

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

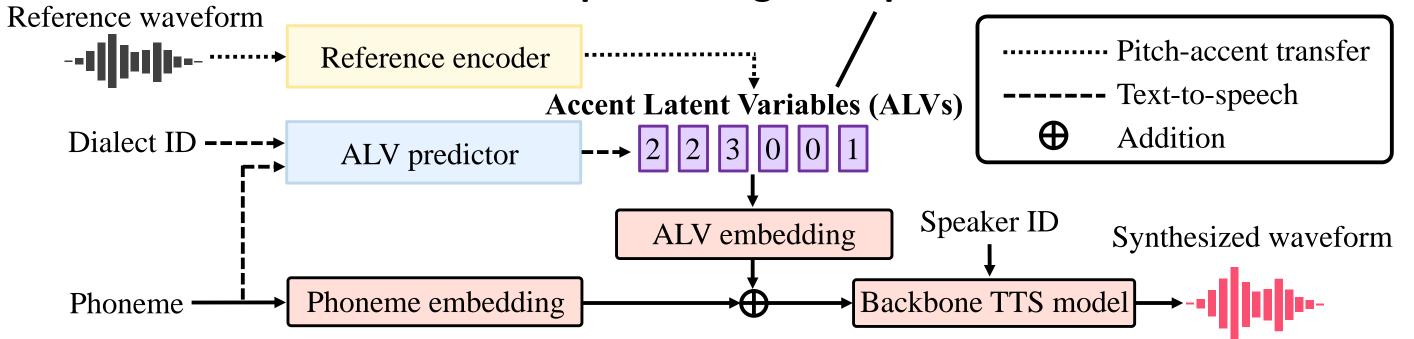
Kazuki Yamauchi, Yuki Saito, Hiroshi Saruwatari

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)
  - O Synthesize speech in a dialect different from the target speaker's native dialect R
  - O Different from intra-dialect TTS (ID-TTS), synthesizing speech in the native dialect
  - Localize TTS systems by adapting the pitch-accent to regional dialects
- <u>Contribution2</u>: Propose a novel TTS model for CD-TTS
  - O 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

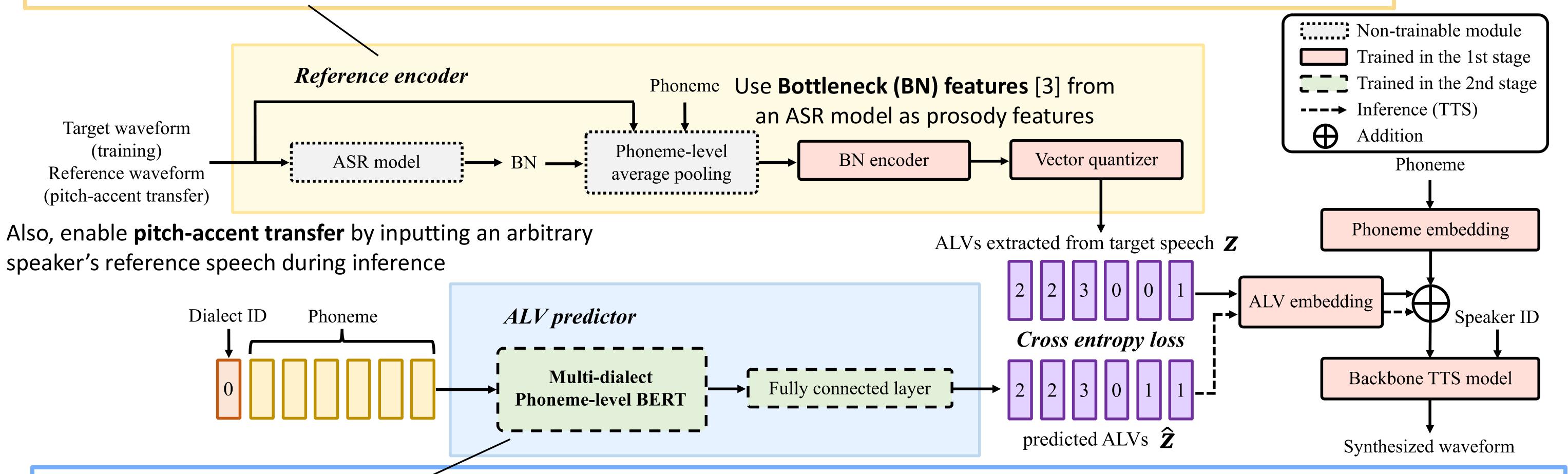
Alternative to accent labels not available in dialects  $\approx$  pseudo high-low pitch-accent labels



