USING LOCAL PHRASE DEPENDENCY STRUCTURE INFORMATION IN NEURAL SEQUENCE-TO-SEQUENCE SPEECH SYNTHESIS

Nobuyoshi Kaiki†, Sakriani Sakti†‡
and Satoshi Nakamura†‡

† Nara Institute of Science and Technology, Japan,
‡ RIKEN AIP, Japan

Background

[Background]

→ Neural networks have made it possible to produce speech synthesis with very high quality

→ Naturalness of the prosody is insufficient when using the speech synthesis for reading novels or storybooks

[Objective]

Generate more natural prosody that reflects the speaker's intention using local dependency phrase structure

Input: 警官は 走って 逃げる 泥棒を 追いかけた。

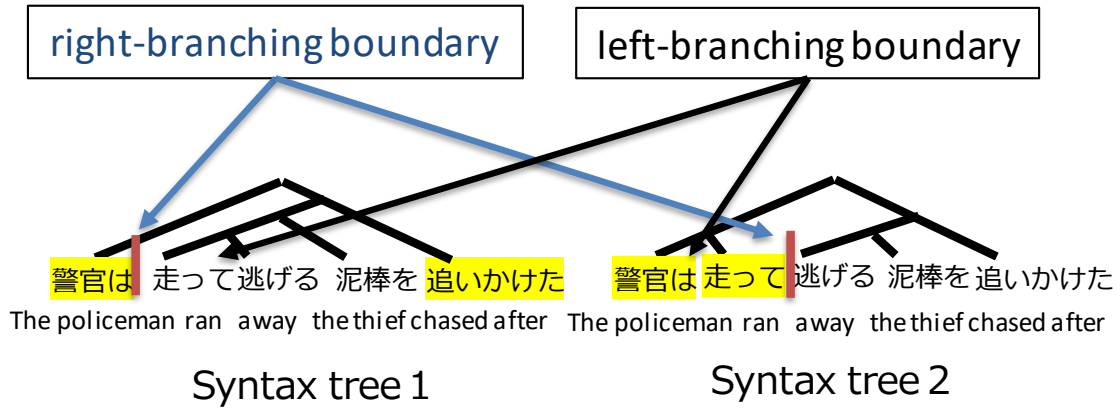
phoneme + accent information:

k e ^ e k a N w a # h a ^ sh i ! cl t e # n i ^ g e ! r u #
d o ^ r o b o o o # o ^ i k a k e ! t a (

Table 1 Prosodic Symbols (Accent) [1]

	Prosodic symbols
Initial rising	^
Accent nucleus	!
Accent phrase boundary	#
Sentence end（Declarative）	(
Sentence end（Interrogative）	?
Pause	_

This sentence has two syntax tree and different intonations



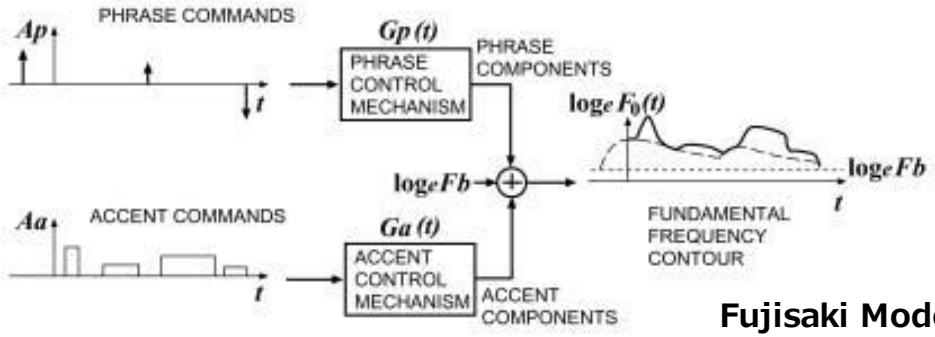
[1] K. Kurihara, N. Seiyama, T. Kumano, “Prosodic Features Control by Symbols as Input of Sequence-to-Sequence Acoustic Modeling for Neural TTS,” IEICE Trans. Inf. & Syst., Vol.E104–D, no2, Feb. 2021

Proposed Approach

[Proposed Approach] Using local phrase dependency

→ **Proposed 1**: With prosodic symbols that represent the depth at phrase boundaries

