

1. 背景・目的

- ▶ **アイヌ語**: 日本の先住民族言語。母語話者は激減した一方、再活性化運動により学習者は増加傾向
- ▶ 音声合成 (TTS) は発音の提示や聴取素材の作成を通じて言語の**学習・継承**を支える
- ▶ 表記と音素は概ね対応しているが、音節末 /r/ の異音 (e.g. 「良い」/pirka/ [pirika]) や無解放閉鎖音 (e.g. 「浜へ下りる」/sap/ [sap̚], 「夏」/sak/ [sak̚]) は表記に現れず、音声のみから学習する必要あり

- RQ1: 限られたデータでどの品質まで到達できるか
- RQ2: ファインチューニング (FT) の効果はモデルによらず得られるか
- RQ3: 自動評価指標は主観評価の代わりになりうるか

2. 手法・実験設定

- ▶ **利用データ**: 3コーパス・5話者 (沙流・千歳・静内方言)・計**12.85時間**を24kHz・モノラルPCMに統一 (韻文は除外)
- ▶ **入力**: ラテン文字表記をそのまま文字単位で入力 (G2Pなし)
- ▶ **モデルと訓練**: Matcha-TTS / VITS (Piper) / F5-TTS を比較。Matcha・VITSはFTとscratchを比較、F5はFTのみ。5分割交差検証を全モデル共通の分割で行い、検証MCD最小のckptを選択 (学習率等は各実装の既定値)
- ▶ **評価**: 自動評価はMCD (音色・分節音の近さ) / F0 RMSE (基本周波数誤差) / VUV F1 (有声無声一致率) / SECS (話者性の近さ)。主観評価は自然性5段階のMOS (30発話、構成を伏せて呈示、評価者は上級学習者1名)

3. 結果・考察

モデル	学習法	MCD↓	F0 RMSE↓	VUV F1↑	SECS↑	MOS↑
自然音声	—	—	—	—	—	4.23
Matcha-TTS	scratch	6.212	24.97	0.750	0.650	2.60
	FT	5.937	24.71	0.749	0.688	3.28
VITS	scratch	6.206	23.86	0.754	0.639	2.28
	FT	6.254	23.07	0.761	0.649	2.00
F5-TTS	FT	9.171	40.63	0.619	0.399	1.13

表: 評価結果 (5 fold平均。F0 RMSEはHz)

[demo] デモで流す合成音声の例文

(1) asirpa ekimne patek ki wa menoko monrayke eaykap.
‘アシリパは山に入ってばかりで女の仕事ができない’
(『ゴールデンカムイ』第2巻11話より)

(2) tanto earkinne sirsesekek humi.
‘今日は本当に暑いなあ’

- ▶ **FTしたMatcha-TTSが最良** (MCD・SECS・MOS)。自然音声 (MOS 4.23) に最も近い品質 (RQ1)
 - ▶ **FTの効果はモデル依存** (RQ2): Matchaは改善、VITSはscratchと同等、F5は事前学習があっても最下位
 - ▶ **自動評価と主観評価の順位は概ね一致** (Matcha > VITS > F5)。粗い判断なら自動評価で代替可 (RQ3)
- ▶ **共通する誤り**: 音節末/r/の脱落が全構成で発生。語末の無解放閉鎖音も不安定
 - ▶ **解釈**: Matchaは音響モデルのみ、VITSは全体の再適応が必要 → 構造と事前学習の内容まで含めた選択が要る
 - ▶ **声質変換 (VC)**: 元話者を特定可能なためSeedVCで匿名化 → 利用データのノイズが変換に影響

4. まとめ

- ▶ FTしたMatcha-TTSが自然音声に最も近い品質。ただしFTの効果はモデルの構造に依存し、モデル選択は事前学習の内容まで含めて行う必要がある
- ▶ 自動評価と主観評価の順位は概ね一致し、評価者の確保が困難な言語でも自動指標は開発の指針として見なせる
- ▶ 学習データの増強、G2Pを用いた入力表現の変換、元データのノイズ除去、匿名化手法の整備など、今後の課題は多い