P4-19-TTS

Cross-Dialect Text-To-Speech in Pitch-Accent Language Incorporating Multi-Dialect Phoneme-Level BERT

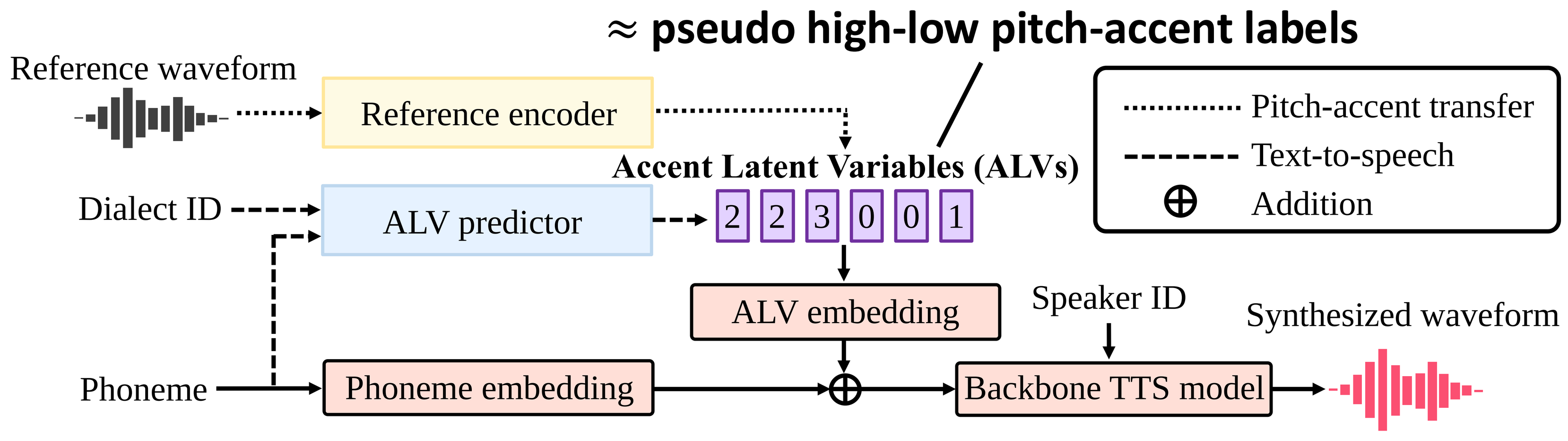
Kazuki Yamauchi, Yuki Saito, Hiroshi Saruwatari

The University of Tokyo, Japan

Contributions: Toward natural speech communication with computers across regions

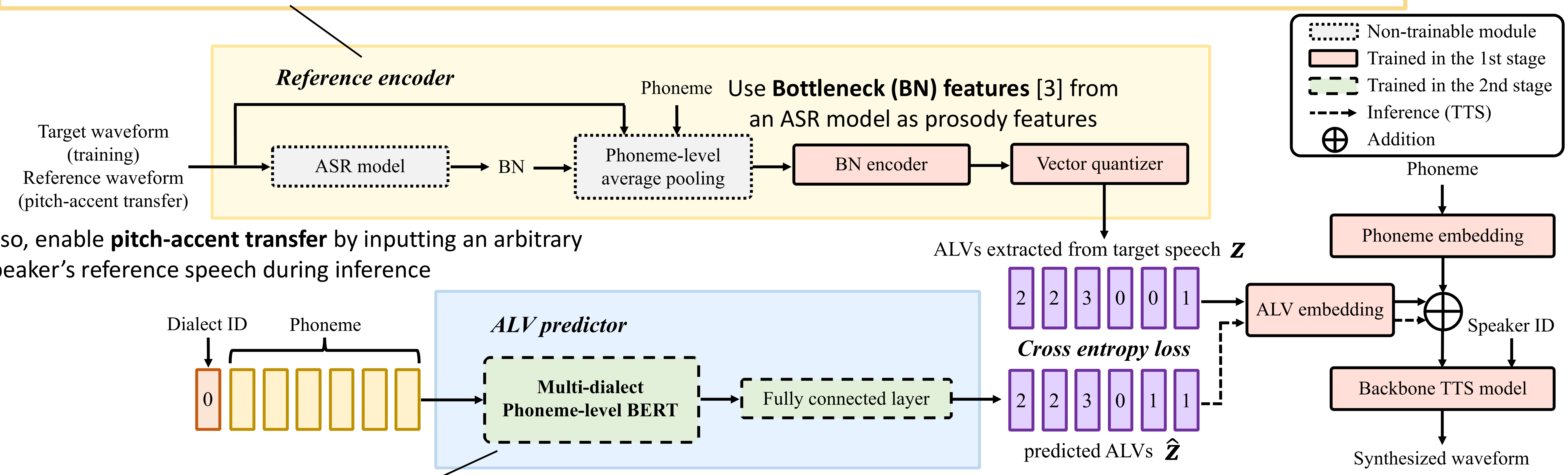
- **Contribution1**: Explore a novel task, **cross-dialect text-to-speech (CD-TTS)**
  - Synthesize speech in a **dialect different from the target speaker's native dialect**
  - Different from intra-dialect TTS (ID-TTS), synthesizing speech in the native dialect
  - Localize TTS systems by **adapting the pitch-accent to regional dialects**
- **Contribution2**: Propose a novel TTS model for CD-TTS
  - Automatically predict **accent latent variables (ALVs)** [1] tailored to each dialect
    - ALV: Phoneme-level quantized latent variables, acquired from speech in data-driven **without relying on accent dictionaries not available in dialects**
  - Propose a dialect-adapted version of phoneme-level BERT (PL-BERT) [2], **multi-dialect (MD)-PL-BERT**, to improve the accuracy of ALV prediction

