

☆山内 一輝, 齋藤 佑樹, 猿渡 洋 (東京大学)

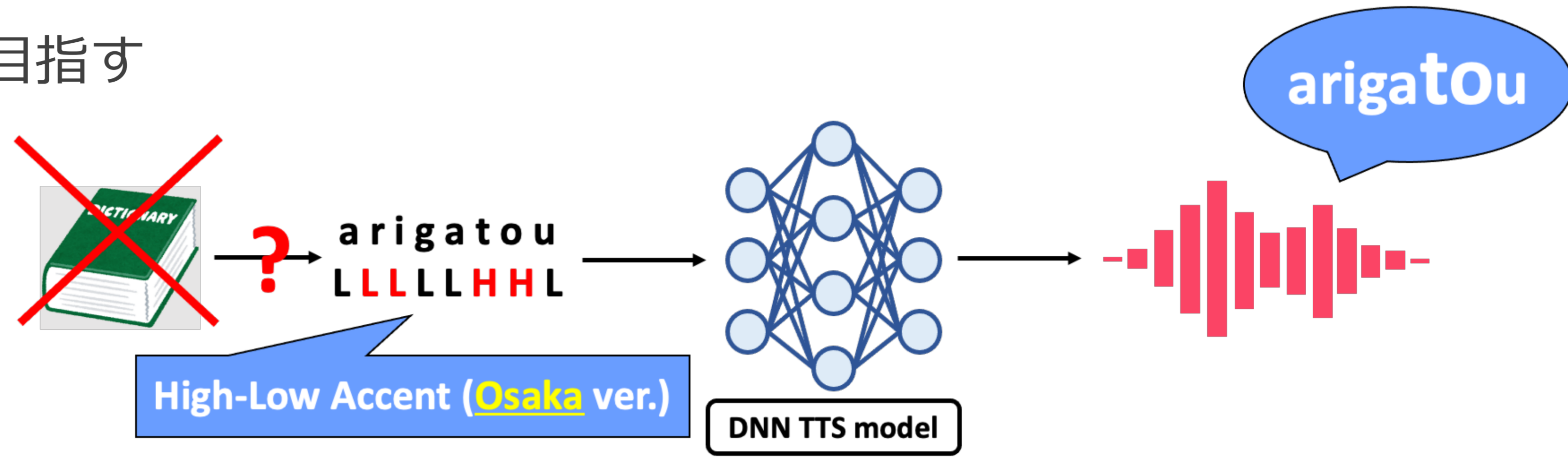
概要：方言音声合成の課題 & 提案手法

方言音声合成

- 標準語と異なる韻律体系をもつ方言の音声合成を目指す
- 課題 1: 話者数が限られた方言の**アクセント辞書不足**
- 課題 2: 十分な品質の方言音声収録は**困難**

提案手法

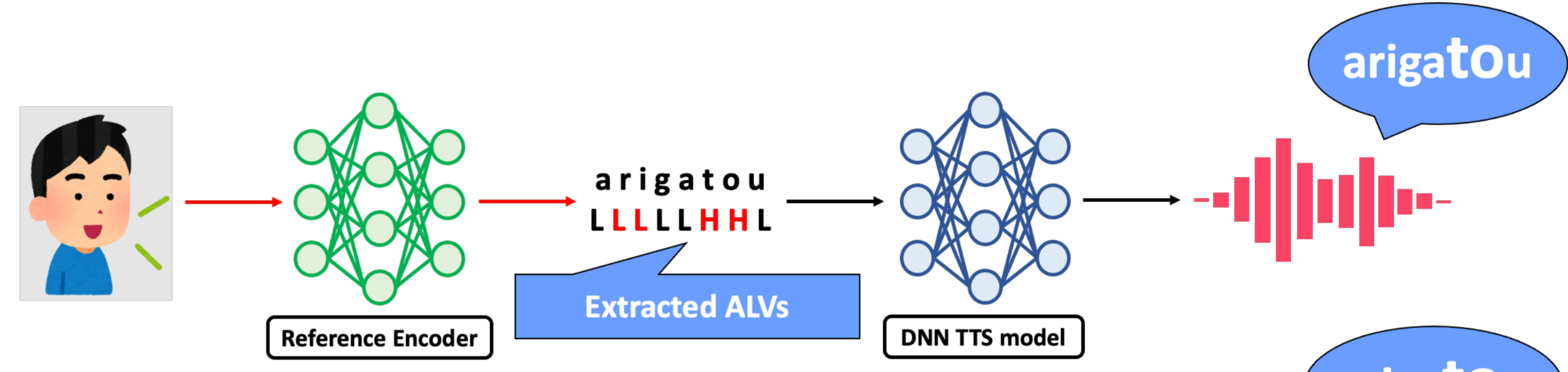
- テキストのみからの**アクセント潜在変数(ALV)予測**
- 音声からの**ALV自動抽出**による合成音声の韻律制御



