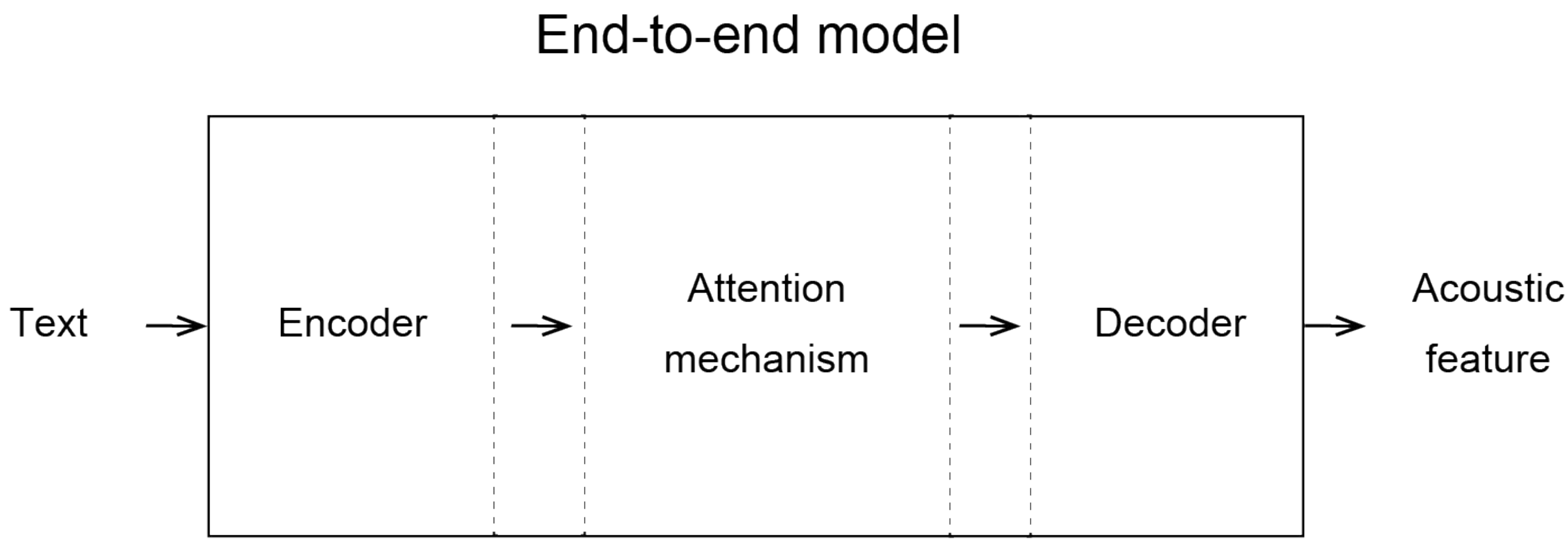
Initial investigation of an encoder-decoder end-to-end TTS framework using marginalization of monotonic hard latent alignments

Yusuke Yasuda, Xin Wang, Junichi Yamagishi
NII, Japan

14:00, 9/22, SSW10, 2019
Oral Session 6: Sequence to sequence model

End-to-end text-to-speech synthesis



Proposed end-to-end TTS methods

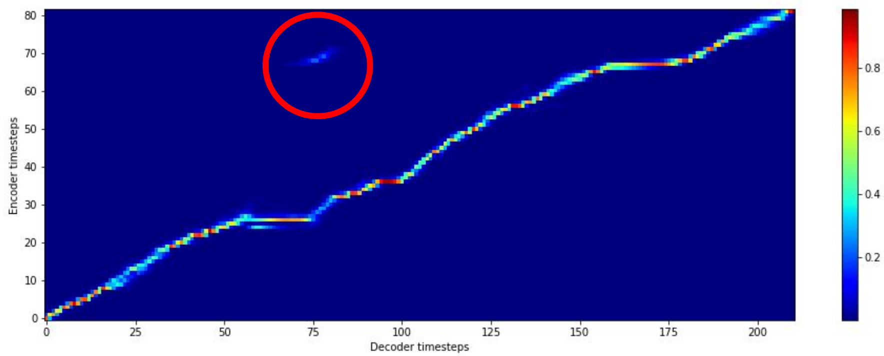
System	Network	Alignment	Decoder output	Post-net output
Char2Wav [1]	RNN	GMM	Vocoder	-
Tacotron [2]	RNN	Additive	Mel	Linear
VoiceLoop [3]	Memory buffer	GMM	Vocoder	-
Deep Voice 3 [4]	CNN	Dot-product	Mel	Linear/Vocoder
Tacotron 2 [5]	RNN	Location-sensitive	Mel	Mel
Transformer [6]	Self-attention	Dot-product	Mel	Mel

All methods use soft attention

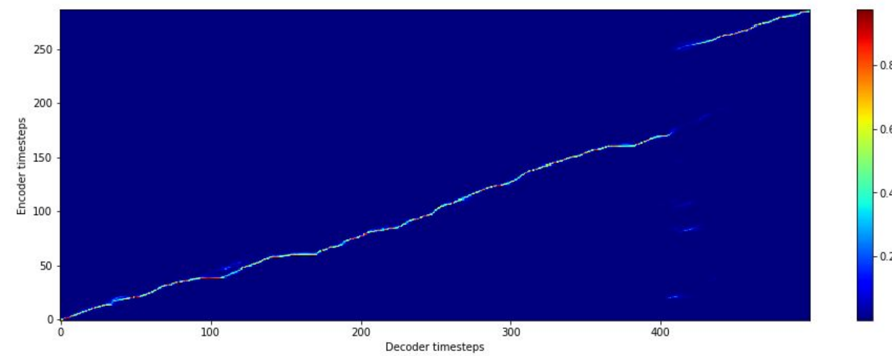
[1] J. Sotelo et al., ICLR, 2017.
[2] Y. Wang et al., Interspeech, 2017.
[3] Y. Taigman et al., ICLR, 2018.
[4] W. Ping et al., ICLR, 2018.
[5] J. Shen et al., ICASSP, 2018.
[6] N.Li et al., CoRR, vol. abs/1809.08895, 2018.