Alternative to accent labels not available in dialects ≈ **pseudo high-low pitch-accent labels**



Proposed method: Comprising of backbone TTS model, reference encoder, and ALV predictor

Reference encoder extracts **ALVs** from reference speech
- Quantize speech prosody features into four classes (Note: Japanese pitch-accent is considered to have four levels)



Also, enable **pitch-accent transfer** by inputting an arbitrary speaker's reference speech during inference

ALV predictor predicts **ALVs** tailored to a target dialect from an input text, leveraging **multi-dialect phoneme-level BERT (MD-PL-BERT)**
- Construct a multi-dialect text corpus by leveraging the **data augmentation through dialect translation using an LLM**
- Train PL-BERT on the large-scale multi-dialect text corpus, **conditioning it on dialect ID**

Experimental evaluation: Synthesizing speech in Osaka-dialect, one of Japanese dialects

Experimental conditions

Dataset	• **JSUT** [4]: Tokyo-dialect speech corpus • **JMD** [5]: Osaka-dialect speech corpus • **CPJD** [6]: Osaka-dialect speech corpus (CPJD is used only for evaluation)
Pre-training MD-PL-BERT	• Japanese Wikipedia corpus • Transcriptions in ReazonSpeech
Model settings	• TTS model: FastSpeech 2 [7] • Vocoder: HiFi-GAN [8] UNIVERSAL V1 • ASR model: Whisper large-v2 [9] • LLM: Japanese Llama 2, Swallow 13B
Compared methods	• **FS2**: Original FastSpeech 2 • **FS2-AP**: Proposed method using ALVs **predicted by the ALV Predictor** • **FS2-REF**: Proposed method using ALVs **extracted by the reference encoder**
Evaluation metrics	• **N-MOS**: Naturalness of speech • **D-MOS**: **Dialectal naturalness (dialectality) of pitch-accent**

Results of subjective evaluation

ID-TTS: Osaka-dialect speech by **Osaka-dialect speaker**

Method	Speaker	N-MOS (↑)	D-MOS (↑)
FS2	JMD (Osaka)	3.30 ± 0.12	3.22 ± 0.13
FS2-AP	JMD (Osaka)	3.31 ± 0.13	3.26 ± 0.12
FS2-REF	JMD (Osaka)	3.23 ± 0.12	3.30 ± 0.12
REF	CPJD (Osaka)	3.89 ± 0.14	4.38 ± 0.09

CD-TTS: Osaka-dialect speech by **Tokyo-dialect speaker**

Method	Speaker	N-MOS (↑)	D-MOS (↑)
FS2	JSUT (Tokyo)	3.57 ± 0.13	2.62 ± 0.13
FS2-AP	JSUT (Tokyo)	3.52 ± 0.13	3.00 ± 0.15
FS2-REF	JSUT (Tokyo)	3.58 ± 0.12	3.05 ± 0.14
REF	CPJD (Osaka)	4.39 ± 0.10	4.32 ± 0.13

35 native Japanese speakers each evaluated 24 samples

- Results:
  - **No performance degradation** in ID-TTS
  - **D-MOS significantly improved** in CD-TTS
  - Pitch-accent transfer by an **unseen speaker improved D-MOS**

Prosody feature: F0 vs. BN

F0 can be used as a prosody feature for ALV extraction instead of BN features

A vs. B	Naturalness	Dialectality
F0 vs. BN	0.400 vs. 0.600	0.424 vs. 0.576

- **BN features** outperformed **F0**
  - F0: ***acoustic feature***
  - BN: ***linguistic feature*** acquired through the ASR task

Effectiveness of MD-PL-BERT

Compare FS2-AP with MD-PL-BERT and original PL-BERT w/o the data augmentation

A vs. B	Naturalness	Dialectality
PL-BERT vs. MD-PL-BERT	0.491 vs. 0.509	0.343 vs. 0.657

MD-PL-BERT significantly improved the dialectality of synthetic speech

References

[1] K. Yufune et al., in Proc. SSW, 2021. [2] Y. A. Li et al., in Proc ICASSP, 2023. [3] L. Ma et al., in Proc NCMMSC, 2023. [4] S. Takamichi et al., AST, 2020. [5] S. Takamichi et al., 2021. [6] S. Takamichi et al., in Proc LREC, 2018. [7] Y. Ren et al., in Proc. ICLR, 2021. [8] J. Kong et al., in Proc NeurIPS, 2020. [9] A. Radford et al., in Proc. ICML, 2023.