関連研究 & 提案手法のコンセプト

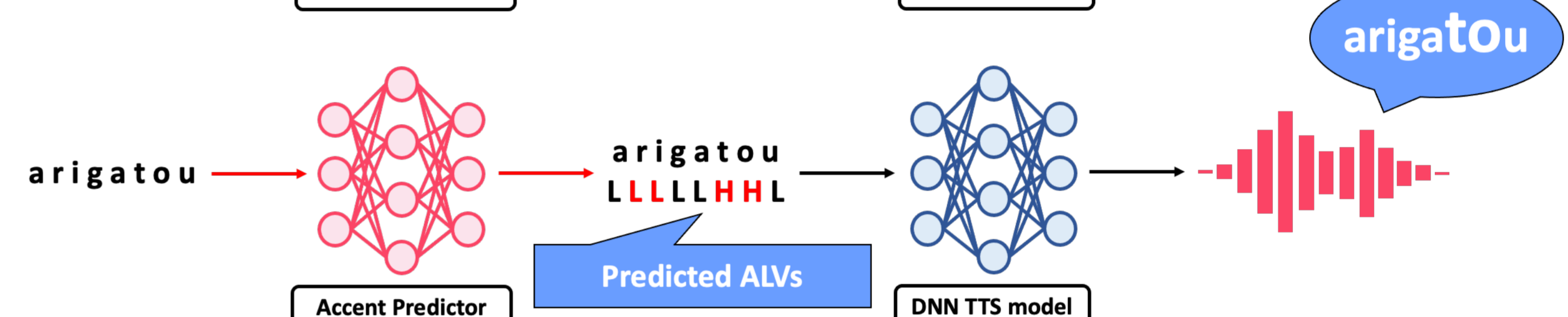
アクセント潜在変数(Accent Latent Variable; ALV)[1]

- 音声から自動でアクセント情報を抽出
- VQ-VAEで音声のF0を量子化された潜在変数(ALV)にエンコード
- ➡参照音声入力による**韻律制御(Prosody Transfer)**に利用



テキストのみからのアクセント予測

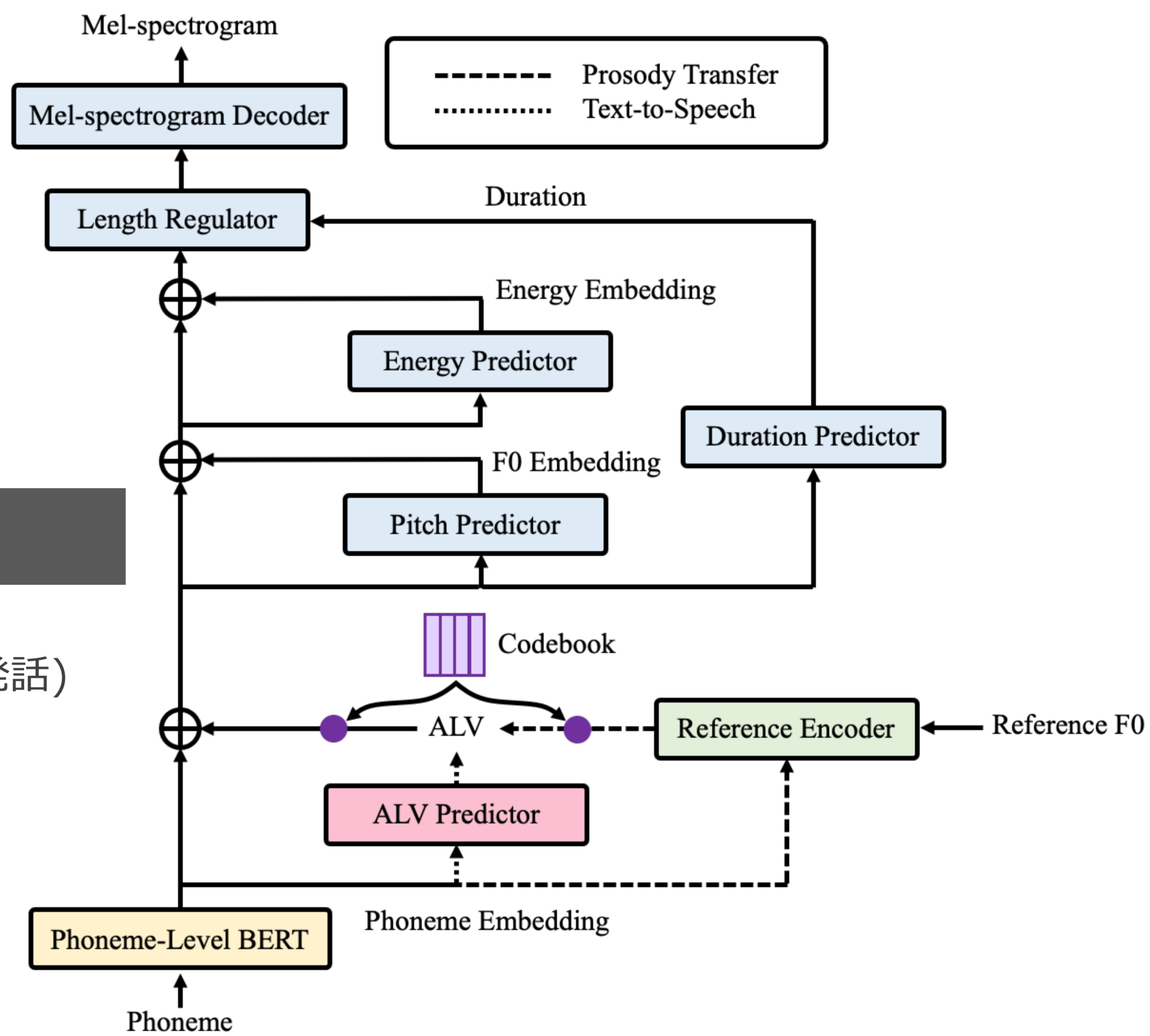
- 十分な語彙を含む学習データが必要
- 現状の方言音声コーパスのサイズは**限定的**
- ➡事前学習モデル(**Phoneme-Level BERT**[2])を活用