→ **Proposed 2**: With prosodic symbols that reflects a folded model of phrase & accent components based on prosodic generation control mechanism



Fujisaki Model[2]

Baseline

	Prosodic symbols
Initial rising	^
Accent nucleus	!
Accent phrase boundary	#
Sentence end (Declarative)	(
Sentence end (Interrogative)	?
Pause	_

Proposed 1

	Prosodic symbols
Initial rising	^
Accent nucleus	!
Accent phrase boundary (syntactic dependency distance)	**#1, #2, #3, #4, #5, #6**
Punctuation mark	,

Proposed 2

	Prosodic symbols
Accent command (rising)	/
Accent command (falling)	¥
Phrase command (syntactic dependency distance)	**#2, #3, #4, #5, #6**
Punctuation mark	,

Input: 警官は 走って 逃げる 泥棒を 追いかけた

Phoneme: k e e k a N w a h a sh i cl t e n i g e r u d o r o b o o o o i k a k e t a

Baseline: k e ^ e k a N w a # h a ^ sh i ! cl t e # n i ^ g e ! r u # d o ^ r o b o o o # o ^ i k a k e ! t a (

Proposed1 (accent + local phrase dependency structure):

Syntax tree 1: k e ^ e k a N w a **#4** h a ^ sh i ! cl t e **#1** n i ^ g e ! r u **#1** d o ^ r o b o o o **#1** o ^ i k a k e ! t a

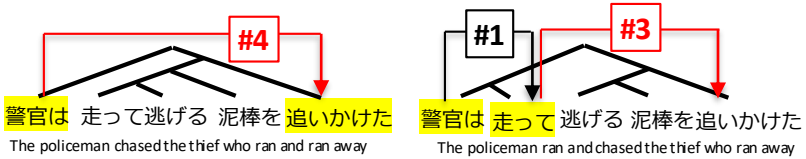
Syntax tree 2: k e ^ e k a N w a #1 h a ^ sh i ! cl t e **#3** n i ^ g e ! r u #1 d o ^ r o b o o o **#1** o ^ i k a k e ! t a

Proposed2 (based on the processes of generating the F0 contour):

Syntax tree 1: k e / e k a N w a ¥ **#4** h a / sh i ¥ cl t e n i / g e ¥ r u d o / r o b o o o ¥ o / i k a k e ¥ t a

Syntax tree 2: k e / e k a N w a ¥ h a / sh i ¥ cl t e **#3** n i / g e ¥ r u d o / r o b o o o ¥ o / i k a k e ¥ t a

syntactic dependency distance:



[2] H. Fujisaki, S. Nagashima, "A model for Synthesis of pitch contours of connected speech," Annual Report of Engineering Research Institute, University of Tokyo, Vol.28, pp.53-60, 1969

Speech Database and Pre-processing

[Speech Database]

→ **Arabian Nights** (reading a story)
→ Single speaker
→ 11,615 sentences
→ 26 hours 26 minutes

[Text processing]

Text ⇒ **Open Jtalk**[3] ⇒ Identification of phoneme,
Accent phrases、
Morphological analysis results

Morphological Analysis Results ⇒ **Chabocha**[4] ⇒ Syntax tree

[Speech processing]

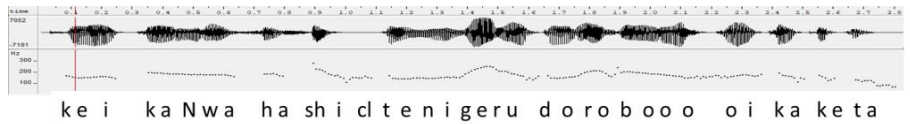
Sentences ⇒ **CTC Segmentation**[5] ⇒ split sentence
Sentences ⇒ **Montreal-Forced-Aligner**[6] ⇒ phoneme segmentation

[3] "Open JTalk," http://open-jtalk.sourceforge.net/
[4] “CaboCha/南瓜: Yet Another Japanese Dependency Structure Analyzer,“http://taku910.github.io/cabocha/
[5] J. Nishitoba, "Introduction to CTC Segmentation," https://tech.retrieva.jp/entry/2020/10/02/143338 (in Japanese)
[6] "Montreal Forced Aligner," https://montrealcorpustools. github.io/Montreal-Forced-Aligner/

