

VQ-VAEに基づく解釈可能な アクセント潜在変数を用いた多方言音声合成

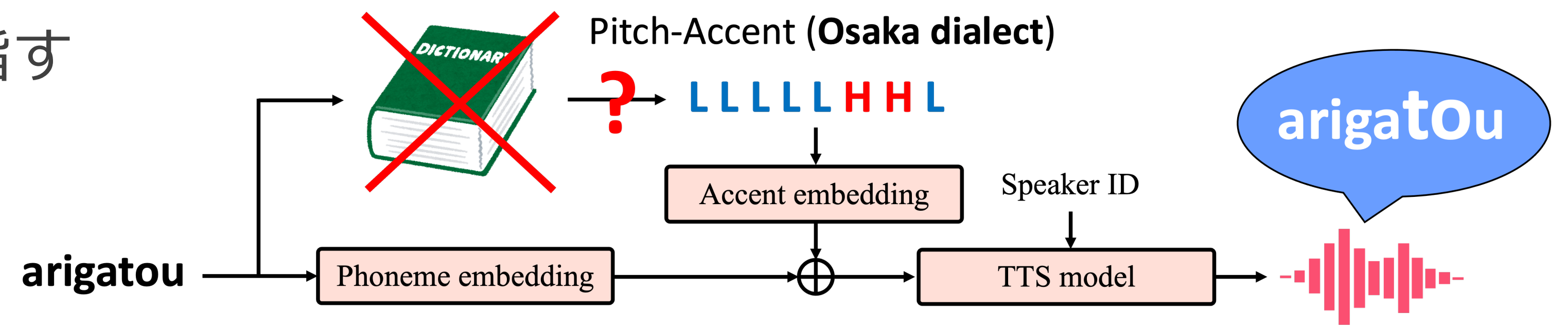


山内 一輝, 齋藤 佑樹, 猿渡 洋 (東京大学)

音声サンプルはこちら↑

概要：方言音声合成の課題 & 提案手法

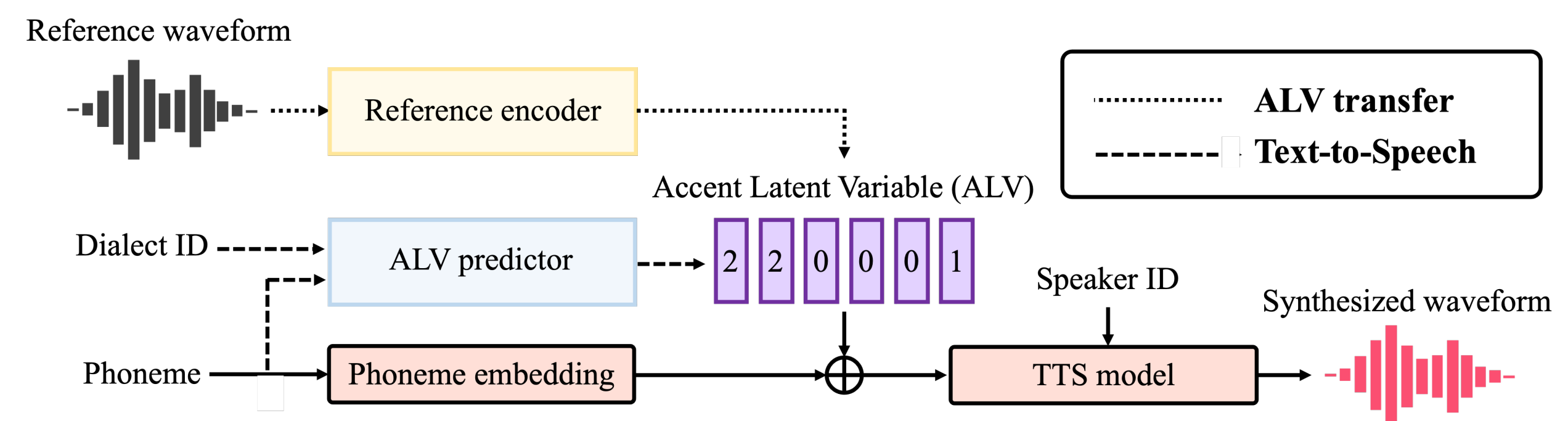
- 方言音声合成
 - 標準語と異なる韻律体系をもつ方言の音声合成を目指す
 - 課題：話者数が限られた方言の**アクセント辞書不足**
- 提案手法
 - 方言に応じた**アクセント潜在変数(ALV)**予測
 - 任意話者による参照音声を用いた **ALV transfer**



