

第23回音声言語シンポジウム

VQVAEによって獲得された キャラクター演技スタイルに基づく 多話者オーディオブック音声合成

○中田亘 郡山知樹 高道慎之介 齋藤佑樹 (東大)
井島勇祐 増村亮 (NTT)
猿渡洋 (東大)

オーディオブックとは

オーディオブック

- 朗読を聞く形式の本
- 急速な市場の拡大が見込まれている*

代表的なオーディオブックサービス

- Amazon Audible
- audiobook.jp

オーディオブックの利点

- 両手が空くため、ながら読書が可能
- 複数人で同期的にコンテンツを楽しめる

*日本能率総合研究所プレスリリース <https://prtimes.jp/main/html/rd/p/000000015.000035568.html>

研究背景

既存のオーディオブック作成方法

- プロ声優の録音音声
- 長時間の録音となるため、多大な時間と資金が必要
- **オーディオブック音声合成による負担の軽減**が望まれる

オーディオブック音声合成とは

- 声優による録音を合成音声に置き換える
- 時間的，人的コストの削減が期待される

オーディオブック音声合成における課題

- 複数文に渡る長い文脈を考慮した韻律の実現[Nakata+21]
- **キャラクター演技の実現**（本研究）

オーディオブック内のキャラクター演技

ありくん
主人公



ありの女の子
脇役



かえるくん
ありくんのライバル



キャラクターによって異なる演技スタイルが
実現されている

出典
音声:

J-KACコーパス[Nakata+21]

紙芝居:

ありくんとかえる
作/絵 ようふゆか
制作 教育画劇

本研究の目指すシステム



多大な時間, 人的コストが発生

大幅なコストの削減
声優の演技の幅を超えた音声
多くの話者を実現可能

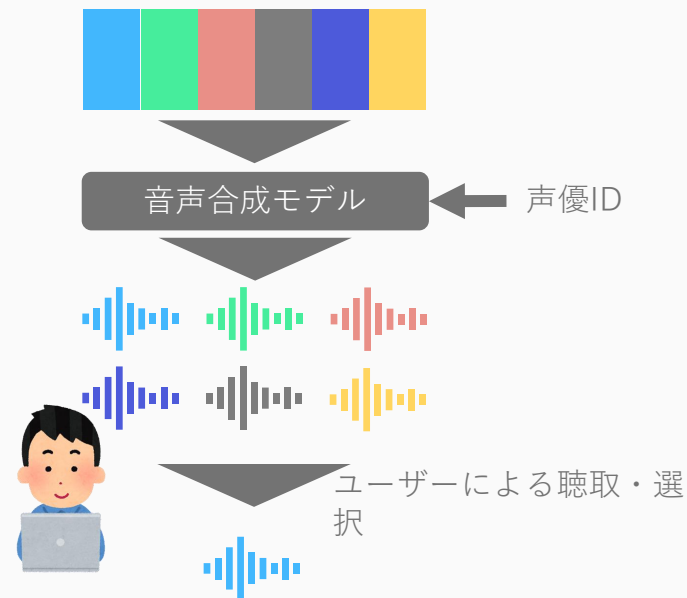
具体的なアプローチ

キャラクター演技スタイルに焦点を当てる

離散的な演技スタイル

人間による解釈が比較的容易
合成 -> 選択が可能

離散的なキャラクター演技スタイル



発表概要

目的

多話者オーディオブック音声合成におけるキャラクター演技スタイルの獲得・制御

手法

VQVAE [van den Oord+17] を用いてキャラクター演技スタイルの離散表現を獲得
話者不変学習 [Meng+18] を用いて離散表現の話者非依存性を確保

結果

話者により自然性が劣化

多様なキャラクター演技スタイルが実現可能

話者間でキャラクター演技スタイルの転写が可能

関連する研究

Learning latent representations for style control and transfer in end-to-end speech synthesis[Zhang+19]

