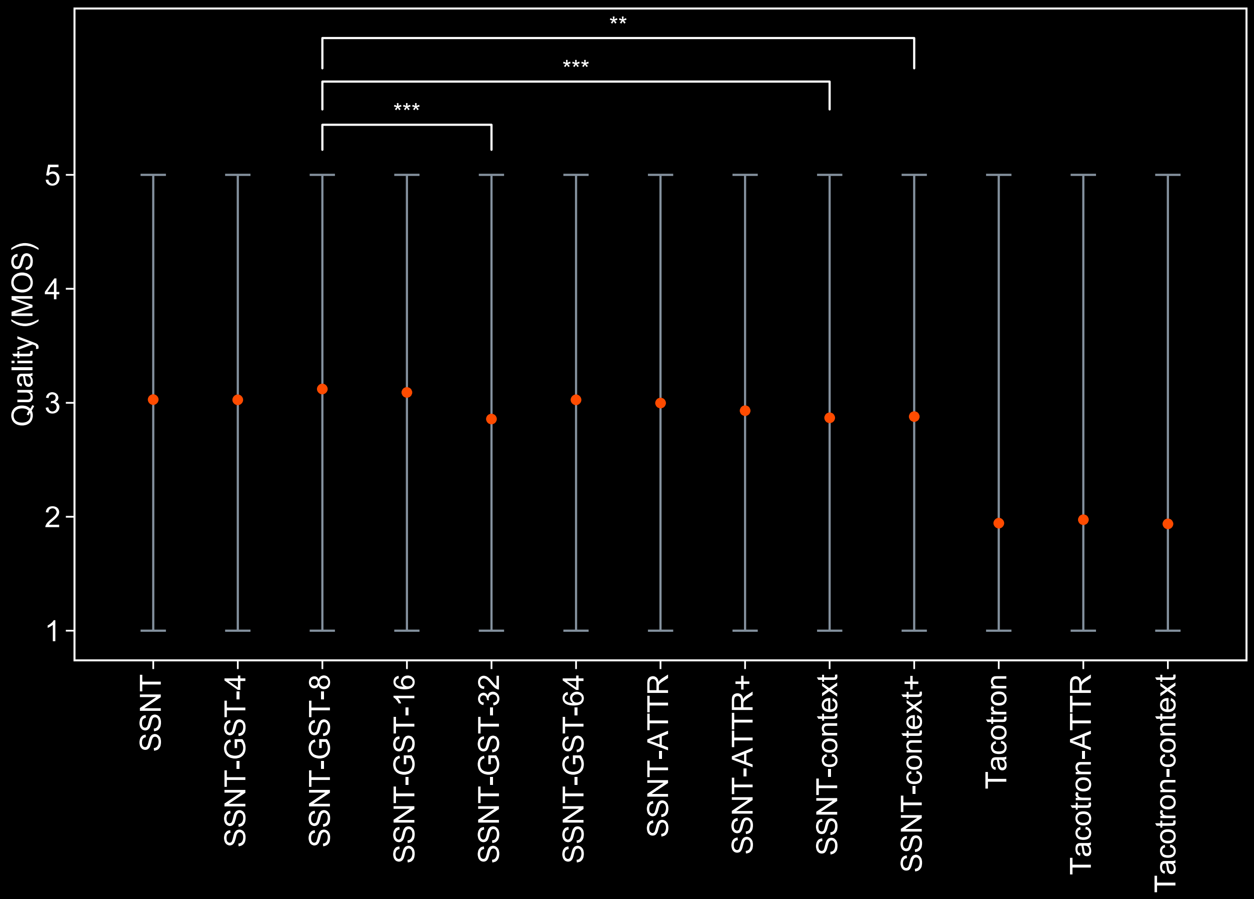
🔊 Synthesized 🔊 Natural

Young man	Oh, look, look, look!
Young man	This crab … this crab looks strange. Crabs walk sideways, do they? It walks straight. What happened?
Performer	Then the crab raises its face and says:
Crab	Excuse me. I’m drunk now.