提案手法：方言に応じたALV予測と任意話者によるALV transfer

Reference encoder

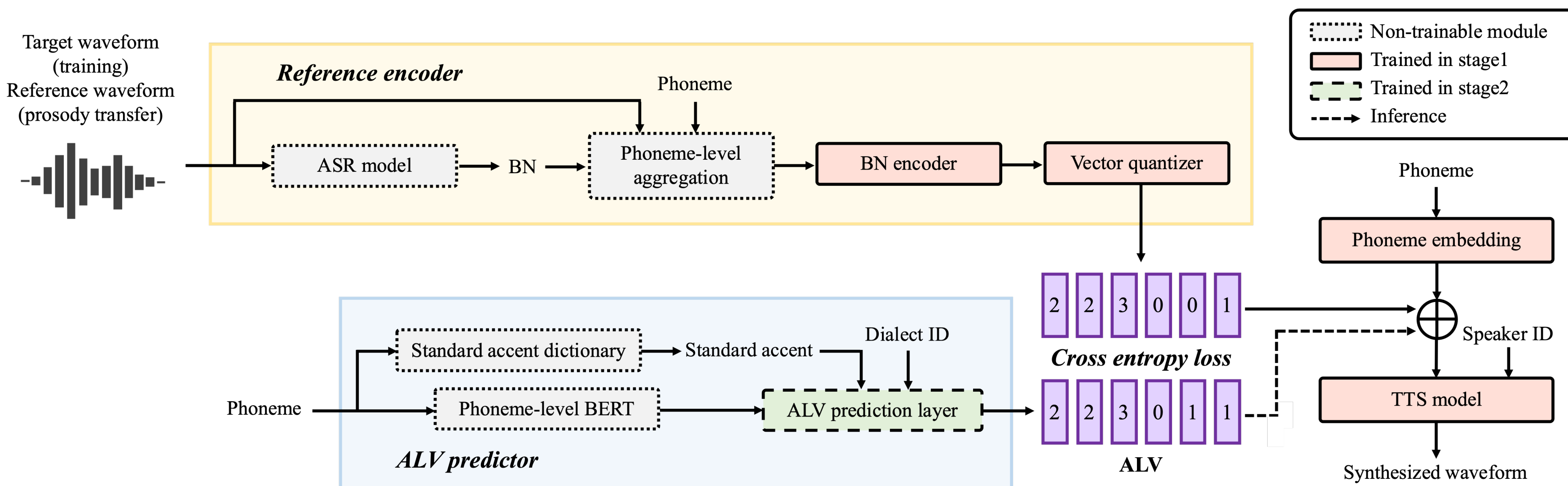
- 参照音声から**アクセント潜在変数(ALV)**[1]を抽出
 - 音声の韻律特徴量を**4クラスに量子化**
 - 日本語のアクセントは4段階と考えられている
- Bottleneck (BN) 特徴量**[2]を利用
 - 単語の弁別には**アクセント情報が必要**
 - 話者性に関する情報をあまり含まない



提案モデルの概略図。
推論時は、参照音声を入力する ALV transfer か、テキストからのALV予測 (TTS) が可能。

ALV predictor

- テキストのみから方言に応じて ALV を予測
 - Phoneme-Level BERT**[3]を活用
 - 現状の方言音声コーパスのサイズは**限定的**
 - テキスト(Wikipediaコーパス)で事前学習
 - 書記素(単語)予測タスクで事前学習
 - 方言IDを入力することで目的方言を指定



提案モデルのアーキテクチャ。学習は2段階に分けられる。
Stage1でReference encoderとTTSモデルが、Stage2でALV predictorが学習される。

実験的評価とALVの分析

実験条件

データセット	■JSUT[4]: 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) ■CPJD[6]: 多方言音声コーパス(各250発話)
モデル設定	■TTSモデル: FastSpeech2[7] ■Vocoder: HiFi-GAN[8] UNIVERSAL V1 ■ASRモデル: Whisper-v2[9]
比較モデル	■FS2: Original FastSpeech2 ■FS2-AP: ALV PredictorでALV予測 ■FS2-REF: 参照音声からALVを抽出
主観評価	■N-MOS: 音声の自然性(5段階) ■D-MOS: アクセントの目的方言らしさ (5段階)

実験結果

Intra-dialect TTS: 目的方言が目的話者の母方言と同じ

手法	話者(方言)	N-MOS	D-MOS
FS2	JMD(大阪)	2.91 ± 0.120	3.15 ± 0.145
FS2-AP	JMD(大阪)	2.91 ± 0.120	3.15 ± 0.151
FS2-REF	JMD(大阪)	2.88 ± 0.131	3.26 ± 0.153
REF	CPJD(大阪)	4.39 ± 0.105	4.18 ± 0.132

Cross-dialect TTS: 目的方言が目的話者の母方言と異なる

手法	話者(方言)	N-MOS	D-MOS
FS2	JSUT(標準語)	3.48 ± 0.114	2.46 ± 0.141
FS2-AP	JSUT(標準語)	3.44 ± 0.100	3.04 ± 0.156
FS2-REF	JSUT(標準語)	3.49 ± 0.104	3.11 ± 0.154
REF	CPJD(大阪)	4.08 ± 0.114	4.10 ± 0.130

※目的方言は**大阪方言**。MOS 評価の受聴者数は35人、1人の評価回数は24

評価結果まとめ

- Intra-dialect TTS**において、アクセントのカスケードモデリングによる**性能劣化は起きなかった**
- Cross-dialect TTS**において、合成音声の**大阪方言らしさが向上**
- 未知話者**による音声を用いた**ALV transfer**により合成音声の**大阪方言らしさが向上**

