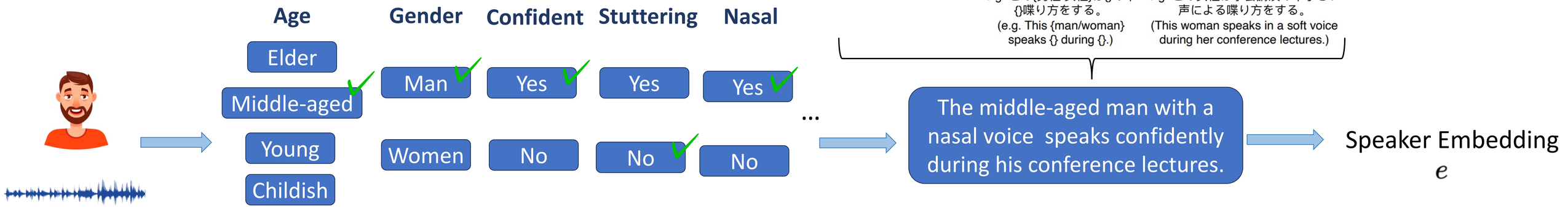
Generating Speakers by Prompting Listener Impressions for Pre-trained Multi-Speaker Text-to-Speech Systems

Zhengyang Chen[1,2], Xuechen Liu[2], Erica Cooper[2], Junichi Yamagishi[2], Yanmin Qian[1]

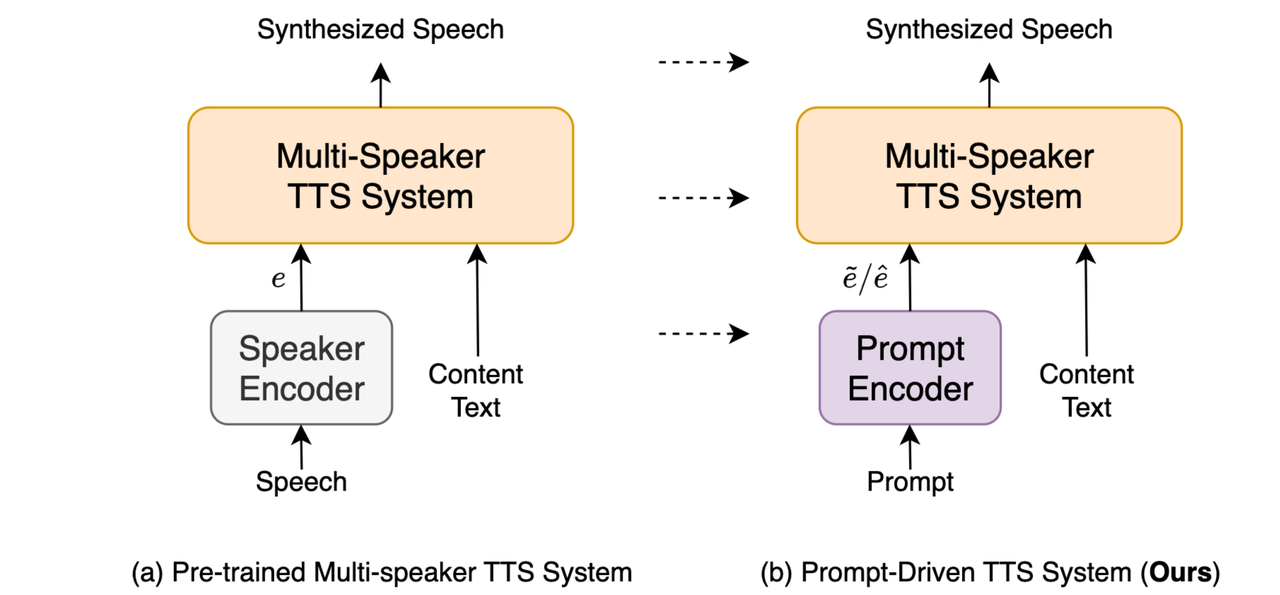
[1]AudioCC Lab, CS Dept, Shanghai Jiao Tong University, Shanghai, China

