Applying Syntax–Prosody Mapping Hypothesis and Prosodic Well-Formedness Constraints to Neural Sequence-to-Sequence Speech Synthesis

Kei Furukawa[1]*, Takeshi Kishiyama*[2]*, and Satoshi Nakamura*[1]

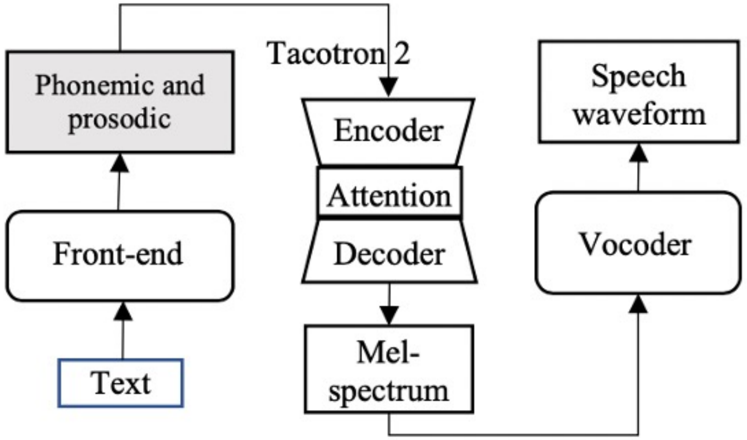
[1]Nara Institute of Science and Technology, Japan

[2]Graduate School of Arts and Sciences, The University of Tokyo, Japan

2022_09_18-22

Backgrounds of TTS

- Advances of End-to-end text-to-speech synthesis (TTS) [1]
- Japanese has a huge number of characters, and the reading of each character is not consistent
- The introduction of phoneme sequences and accent symbols as inputs of TTS improves the naturalness of speech synthesis [2, 3, 4]



cited from Kaiki et al. 2021

[1] Shen et al., 2018
[2] Yasuda et al., 2019
[3] Fujimoto et al., 2019
[4] Kurihara et al., 2021

Backgrounds of TTS

- Other studies:
  - Incorporated information of the post-lexical level, such as syntactic structure and syntactic dependency information [5, 6].
- However,
  - Not objectively examined whether they can reproduce pitch patterns of phonological phenomena

[5] Guo et al., 2019
[6] Kaiki et al., 2021

Motivation

- Reproduce speech sounds with syntactic and phonological phenomena by applying linguistic theories to neural sequence-to-sequence speech synthesis.

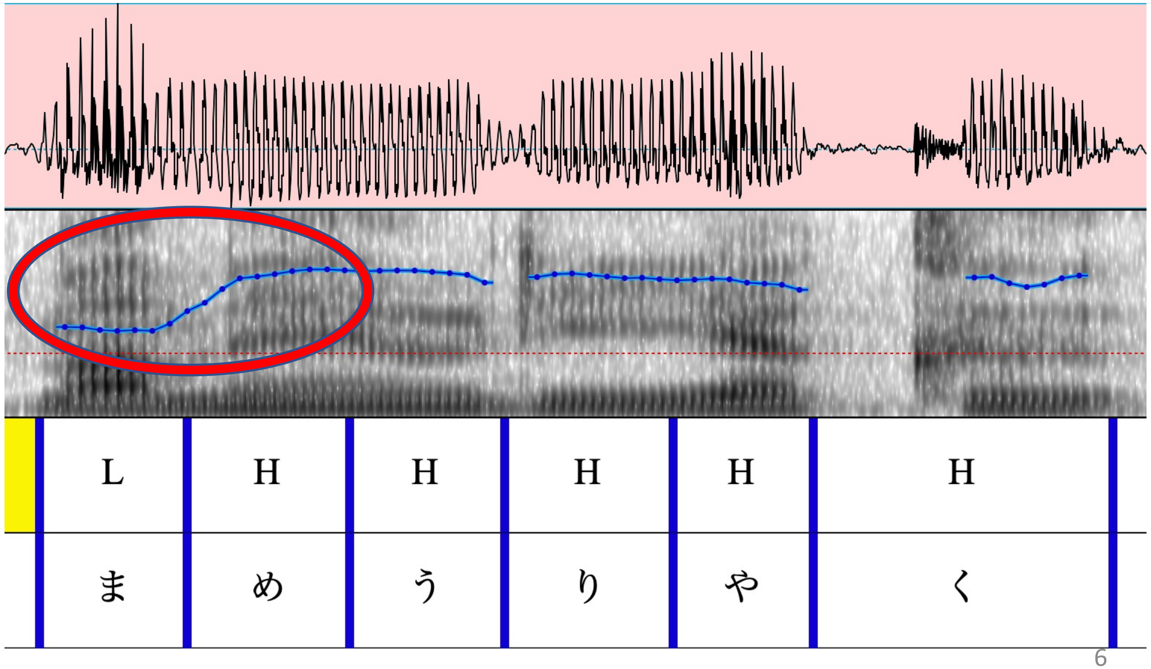
- Target Phenomena:
  - **Phenomenon 1. initial lowering**
  - **Phenomenon 2. rhythmic boost**

Outline of this talk

- Backgrounds and problems in current text-to-speech synthesis (TTS) systems: two phenomena
- **Phenomenon 1: the degree of initial lowering**
  - Proposed model 1: Syntax–Prosody Mapping Hypothesis
  - Experiment 1
- **Phenomenon 2: rhythmic boost**
  - Proposed model 2: Prosodic Well-Formedness Constraints
  - Experiment 2
- Discussions and Conclusion

Phenomenon 1: the degree of initial lowering

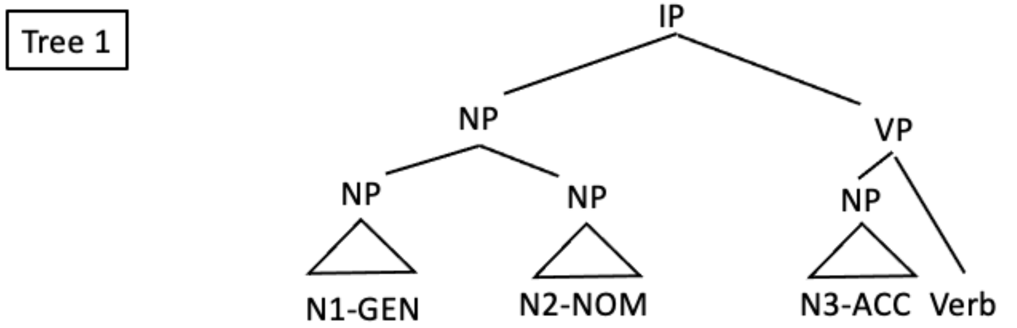
- The initial lowering is the F0 rise at the beginning of a PPhrase [7, 8]
- The degree of F0 rise in initial lowering varies in response to syntactic structure [9]