Problems of soft attention: Fatal alignment errors

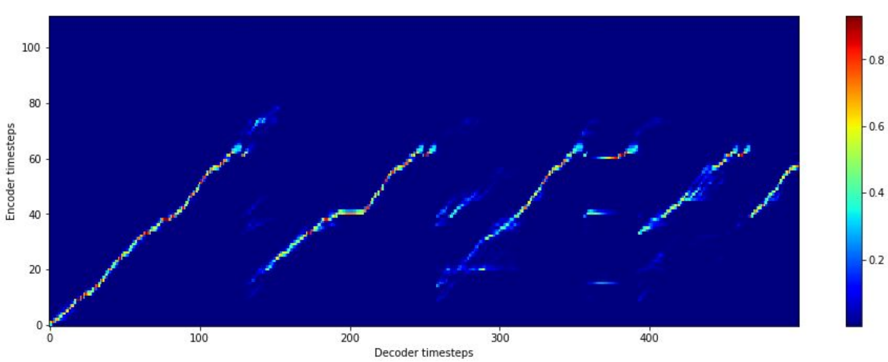
Mode split



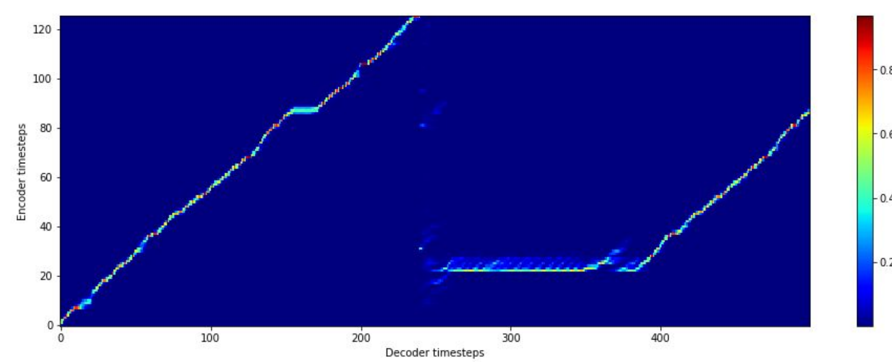
Skip



Repeat



Late termination



Problems of soft attention: Fatal alignment errors

DeepVoice3 [4]

Text Input	**Attention**	**Inference constraint**	**Repeat**	**Mispronounce**	**Skip**
Characters-only	Dot-Product	Yes	3	35	19
Phonemes & Characters	Dot-Product	No	12	10	15
Phonemes & Characters	**Dot-Product**	**Yes**	**1**	**4**	**3**
Phonemes & Characters	Monotonic	No	5	9	11

Transformer TTS [7]

Method	Repeats	Skips	Error Sentences	Error Rate
Transformer TTS	7	15	17	34%
FastSpeech	0	0	0	0%

[4] W. Ping et al., ICLR, 2018. [7] Y. Ren et al., CoRR, vol. abs/1905.09263, 2019

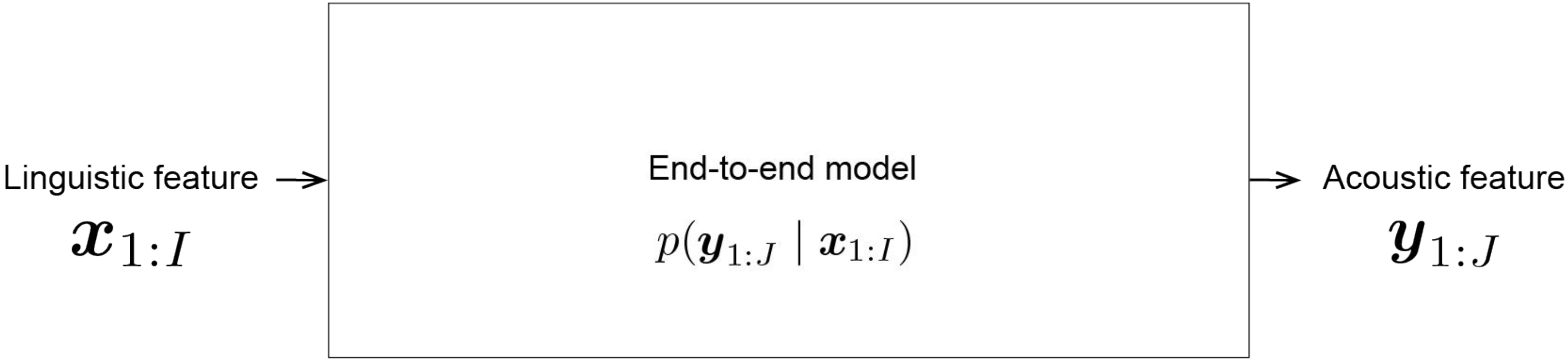
Design of the proposed method: SSNT based TTS

- Alignment structure is designed to be **monotonic**
- Alignment method is **hard attention**, instead of soft
- Alignment is a **latent variable**, part of probabilistic model

- Based on **SSNT** (Segment-to-Segment Neural Transduction) [8]
- Output distribution is continuous, instead of discrete

[8] L.Yu et al., EMNLP, 2016..

