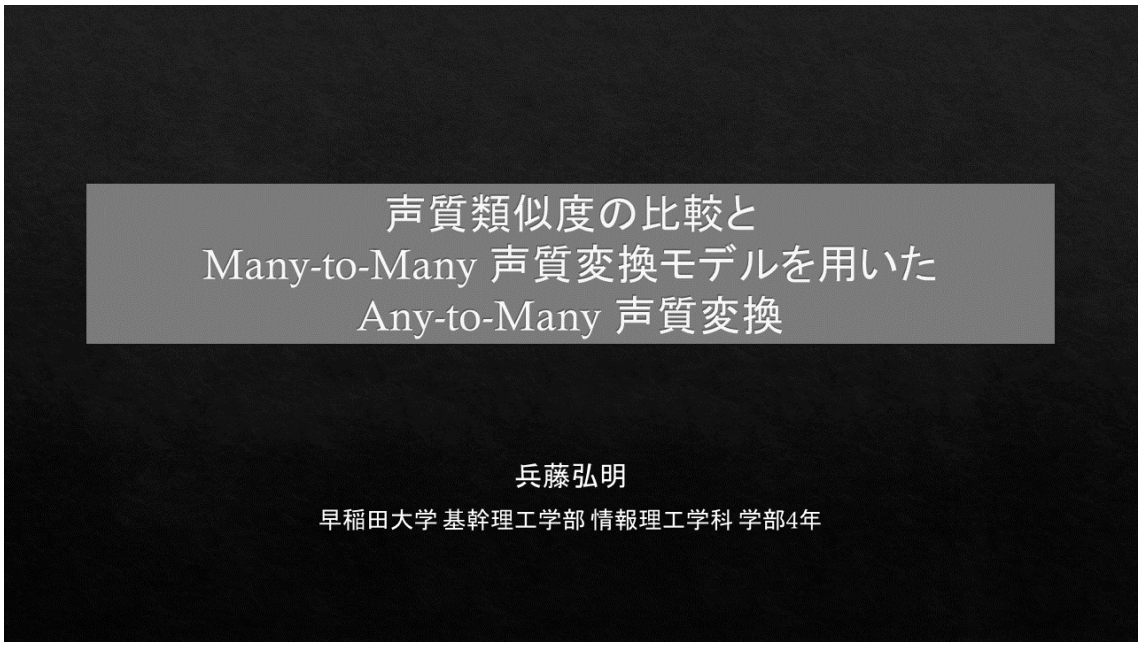
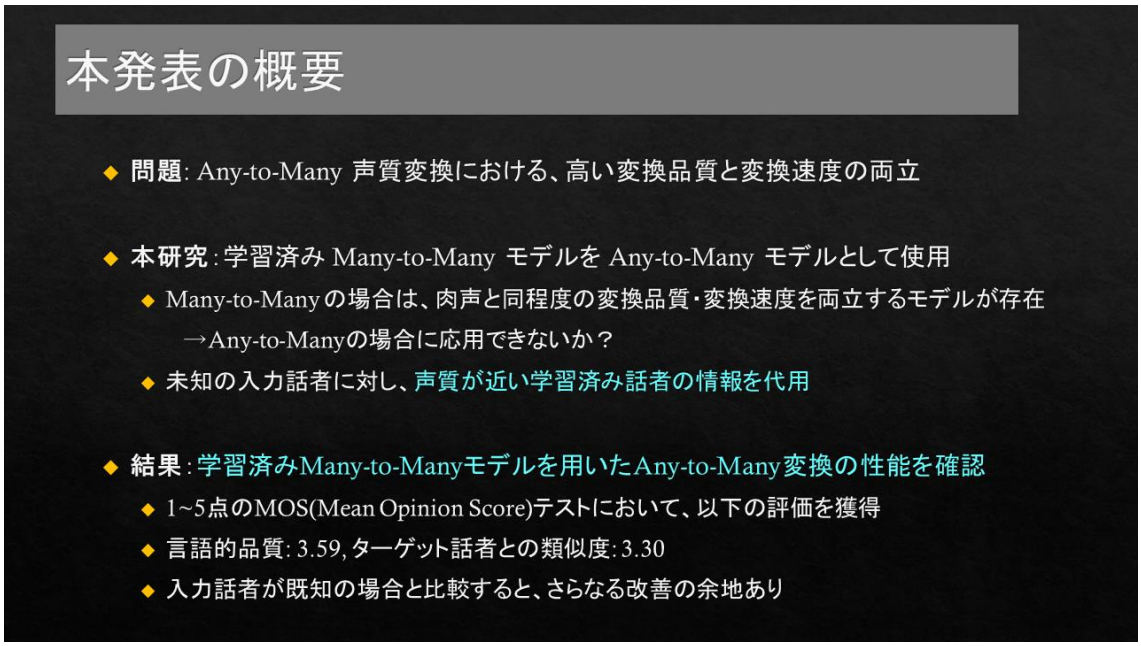
B4 の兵藤です。
2023 年 2 月 28 日, 3 月 1 日に沖縄県立博物館・美術で行われた“第 9 回 音声・音響・信号処理ワークショップ（SPEASIP 2023）”に参加し、“声質類似度の比較と Many-to-Many 声質変換モデルを用いた Any-to-Many 声質変換”という題目でショート・オーラル発表を行いました。