RNNを内包するReference Encoder+VAEを用いて連続的な音声スタイル表現を
獲得・制御

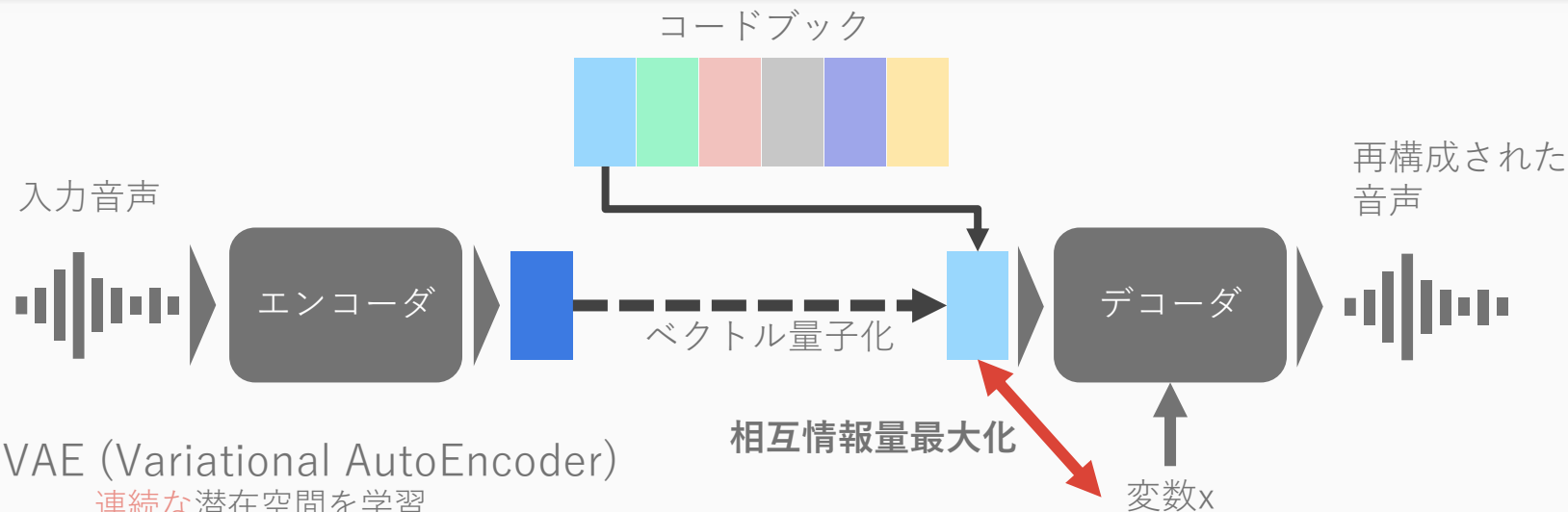
連続な潜在空間は人間による解釈が難しい
大量の合成音声を用意する点では不向き

本研究

ResCNN[Li+17]を用いて発話内で時間変化しない音声特徴量（キャラクター演技
スタイル）に着目

VQVAEを用いて離散的な表現を獲得し制御を容易にする

VQVAEとは



VAE (Variational AutoEncoder)

連続な潜在空間を学習

VQVAE (Vector Quantized Variational AutoEncoder)

離散的な潜在ベクトルを学習

音声合成における利用

アクセントモデリング[Yufune+21]

音声波形再構成[Zhao+20]