Synthesized Speech: Pause Generation

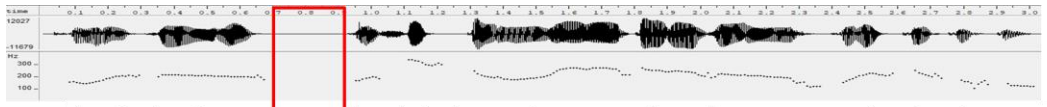
Comparison of phonemes and prosodic symbols based on two syntax tree candidates

Baseline:

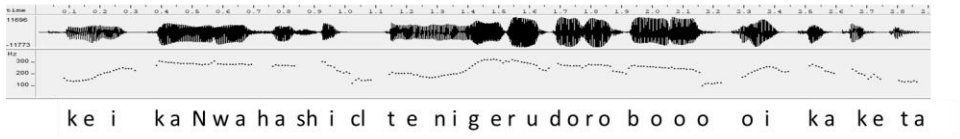


Proposed 1 :

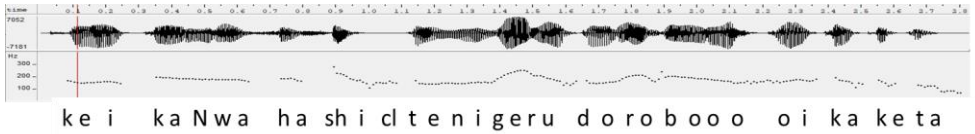
Syntax tree 1



Syntax tree 2

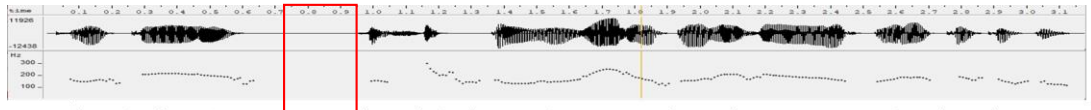


Baseline:

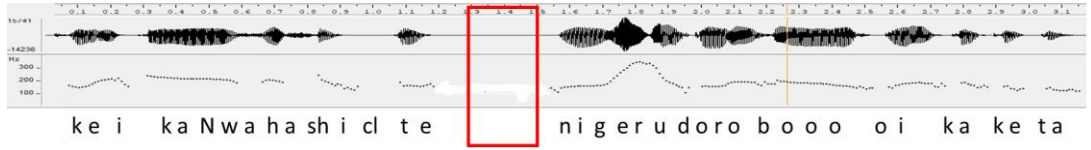


Proposed 2:

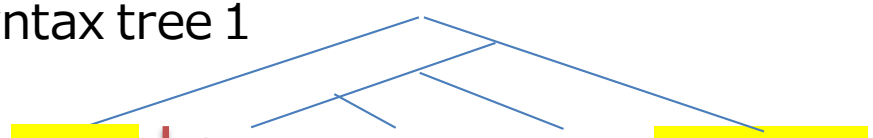
Syntax tree 1



Syntax tree 2



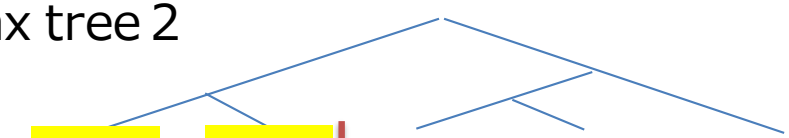
Syntax tree 1



[The policeman] [ran] [run away] [the thief] [chased after]

→ The policeman chased the thief who ran and ran away

Syntax tree 2

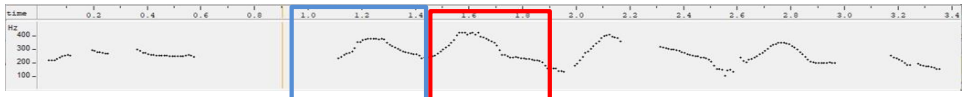


[The policeman] [ran] [run away] [the thief] [chased after]

→ The policeman ran and chased the thief who ran away

Synthesized Speech: F0 Resetting

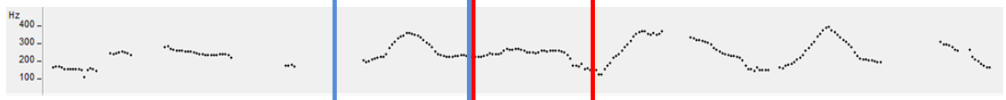
Baseline:



Proposed 1 :

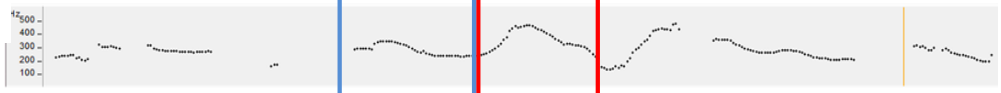
Syntax tree 1

left-branching boundary: **lower F0** than the previous phrase



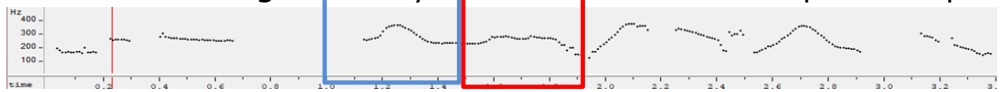
Syntax tree 2

right-branching boundary: **higher F0** than the previous phrase



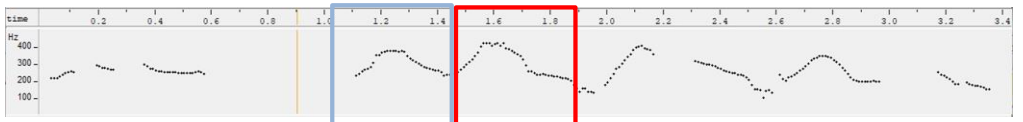
Syntax tree 3

left-branching boundary: **lower F0** than the previous phrase



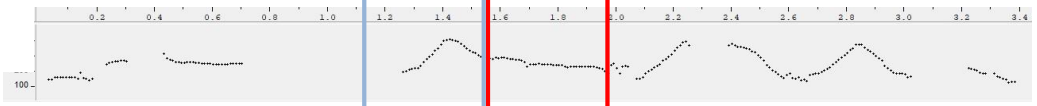
wa ta shi wa shi ro i ya ne no o o ki i i e ga sU ki da

Baseline:

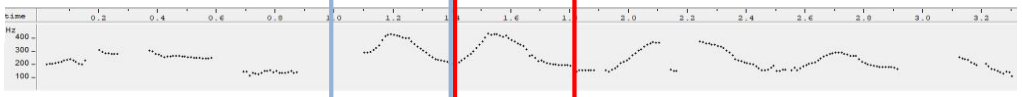


Proposed2:

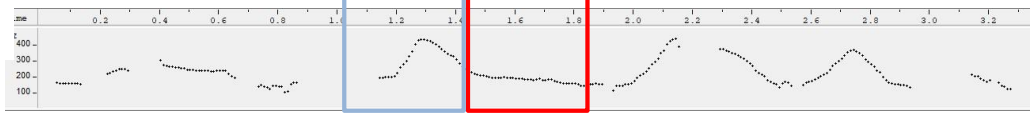
Syntax tree 1



Syntax tree 2

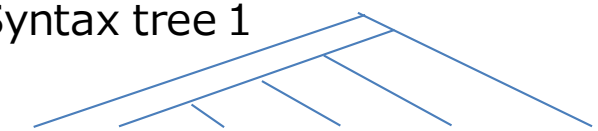


Syntax tree 3



wa ta shi wa shi ro i ya ne no o o ki i i e ga sU ki da

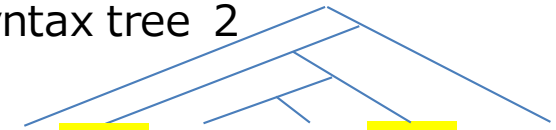
Syntax tree 1



私は、白い 屋根の大きい 家が 好きだ。

[I] [white] [roof] [big] [house] [like]

Syntax tree 2

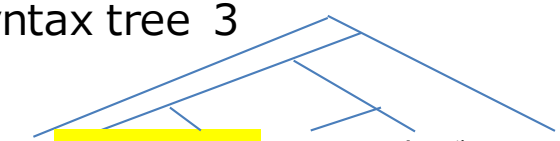


私は、白い屋根の大きい家が 好きだ。

[I] [white] [roof] [big] [house] [like]

→ I like white houses with big roofs.