[7] Pierrehumbert &. Beckman, 1988
[8] Igarashi, 2015
[9] Selkirk et al., 2013

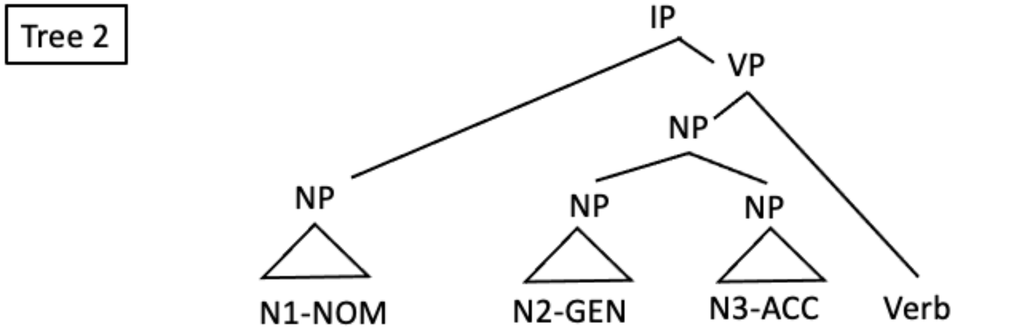
Phenomenon 1: the degree of initial lowering

- Initial lowering is greater at **position B than A in tree 1**, while the initial lowering is greater **at position A than B in tree 2** [9]
- The results can be explained via
  - Syntax–prosody mapping hypothesis [10]
  - Edge boost hypothesis (proposal)



a. Syntax: {$_{\text{IP}}$ [$_{\text{NP}}$ [$_{\text{NP}}$ N1-GEN] **A** [$_{\text{NP}}$ N2-NOM]] **B** [$_{\text{VP}}$ [$_{\text{NP}}$ N3-ACC] Verb]}

b. Phonology: {$_{\text{PC}}$ [$_{\text{PP}}$ [$_{\text{PP}}$ N1-GEN] **A** [$_{\text{PP}}$ N2-NOM]] **B** [$_{\text{PP}}$ [$_{\text{PP}}$ N3-ACC] Verb]}



a. Syntax: {$_{\text{IP}}$ [$_{\text{NP}}$ N1-NOM] **A** [$_{\text{VP}}$ [$_{\text{NP}}$ [$_{\text{NP}}$ N2-GEN] **B** [$_{\text{NP}}$ N3-ACC]] Verb]}

b. Phonology: {$_{\text{PC}}$ [$_{\text{PP}}$ N1-NOM] **A** [$_{\text{PP}}$ [$_{\text{PP}}$ [$_{\text{PP}}$ N2-GEN] **B** [$_{\text{PP}}$ N3-ACC]] Verb]}

[9] Selkirk et al., 2013
[10] Selkirk, 2011

Phenomenon 1: the degree of initial lowering

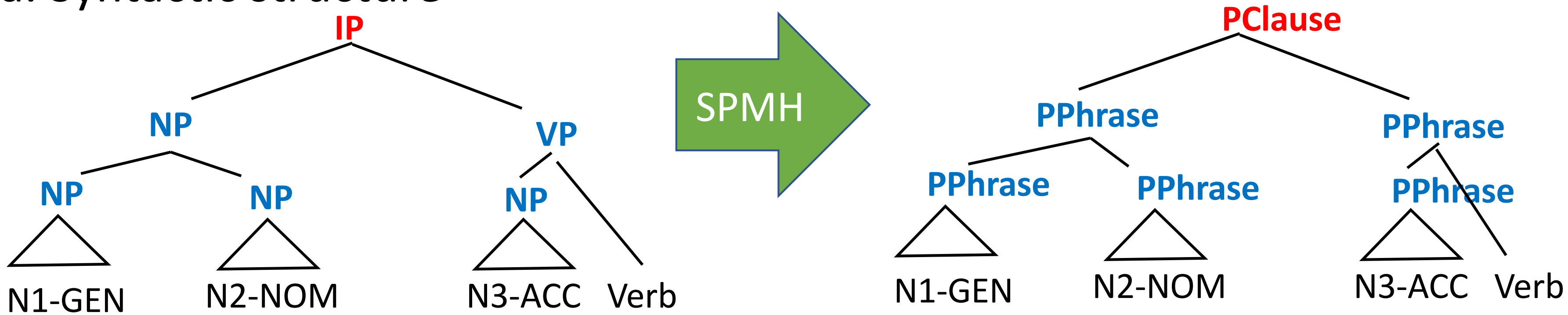
- Syntax–Prosody Mapping Hypothesis (SPMH) [10] require that syntactic categories be mapped to their corresponding phonological counterparts
  - **Syntactic clause → PClause**
  - **Syntactic phrase such as NP, VP → PPhrase**