Result of the listening tests 1) Naturalness

Very natural

Very unnatural

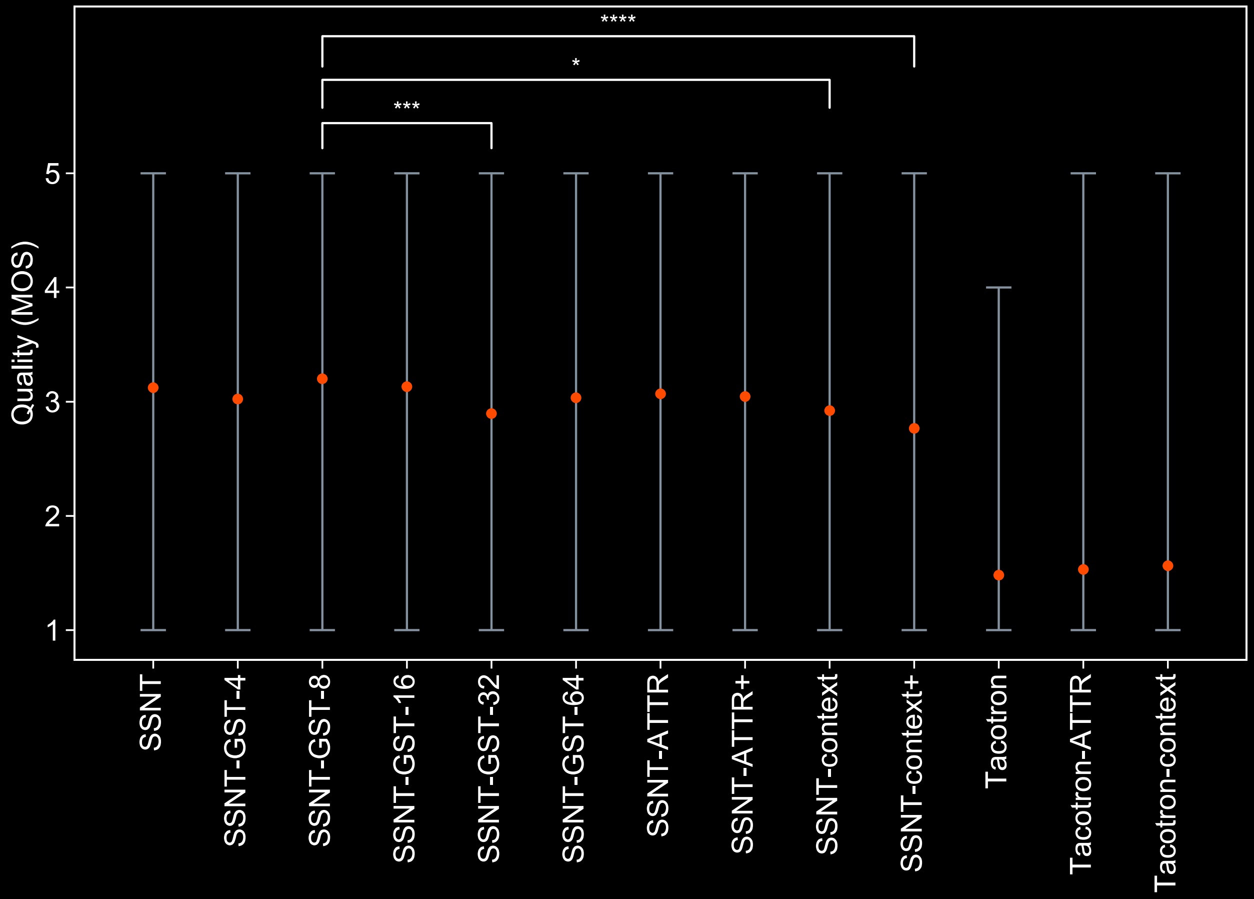


$*: p < 0.05, **: p < 0.01, ***: p < 0.005, ****: p < 0.001$

Result of the listening tests 2) Distinction of each character

Can distinguish very well

Cannot distinguish at all

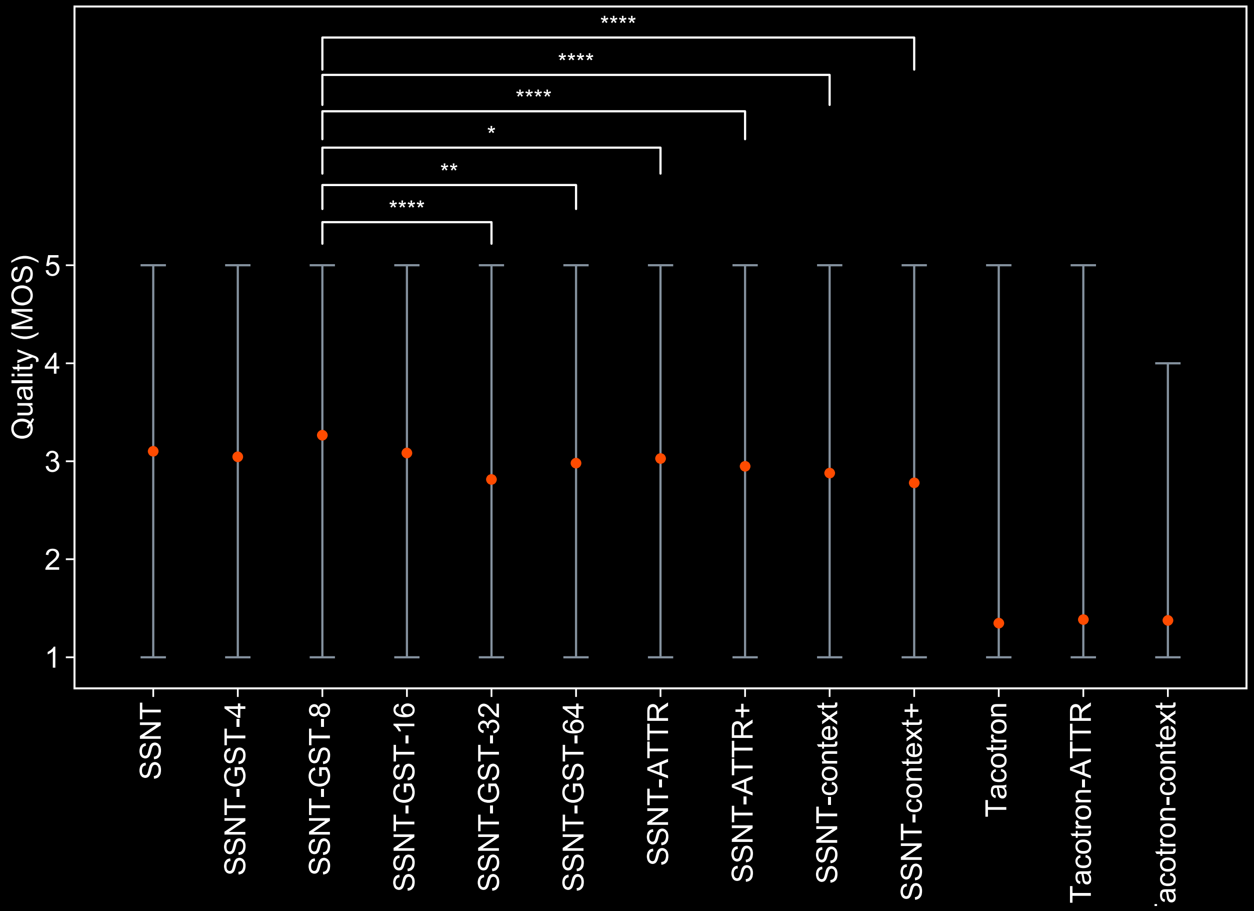


$*: p < 0.05$, $**: p < 0.01$, $***: p < 0.005$, $****: p < 0.001$

Result of the listening tests
3) Understanding the content

Can understand very well

Cannot understand at all



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

Discussions & conclusions

- This work is the first step of building TTS that entertains the audience.
- **SSNT-based TTS can synthesize rakugo speech** in which listeners can distinguish each character and understand the contents to a certain degree.
- Listening test results for naturalness, distinction of each character, and understanding the content were similar to each other. They seem to be closely correlated.
- Too many attention heads of GST and many context labels seem to cause overfitting.
- MOS for the best system was around 3, so **SSNT-based TTS should be improved more**.

Very important additional information

After SSW submissions, we have significantly refined our implementation of Tacotron, and it can now model alignment of rakugo speech much more accurately.

We are writing a journal paper!

