Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language

Yusuke YASUDA (NII, SOKENDAI), Xin WANG (NII), Shinji Takaki (NII), Junichi YAMAGISHI (NII, SOKENDAI, University of Edinburgh)

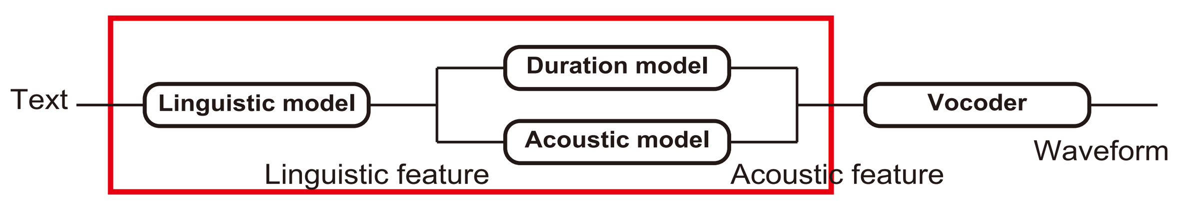
Abstract

- The end-to-end approach has not been fully investigated in languages other than English.
- We applied Tacotron to the Japanese language (pitch-accented language).
- To handle its pitch accent, accentual label embedding is introduced.
- A new architecture with self-attention is proposed to capture long term dependencies.
- We conducted a listening test with various systems and conditions.
- Our proposed systems showed the effectiveness of self-attention.

Background

Difference between pipeline and Tacotron

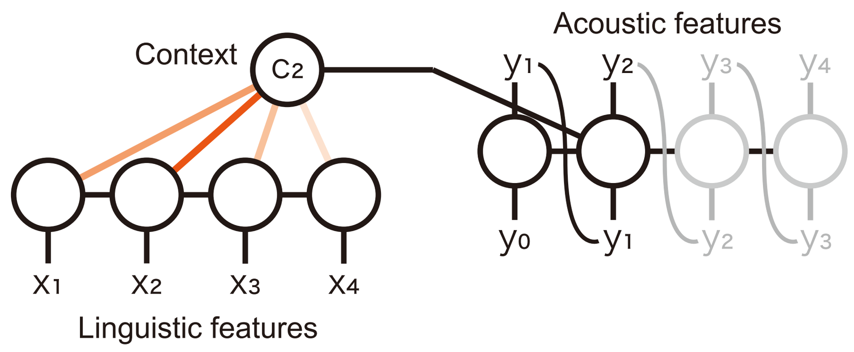
Traditional pipeline system



Tacotron handles the three components as a single model.

Implicit duration modeling with attention

Attention mechanism relates input and output.



Experiments

Three condition variants:

- Acoustic feature: M M mel-spectrogram (12.5 / 5ms frame shift); V vocoder parameters
- Accentual type: ✓ included; N/A excluded; C corrupted※
- Alignment: P predicted; F force aligned

※Corrupted accent is randomly modified accent.

はし

箸 (chopsticks)　橋 (bridge)

Baseline: pipeline systems [5,6]

Overview of Ximera dataset [7]