End-to-end TTS as a probabilistic model



Alignment as a latent variable

Linguistic feature $\rightarrow$

$\boldsymbol{x}_{1:I}$

End-to-end model

$$p(\boldsymbol{y}_{1:J} \mid \boldsymbol{x}_{1:I}) = \sum_{\forall \mathbf{z}} p(\boldsymbol{y}_{1:J}, \mathbf{z} \mid \boldsymbol{x}_{1:I})$$

Marginalization of latent variable

Alignment latent variable

$\rightarrow$ Acoustic feature

$\boldsymbol{y}_{1:J}$

Factorization for joint probability of alignment and output

$$p(\boldsymbol{y}_{1:J} \mid \boldsymbol{x}_{1:I}) = \sum_{\forall \mathbf{z}} p(\boldsymbol{y}_{1:J}, \mathbf{z} \mid \boldsymbol{x}_{1:I})$$

Factorization for joint probability

of alignment and output

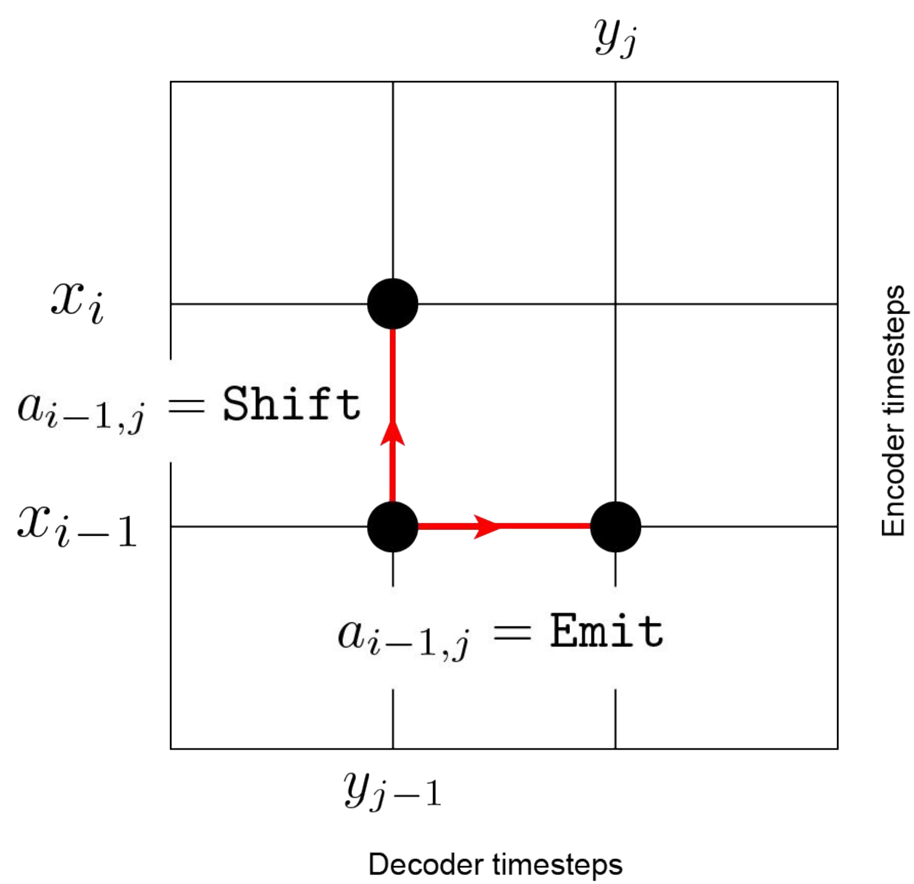
$$p(\boldsymbol{y}_{1:J}, \mathbf{z} \mid \boldsymbol{x}_{1:I}) \approx \prod_{j=1}^{J} \underbrace{p(z_j \mid z_{j-1}, \boldsymbol{y}_{1:j-1}, \boldsymbol{x}_{1:I})}_{\text{Alignment probability}} \underbrace{p(\boldsymbol{y}_j \mid \boldsymbol{y}_{1:j-1}, z_j, \boldsymbol{x}_{1:I})}_{\text{Output probability}}$$

Definition of alignment transition variables

$$\prod_{j=1}^{J} \underbrace{p(z_j \mid z_{j-1}, \boldsymbol{y}_{1:j-1}, \boldsymbol{x}_{1:I})}_{\text{Alignment probability}} \underbrace{p(\boldsymbol{y}_j \mid \boldsymbol{y}_{1:j-1}, z_j, \boldsymbol{x}_{1:I})}_{\text{Output probability}}$$

