1R-02

声質変換を用いたデータ拡張に基づく咽喉マイク音声認識

塚越　駿大† 西田　昌史† 西村　雅史†‡

静岡大学† 愛知産業大学‡

1. はじめに

スマートフォンや音声アシスタントの普及により，多様な場面で音声認識技術が活用されるようになったが，多人数会話など発話重畳が頻発する環境や，騒音環境下において高い音声認識精度を保つことは困難であった．

外部雑音の影響を抑制する手法として，咽喉マイク（図1）の活用が提案されている[1]．咽喉マイクは，話者の咽喉部分に装着し，音声を皮膚の振動を介して収録する．外部の空気を媒介しないため，騒音に対して頑健であるという特徴を持つ．一方で，咽喉マイクは音声の高周波成分が大きく欠落する特徴（図2）を持っており，従来の音声認識システムに入力すると音響ミスマッチが原因となり認識精度が著しく低下してしまう．

音声認識モデルを咽喉マイクに適応させることで解決が期待されるが，利用可能な咽喉マイク収録音声データは限られており，既存の音声認識モデルを十分に学習させることができない．それに対して，鈴木ら[1]は，LSTM を用いた特徴マッピングによるデータ拡張と知識蒸留を用いて咽喉マイクに関する音声認識精度を向上させたが，接話マイク音声の認識精度には及ばなかった．

そこで本研究では TTS（Text-to-speech）の代表的な手法である VITS モデルを応用した声質変換モデル[2][3]に着目した．本来は話者情報の変換に用いられる声質変換技術を，接話マイク音声から咽喉マイク音声へ変換する技術として用い，咽喉マイク音声の学習データを拡張する手法を提案する．

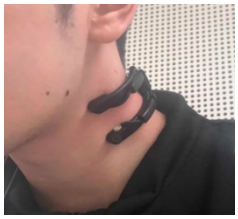

図1 咽喉マイク

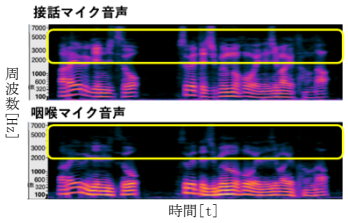

図2 マイク間の音響特性の違い

2. 提案手法

提案する声質変換を用いたデータ拡張に基づく咽喉マイク音声認識の概要を図3に示す．以後，接話マイクを CM，咽喉マイクを TM と記す．

学習は 2 段階にて行う．一段階目は声質変換部の学習であり，2 段階目は音声認識部の学習である．1 段階目で学習した声質変換モデルを用いて，大規模な日本語音声コーパスから擬似 TM 音声コーパスを作成する．作成した大量の擬似 TM 音声と実収録した少量の TM 音声を利用して自己教師あり学習モデルを多段階で Fine-tuning することで音声認識部の学習を行う．

Throat Microphone Speech Recognition Based on Data Augmentation Using Voice Conversion
Toshihiro Tsukagoshi† Masafumi Nishida† Masafumi Nishimura†‡
† Shizuoka University
‡ Aichi Sangyo University

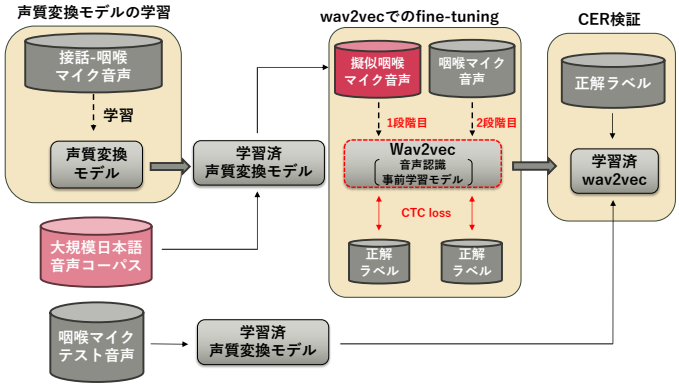

図3 咽喉マイク音声認識の概要

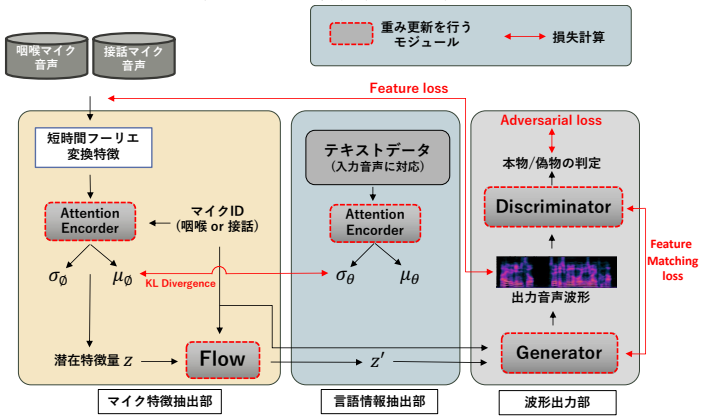

図4 声質変換モデルの学習プロセス

2.1. 声質変換部の学習

声質変換部には MMVC モデル[3]を採用した．MMVC は従来の音声合成における，音声特徴量の抽出＋ボコーダの学習による 2 ステージモデルの学習とは異なり，1 段階で波形を再構成するように学習を行うため，より高速に自然な音声の生成が可能となる．声質変換モデルの学習プロセスを図4に示す．入力は CM，TM マイクとそれに対応するテキスト，さらにマイク種別を表す ID である．音声は STFT 変換特徴を抽出し，Attention Encorder へと入力する．テキストは埋め込み表現を抽出した後に Attention Encoder へと入力する．それぞれの出力分布が一致するように，KL Divergence を最小化させる．また，出力音声の生成品質を担保するために，GAN による Adversarial loss と Generator と Discriminator の中間特徴量損失を計算する Feature Matching loss を導入している．さらに出力波形のスペクトラム損失を計算する Feature loss と，変換，逆変換を通して元の音声に再合成するかを計算した Voice Conversion loss の合計 5 つの loss の和が収束するように重み更新を行う．Flow は逆変換可能なモデルである．推論時には CM マイクの ID を入力として音声情報から言語情報を抜き出した後，TM マイク ID を入力として言語情報にマイク特性を付与することで CM→TM へとマイク特性を変換することが可能となる．