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

### **Reference encoder** extracts **ALVs** from reference speech

O Quantize speech prosody features into four classes (Note: Japanese pitch-accent is considered to have four levels)



ALV predictor predicts ALVs tailored to a target dialect from an input text, leveraging multi-dialect phoneme-level BERT (MD-PL-BERT)

- o 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

<b>Experimental conditions</b>			ective eva	luation	Prosody feature: F0 vs. BN		
<ul> <li>Dataset</li> <li>JSUT [4]: Tokyo-dialect speech corpus</li> <li>JMD [5]: Osaka-dialect speech corpus</li> <li>CPJD [6]: Osaka-dialect speech corpus (CPJD is used only for evaluation)</li> </ul>	ID-TTS: Osaka-dialect speech by Osaka-dialect speakerMethodSpeakerN-MOS (1)D-MOS (1)				F0 can be used as a prosody feature for ALV extraction instead of BN features		
	FS2 FS2-AP				A vs. B	Naturalness Dialect	,
<ul> <li>Japanese Wikipedia corpus</li> <li>Transcriptions in ReazonSpeech</li> </ul>					<ul> <li>BN features outperformed F0</li> <li>F0: acoustic feature</li> <li>BN: linguistic feature acquired through the ASR task</li> </ul>		
<ul> <li>Model settings</li> <li>TTS model: FastSpeech 2 [7]</li> <li>Vocoder: HiFi-GAN [8] UNIVERSAL V1</li> <li>ASR model: Whisper large-v2 [9]</li> <li>LLM: Japanese Llama 2, Swallow 13B</li> </ul>	Method	Speaker	N-MOS (1)	D-MOS (↑)			
	FS2-AP	JSUT (Tokyo)	$3.52 \pm 0.13$	3.00 ± 0.15			
<ul> <li>FS2: Original FastSpeech 2</li> <li>FS2-AP: Proposed method using ALVs predicted by the ALV Predictor</li> <li>FS2-REF: Proposed method using ALVs extracted by the reference encoder</li> </ul>	REF	CPJD (Osaka)	$4.39 \pm 0.10$	4.32 ± 0.13			
					A vs. B PL-BERT vs.		ctality
<ul> <li>N-MOS: Naturalness of speech</li> <li>D-MOS: Dialectal naturalness (dialectality) of pitch-accent</li> </ul>	<ul> <li>D-MOS significantly improved in CD-TTS</li> <li>Pitch-accent transfer by an unseen speaker improved D-MOS</li> </ul>				MD-PL-BERT <b>Significantly improved the</b> <b>dialectality</b> of synthetic speech		
	<ul> <li>JSUT [4]: Tokyo-dialect speech corpus</li> <li>JMD [5]: Osaka-dialect speech corpus</li> <li>CPJD [6]: Osaka-dialect speech corpus (CPJD is used only for evaluation)</li> <li>Japanese Wikipedia corpus</li> <li>Transcriptions in ReazonSpeech</li> <li>TTS model: FastSpeech 2 [7]</li> <li>Vocoder: HiFi-GAN [8] UNIVERSAL V1</li> <li>ASR model: Whisper large-v2 [9]</li> <li>LLM: Japanese Llama 2, Swallow 13B</li> <li>FS2: Original FastSpeech 2</li> <li>FS2-AP: Proposed method using ALVs predicted by the ALV Predictor</li> <li>FS2-REF: Proposed method using ALVs extracted by the reference encoder</li> <li>N-MOS: Naturalness of speech</li> <li>D-MOS: Dialectal naturalness</li> </ul>	<ul> <li>JSUT [4]: Tokyo-dialect speech corpus</li> <li>JMD [5]: Osaka-dialect speech corpus (CPJD [6]: Osaka-dialect speech corpus (CPJD is used only for evaluation)</li> <li>Japanese Wikipedia corpus</li> <li>Transcriptions in ReazonSpeech</li> <li>TTS model: FastSpeech 2 [7]</li> <li>Vocoder: HiFi-GAN [8] UNIVERSAL