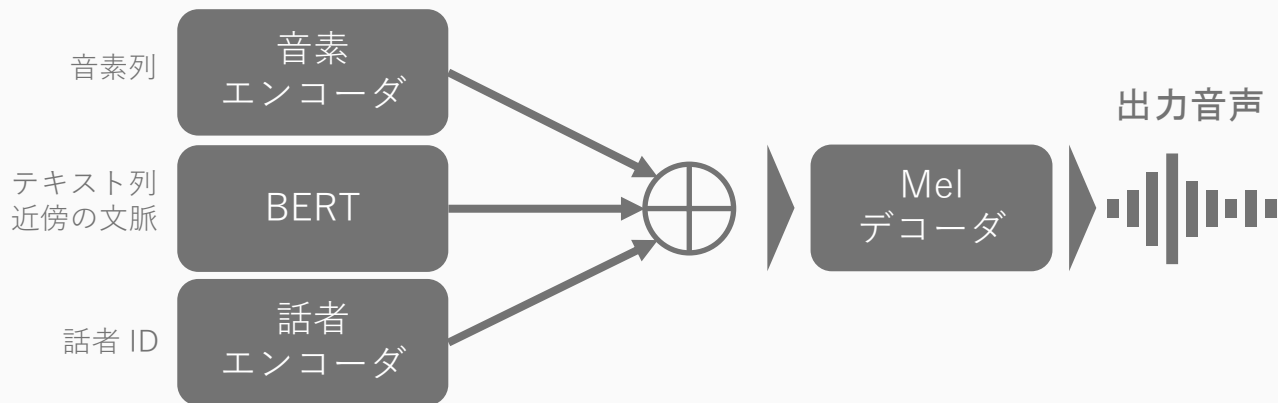
ベースライン

MultiSentences[Nakata+21]を使用

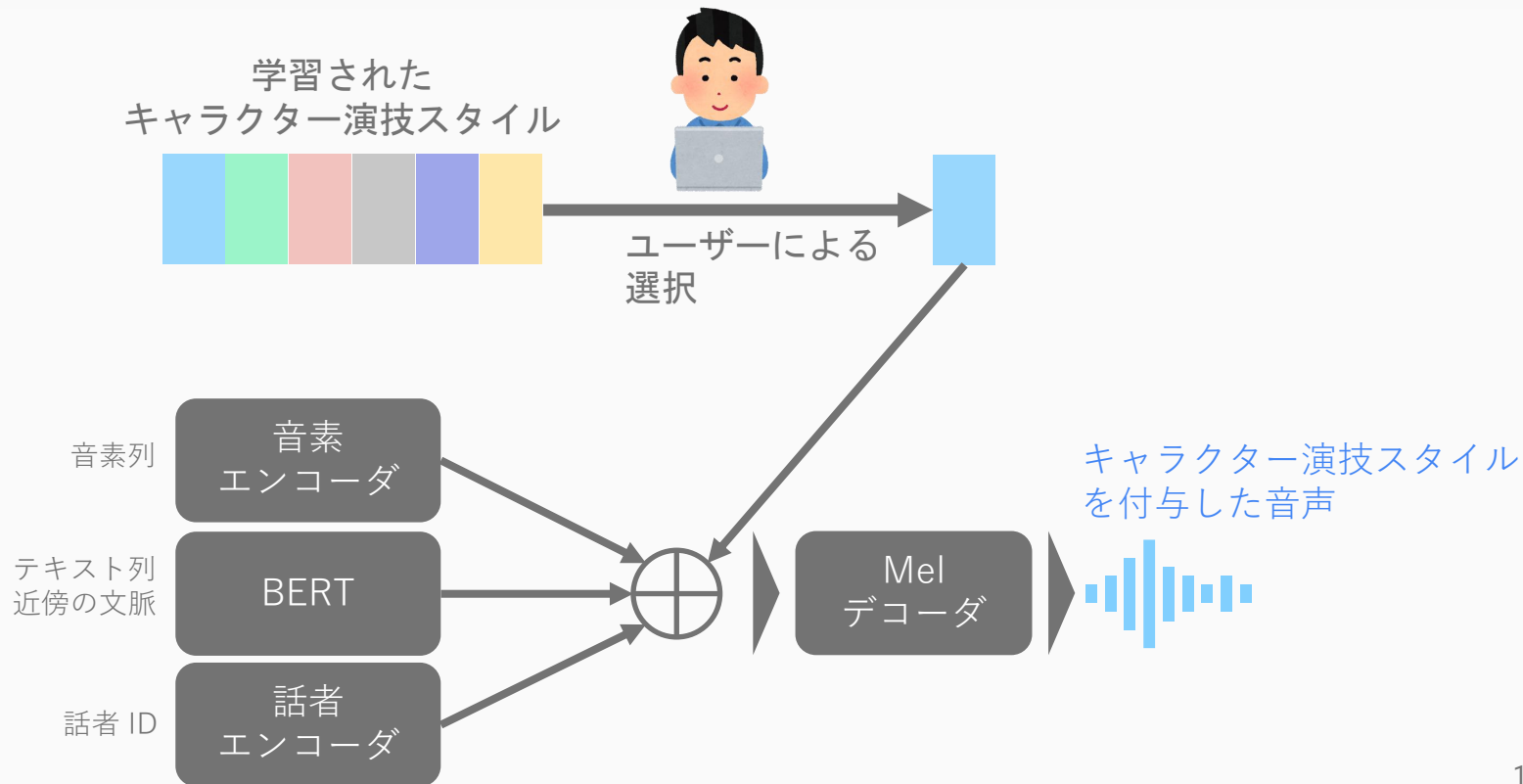
複数文から得られる言語特徴を利用

文脈を考慮した音声合成

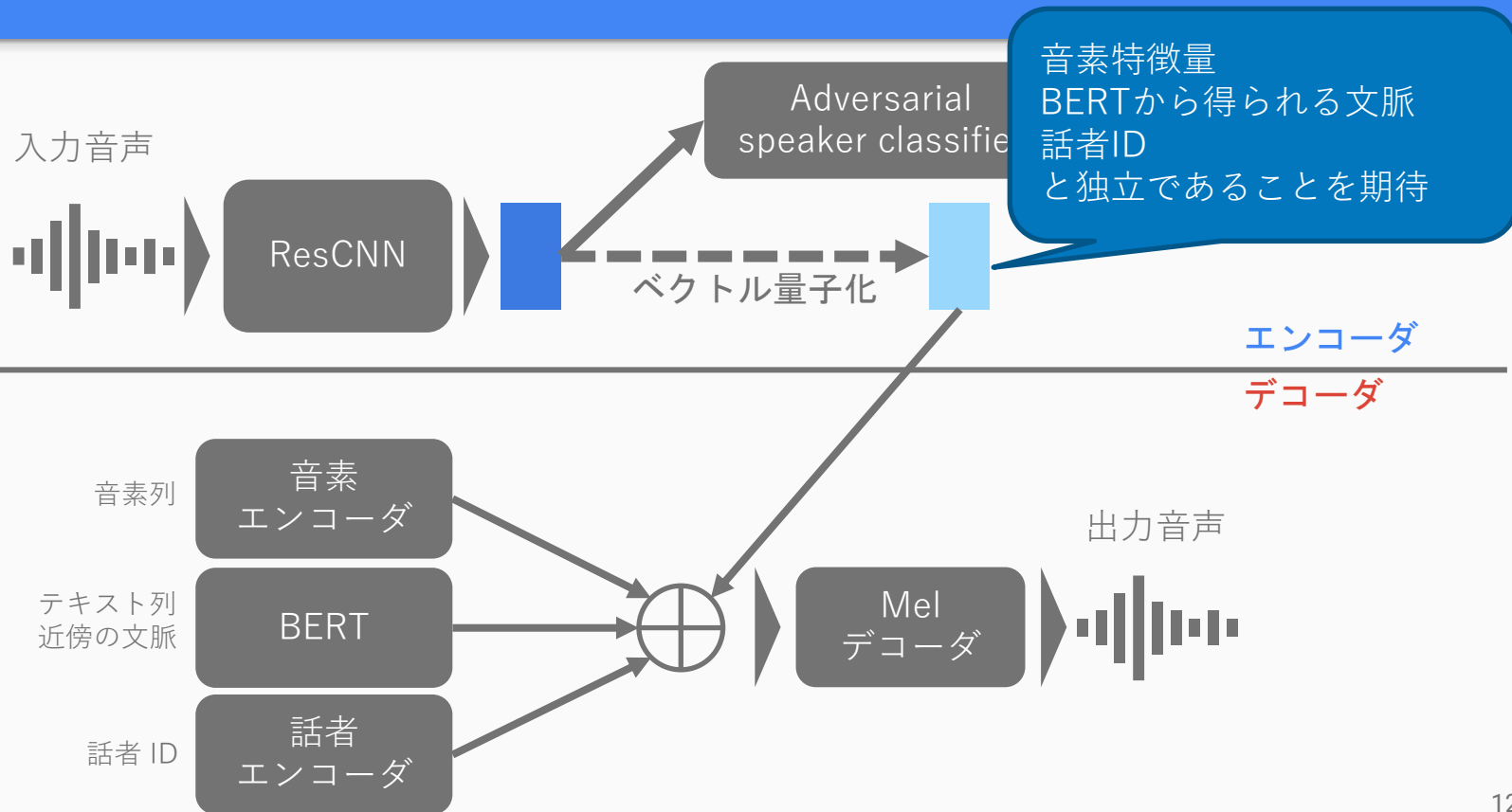
多話者音声合成に適用するために話者エンコーダーを追加



提案する音声合成モデル – 推論時



提案する音声合成モデル – 訓練時



実験条件

データセット	J-KAC[Nakata+21] 単一話者紙芝居・オーディオブック音声 男性1名 9時間 J-MAC https://doi.org/10.32130/src.J-MAC 多話者オーディオブック音声 男性23名 女性16名 25.5時間
音素エンコーダ入力	音素列 アクセント列
BERT[Delvin+18] 事前学習済みモデル	日本語Wikipedia学習済みモデル
コードブック次元数	256
コードブックサイズ	64 (学習後に実際に使用されるのは21)
ボコーダー	HiFi-GAN[Kong+20]