[10] Selkirk, 2011

Tree 1

a. Syntactic structure

b. Phonological structure



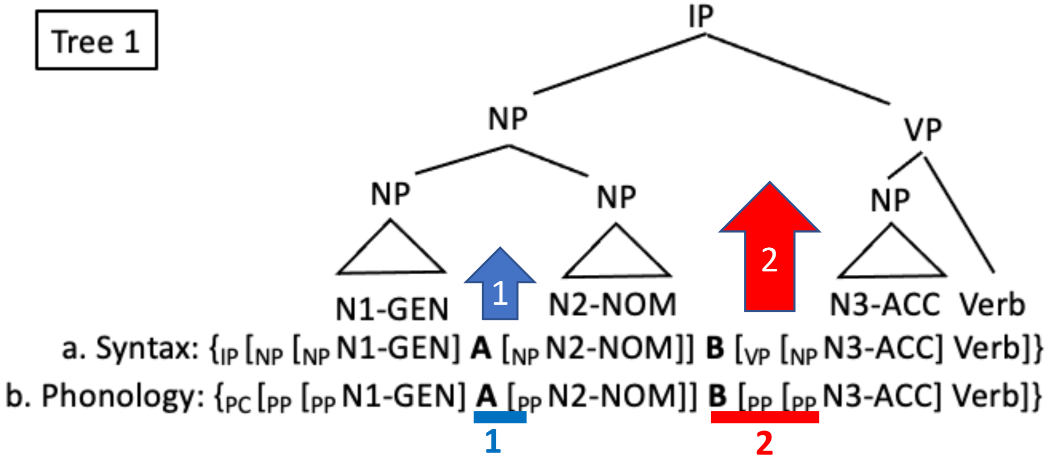
{$_{\text{IP}}$ [$_{\text{NP}}$ [$_{\text{NP}}$ N1-GEN] **A** [$_{\text{NP}}$ N2-NOM]] **B** [$_{\text{VP}}$ [$_{\text{NP}}$ N3-ACC] Verb]}

{$_{\text{PC}}$ [$_{\text{PP}}$ [$_{\text{PP}}$ N1-GEN] **A** [$_{\text{PP}}$ N2-NOM]] **B** [$_{\text{PP}}$ [$_{\text{PP}}$ N3-ACC] Verb]}

Phenomenon 1: the degree of initial lowering

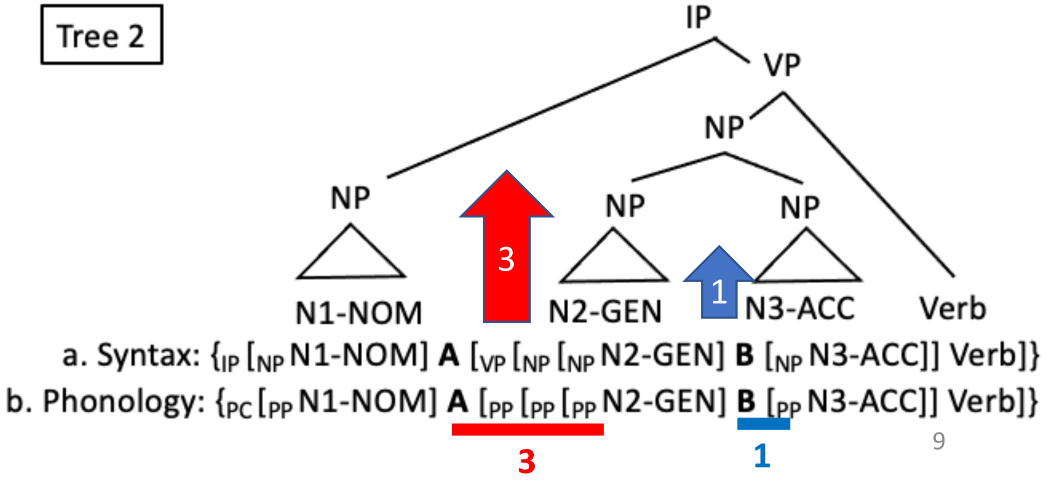
- Initial lowering is greater at **position B than A in tree 1**, while the initial lowering is greater at **position A than B in tree 2** [9]
- Edge boost hypothesis (proposal)
  - the number of edges in the PPhrases is proportional to the degree of the pitch increase in the initial lowering

[9] Selkirk et al., 2013



a. Syntax: {$_{IP}$ [$_{NP}$ [$_{NP}$ N1-GEN] **A** [$_{NP}$ N2-NOM]] **B** [$_{VP}$ [$_{NP}$ N3-ACC] Verb]}

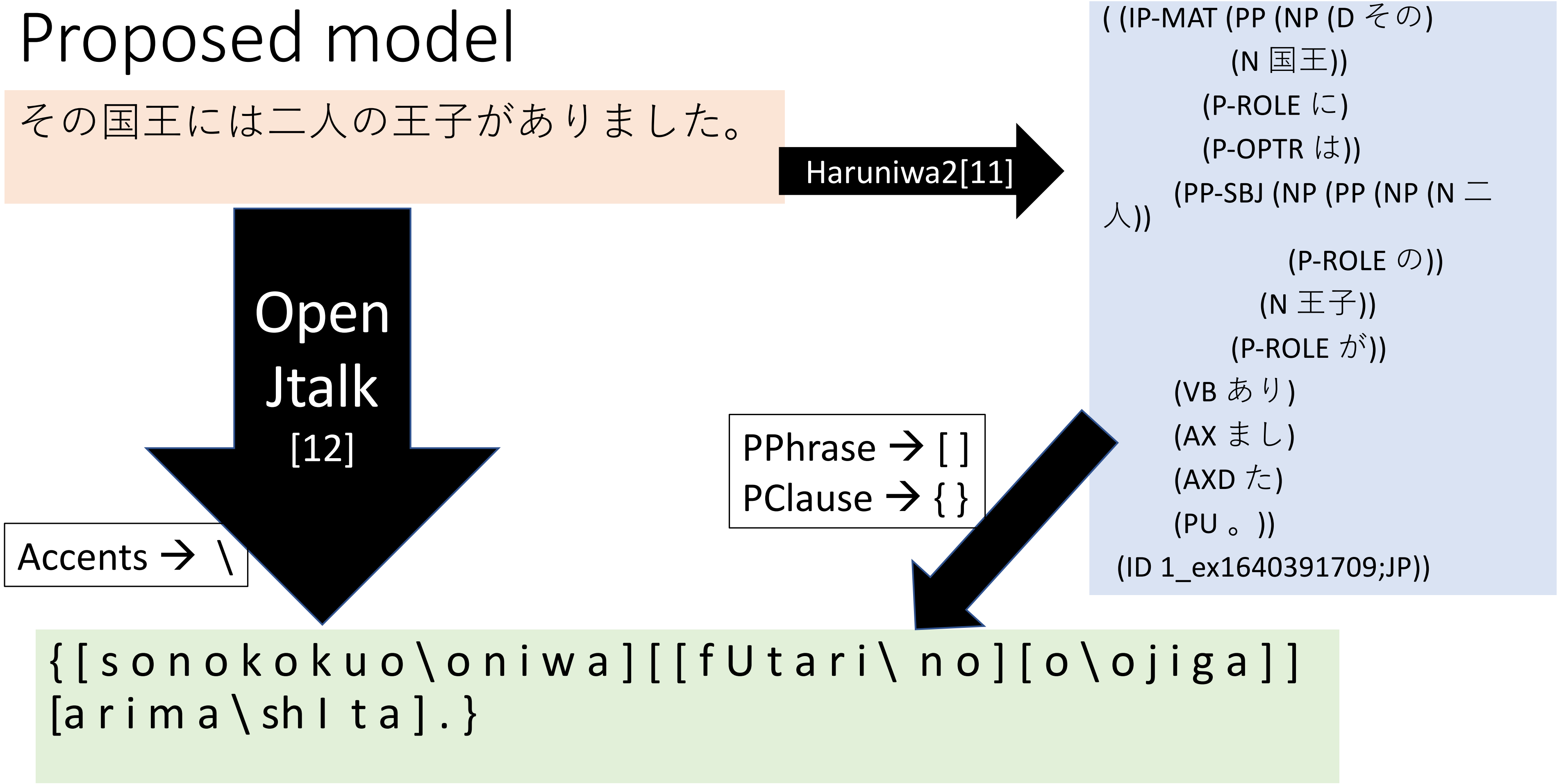
b. Phonology: {$_{PC}$ [$_{PP}$ [$_{PP}$ N1-GEN] **A** [$_{PP}$ N2-NOM]] **B** [$_{PP}$ [$_{PP}$ N3-ACC] Verb]}