V1</li> <li>ASR model: Whisper large-v2 [9]</li> <li>LLM: Japanese Llama 2, Swallow 13B</li> <li>FS2-AP: Proposed method using ALVs predicted by the ALV Predictor</li> <li>FS2-REF: Proposed method using ALVs extracted by the reference encoder</li> <li>N-MOS: Naturalness of speech</li> <li>D-MOS: Dialectal naturalness</li> <li>Ib-TTS: O</li> <li>Method</li> <li>FS2</li> <li>FS2-REF</li> <li>REF</li> <li>O</li> <li>O</li> <li>Pir</li> </ul>	<ul> <li>JSUT [4]: Tokyo-dialect speech corpus</li> <li>JMD [5]: Osaka-dialect speech corpus</li> <li>(CPJD [6]: Osaka-dialect speech corpus</li> <li>(CPJD is used only for evaluation)</li> <li>Japanese Wikipedia corpus</li> <li>Transcriptions in ReazonSpeech</li> <li>TTS model: FastSpeech 2 [7]</li> <li>Vocoder: HiFi-GAN [8] UNIVERSAL V1</li> <li>ASR model: Whisper large-v2 [9]</li> <li>LLM: Japanese Llama 2, Swallow 13B</li> <li>FS2-AP: Proposed method using ALVs predicted by the ALV Predictor</li> <li>FS2-REF: Proposed method using ALVs predicted by the ALV Predictor</li> <li>FS2-REF: Proposed method using ALVs extracted by the reference encoder</li> <li>N-MOS: Naturalness of speech</li> <li>D-MOS: Dialectal naturalness</li> <li>ID-TTS: Osaka-dialect speech corpus</li> <li>Method Speaker</li> <li>FS2-REF JMD (Osaka)</li> <li>FS2-REF JMD (Osaka)</li> <li>CD-TTS: Osaka-dialect speech</li> <li>Method Speaker</li> <li>FS2-AP JSUT (Tokyo)</li> <li>FS2-REF JSUT (Tokyo)</li> <li>REF CPJD (Osaka)</li> <li>35 native Japanese speak</li> <li>Results:</li> <li>No performance</li> <li>D-MOS: Dialectal naturalness</li> <li>Wethod Speaker</li> <li>Pitch-accent trar</li> </ul>	<ul> <li>JSUT [4]: Tokyo-dialect speech corpus</li> <li>JMD [5]: Osaka-dialect speech corpus (CPJD [6]: Osaka-dialect speech corpus (CPJD is used only for evaluation)</li> <li>Japanese Wikipedia corpus</li> <li>Transcriptions in ReazonSpeech</li> <li>TTS model: FastSpeech 2 [7]</li> <li>Vocoder: HiFi-GAN [8] UNIVERSAL V1</li> <li>ASR model: Whisper large-v2 [9]</li> <li>LLM: Japanese Llama 2, Swallow 13B</li> <li>FS2-AP: Proposed method using ALVs predicted by the ALV Predictor</li> <li>FS2-REF: Proposed method using ALVs predicted by the ALV Predictor</li> <li>FS2-REF: Proposed method using ALVs extracted by the reference encoder</li> <li>N-MOS: Naturalness of speech</li> <li>D-MOS: Dialectal naturalness</li> <li>IB-TTS: Osaka-dialect speech by Osaka-dialect method</li> <li>Speaker</li> <li>Method</li> <li>Speaker</li> <li>N-MOS (1)</li> <li>FS2 JSUT (Tokyo)</li> <li>Staturalness of speech</li> <li>Pitch-accent transfer by an unstruction</li> </ul>	<ul> <li>JSUT [4]: Tokyo-dialect speech corpus JMD [5]: Osaka-dialect speech corpus (CPJD [6]: Osaka-dialect speech corpus (CPJD is used only for evaluation)</li> <li>Japanese Wikipedia corpus Transcriptions in ReazonSpeech</li> <li>TTS model: FastSpeech 2 [7]</li> <li>Vocoder: HiFi-GAN [8] UNIVERSAL V1 ASR model: Whisper large-v2 [9]</li> <li>LLM: Japanese Llama 2, Swallow 13B</li> <li>FS2-AP: Proposed method using ALVs predicted by the ALV Predictor</li> <li>FS2-REF: Proposed method using ALVs predicted