■**今後の展望**: Human Feedbackを用いた ALV predictorの継続改善

■**謝辞**: 本研究は、JST, ACT-X, JPMJAX23CB の支援を受けたものである

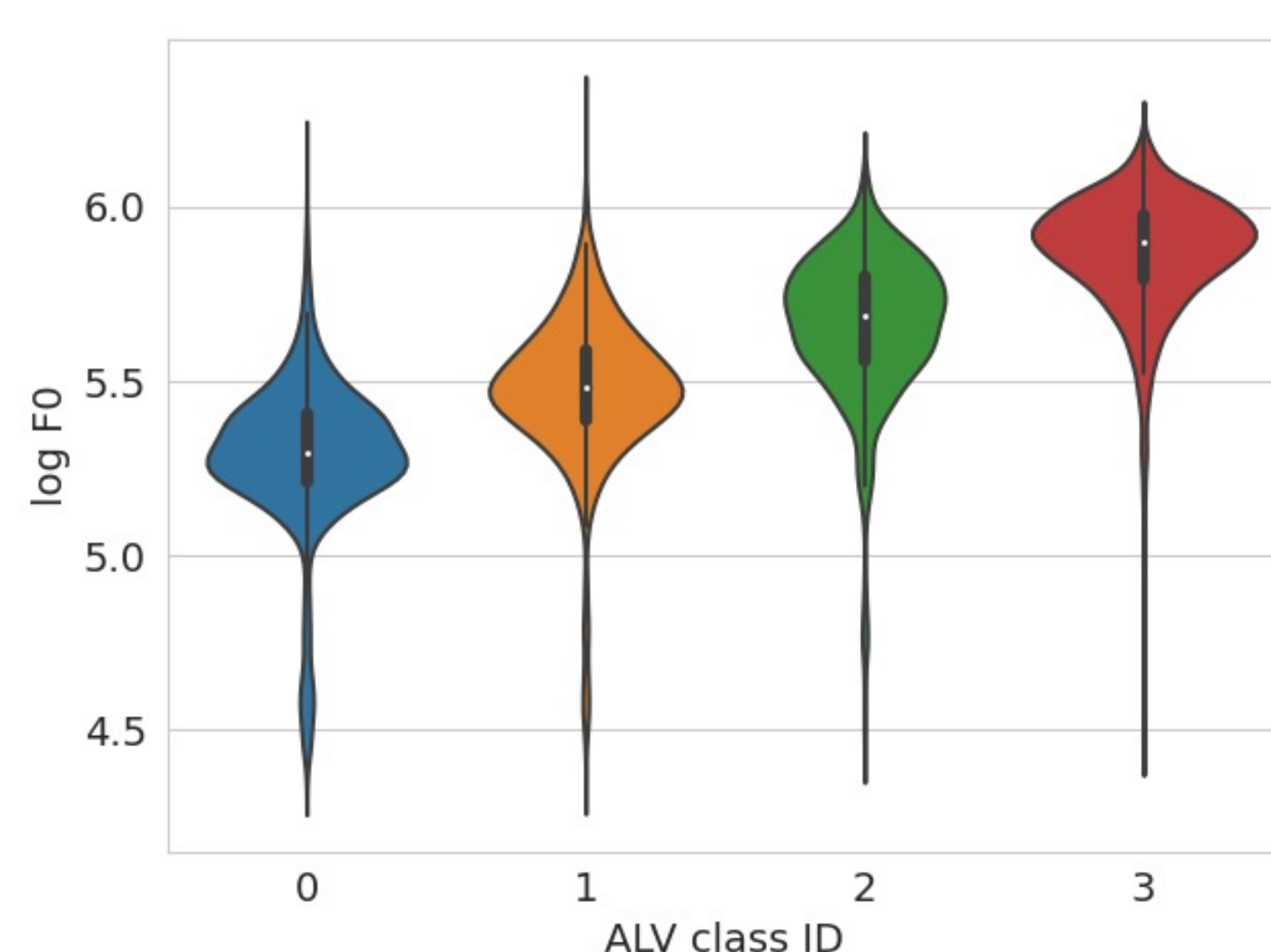
韻律特徴量: F0 vs. BN

特徴	N-MOS	D-MOS
F0	3.33 ± 0.109	2.87 ± 0.143
BN	3.49 ± 0.104	3.11 ± 0.154

- BN特徴量の方がN-MOS, D-MOSともに**高い**
- F0は話者非依存な特徴量にするため**発話単位で正規化**
→ **性能劣化**の要因の1つ

ALVの分析

ALVクラスID毎のlog F0の分布



※コードブック崩壊解消の工夫後

参考文献

[1] K. Yufune et al., in Proc. SSW, 2021. [2] L. Sun et al., in Proc ICME, 2016. [3] Y. A. Li et al., in Proc ICASSP, 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., arXiv:2212.04356, 2022.

