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
furukawa.kei.fi4@is.naist.jp, kishiyama.t@gmail.com, s-nakamura@is.naist.jp

Summary

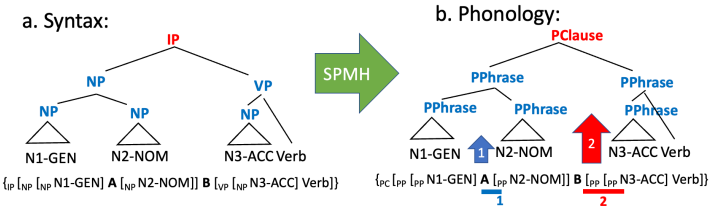
- We applied linguistic theories to neural sequence-to-sequence speech synthesis
- The proposed method was able to reproduce not only syntactic but also phonological phenomena
  - Phenomenon 1. initial lowering
  - Phenomenon 2. rhythmic boost

Backgrounds

- Advances of End-to-end text-to-speech synthesis (TTS) [1]
- Other studies
  - Incorporated information of the post-lexical level, such as syntactic structure and syntactic dependency information [2]
- However,
  - Not objectively examined whether they can reproduce pitch patterns of phonological phenomena
- This study aims to reproduce speech sounds of syntactic and phonological phenomena

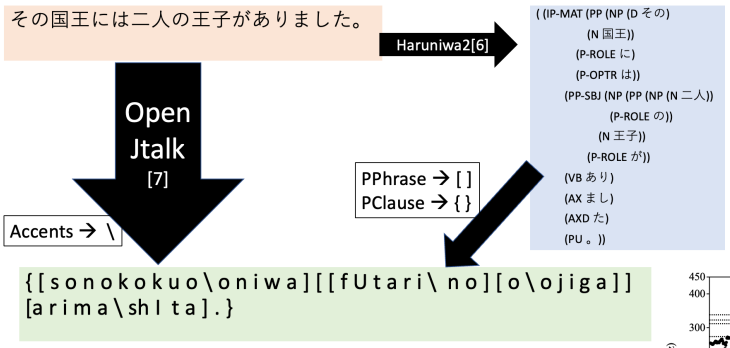
Phenomenon 1. initial lowering

- The initial lowering is the F0 rise at the beginning of a PPhrase [3]
- The degree of pitch increase in initial lowering varies in response to syntactic structure [4]



- Syntax–prosody mapping hypothesis (SPMH) [5]
  - **Syntactic clause** → be mapped to **PClause**
  - **Syntactic phrase such as NP, VP** → be mapped to **PPhrase**
- Edge boost hypothesis (proposal): assuming that the number of edges in the PPhrases is proportional to the degree of the pitch increase in the initial lowering

Proposed method



tree 1	N1	A	N2	B	N3	
item	imamurasan-no		imagawayaki-ga		omoide-o	tsukurimashita.
tone	LHHHHH		LHHHHH		LHHH	LHHH*LL
gloss	Mr. Imamura-GEN		Japanese.muffin-NOM		memory-ACC	made
	'Mr. Imamura's muffin made memories.'					

tree 2	N1	A	N2	B	N3	
item	imamurasan-ga		imagawayaki-no		omoide-o	tsukurimashita.
tone	LHHHHH		LHHHHH		LHHH	LHHH*LL
gloss	Mr. Imamura-NOM		Japanese.muffin-GEN		memory-ACC	made
	'Mr. Imamura made a memory about muffins.'					

baseline 1 (phonemes and accents)
tree 1: i m a m u r a s a N n o i m a g a w a y a k i g a o m o i d e o t s U k u r i m a sh I t a .
tree 2: i m a m u r a s a N g a i m a g a w a y a k i n o o m o i d e o t s U k u r i m a sh I t a .
baseline 2 (phonemes, accents, initial lowering, and dependency length)
tree 1: i / m a m u r a s a N n o #1 i / m a g a w a y a k i g a #2 o / m o i d e o #1 t s U / k u r i m a \ sh I t a .
tree 2: i / m a m u r a s a N g a #3 i / m a g a w a y a k i n o #1 o / m o i d e o #1 t s U / k u r i m a \ sh I t a .
proposed (phonemes, accents, and phonological structures)
tree 1: {[[i m a m u r a s a N n o] [i m a g a w a y a k i g a]] [[o m o i d e o] [t s U k u r i m a \ sh I t a]] .}
tree 2: {[i m a m u r a s a N g a] [[[i m a g a w a y a k i n o] [o m o i d e o]] [t s U k u r i m a \ sh I t a]] .}

Experimental settings

- The database consists of an oral transcription of the Arabian Nights and its reading voice by a single speaker
- Japanese Tacotron 2 [8] generated a mel-spectrum, which is converted to waveforms via Griffin-Lim in ESPNet2 [9]
- 5,453 sentences for training, and 250 each for validation and testing