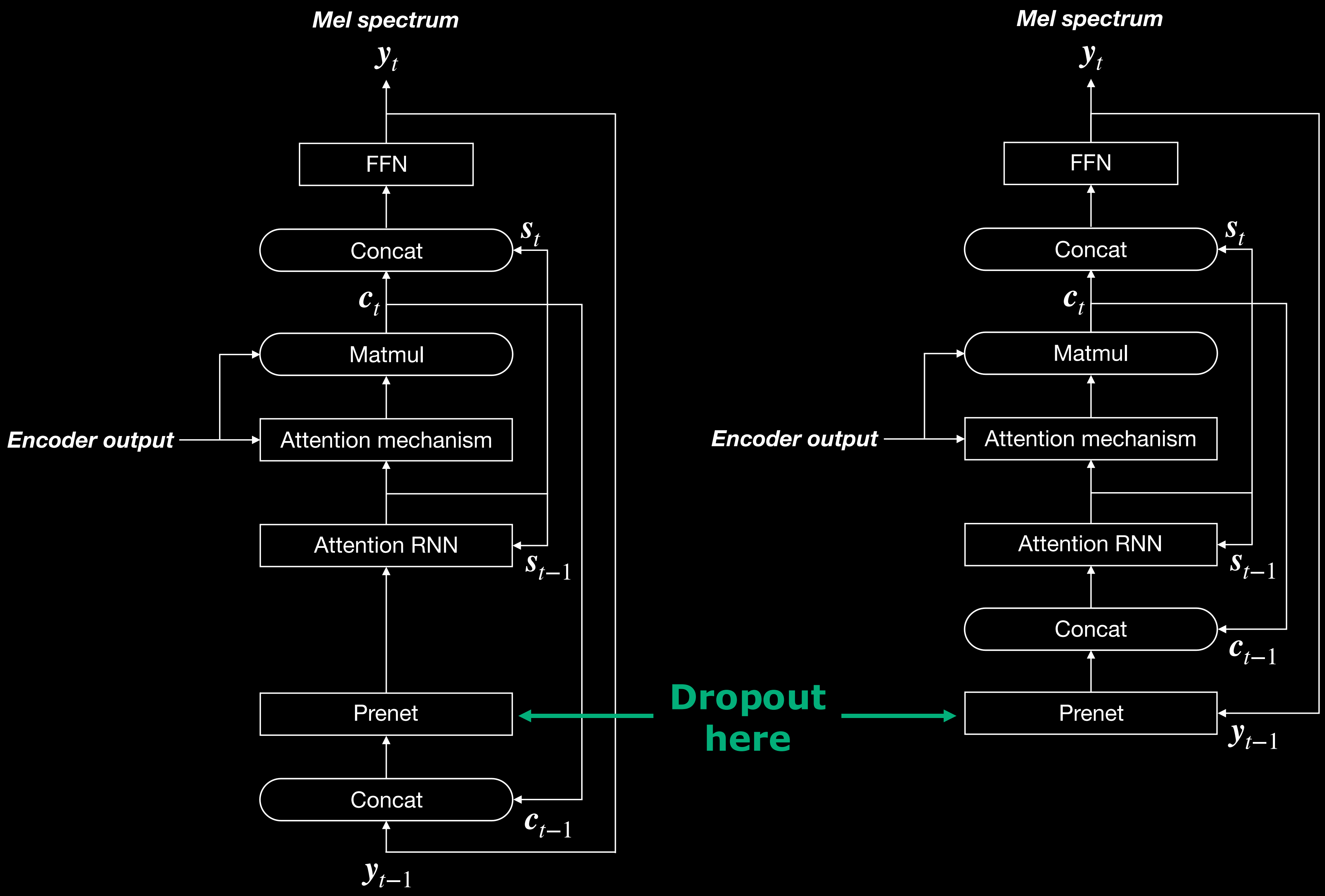
Visit our website: nii-yamagishilab.github.io ← Audio samples will be available here.

Follow us on Twitter: @yamagishilab 🐦

↑ **Our newest rakugo audio sample is available here!**

Acknowledgement

This work was partially supported by a JST CREST Grant (JPMJCR18A6, VoicePersonae project), Japan, and by MEXT KAKENHI Grants (16H06302, 17H04687, 18H04120, 18H04112, 18KT0051), Japan. The numerical calculations were carried out on the TSUBAME 3.0 supercomputer at the Tokyo Institute of Technology.



Mel spectrum

y_t

FFN

Concat

s_t

c_t

Matmul

Encoder output

Attention mechanism

Attention RNN

s_{t-1}

Prenet

Concat

c_{t-1}

y_{t-1}

Dropout here

Mel spectrum

y_t

FFN

Concat

s_t

c_t

Matmul

Encoder output

Attention mechanism

Attention RNN

s_{t-1}

Concat

c_{t-1}

Prenet

y_{t-1}

Wrong

Correct

New audio sample 🔊 🔊

Tacotron-based Natural

Young man Oh, look, look, look!

Young man This crab … this crab looks strange. Crabs walk sideways, do they? It walks straight. What happened?

Performer Then the crab raises its face and says:

Crab Excuse me. I’m drunk now.