Binary alignment transition variable

$$a_{i,j} \in \{\mathtt{Emit}, \mathtt{Shift}\}$$

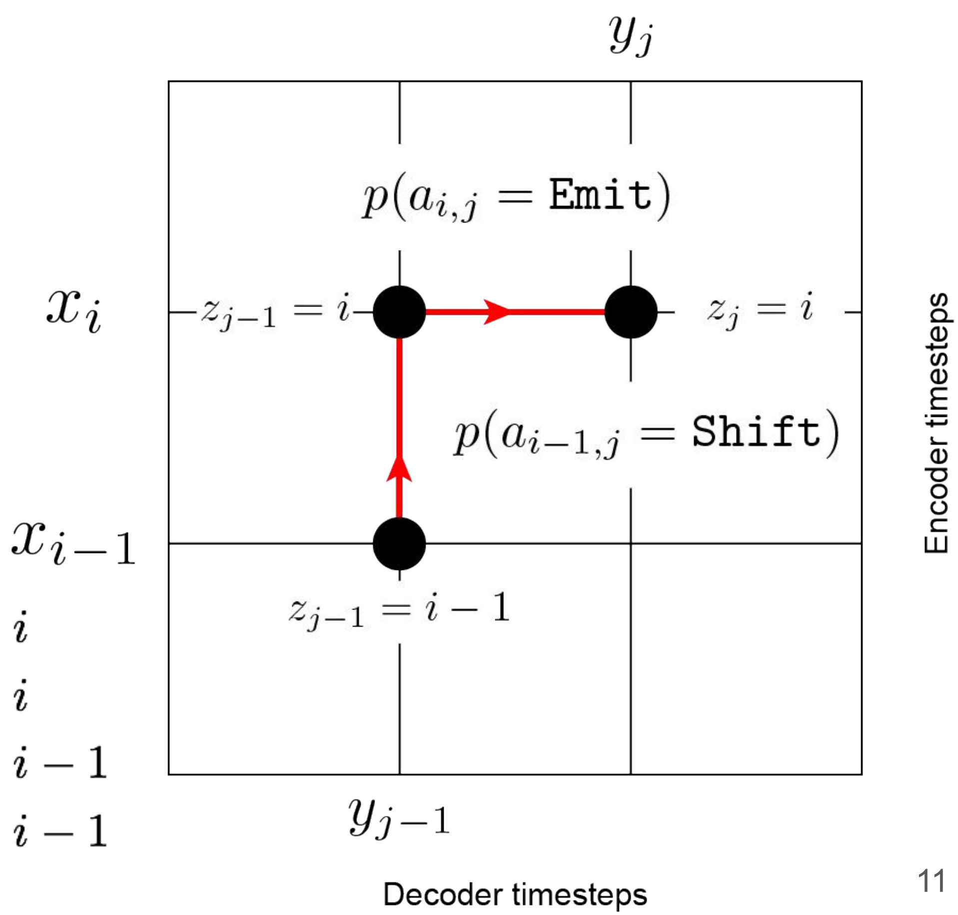


Definition of alignment probability

$$\prod_{j=1}^{J} \underbrace{p(z_j \mid z_{j-1}, \boldsymbol{y}_{1:j-1}, \boldsymbol{x}_{1:I})}_{\text{Alignment probability}} \underbrace{p(\boldsymbol{y}_j \mid \boldsymbol{y}_{1:j-1}, z_j, \boldsymbol{x}_{1:I})}_{\text{Output probability}}$$

Probability when an alignment reaches input position *i* at timestep *j*

$$p(z_j = i \mid z_{j-1}, \boldsymbol{y}_{1:j-1}, \boldsymbol{x}_{1:I}) =$$
$$\begin{cases} 0 & z_{j-1} > i \\ p(a_{i,j} = \texttt{Emit}) & z_{j-1} = i \\ p(a_{i-1,j} = \texttt{Shift})p(a_{i,j} = \texttt{Emit}) & z_{j-1} = i-1 \\ 0 & z_{j-1} < i-1 \end{cases}$$



Definition of output probability

$$\prod_{j=1}^{J} \underbrace{p(z_j \mid z_{j-1}, \boldsymbol{y}_{1:j-1}, \boldsymbol{x}_{1:I})}_{\text{Alignment probability}} \underbrace{p(\boldsymbol{y}_j \mid \boldsymbol{y}_{1:j-1}, z_j, \boldsymbol{x}_{1:I})}_{\text{Output probability}}$$

We used isotropic Gaussian distribution.

$$p(\boldsymbol{y}_j \mid \boldsymbol{y}_{1:j-1}, z_j, \boldsymbol{x}_{1:I}) = \mathcal{N}(\boldsymbol{y}_j; \boldsymbol{\mu}, \sigma^2 \mathbf{I})$$

Training (1): Objective function

Maximize $p(\boldsymbol{y}_{1:J} \mid \boldsymbol{x}_{1:I})$

$$\mathcal{L}(\boldsymbol{\theta}) = -\log p(\boldsymbol{y}_{1:J} \mid \boldsymbol{x}_{1:I}; \boldsymbol{\theta})$$