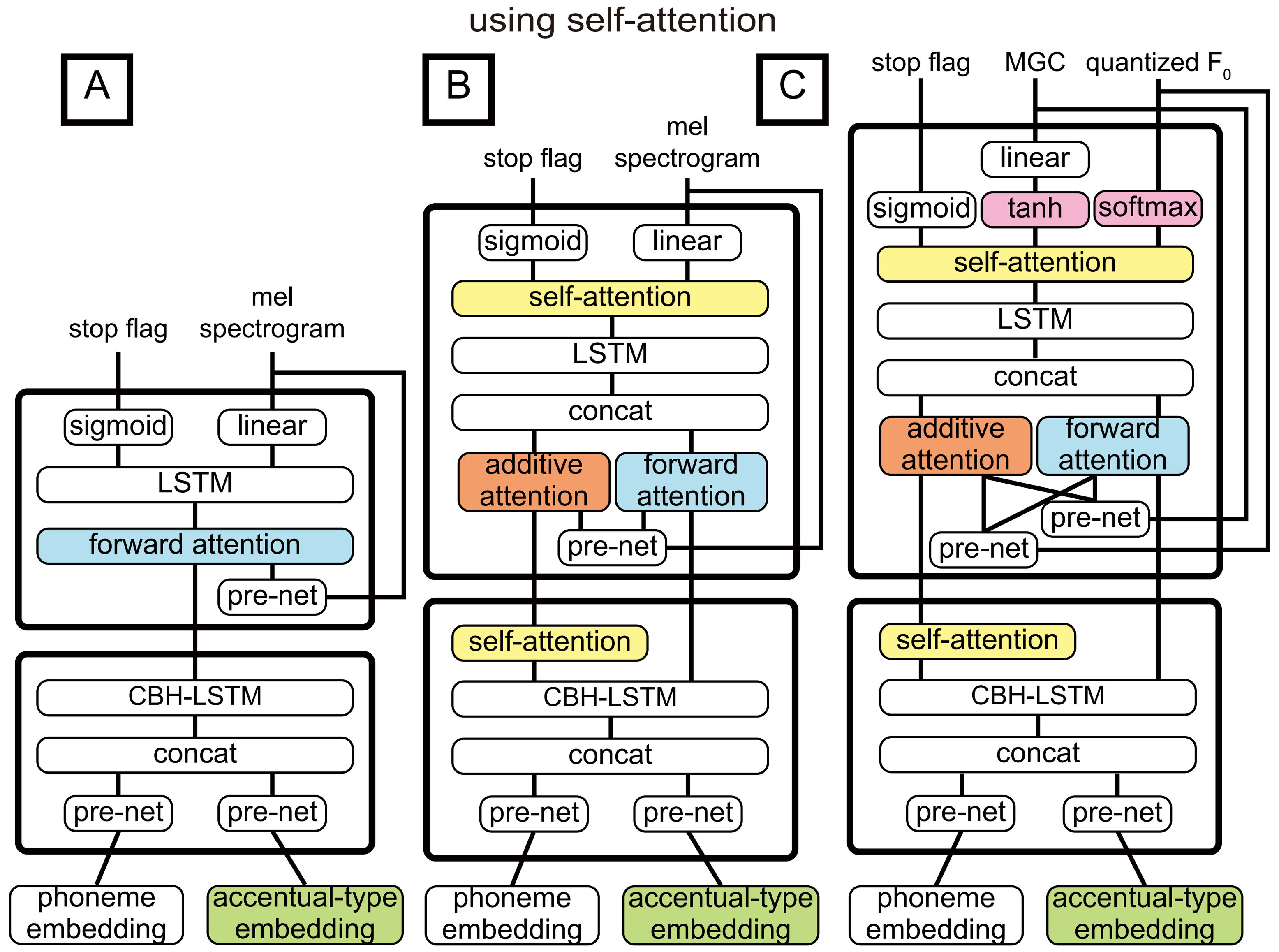
Speaker	Single female
Number of utterances	28,959
Total duration	46.9h (33.5h w/o sil)
Phoneme classes	58
Training/validation/test set	27,999/480/142

Proposed methods

Proposed architectures

We investigate the effect of :

- accentual type label
- long-term dependency modeling using self-attention
- explicit F0 modeling

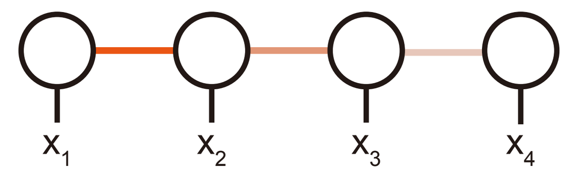


A: Japanese Tacotron
- Accentual-type embedding
- Forward attention [3]
- Zoneout [4]

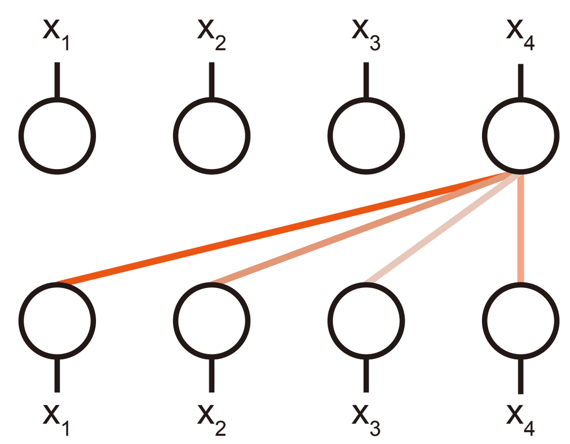
B: Self-attention Tacotron
- Self-attention
  - Encoder & Decoder
- Dual source attention
  - Forward attention [3]
  - Additive attention [8]

C: Self-attention Tacotron that outputs vocoder parameters
- Outputs
  - quantized F0
  - MGC