a. Syntax: {$_{IP}$ [$_{NP}$ N1-NOM] **A** [$_{VP}$ [$_{NP}$ [$_{NP}$ N2-GEN] **B** [$_{NP}$ N3-ACC]] Verb]}

b. Phonology: {$_{PC}$ [$_{PP}$ N1-NOM] **A** [$_{PP}$ [$_{PP}$ [$_{PP}$ N2-GEN] **B** [$_{PP}$ N3-ACC]] Verb]}

Proposed model



その国王には二人の王子がありました。

Haruniwa2[11]

```
( (IP-MAT (PP (NP (D その)
              (N 国王))
           (P-ROLE に)
           (P-OPTR は))
       (PP-SBJ (NP (PP (NP (N 二人))
                     (P-ROLE の))
                  (N 王子))
               (P-ROLE が))
       (VB あり)
       (AX まし)
       (AXD た)
       (PU 。))
  (ID 1_ex1640391709;JP))
```

Open Jtalk [12]

Accents → \

PPhrase → [ ]
PClause → { }

{[sonokokuo\oniwa][[fUtari\ no][o\ojiga]] [arima\shI ta].}

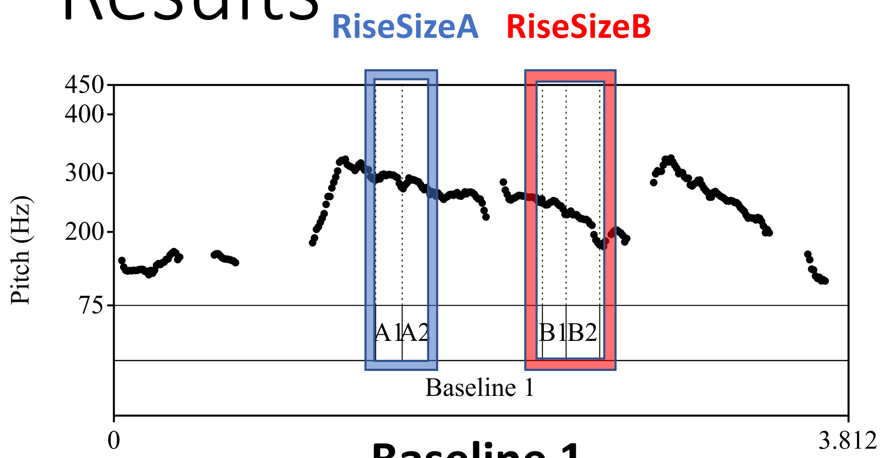
5,453 sentences were used for training, and 250 each were used for validation and testing.

Proposed model

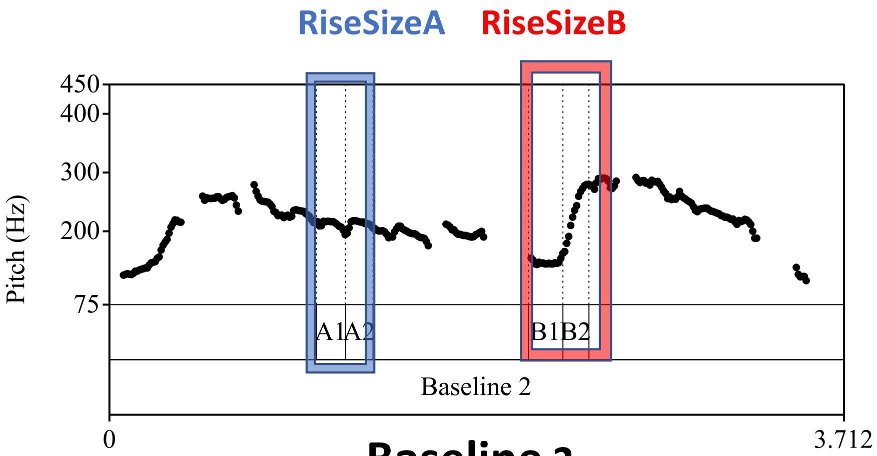
- The database consists of an oral transcription of the Arabian Nights and its reading voice by a single speaker [13].
- Japanese Tacotron 2 [14] generated a mel-spectrum, which is converted to waveforms via Griffin-Lim in ESPNet2 [15].
- 5,453 sentences were used for training, and 250 each were used for validation and testing.