評価指標

ベースライン：MultiSentences[Nakata+21]と比較

合成音声の自然性

キャラクター演技スタイルを付与する事により品質が劣化していないか

話者類似性

キャラクター演技スタイルを付与することにより話者性が変化していないか

音声の多様性

キャラクター演技スタイルを付与することにより多様な音声合成ができるか

話者間のキャラクター演技スタイルの転写

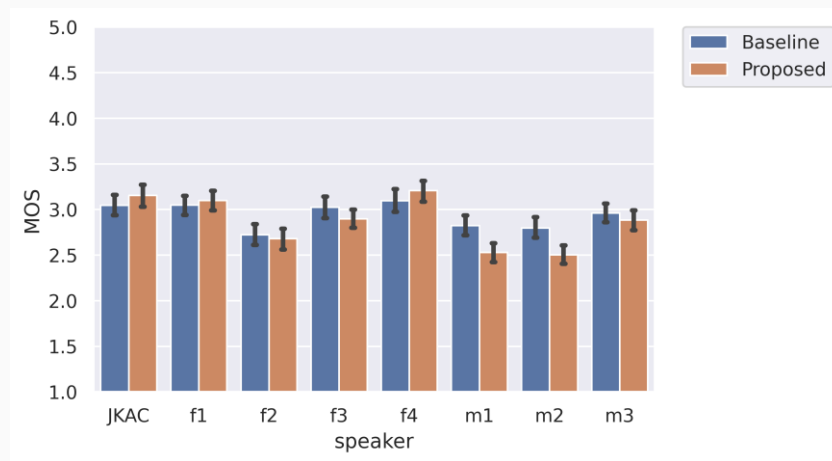
応用例として、ある話者から得られた演技スタイルを他の話者に転写可能か

音声の自然性

5段階の自然性MOSによる主観評価

原音声から得られたキャラクター演技スタイルを用いて合成

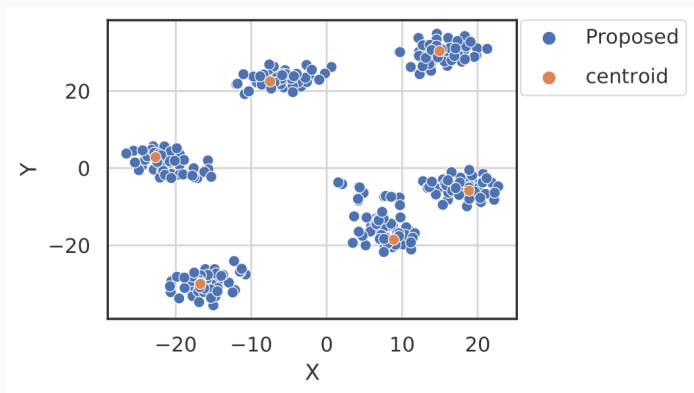
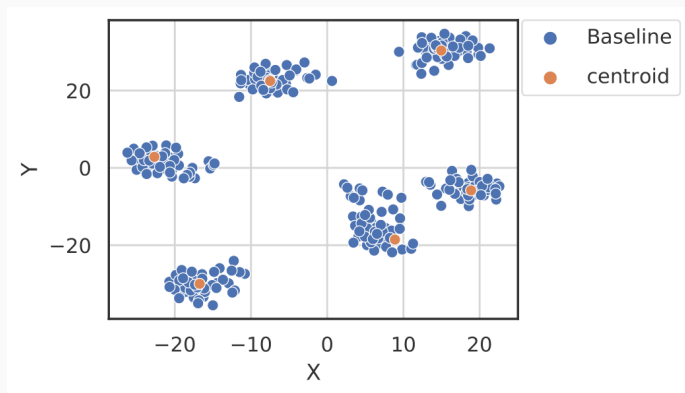
手法	平均
Baseline	2.943
Proposed	2.862