Self-attention vs RNN



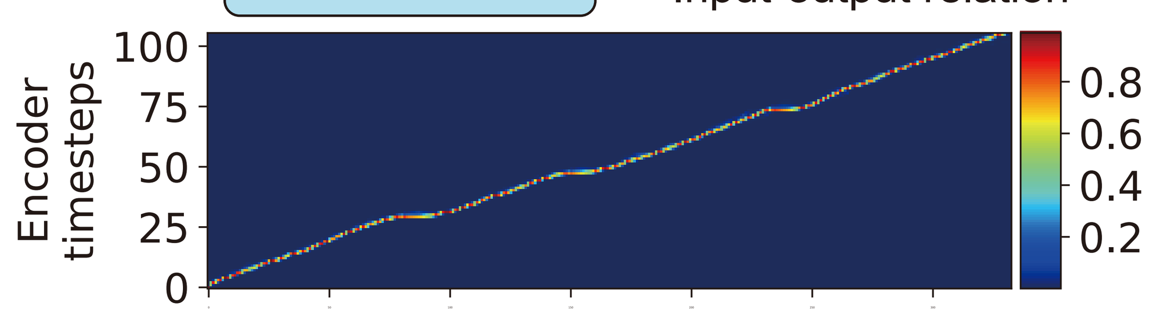
The RNN propagates input information gradually across states. It is hard to reach distant states.



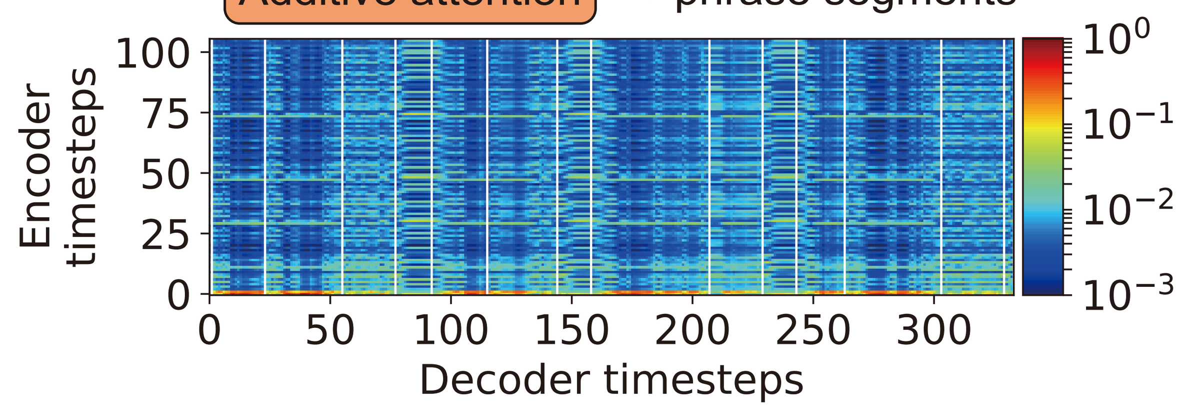
Self-attention directly connects inputs regardless of their distance. Possible to learn relationships between inputs

What do the two attentions capture?