Linguistic feature $\boldsymbol{x}_{1:I}$ →

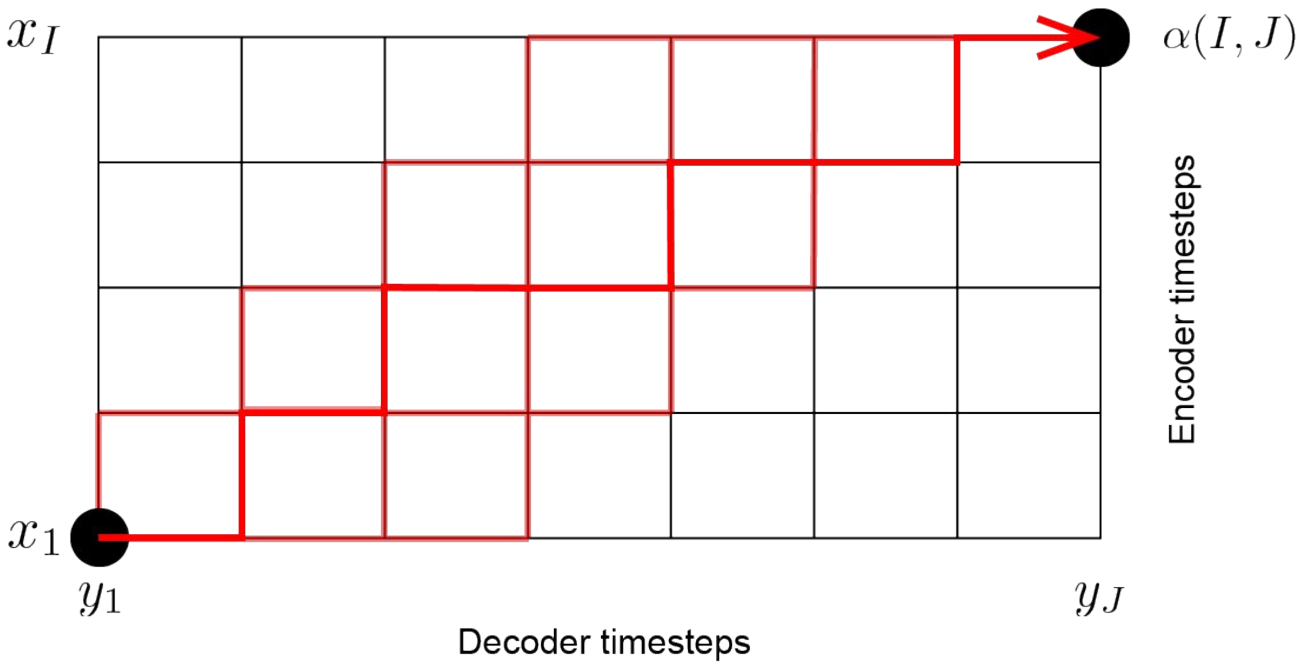
End-to-end model

$$p(\boldsymbol{y}_{1:J} \mid \boldsymbol{x}_{1:I}) = \sum_{\forall \mathbf{z}} p(\boldsymbol{y}_{1:J}, \mathbf{z} \mid \boldsymbol{x}_{1:I})$$

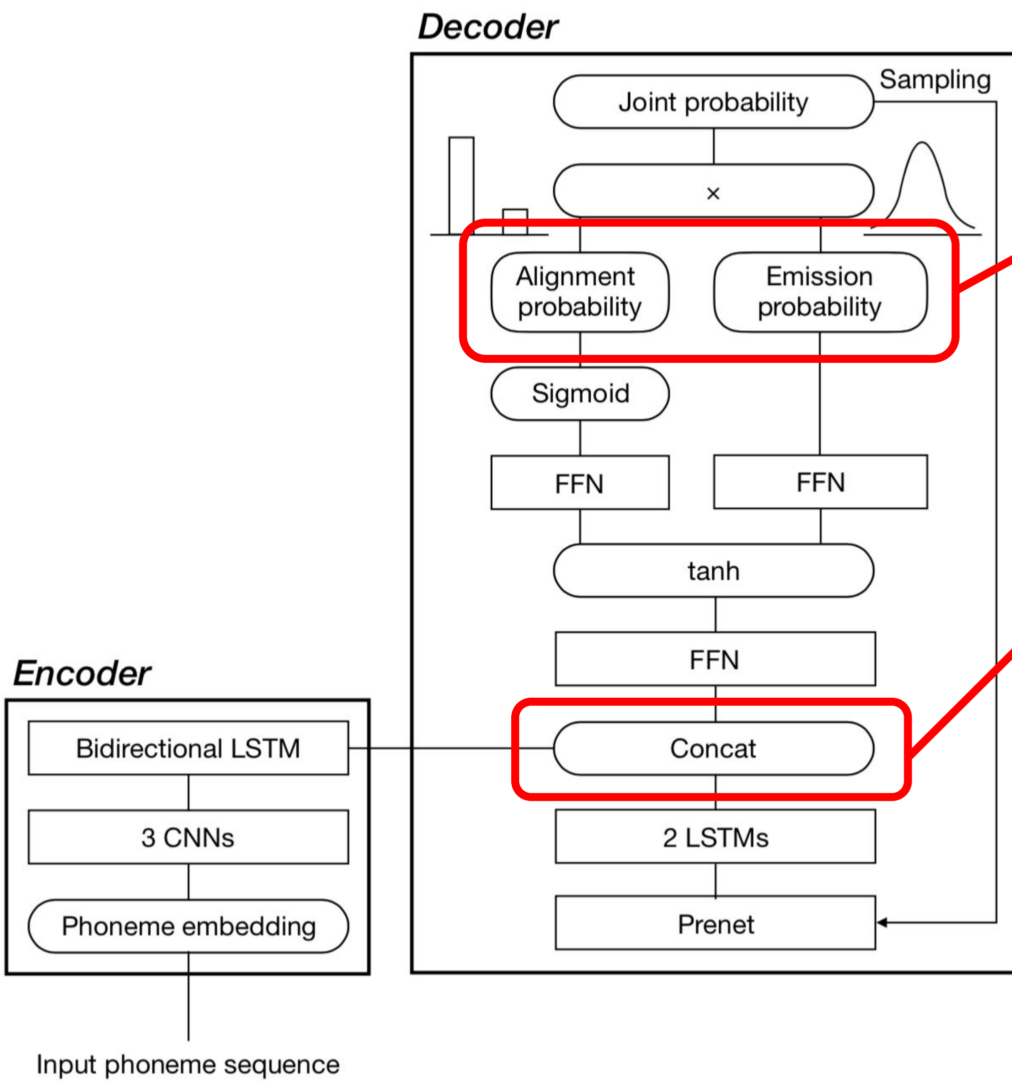
→ Acoustic feature $\boldsymbol{y}_{1:J}$