エラーバー：95%信頼区間

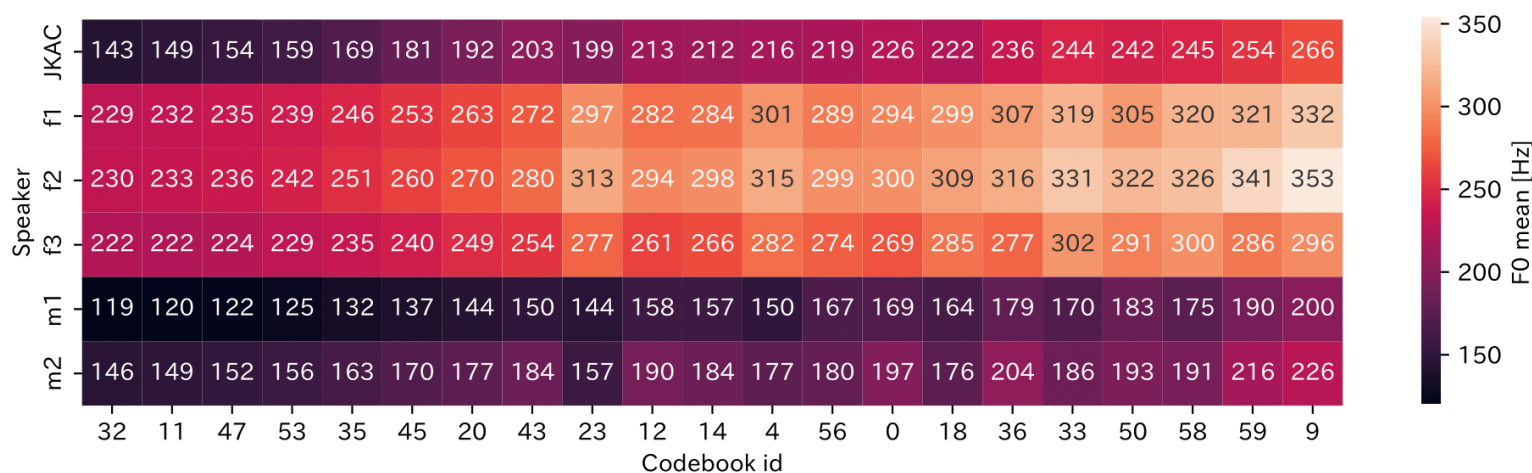
話者類似性

Resemblyzerを用いて抽出したd-vectorの分布を比較
原音声から得られるキャラクター演技スタイルを使用



キャラクター演技スタイルを導入したことによる
大きな差異は見られない
話者類似性は変わっていない

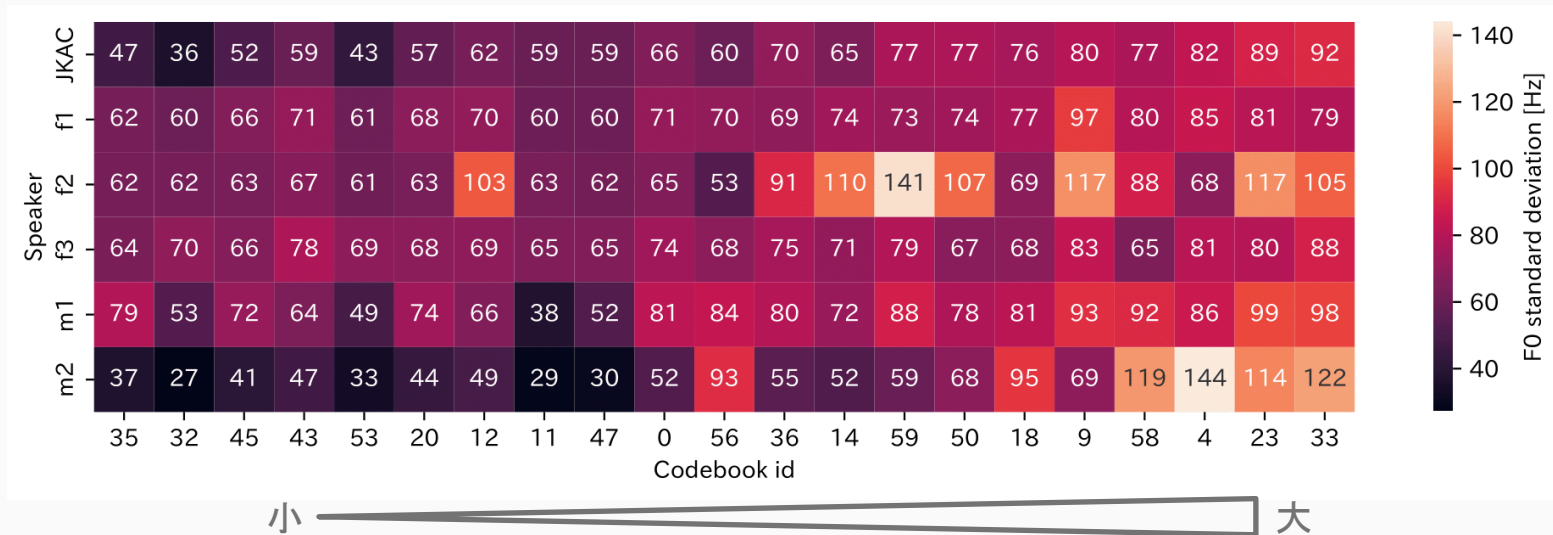
多様性 – ピッチの平均



小 ————— 大

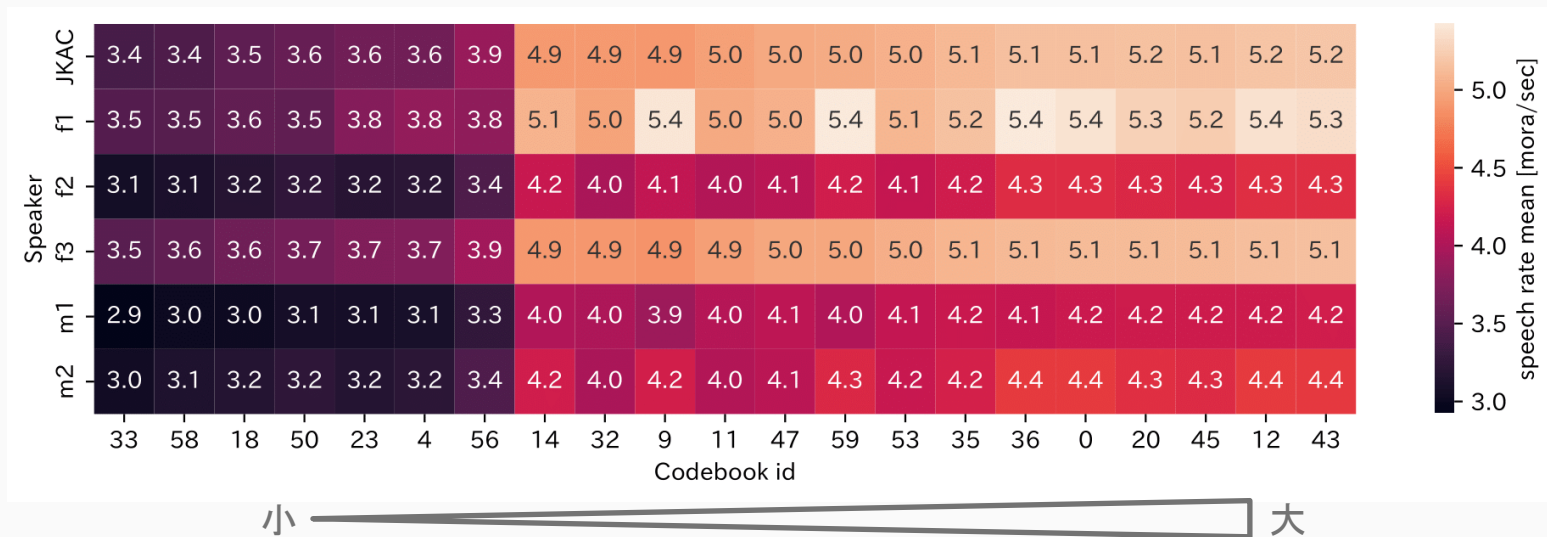
コードブックを変化させることによりピッチが変化
話者非依存

多様性 – ピッチの標準偏差



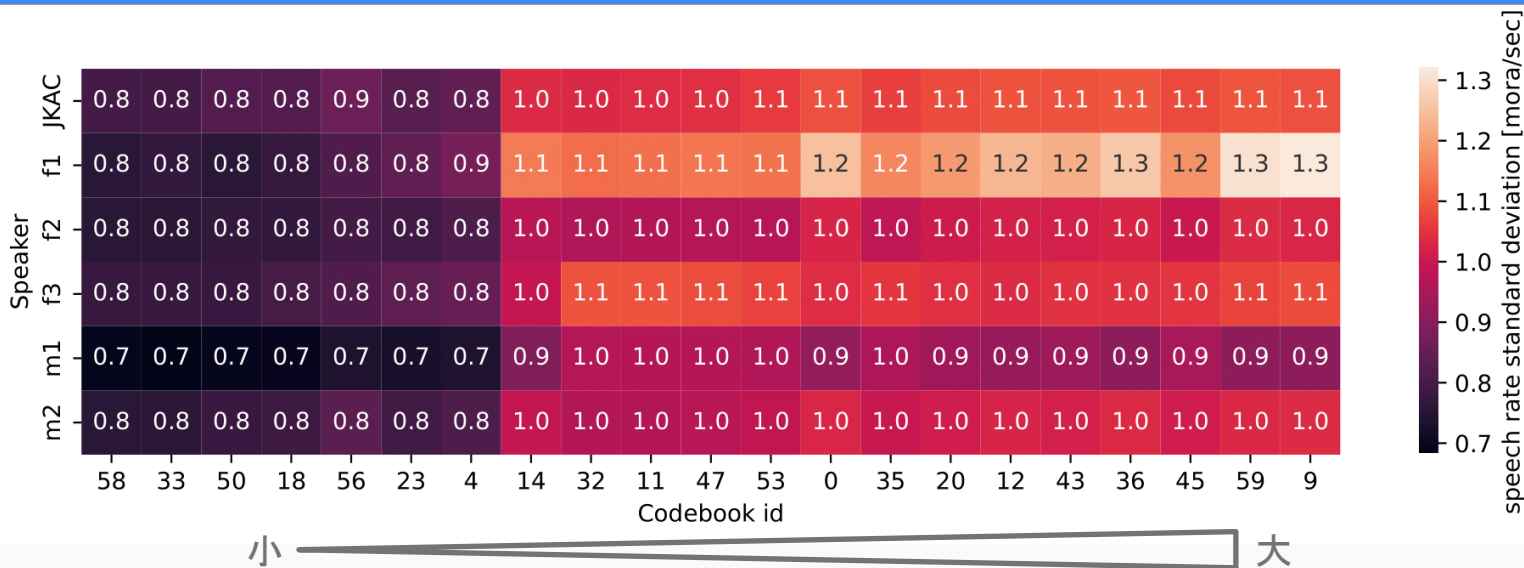
ピッチの標準偏差に関しても同様に変化
話者非依存

多様性 – 話速の平均



コードブックを変化させることにより話速が変化
話者非依存

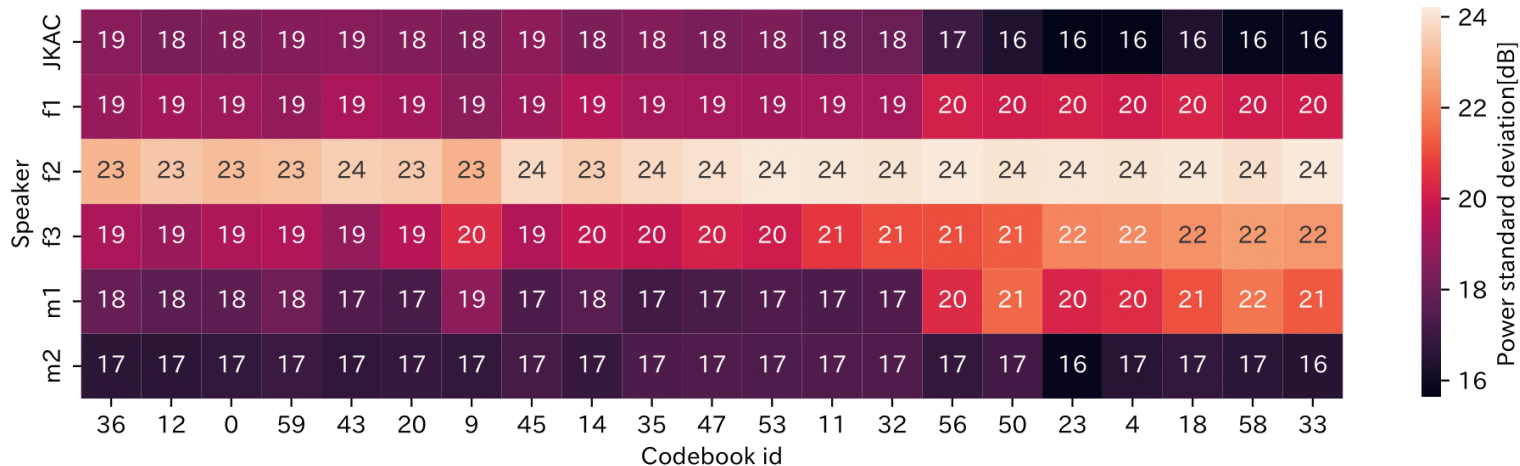
多様性 – 話速の標準偏差



標準偏差についても同様の変化が見られる
話者非依存

* 原稿の結果に誤りあり

多様性 - パワーの標準偏差



小 ————— 大