by the ALV Predictor</li> <li>FS2-REF: Proposed method using ALVs extracted by the reference encoder</li> <li>N-MOS: Naturalness of speech</li> <li>D-MOS: Dialectal naturalness</li> <li>ID-TTS: Osaka-dialect speech by Osaka-dialect speaker</li> <li>Method Speaker N-MOS (↑) D-MOS (↑)</li> <li>FS2-REF JMD (Osaka) 3.30 ± 0.12 3.30 ± 0.12</li> <li>REF CPJD (Osaka) 3.89 ± 0.14 4.38 ± 0.09</li> <li>CD-TTS: Osaka-dialect speech by Tokyo-dialect speaker</li> <li>Method Speaker N-MOS (↑) D-MOS (↑)</li> <li>FS2-AP JSUT (Tokyo) 3.57 ± 0.13 2.62 ± 0.13 3.00 ± 0.15</li> <li>FS2-AP: JSUT (Tokyo) 3.52 ± 0.13 3.00 ± 0.15</li> <li>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</li> <li>Results:</li> <li>N-MOS: Naturalness of speech</li> <li>D-MOS: Dialectal naturalness</li> <li>O No performance degradation in ID-TTS O D-MOS significantly improved in CD-TTS</li> <li>Pitch-accent transfer by an unseen speaker</li> </ul>	JSUT [4]: Tokyo-dialect speech corpusID-TTS: Osaka-dialect speech by Osaka-dialect speekerF0 can be uJMD [5]: Osaka-dialect speech corpusMethod SpeakerN-MOS (†)D-MOS (†)extraction i(CPJD [6]: Osaka-dialect speech corpusF52JMD (Osaka)3.30 ± 0.123.22 ± 0.13A vs. B(CPJD is used only for evaluation)F52JMD (Osaka)3.23 ± 0.123.30 ± 0.12A vs. BJapanese Wikipedia corpusF52-APJMD (Osaka)3.23 ± 0.123.30 ± 0.12BN featuTranscriptions in ReazonSpeechCD-TTS: Osaka-dialect speech by Tokyo-dialect speakerO F0: auOcrtTS: Model: FastSpeech 2 [7]Wethod SpeakerN-MOS (†)D-MOS (†)BN featuVocoder: HiFi-GAN [8] UNIVERSAL V1ASR model: Whisper large-v2 [9]ES2-APJSUT (Tokyo)3.57 ± 0.132.62 ± 0.13the AFS2: Original FastSpeech 2F52-AP: Proposed method using ALVs extracted by the ALV PredictorFS2-REFJSUT (Tokyo)3.58 ± 0.123.05 ± 0.14Compare F3Stative Japanese of speechN-MOS: Naturalness of speechNo performance degradation in ID-TTSA vs. BPL-BERT vs. MD-PL-BERTN-MOS: Dialectal naturalnessO Pitch-accent transfer by an unseen speakerMD-PL-BERT vs.	ID-TTS: Osaka-dialect speech corpus JMD [5]: Osaka-dialect speech corpus (CPID [6]: Osaka-dialect speech corpus (CPID is used only for evaluation)F0 can be used as a prosody feature extraction instead of BN featuresJapanese Wikipedia corpus Transcriptions in ReazonSpeechF52: JMD (Osaka)3.30 ± 0.123.20 ± 0.12A vs. BNaturalnessNaturalnessDialectTTS model: FastSpeech 2 [7] Vocoder: HIFI-GAN [8] UNIVERSAL V1 ASR model: Whisper large-v2 [9] LLM: Japanese Llama 2, Swallow 13BEF2 -APJSUT (Tokyo)3.57 ± 0.132.62 ± 0.130.0 ± 0.10SPeakerN-MOS (1)P-MOS (1)F52: Original FastSpeech 2 F52-AP: Proposed method using ALVs extracted by the ALV PredictorSpeaker IS2-AFE: Proposed method using ALVs extracted by the reference encoderSpeaker N-MOS significantly improved in CD-TTS oPL-BERT vs. 0.491 vs. 0.5090.343 vsN-MOS: Naturalness of speech D-MOS: Dialectal naturalnessO No performance degradation in ID-TTS oPitch-accent transfer by an unseen speakerMD-PL-BERT significantly improved

#### 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.

Saruwatari Lab., The University of Tokyo, Japan.

IEEE SLT@Macau, China.

©Kazuki Yamauchi, Dec. 2024.