[13] Takehazuchi
[14] Wang et al., 2017
[15] Watanabe et al., 2018

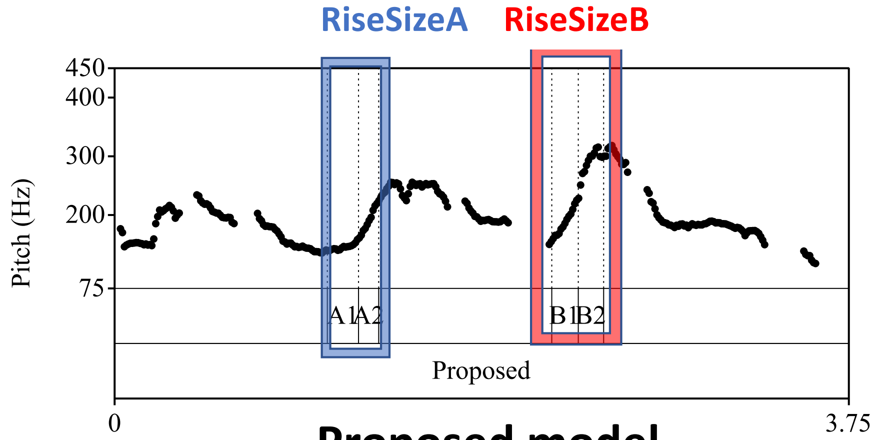
Results



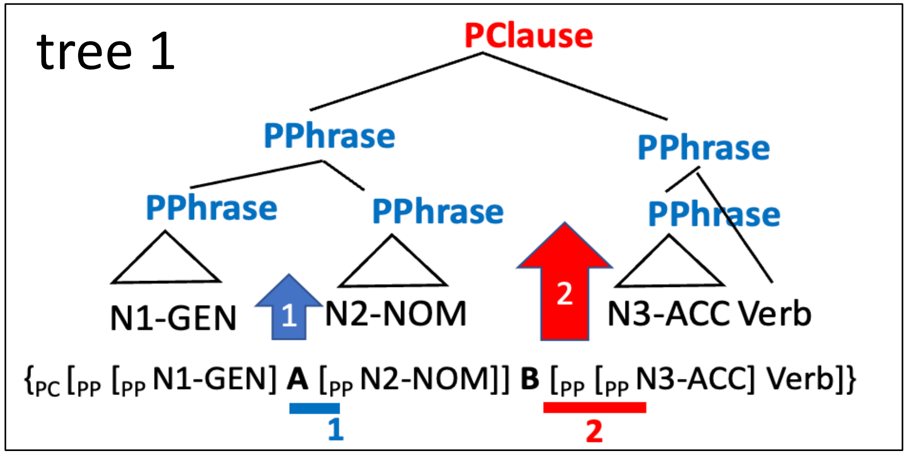
Baseline 1



Baseline 2



Proposed model



Results

- In natural prosody, RiseSizeB is greater than RiseSizeA in tree 1, while RiseSizeA is greater than RiseSizeB in tree 2 [9]
- The proposed model and Baseline 2 showed the same pattern as the natural prosody reported earlier [9]

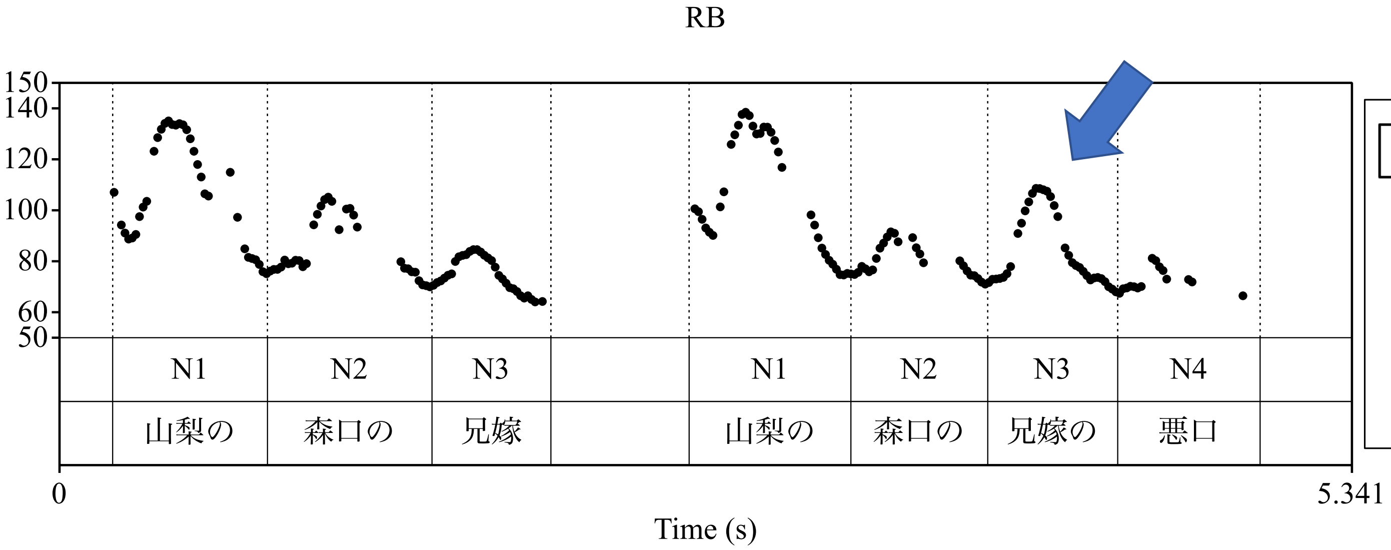
model	sentence	cond	RiseSizeA	RiseSizeB	Same pattern as natural prosody?
baseline 1	1	tree 1	0.68	-0.26	No
baseline 2	1	tree 1	1.75	12.12	Yes
proposed	1	tree 1	8.17	11.84	Yes
baseline 1	1	tree 2	0.72	0.51	Yes
baseline 2	1	tree 2	14.17	3.50	Yes
proposed	1	tree 2	11.96	1.56	Yes
baseline 1	2	tree 1	2.17	8.06	Yes
baseline 2	2	tree 1	4.32	10.58	Yes
proposed	2	tree 1	6.12	9.39	Yes
baseline 1	2	tree 2	0.50	-2.27	No
baseline 2	2	tree 2	12.35	2.71	Yes
proposed	2	tree 2	9.30	9.07	Yes