Forward attention → Input-output relation

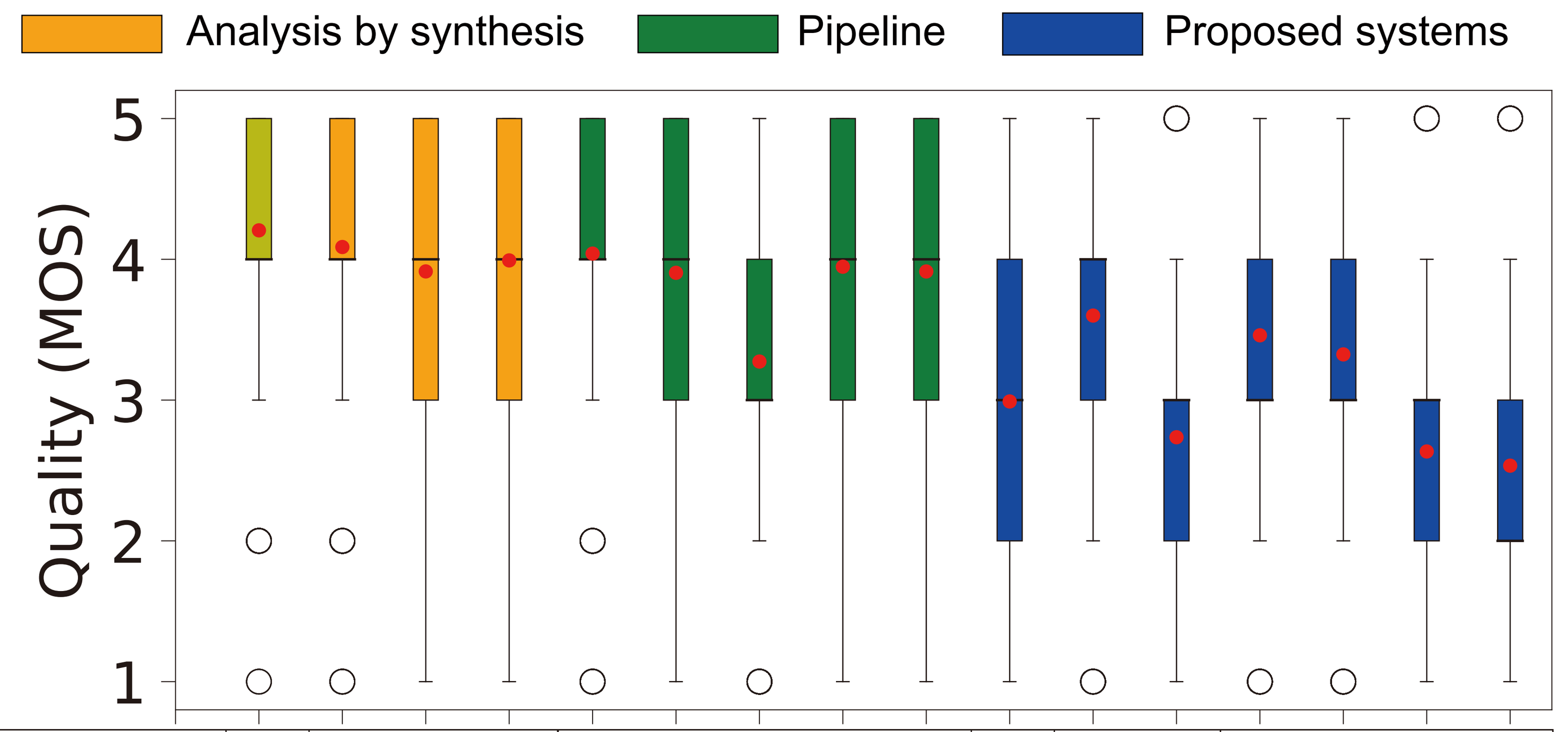


Additive attention → phrase segments



Result

Subjective evaluation



System	NAT	ABS			Pipeline					C	B		A			
Acoustic feature		V	M	M	V	V	V	M	M	V	M	M	M	M	M	M
Accent					✓	✓	C	✓	✓	✓	✓	N/A	✓	✓	N/A	N/A
Alignment					F	P	F	F	P	P	P	P	P	F	P	F

Listener's reaction to corrupted accents:
- Listeners are very sensitive to wrong accents

Accentual type label:
- Accentual type label helps to generate correct accents.

Self-attention:
- Self-attention help to improve naturalness.

Mel spectrogram vs Vocoder parameters:
- Vocoder parameters are better feature for pipelines
- Vocoder parameters are hard to predict for Tacotron.

Predicted vs Forced alignment:
- Forced alignment is better for pipeline
- Forced alignment causes unnaturalness for Tacotron

Proposed systems vs pipelines:
- Pipeline won.
- Possible reason: linguistic feature and model's parameter size limitation

Conclusion

- Three Japanese Tacotron architectures are proposed.
- The system using a mel-spectrogram with accentual type label and self-attention was the best system among the proposed systems.
- The vocoder parameter is not the right choice for Tacotron.
- Additional improvement is required to reach the quality of pipelines.

Bibliography

[1] Wang et al., in Proc. Interspeech, 2017, pp. 4006-4010.

[2] Shen et al., in Proc ICASSP, 2018, pp. 4779-4783.

[3] Zhang et al., ICASSP, 4789-4793, 2018.

[4] Krueger et al., arXiv, arXiv:1606.01305, 2017.

[5] Luong et al., Interspeech, 1227, 2018.

[6] Lorenzo-Trueba et al., Odyssey 240-247, 2018.

[7] Kawai et al., in Proc. SSW5, 179–184, 2004.

[8] Bahdanau et al.,ICLR, 2015.

SOKENDAI NII 国立情報学研究所 National Institute of Informatics Github: Email: yasuda@nii.ac.jp