今回の発表は私にとって初めての学会発表でしたが、酒井先生をはじめ多くの方に研究を支援していただき、無事に発表を終えることができました。発表を通じて知り合った他の参加者の方と様々な議論や意見交換を行うことができ、大きく成長できたと感じています。酒井先生をはじめ、研究を支えてくださった研究室のメンバーにこの場を借りて感謝申し上げます。

発表で使用したスライドは以下の通りです。

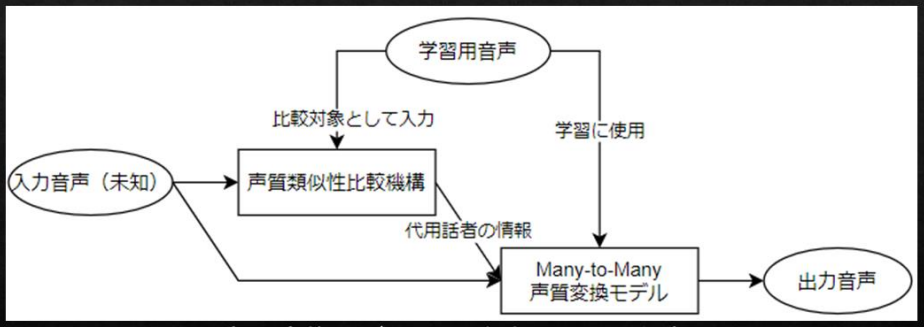




リアルタイム声質変換の従来法

- Many-to-Many 声質変換
  - [1] Kim et al.の音声合成・声質変換モデル(VITS)
    - 肉声と同程度の品質の音声を合成可能
    - 学習時にパラレルな音声データが不要
- Any-to-Many 声質変換
  - [2] Liu et al. の声質変換モデル
    - 変換音声は高品質だが、肉声と比較すると改善の余地あり
- Any-to-Many でも、肉声と同程度の品質の音声への変換を可能にしたい

提案法の概要



- Many-to-Many 声質変換モデルには前述の VITS を使用
  - 高品質なリアルタイム変換が可能であり、non-parallel なデータで学習可能であるため
- 声質の類似性は メルスペクトログラムの平均絶対誤差で評価

品質評価実験

- 実験条件
  - モデル学習用データセット: VCTK corpus [3]
  - 評価用データセット: LibriTTS [4]
- 変換音声に対し、MOS(Mean Opinion Score)テストによる主観評価実験を実施

話者の性別 (入力-ターゲット)	未知の入力話者（提案手法）		既知の入力話者（比較対象）	
	自然性	類似性	自然性	類似性
M-M	4.06	3.67	4.61	4.67
M-F	3.70	3.31	4.92	4.89
F-M	3.25	3.03	4.69	4.78
F-F	3.33	3.17	4.91	4.86

品質評価実験

- 実験条件
  - モデル学習用データセット: VCTK corpus [3]
  - 評価用データセット: LibriTTS [4]
- 変換音声に対し、MOS(Mean Opinion Score)テストによる主観評価実験を実施

話者の性別 (入力-ターゲット)	未知の入力話者（提案手法）		既知の入力話者（比較対象）	
	自然性	類似性	自然性	類似性
M-M	4.06	3.67	4.61	4.67
M-F	3.70	3.31	4.92	4.89
F-M	3.25	3.03	4.69	4.78
F-F	3.33	3.17	4.91	4.86

品質評価実験

- 実験条件
  - モデル学習用データセット: VCTK corpus [3]
  - 評価用データセット: LibriTTS [4]
- 変換音声に対し、MOS(Mean Opinion Score)テストによる主観評価実験を実施

話者の性別 (入力-ターゲット)	未知の入力話者（提案手法）		既知の入力話者（比較対象）	
	自然性	類似性	自然性	類似性
M-M	4.06	3.67	4.61	4.67
M-F	3.70	3.31	4.92	4.89
F-M	3.25	3.03	4.69	4.78
F-F	3.33	3.17	4.91	4.86

まとめ・今後の方針

- 目的: 肉声と同程度の品質の音声に変換でき、
  リアルタイムに動作する Any-to-Many 声質変換モデルの実現
- 提案法: 事前学習済みMany-to-Many声質変換モデル + 声質類似性比較機構
- 評価結果:
  - 言語的品質: 3.59, ターゲット話者との類似度: 3.30
  - 入力話者が既知の場合と比較すると、さらなる改善の余地あり
- 今後の改善点
  - 代用話者を選ぶ機構の改良

引用文献

[1] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In International Conference on Machine Learning, pages 5530–5540. PMLR, 2021.

[2] Songxiang Liu, Yuewen Cao, Disong Wang, Xixin Wu, Xunying Liu, and Helen Meng. Any-to-many voice conversion with location-relative sequence-to-sequence modeling. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:1717–1728, 2021.

[3] Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald. English multi-speaker corpus for cstr voice cloning toolkit. ac. uk/jyamagis/page3/page58/page58. html,[Jan. 9, 2017], 2017.

[4] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. arXiv preprint arXiv:1904.02882, 2019.