Syntax tree 3



私は、白い屋根の大きい家が 好きだ。

[I] [white] [roof] [big] [house] [like]

→ I like big houses with white roofs.

Subjective Evaluation of Naturalness

[Participants]

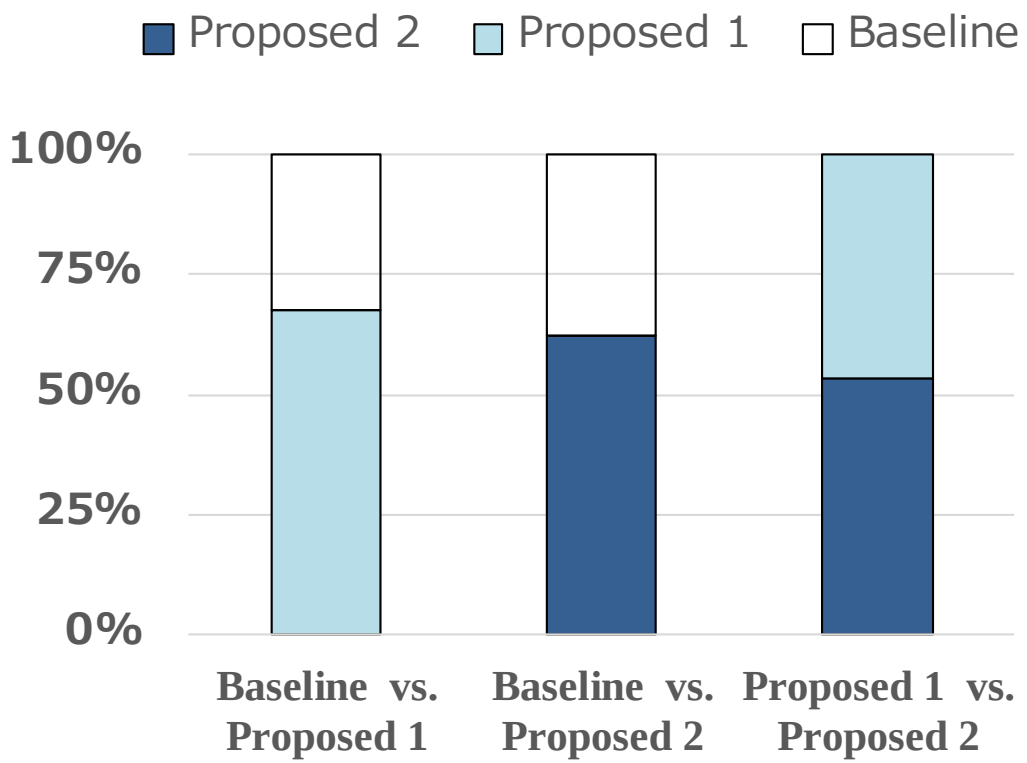
→ 13 native speakers of Japanese
→ Listened to the synthesized speech and judged naturalness of prosody

[Evaluation sentences]

→ Select 20 sentences from 250 sentences in the created DB
→ AB Test: 60 pair speech utterances
Baseline vs Proposed 1: 20 sentences * 2
Baseline vs Proposed 2: 20 sentences * 2
Proposed1 vs Proposed 2: 20 sentences * 2

[Results]

→ Proposed methods were judged to be significantly more natural than the baseline
・Proposed1 ＞ Baseline 68% ＞ 32%
・Proposed2 ＞ Baseline 62% ＞ 38%
→ No significant difference between proposed model 1 and 2 (significant difference 5%)



Conclusion

[Purpose]

To synthesize more natural prosody

→ Incorporated new prosody symbol of syntactic dependency for neural end-to-end TTS

[**Method**]

Proposed two models:

1) a model with prosodic symbols representing the syntactic dependency distance at the phrase boundaries
2) a model with prosodic symbols that reflect a prosodic generation control mechanism

[Results]

→ Both proposed models could successfully synthesize speech sounds that reflect syntactic structures

1) pause insertion that indicates the phrase boundary
2) F0 resetting at the right-branching boundaries

→ Subjective evaluation

Synthesize speech from two proposed models were more natural than the baseline