Training (2): Marginalization of alignments by forward probability

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= -\log p(\boldsymbol{y}_{1:J} \mid \boldsymbol{x}_{1:I}; \boldsymbol{\theta}) \\
&= -\sum_{\forall \mathbf{z}} p(\boldsymbol{y}_{1:J}, \mathbf{z} \mid \boldsymbol{x}_{1:I}) \\
&= -\log \alpha(I, J)
\end{aligned}
$$



Network structure



- Encoder-Decoder
- Decoder calculates alignment probability and output probability
- Decoder output is concatenated with encoder output to form trellis

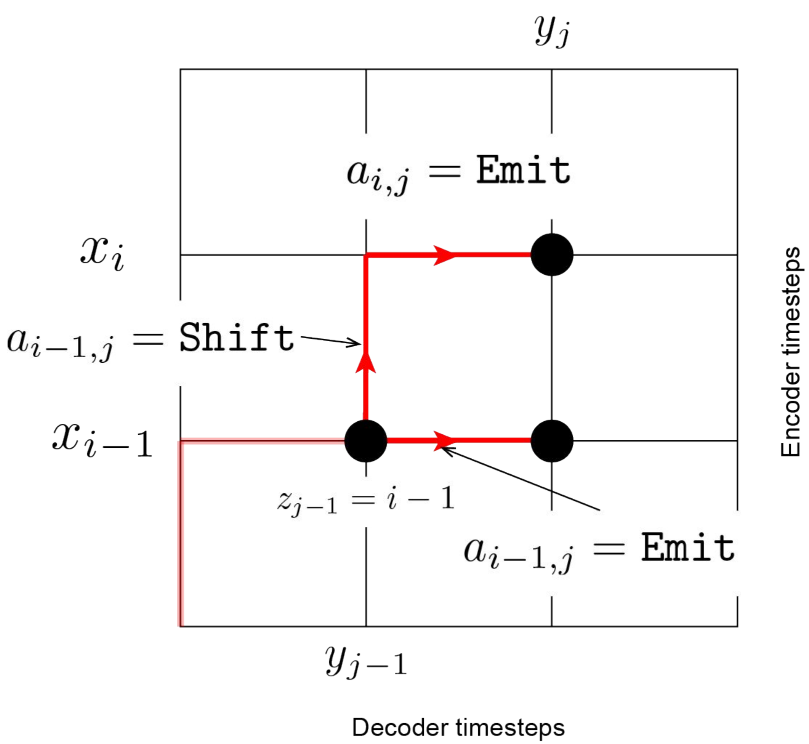
Inference (1): Alignment prediction

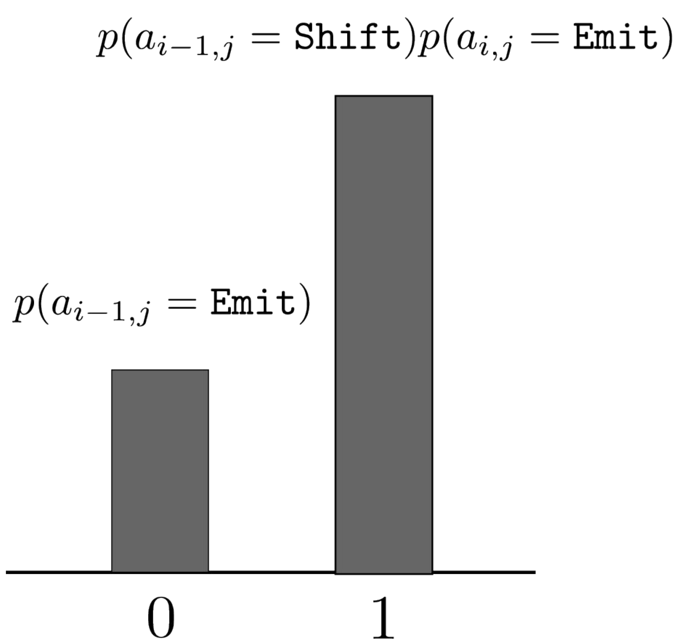
$$z_j = z_{j-1} + k$$

- Greedy decode $k = \mathrm{argmax}\left(p(z_j = i \mid z_{j-1}, \boldsymbol{y}_{1:j-1}, \boldsymbol{x}_{1:I})\right)$