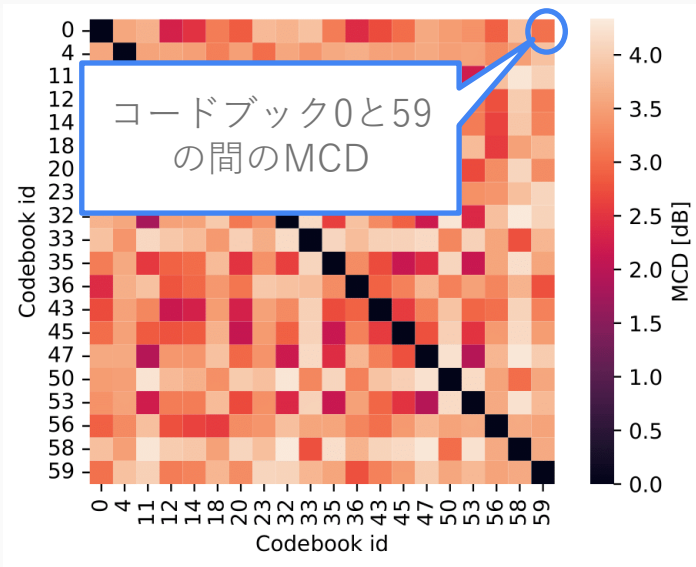
コードブックを変える事によりパワーの標準偏差が
大きく/小さくなる

多様性 – 合成音声間のMCD

異なるコードブック（キャラクター演技スタイル）で合成された音声間のメルケプストラム歪（MCD）を可視化

最低値は1.7[dB]

多様な音声の実現されている

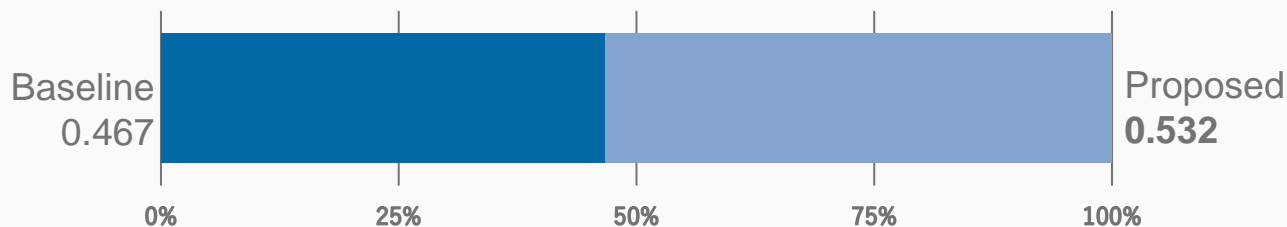


コードブックを変化させることにより合成音声に変化することを確認
多様な音声の実現されている

話者間のキャラクター演技スタイルの転写

J-KACから抽出したキャラクター演技スタイルを他の話者（男性6名女性4名）に転写

評価基準：「以下の音声は参照音声のキャラクター演技スタイルを元に他の話者に転写したものです。どちらがより参照音声のキャラクター演技スタイルに近いですか？」



p値 0.022










提案法は参照音声に近い
キャラクター演技スタイルを実現できることを確認

音声サンプル - ベースラインとの比較

連続した文章の音声

原音声 (JKAC)	ベースライン	提案法
		
		
		

音声サンプル – キャラクター演技スタイル