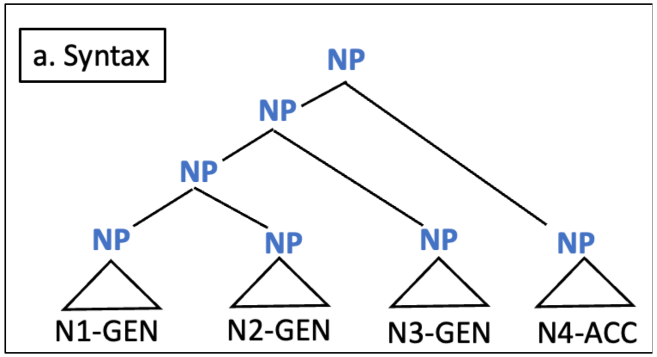
[9] Selkirk et al., 2013

Phenomenon 2: Rhythmic boost

- Rhythmic boost
  - F0 is boosted on the third word in four-word sequences
  - but not in three-word sequences [16, 17]

[16] Kubozono, 1989
[17] Shinya et al., 2004

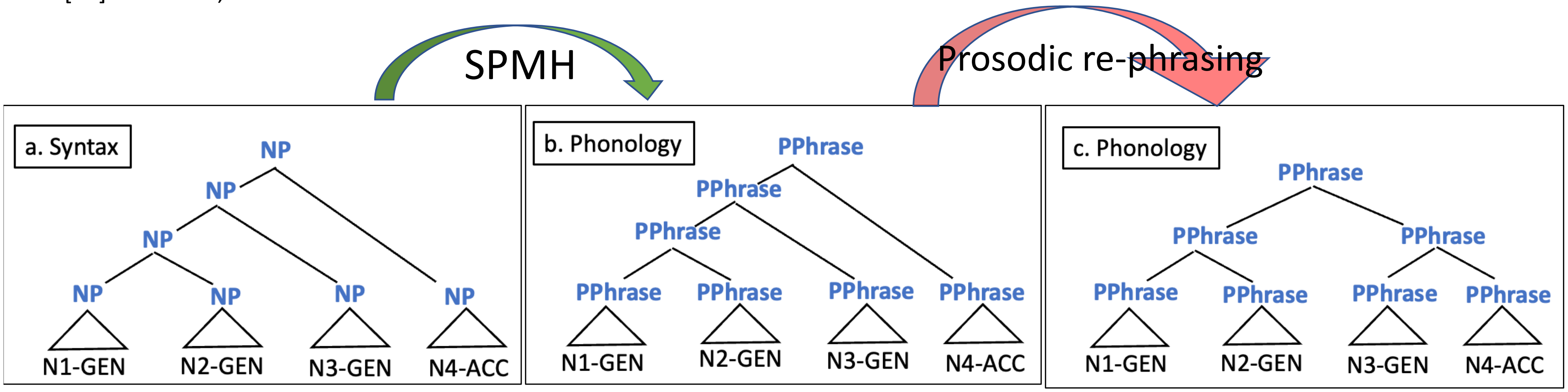




Phenomenon 2: Rhythmic boost

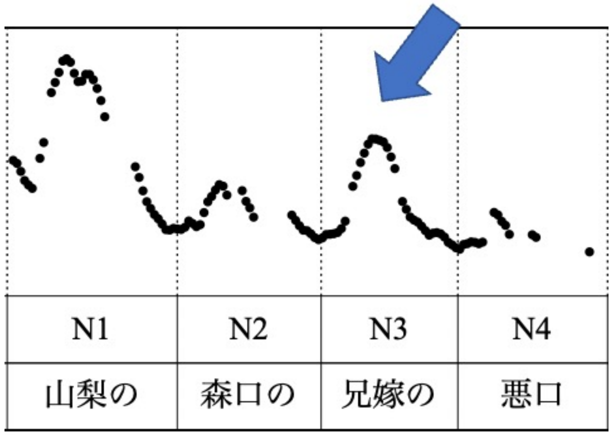
- Due to Syntax-Prosody Mapping Hypothesis, a left-branching phonological structure is predicted
- However, a prosodic well-formedness constarint triggers phonological re-phrasing [16]
- prosodically re-phrased as two intermediate PPhrases (MiPs) recursively dominating two minimal PPhrases (PPs) each

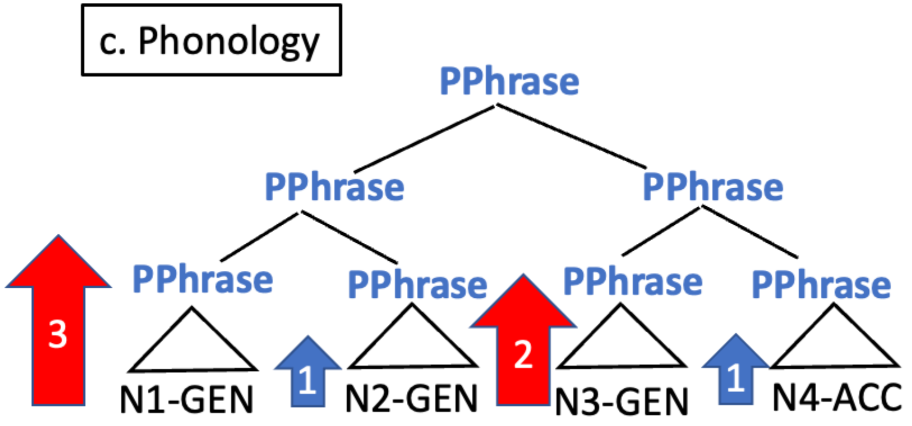
[16] Kubozono, 1989



Phenomenon 2: rhythmic boost

- F0 is boosted on the third word in four-word sequences [16, 17]
- The results can be explained via
  - syntax–prosody mapping hypothesis
  - phonological re-phrasing
  - Edge boost hypothesis (proposal): assuming that **<u>the number of edges</u>** in the PPhrases is proportional to the degree of the pitch increase in the initial lowering