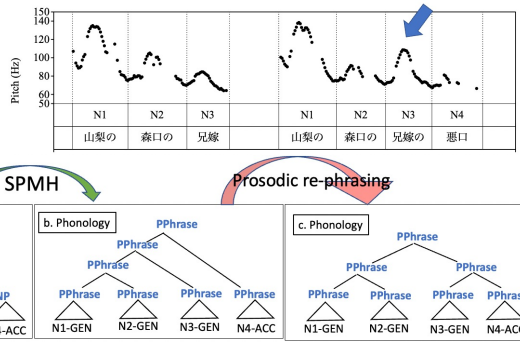
Result of Exp.1

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

- The proposed model and Baseline 2 showed the same pattern as the natural prosody reported earlier [9]

Phenomenon 2. Rhythmic boost

- Rhythmic boost
  - F0 is boosted on the third word in four-word sequences [10]
  - But not in three-word sequences [10, 11]
- A prosodic well-formedness constarint triggers phonological re-phrasing [10]

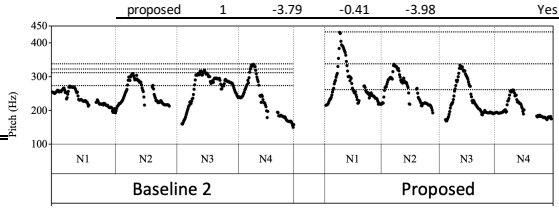


Proposed method and Results of Exp.2

input		N1	N2	N3	N4		
item	kinou	yamanashi-no	moriguchi-no	aniyome-no	waruguchi-o	kouen-de	tsutaeta
tone	LHH	LH*LLL	LH*LLL	LH*LLL	LH*LLL	HHHHH	LHHHH
gloss	yesterday	Yamanashi-GEN	Moriguchi-GEN	sister.in.law-GEN	bad things-ACC	park-in	tell
	'Yesterday, I said the bad things about the sister-in-law of Moriguchi in Yamanashi in the park.'						

baseline 1 (phonemes and accents)
4N: k i n o o y a m a \ n a sh i n o m o r i \ g u ch i n o a n i \ y o m e n o w a r u \ g u ch i o k o o e N d e t s U t a e t a .
baseline 2 (phonemes, accents, initial lowering, and dependency length)
4N: k i / n o o #6 y a / m a \ n a sh i n o #1 m o / r i \ g u ch i n o #1 a / n i \ y o m e n o #1 w a / r u \ g u ch i o #2 k o / o e N d e #1 t s U / t a e t a .
proposed (phonemes, accents, and phonological structures)
4N: {[k i n o o] [[[[y a m a \ n a sh i n o] [m o r i \ g u ch i n o]] [[a n i \ y o m e n o] [w a r u \ g u ch i o]]] [k o o e N d e] [t s U t a e t a]] .}

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



- Only the proposed model showed the same patterns as those of natural language [11]

General Discussions

- The proposed method was able to reproduce not only syntactic but also phonological phenomena
- The proposed model efficiently synthesizes phonological phenomena in the test data that were not explicitly included in the training data
- The proposed method is applicable to other languages

Selected References

1. Shen et al., "Natural TTS synthesis by conditioning wavenet on Mel Spectrogram predictions," *ICASSP*, 2018.
2. Kaiki et al., "Using local phrase dependency structure information in neural sequence-to-sequence speech synthesis," *O-COCOSDA*, 2021.
3. Pierrehumbert and M. Beckman, "Japanese tone structure," Cambridge: MIT Press, 1988.
4. Selkirk et al., "Degree of initial lowering in Japanese as a reflex of prosodic structure organization," ICPhS, 2003.
5. Selkirk, "The syntax-phonology interface," *The Handbook of Phonological Theory*, pp. 435-484, 2011.
6. Horn et al., "Keyaki Treebank segmentation and part-of speech labelling," In Proceedings of the 23th Meeting of the Association for Natural Language Processing 2017.
7. Open JTalk. [Online]. Available: http://open-jtalk.sourceforge.net/. [Accessed: 20-Mar-2022].
8. Wang et al., "Tacotron: Towards end-to-end speech synthesis," *Interspeech 2017*, 2017.
9. Watanabe et al., "ESPnet: End-to-end speech processing toolkit," *Interspeech 2018*, 2018.
10. Kubozono, "Syntactic and rhythmic effects on downstep in Japanese," *Phonology*, 6(1), 1989.
11. Shinya et al., "Rhythmic boost and recursive minor phrase in Japanese," In *Speech Prosody 2004, International Conference*, 2004.