Thu-P-OS-9-1

Predicting VQVAE-based Character Acting Style from Quotation-Annotated Text for Audiobook Speech Synthesis

Wataru Nakata[1], Tomoki Koriyama[1], Shinnosuke Takamichi[1], Yuki Saito[1], Yusuke Ijima[2], Ryo Masumura[2], Hiroshi Saruwatari[1]

1. The University of Tokyo, 2. NTT Corporation

Synopsis: Speech synthesis model which can "act" as a charactor

- **Audiobook speech synthesis**
  - Aims to automate the audiobook creation
  - Reduces the cost of the creation

Ant: "..."
Ant girl: "..."
Ant: "..."
Character's voice
Ant: "..."
Ant girl: "..."
Ant: "..."
Speech synthesis model

- **Challenges in audiobook speech synthesis**
  - Reflecting context from multiple sentences [1]
  - **Achieving character acting (this work)**

- **Our proposed model**
  - Extracts character acting style using VQVAE-based model [2]
  - Predicts character acting style using attention-based model
  - Introduces use of character embedding [3] in speech synthesis

- **Experimental results**
  - Comparable naturalness at MOS on chapter level samples
  - Significant improvement on character distinction

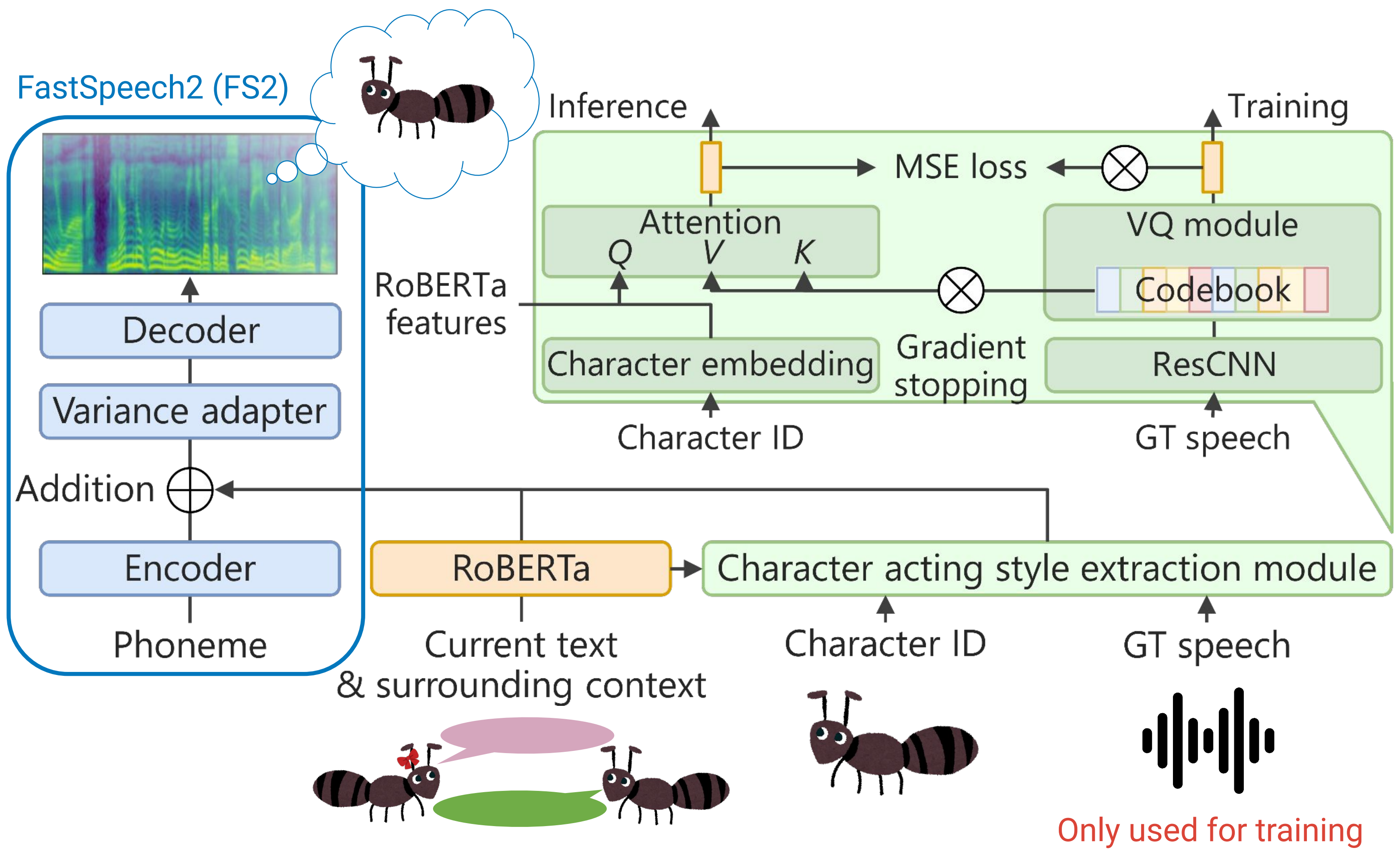
Speech samples😍

https://wataru-nakata.github.io/is2022-audiobook/

Proposed model

Input: quotation-annotated text

Character	Phrase
(Narration)	The foremost ant said.
Ant	"The other day, there were chocolates, and ice cream..."
Ant girl	"Yes. Human children from school had dropped them..."
Ant	"I wonder if there were any treats today, too."



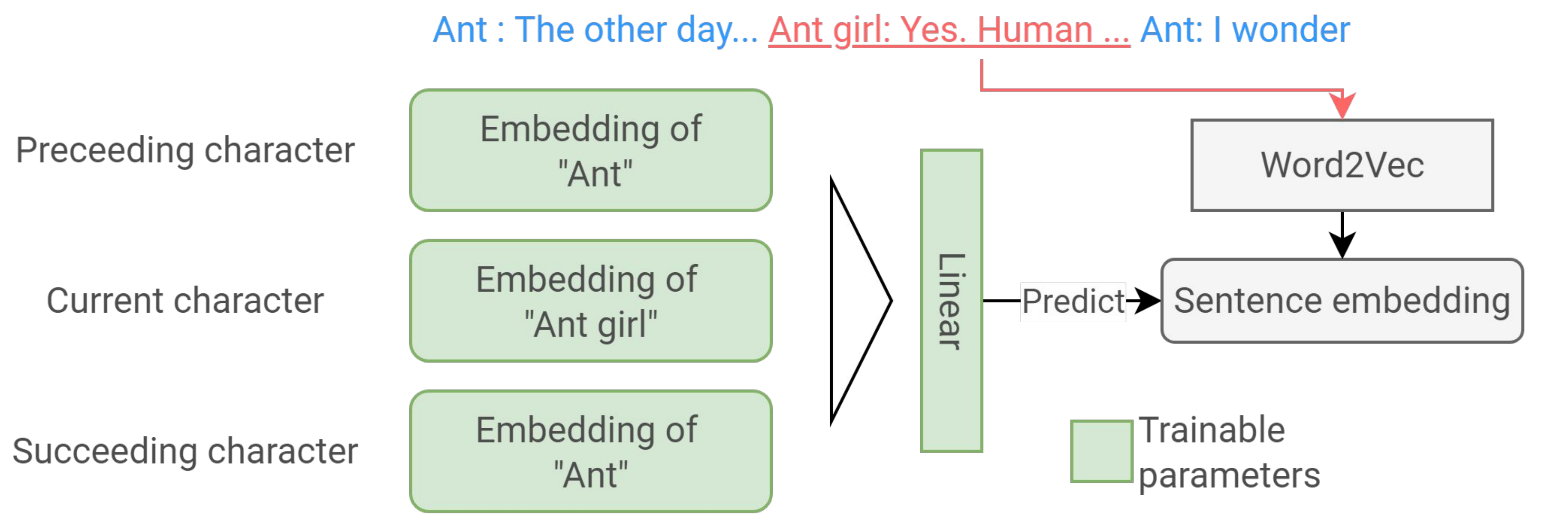
Only used for training

Character acting style extraction module (Green box)

- Extracts time-invariant feature using ResCNN [4]
- Learns discrete set of character acting styles as VQ codebook
- Predicts proper character acting from RoBERTa [5] features and character embedding using attention

Character embedding (Skip-Gram) [3]

- Proposed in NLP field to represent movie characters
- Trains character embedding from quotation-annotated text