[2]National Institute of Informatics, Tokyo, Japan

➢ Prompting Listener Impressions

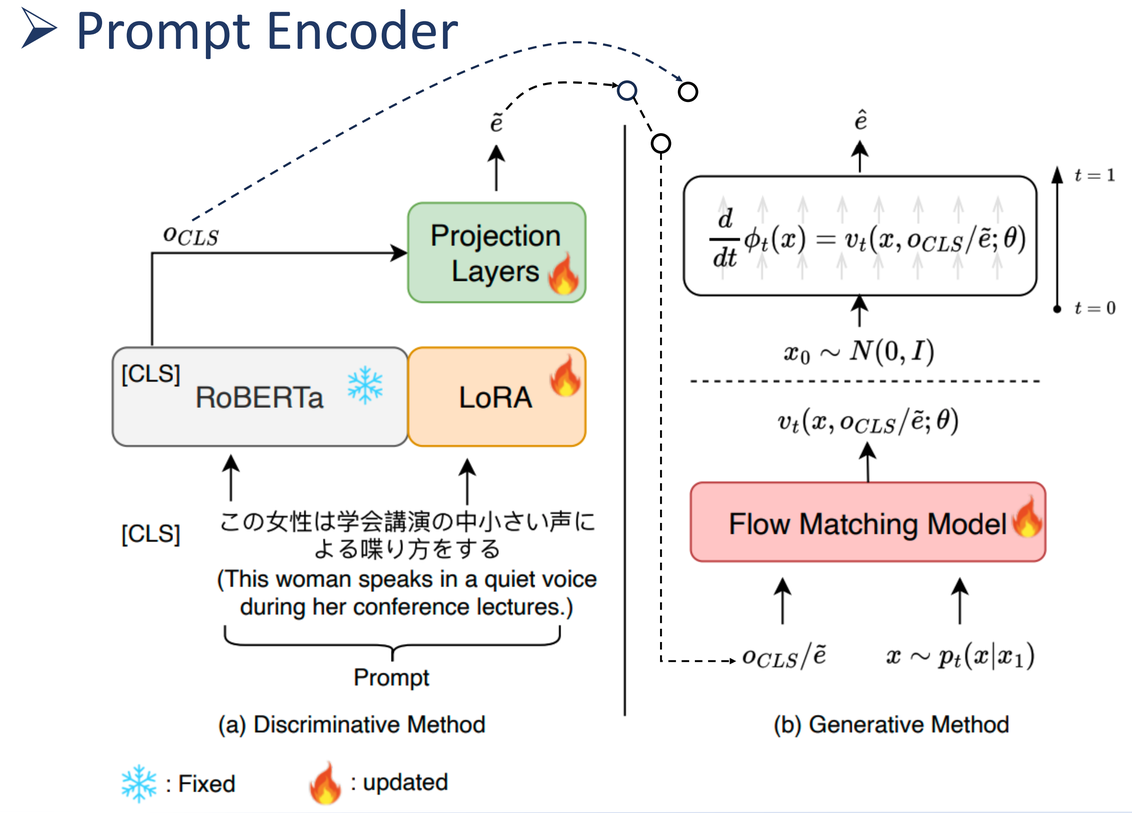


➢ System Overview



(a) Pre-trained Multi-speaker TTS System (b) Prompt-Driven TTS System (**Ours**)

➢ Prompt Encoder



(a) Discriminative Method (b) Generative Method

➢ Experimental Setup

- ❑ Dataset: Japanese CSJ dataset
  - ❑ Speaker attributes #: 26
  - ❑ Train: 2672 speakers
  - ❑ Test: 30 speakers
- ❑ Objective Evaluation:
  - ❑ FAD score
- ❑ Subjective Evaluation:
  - ❑ Naturalness MOS
  - ❑ Speaker attribute impression MOS on a 5-point rating scale (each sample is rated 8 times by different raters)

➢ Objective Evaluation and Naturalness Moss

Table 2: ***FAD score and Naturalness MOS results on the CSJ evaluation set.***

System	FAD Score	Naturalness MOS
ground-truth	-	4.06 ± 0.25
Discriminative (w/o LoRA)	11.217	3.15 ± 0.25
Discriminative (w/ LoRA)	5.244	3.45 ± 0.19
Flow-Matching (w/ LoRA)	3.559	3.52 ± 0.26
Discriminative + Flow-Matching	3.126	3.50 ± 0.24

- ❑ The generative method significantly contributes to producing high-fidelity speech

➢ Correlation between the Impression MOS from Reference Speech and Generated Speech

Table 1: ***Spearman Rank Correlation Coefficient (SRCC) between MOS scores from reference and synthesized speech.***

Scenario	System	Speaker Attribute									
		expressiveness	confidence	relaxation	voice_depth	age	energy	pitch	speed	clarity	Avg
Seen	Discriminative (w/o LoRA)	0.72	0.53	0.48	0.75	0.86	0.71	0.89	0.89	0.23	0.67
	Discriminative (w/ LoRA)	0.71	0.69	0.65	0.83	0.90	0.76	0.94	0.85	0.37	**0.74**
	Flow-Matching (w/ LoRA)	0.68	0.53	0.66	0.76	0.79	0.50	0.86	0.38	0.22	0.60
	Discriminative + Flow-Matching	0.74	0.71	0.75	0.87	0.96	0.72	0.90	0.68	0.35	**0.74**
Unseen	Discriminative (w/o LoRA)	0.04	0.05	0.46	0.38	0.67	0.29	0.73	0.57	-0.37	0.31
	Discriminative (w/ LoRA)	0.54	0.38	0.49	0.48	0.77	0.25	0.81	0.36	0.41	**0.50**
	Flow-Matching (w/ LoRA)	-0.10	0.12	0.32	0.42	0.82	0.39	0.74	0.14	0.21	0.34
	Discriminative + Flow-Matching	0.36	0.08	0.49	0.35	0.74	0.34	0.75	0.37	0.20	0.41

underline: The statistical significance (p-value) is less than 0.001, indicating the MOS scores of synthetic speech are significantly correlated with the MOS scores of reference speech.

- ❑ The Discriminative method enables the model to better follow the prompt's description
- ❑ The generative method significantly contributes to producing high-fidelity speech
- ❑ Combining both methods can achieve good results in terms of both speech quality and adherence to the prompt.