or

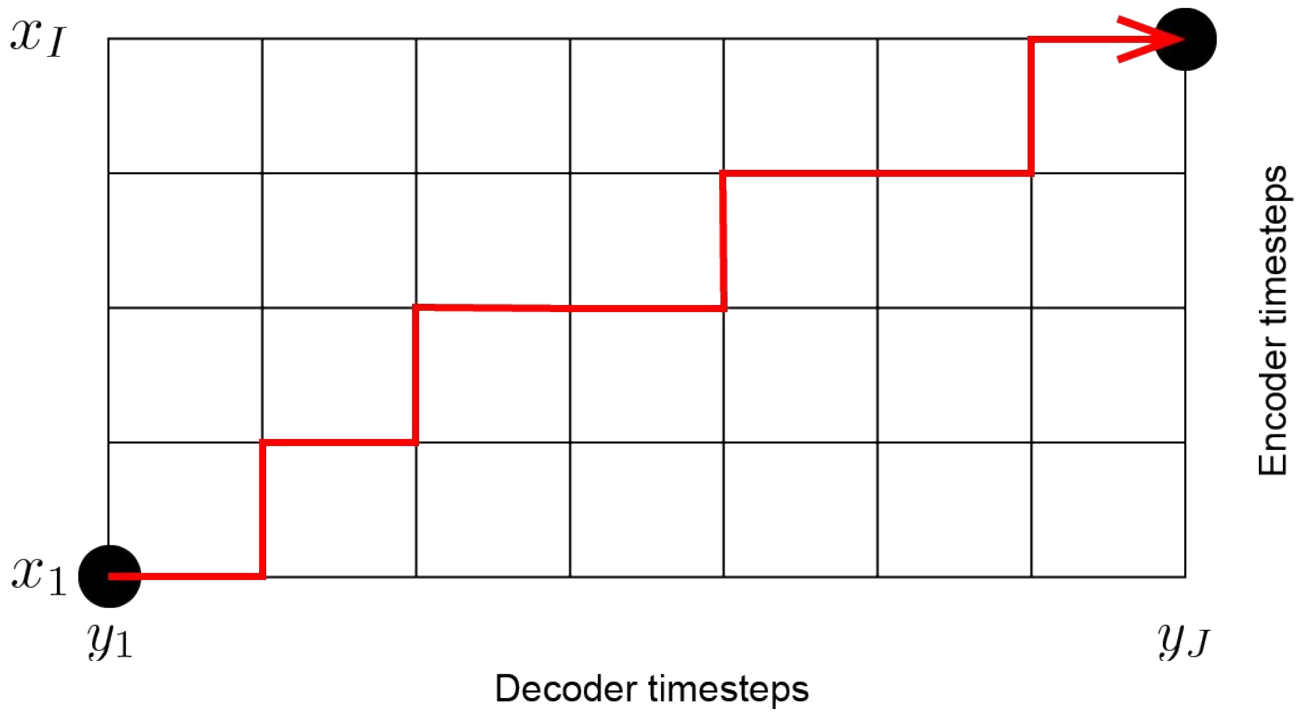
- Random sampling $k \sim \mathrm{Bernoulli}\left(p(z_j = i \mid z_{j-1}, \boldsymbol{y}_{1:j-1}, \boldsymbol{x}_{1:I})\right)$





