Personalized filled-pause generation with group-wise prediction models

MOONSHOT RESEARCH & DEVELOPMENT PROGRAM
AVATAR SYMBIOTIC SOCIETY

Graduate School of Information Science and Technology, The University of Tokyo, Japan.
Yuta Matsunaga, Takaaki Saeki, Shinnosuke Takamichi, and Hiroshi Saruwatari

@YutaMResearch
My page

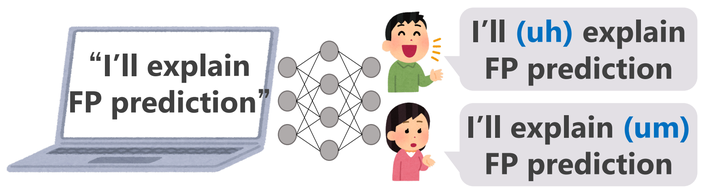
1. Overview

Personalized disfluent text generation

- Application: avatars speaking instead of humans
  - Personalization: reproduce speakers' individuality.
  - Disfluency: reproduce human-like disfluency.
- Recent reading-style speech synthesis:
  - Synthesize only fluent speech.
- Personalized disfluent speech synthesis (our goal)
  - **Personalized disfluent text generation ← this work**
- We focus on one kind of disfluency, filled pauses (FPs).
  - Ex. I'll (uh) explain FP prediction.

Filled pauses (FPs): one kind of disfluency

- Roles: help speech generation [1] and communication [2].
- Diversity:
  - FP words (160 in Japanese [3])
  - Difference among speakers
- Features: position and word



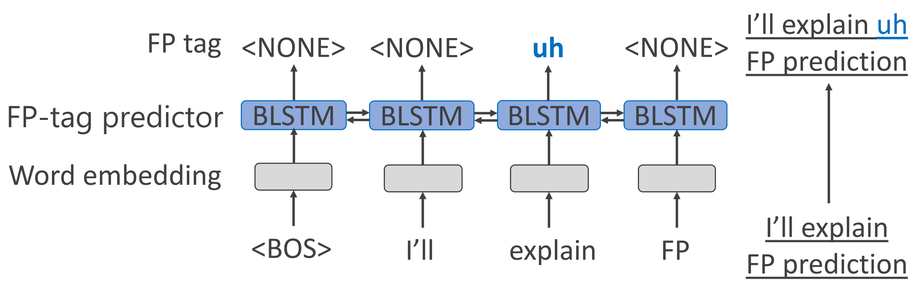
Personalized filled-pause prediction

This work

- Personalized FP generation by grouping speakers
- Improvement of prediction performance

2. Method

Basic architecture of FP prediction model [5]



Weighted cross entropy loss [6]

- Data imbalance:
  - It is harder to predict less frequent words.
- Increase weights of the loss of less frequent FP words.

Rich word representation model

- Use BERT [8] as rich word representation model.

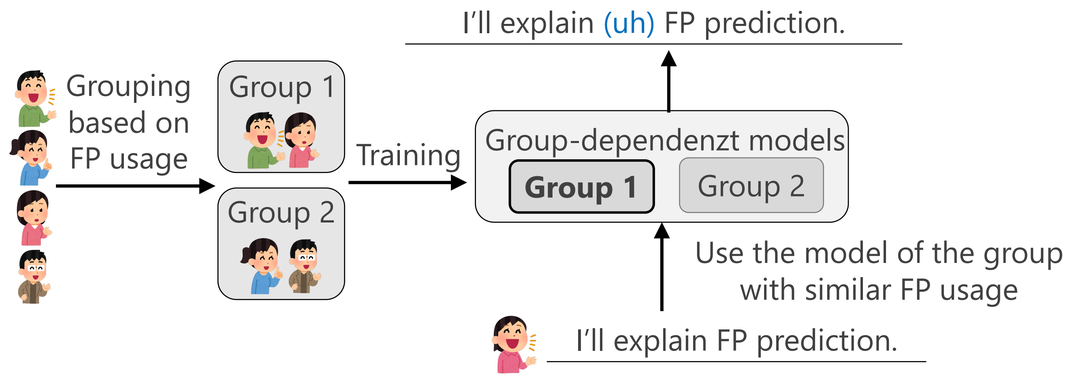
Personalized FP prediction model.

Case 1: group-dependent

- Advantage: not need to train a prediction model for each speaker.
- Training: train a model of each group in multi-speaker spontaneous speech corpus.
- Inference: use the model of the group closest to the target speaker's FP usage.

Case 2: speaker-dependent

- Train a model for each speaker in Japanese lecture spontaneous speech corpus.



3. Experiment

Experimental conditions

- Criteria: precision / recall / F score / (specificity)
- Cross validation

Dataset		137 speakers in CSJ [9]
		JLecSponSpeech [10]
Tokenization		Juman++ [11] (Sudachi [12] for fastText)
Word embedding		BERT (pretrained)
Prediction	Model	BLSTM
	Input	Morphemes
	Output	14 classes (none or 13 FP words)

Weighted cross entropy loss

- Using weighted loss improves the performance.