Experiment

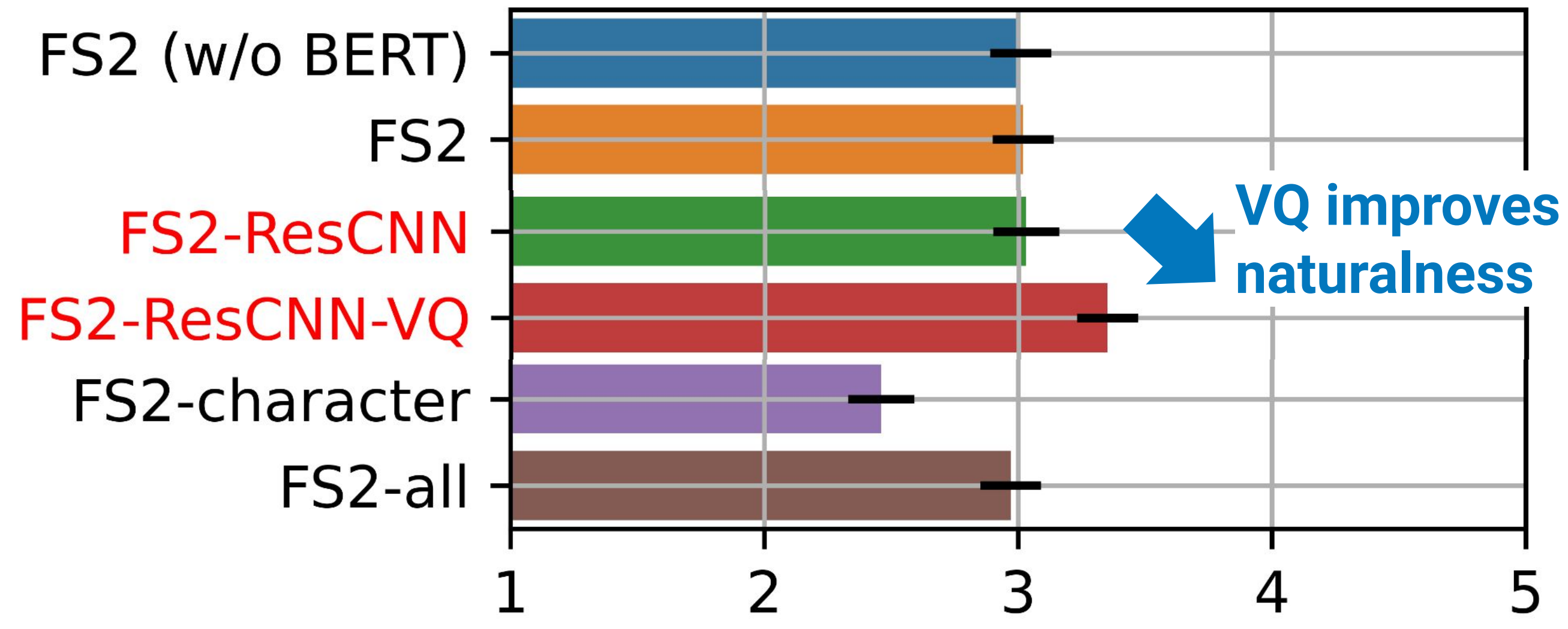
Compared models

Model name	RoBERTa	ResCNN	Vector Quantization	Character embedding
FS2 (w/o BERT)				
FS2	✅			
FS2-ResCNN	✅	✅		
FS2-ResCNN-VQ	✅	✅	✅	
FS2-character	✅			✅
FS2-all (proposed)	✅	✅	✅	✅

- All models were trained using J-KAC corpus [1] (expressive audiobook speech)
- RoBERTa-base model published by rinna Co., Ltd. RoBERTa trained on Japanese
- **Red models** takes ground truth speech at inference

Naturalness MOS of synthetic speech

- 120 raters with each rater evaluated 12 samples
- Chapter level (2-5 sentences) evaluations



Similar naturalness as baseline

Character distinction

- AB tests with 60 raters and each rater evaluated 12 samples
- Chapter level (2-5 sentences) evaluations
- Quotation-annotated texts are provided to the raters
- Raters selected sample with more appropriate distinction

Is ResCNN feature useful? Yes!

FS2	**0.32 vs. 0.68**	FS2-ResCNN

Does VQ worsen the result? No!

FS2-ResCNN-VQ	0.48 vs. 0.51	FS2-ResCNN

Is surrounding context effective? Yes!

FS2 (w/o BERT)	**0.44 vs. 0.56**	FS2

Does predicting character acting style work better? Yes!

FS2-character	**0.43 vs. 0.57**	FS2-all
FS2	**0.40 vs. 0.60**	FS2-all

Is the character acting style prediction perfect? No😢

FS2-all	**0.39 vs. 0.61**	FS2-ResCNN-VQ

Bold values represent significant ($p < 0.05$) results

Predicting character acting style is an effective approach

References

[1] W. Nakata et al., Proc. SSW11, 2021
[2] A. van den Oord et al., Proc. NIPS, 2017
[3] M. Azab et al., Proc. 23rd CoNLL, 2019
[4] C. Li et al., arXiv, 2017
[5] Y. Liu et al., Proc. ICLR, 2020
[6] Y. Ren et al., Proc. ICLR, 2021