	あり君の声 コードブック20	ありの子の声 コードブック0	かえる君の声 コードブック47
あり君のセリフ			
ありの子のセリフ			
かえる君のセリフ			

まとめ

目的

多話者オーディオブック音声合成におけるキャラクター演技スタイルの獲得・制御

手法

VQVAE [Oord+17] を用いてキャラクター演技スタイルの離散表現を獲得

話者不変学習 [Meng+18] を用いて離散表現の話者非依存性を確保

結果

話者により自然性が劣化

多様なキャラクター演技スタイルが実現可能

話者間でキャラクター演技スタイルの転写が可能

今後の課題

オーディオブック音声合成の具体的な評価方法の検討

順位相関係数

	ピッチの標準偏差	話速の平均	話速の標準偏差	パワーの標準偏差
ピッチの平均	0.731	-0.39	-0.10	0.049
ピッチの標準偏差		-0.75	-0.49	0.49
話速の平均			0.84	-0.88
話速の標準偏差				-0.88
パワーの標準偏差				

太字 : $p < 0.05$ 相関がある

話速とパワーの標準偏差の間に強い相関

音声の自然性

5段階の自然性MOSによる主観評価

原音声から得られたキャラクター演技スタイルを用いて合成

特に評価結果が
異なる音声

ベースライン 

提案法 