提案手法

提案モデル

- 概要
 - FastSpeech2[3]をベースモデルに採用
 - Reference Encoder, ALV Predictorを導入
- Reference Encoder
 - 参照音声からALVを抽出
 - VQ-VAEを利用
- ALV Predictor
 - テキストのみからALVを予測
 - 事前学習済みの**Phoneme-Level BERT**を利用



主観評価実験 & 今後の展望

データセット

- JSUT[4]: 単一女性話者による標準語音声コーパス(約7700発話)
- JMD[5]: 多方言音声コーパス(各1300発話)

比較モデル

- FS2 w/o Acc: FastSpeech2にアクセント情報を与えず学習
- FS2 w/ AP: **ALV Predictor**でテキストからALVを予測
- FS2 w/ PT: **Reference Encoder**で音声からALVを抽出

主観評価実験

- 音声の自然性MOS(5段階)と大阪方言らしさMOS(3段階)を評価
- 受聴者数は40人, 1人あたりの評価回数は24
- テキストから予測したALVを使うと音声の自然性と方言らしさが**低下**
- 参照音声から抽出したALVを使うと音声の自然性と方言らしさが**向上**

今後の展望:

- 未知話者によるProsody Transfer
 - 多話者音声コーパスを使って学習, 話者埋め込みを利用など
- ユーザーによるフィードバックを用いて**ALV Predictor**を継続学習
 - アクセント誤り訂正可能な**TTSモデル**[6]の枠組みをALVに応用
 - ALV Predictorを**模倣学習**や**Reinforcement Learning from Human Feedback**などを用いて継続学習

手法	自然性MOS	方言性MOS
JMD	4.57 ± 0.071	2.75 ± 0.065
FS2 w/o Acc	2.95 ± 0.117	2.08 ± 0.081
FS2 w/ AP	2.71 ± 0.102	1.75 ± 0.079
FS2 w/ PT	3.19 ± 0.118	2.28 ± 0.077

合成音声の自然性および大阪方言らしさに関するMOSスコア (±95% 信頼区間)

謝辞

本研究は公益財団法人立石科学技術振興財団2023年度研究助成 (S) による支援を受けたものです。

参考文献

[1] K. Yufune et al., in Proc. SSW, 2021. [2] Y. A. Li et al., arXiv:2301.08810, 2023. [3] Y. Ren et al., in Proc. ICLR, 2021. [4] S. Takamichi et al., Acoustical Science and Technology, vol. 41, no. 5, 2020. [5] S. Takamichi et al., Available: <https://sites.google.com/site/shinnosuketakamichi/>, 2021 [6] K. Fujii et al., in Proc. APSIPA ASC, 2022.