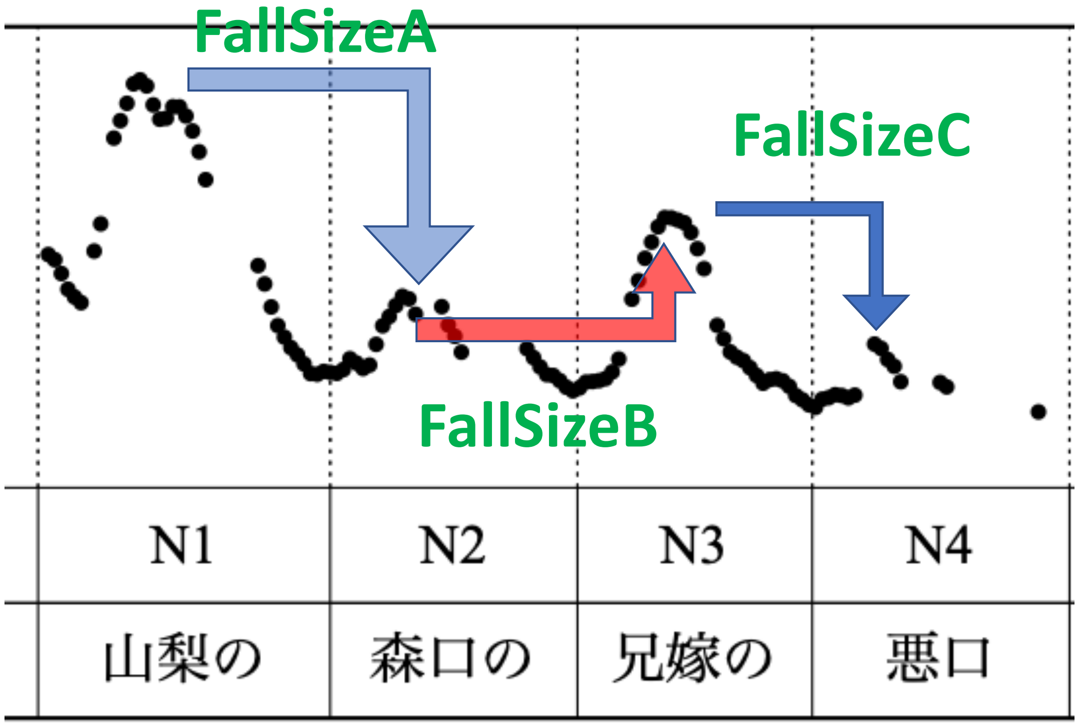


b. Phonology: [$_{\text{PP}}$ [$_{\text{PP}}$ [$_{\text{PP}}$ N1-GEN] [$_{\text{PP}}$ N2-GEN]] [$_{\text{PP}}$ [$_{\text{PP}}$ N3-GEN] [$_{\text{PP}}$ N4-ACC]]]

3 1 2 1

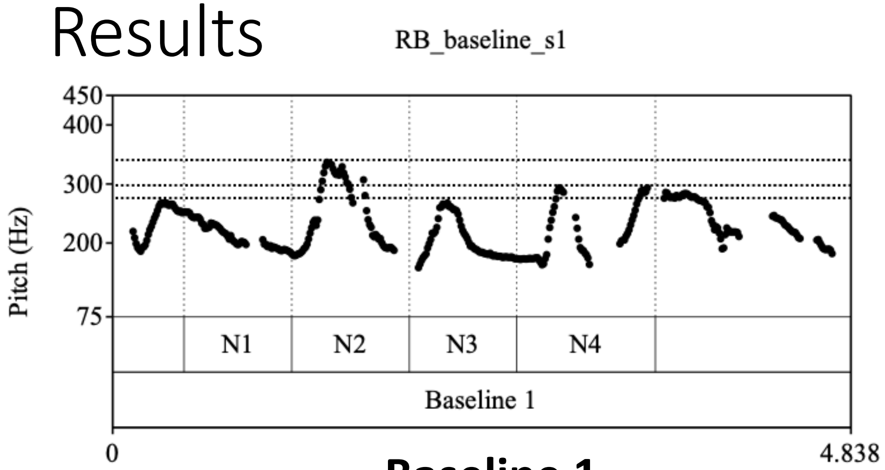
[16] Kubozono, 1989
[17] Shinya et al., 2004

Measurements

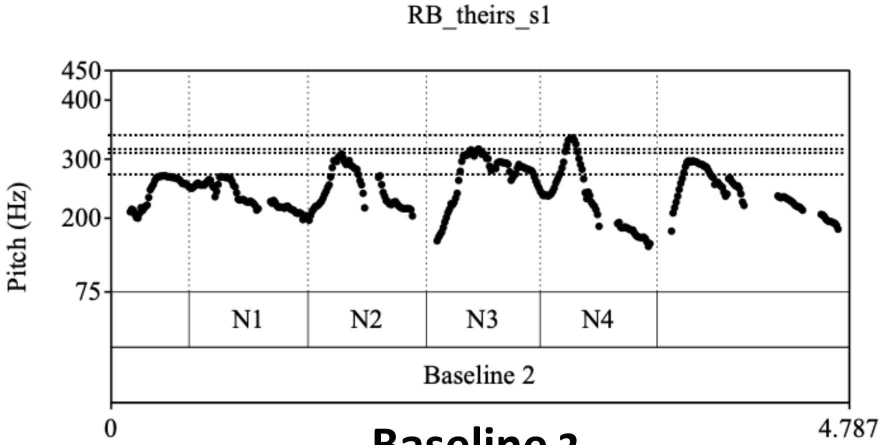


- Measurements (in semitones)
  - FallSizeA = maximum F0 of N2 minus maximum F0 of N1
  - FallSizeB = maximum F0 of N3 minus maximum F0 of N2
  - FallSizeC = maximum F0 of N4 minus maximum F0 of N3
- In natural speech,
  - FallSizeA becomes negative
  - FallSizeB approaches zero or becomes positive
  - FallSizeC becomes negative