Inference (2): Stop criteria

- When alignment reaches the final position of input
- No stop flag prediction



Experiments: listening test about naturalness

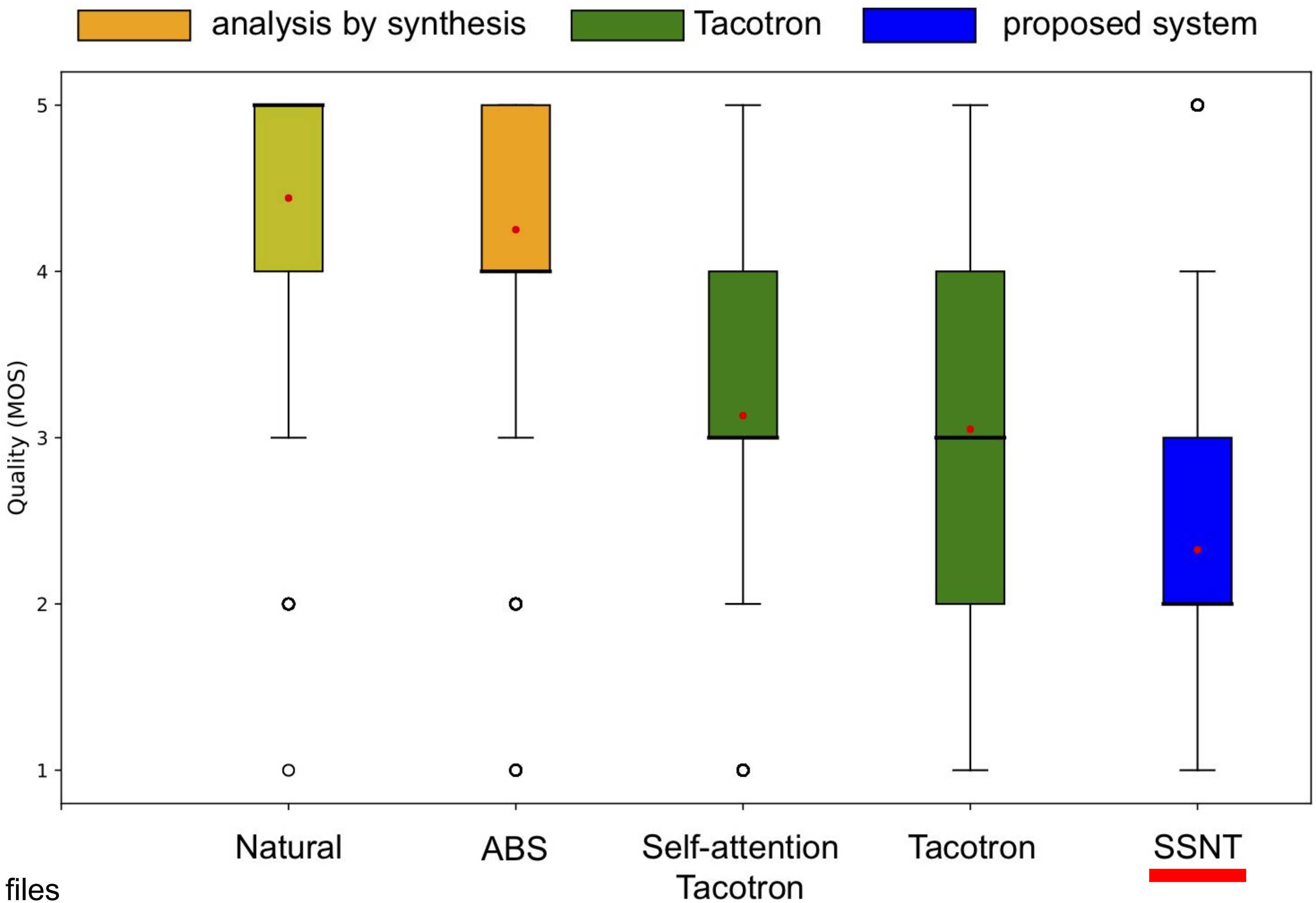
Data

- ATR Ximera (Japanese, Single speaker, 46.9h, 28,959 utterance)
- Linguistic feature: Phoneme (No accentual type label)
- Acoustic feature: Mel spectrogram (12.5 ms frame shift)
- Train/Validation/Test: 27,999/480/480
- Waveform synthesis: WaveNet

Evaluation

- Listening test about naturalness
- Listeners: 104
- Evaluation values: 19,200
- 5 systems
  - Natural, ABS, SA Tacotron, Tacotron, SSNT (Proposed)

Experimental results: underperform baselines



Audio files

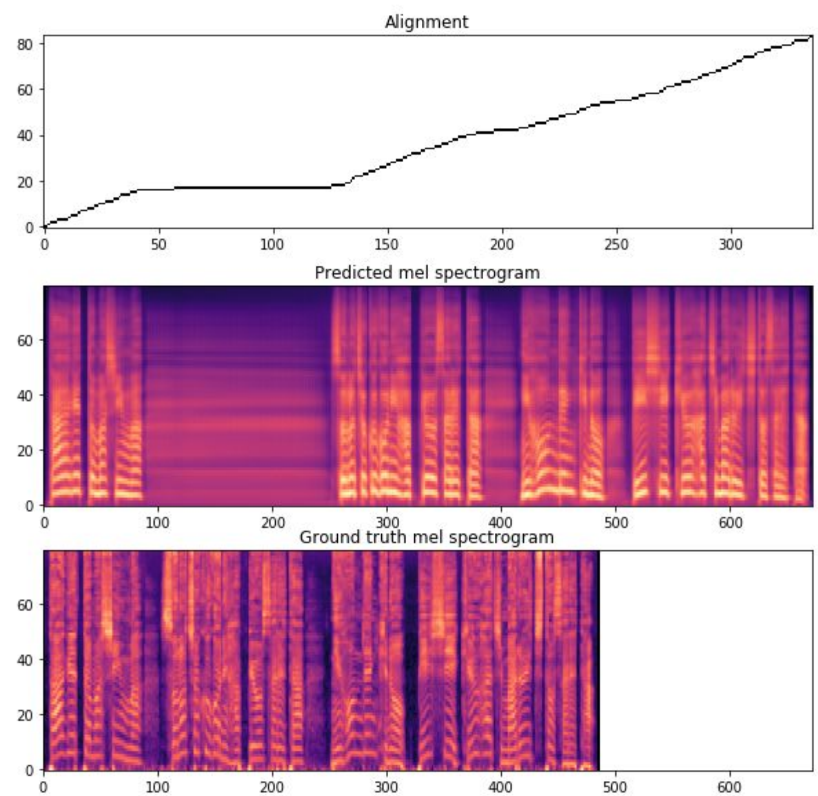
Analysis of generated samples

- Different kinds of alignment errors
  - Underestimation of duration
  - Overestimation of duration

Predicted alignment

Predicted spectrogram

Ground truth spectrogram



Conclusion

- A new end-to-end TTS method
  - Monotonic alignment structure by design
  - Hard attention instead of soft attention
  - Alignment is a latent variable
  - Objective function is likelihood of marginal probability
  - Alignment can be sampled from learned distribution
- Low naturalness of synthetic speech
  - No fatal alignment errors
  - Underestimation and overestimation of duration
- Future perspective
  - Testing various alignment distribution and sampling methods
  - Covariance estimation of output probability

Audio samples

Efficient gradient calculation of objective function

$$
\begin{aligned}
&\frac{\partial \log p(\boldsymbol{y}_{1:J} \mid \boldsymbol{x}_{1:I}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\
&= \frac{1}{p(\boldsymbol{y}_{1:J} \mid \boldsymbol{x}_{1:I}; \boldsymbol{\theta})} \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\partial p(\boldsymbol{y}_{1:J} \mid \boldsymbol{x}_{1:I}; \boldsymbol{\theta})}{\partial \alpha(i,j)} \frac{\partial \alpha(i,j)}{\partial \boldsymbol{\theta}} \\
&= \frac{1}{p(\boldsymbol{y}_{1:J} \mid \boldsymbol{x}_{1:I}; \boldsymbol{\theta})} \sum_{i=1}^{I} \sum_{j=1}^{J} \beta(i,j) \frac{\partial \alpha(i,j)}{\partial \boldsymbol{\theta}}
\end{aligned}
$$

The relationship $\dfrac{\partial p(\boldsymbol{y}_{1:J} \mid \boldsymbol{x}_{1:I}; \boldsymbol{\theta})}{\partial \alpha(i,j)} = \beta(i,j)$ is used.