Rich word representation model

- BERT performs better than fastText.

Group-dependent models

- Higher scores than the universal model for both position and word, except for group 2 for position

Criterion	Universal	Group-dependent model							
		Grouping by word frequency				Grouping by position*			
		1	2	3	4	1	2	3	4
Position	0.376	**0.454**	**0.456**	**0.427**	**0.390**	**0.461**	0.323	**0.413**	**0.444**
Word	0.089	**0.284**	**0.288**	**0.248**	**0.196**	**0.277**	**0.212**	**0.158**	**0.237**

Position:
- Consider 4 positions:

1) head of the sentence,
2) boundary of breath group,
3) middle of breath group,
4) end of the sentence,

- Divide speakers by the ratio of FP at each position out of all FP.

- Prediction for each FP
  - Improves diversity of FP words.

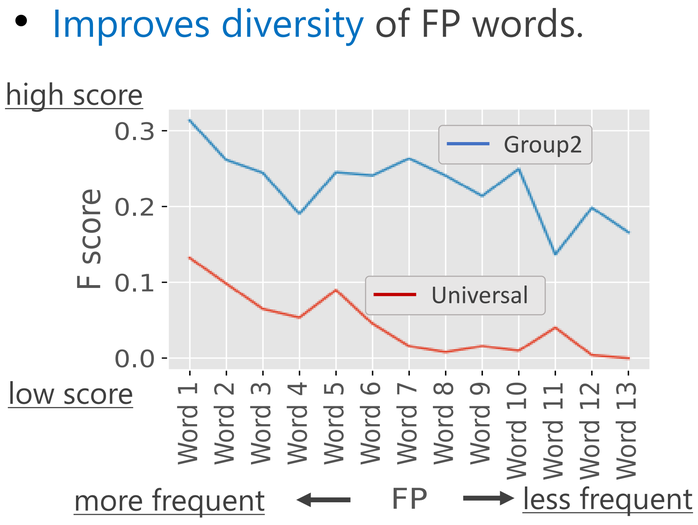


Fig.: F scores of prediction for each FP word

- Prediction for each speaker
  - Differs among speakers.

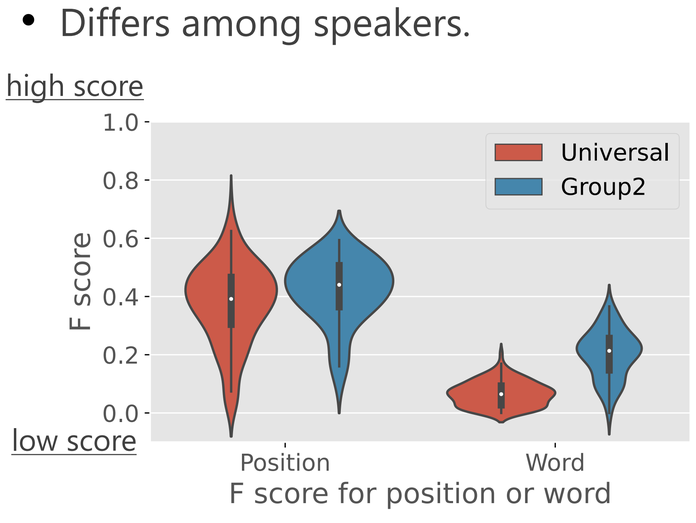


Fig.: Distribution of F scores of prediction for each speaker

Speaker-dependent models

- Speaker models have lower scores.
  - Speaker adaptation is difficult.
- Discussion

★The universal model has higher scores than group-dependent models.

- For the prediction for such speakers, we can use the universal model.

Tab.: F scores of FP position for speakers in Japanese lecture spontaneous speech corpus by FP prediction models

	Universal	Group (word)	Group (position)	Speaker
Spk. A	**0.243**	0.137	0.114	0.146
Spk. B	**0.384**	0.302	0.366	0.212

★

References

[1] W. J. Levelt, Cognition, 1983.
[2] A. Gravano, et al., Computer Speech & Language, 2011.
[3] K. Hirose, et al., in Proc. Speech Prosody, 2006.
[4] K. Ohta, et al., in Proc. INTERSPEECH, 2007.
[5] Y. Yamazaki, et al., in Proc. GCCE, 2020.
[6] Y. Yan, et al., in Proc. INTERSPEECH, 2021.
[7] P. Bojanowski, et al., TACL, 2017.
[8] J. Devlin, et al., arXiv, 2019.
[9] K. Maekawa, in Proc. SSPR, 2003.
[10] Y. Matsunaga, et al., in Proc. SLP, 2022.
[11] H. Morita, et al., in Proc. EMNLP, 2015.