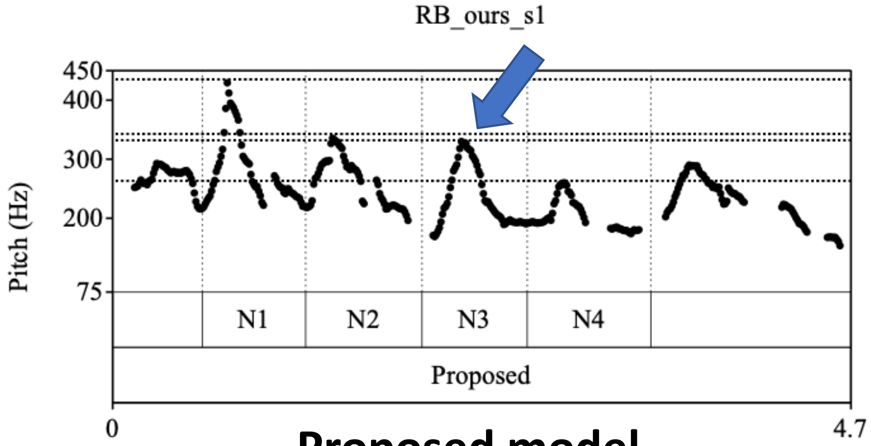
Results



Baseline 1

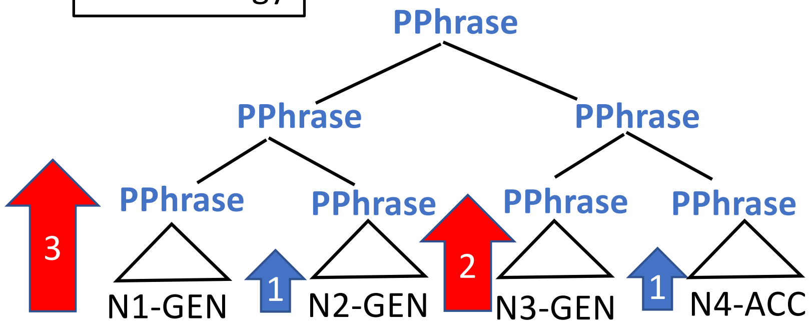


Baseline 2



Proposed model

c. Phonology



b. Phonology: [$_{PP}$ [$_{PP}$ [$_{PP}$ N1-GEN] [$_{PP}$ N2-GEN]] [$_{PP}$ [$_{PP}$ N3-GEN] [$_{PP}$ N4-ACC]]]

Results

- In natural speech,
  - FallSizeA becomes negative
  - FallSizeB approaches zero or becomes positive
  - FallSizeC becomes negative
- Only the proposed model showed the same patterns as those of natural language

model	sentence	FallSizeA	FallSizeB	FallSizeC	Same pattern as natural prosody?
baseline 1	1	4.95	-4.02	1.57	No
baseline 2	1	2.25	0.53	1.00	No
proposed	1	-3.79	-0.41	-3.98	Yes
baseline 1	2	1.29	-2.71	-2.24	No
baseline 2	2	0.62	2.60	-4.54	No
proposed	2	-2.84	1.94	-1.12	Yes

Discussions and conclusion

- We applied linguistic theories to TTS
- The proposed method was able to reproduce not only syntactic but also phonological phenomena
    - **Phenomenon 1. initial lowering**
    - **Phenomenon 2. rhythmic boost**
- The proposed method efficiently synthesizes phonological phenomena in the test data that were not explicitly included in the training data
- The proposed method is applicable to Japanese and other languages

Thank you

Part of this work is supported by JSPS KAKENHI Grant Number JP21H05054.

Selected References

1. J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, “Natural TTS synthesis by conditioning wavenet on Mel Spectrogram predictions,” *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.
2. Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, “Investigation of enhanced Tacotron text-to-speech synthesis systems with self- attention for pitch accent language,” in *Proceedings of ICASSP*, 2019, pp. 6905–6909.
3. T. Fujimoto, K. Hashimoto, K. Oura, Y. Nankaku, K. Tokuda “Impacts of Input Linguistic Feature Representation on Japanese End-to-End Speech Synthesis,” Proc. of 10th ISCA Speech Synthesis Workshop (SSW), pp.166-171, Vienna, Austria, Sep. 2019
4. K. Kurihara, N. Seiyama, and T. Kumano, “Prosodic features control by symbols as input of sequence-to-sequence acoustic modeling for neural tts,” *IEICE Transactions on Information and Systems*, vol. E104.D, no. 2, pp. 302–311, 2021.
5. H. Guo, F. K. Soong, L. He, and L. Xie, “Exploiting syntactic features in a parsed tree to improve end-to-end TTS,” *Interspeech 2019*, 2019.
6. N. Kaiki, S. Sakti, and S. Nakamura, “Using local phrase dependency structure information in neural sequence-to-sequence speech synthesis,” *2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, 2021.
7. J. Pierrehumbert, and M. Beckman, “Japanese tone structure,” Cambridge: MIT Press, 1988.
8. Y. Igarashi, “13 Intonation,” *Handbook of Japanese Phonetics and Phonology*, pp. 525–568, 2015.

Selected References

9. E. Selkirk, T. Shinya, and M. Sugahara, “Degree of initial lowering in Japanese as a reflex of prosodic structure organization,” Proc. 15th ICPhS, Barcelona, Spain, 491-494, 2003.

10. E. Selkirk, “The syntax-phonology interface,” *The Handbook of Phonological Theory*, pp. 435–484, 2011.

11. S. W. Horn, A. Butler, and K. Yoshimoto, “Keyaki Treebank segmentation and part-of speech labelling,” In Proceedings of the 23th Meeting of the Association for Natural Language Processing, pages 414–417, 2017.

12. *Open JTalk*. [Online]. Available: http://open-jtalk.sourceforge.net/. [Accessed: 20-Mar-2022].

13. Takehazuchi, "On the reading of relieving stories Arabian Nights oral translation," https://o-keil.com/okinu-ba- ba/wordpress/?p=818 (in Japanese)

14. Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, “Tacotron: Towards end-to-end speech synthesis,” *Interspeech 2017*, 2017.

15. S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, “ESPnet: End-to-end speech processing toolkit,” *Interspeech 2018*, 2018.

16. H. Kubozono, “Syntactic and rhythmic effects on downstep in Japanese,” *Phonology*, *6*(1), pp. 39-67, 1989.

17. T. Shinya, E. Selkirk, and S. Kawahara, “Rhythmic boost and recursive minor phrase in Japanese,” In *Speech Prosody 2004, International Conference*, 2004.