表1 各モデルの詳細

モデル	音声認識部：Fine-tuning時の学習データ	特徴マッピング	リスコアリング
CSJ_CM_FT	CM音声(230h)	なし	なし
TM_FT	収録TM音声(22h)		
TM2CM_FT	擬似CM音声(22h)	TM→CM	
CSJ_CM2TM_FT	擬似TM音声(325h)	CM→TM	なし
+LM score	+収録TM音声(22h)		あり
CSJ_CM_CM2TM_FT	CM音声(230h)+擬似TM音声(325h)		なし
+LM score	+収録TM音声(22h)		あり

2.2. 音声認識部の学習

音声認識部では,SSL(自己教師あり学習)モデルによって潜在特徴量を抽出し,CTC(Connectionist Temporal Classification)を用いて学習を行う.SSL モデルには wav2vec2.0[4]を利用し,事前学習モデルとして,53 言語 56000 時間の音声データで学習した large モデルを利用した.学習データには大規模な日本語音声コーパスとして CSJ コーパスを利用した. 提案手法では Fine-tuning を 3 段階にて行う. 1 段目では CSJ コーパスの CM 音声(230h)を用いる.2 段階目では CSJ コーパスを疑似咽喉マイク音声に声質変換したデータ(325h)を用いる.3 段階目では実収録した少量の咽喉マイク音声(22h)を用いて Fine-tuning を行う.

2.3. 言語モデルによるリスコアリング

KenLM[5]を利用したリスコアリングの導入も行った.KenLM は軽量・高速な n-gram 言語モデルであり,リアルタイム性が求められる音声認識と相性が良い. Common crawl データから日本語文献を事前学習させることで,音響モデルによるスコアと言語モデルによるスコアに重みをつけて結合し,最終的な認識結果の出力を行う.

3. 評価実験

3.1. データセット

学習データとして,男性話者 29 名分の音素バランス文読み上げ音声(22h)を利用した.CM 音声と TM 音声は同時収録したものであり,パラレルデータとなっている.テストデータには男性話者 10 名による新聞記事読み上げ音声(1h)を利用した.いずれも静かな環境で録音された音声であり,テストデータの中に学習データの話者は含まれていない.TM 音声に関しては,血流振動におけるノイズを除去するため,カットオフ周波数 100Hz でハイパスフィルターを適用した.

3.2. 実験方法

比較手法を含め 5 つの学習済みモデルで検証を行った.表 1 にモデルの詳細をまとめた.ここで,特徴マッピングとは声質変換をする方向である. CM→TM は CM 音声を入力として,擬似 TM 音声を作成したことを指す.また,TM→CM を行う際には,FreeVC[4]を利用した.FreeVC モデルは MMVC モデルとは異なり,マイク ID を用いず,音声から LSTM を用いて動的に生成される話者 ID を利用する.推論時には one-shot で変換したいマイク音声を入力すれば良い.予備実験により MMVC モデルでは TM 音声から話者性が分離されず,適切に CM 音声に変換されない結果を得た.一方で FreeVC モデルは逆の変換(CM→TM)の際に話者の TM 音声を必要とするため,既存の音声コーパスからデータ拡張をすることができない.そのため変換方向に応じて適宜モデルを変更して実験を行った.検証において文字誤り率(Character Error Rate; CER)を算出した.

3.3. 実験結果・考察

実験結果を表 2 に示した.CM 音声に適応した音声認識モデル(CSJ_CM_FT)に対して,TM 音声を入力した際の CER 値を基準とした際の相対的な文字誤り率の削減率を各モデルに対し算出した.CSJ_CM_FT モデルでは,TM 音声を声質変換モデルを利用して CM 音声に変換することで,収録 TM 音声をそのままモデルに入力するのに比べ約 7.5%(削減率 31%)文字誤り率を削減した.また声質変換後の擬似 CM 音声で認識部の Fine-tuning を行うモデル(TM2CM_FT)では,さらに 3.5%(削減率 46%)ほど文字誤り率を削減した.擬似 TM 音声を利用した実験では,CM 音声＋擬似 TM 音声＋収録 TM 音声を利用し,さらに言語モデルによるリスコアリングをすることによって約 17.7%(削減率 74%)の文字誤り率の削減を達成した.また,データ拡張せずに TM 音声を用いて Fine-tuning したモデル(TM_FT)に比べ,約 1.4%(削減率 18%)の文字誤り率を削減した.以上の結果より,咽喉マイク音声認識タスクにおいて,咽喉マイク音声を接話マイク音声に変換するよりも,接話マイク音声を咽喉マイク音声に変換する方が既存の音声認識モデルを活用する上で有効であることが推察できる.

表2 各モデルの音声認識精度

モデル	テストデータ	CER(%)	相対削減率(%)
CSJ_CM_FT	CM音声	3.95	
	TM音声	**24.02**	0.00
	TM音声 (入力時に擬似CMに声質変換)	16.51	31.27
TM2CM_FT		13.05	45.67
TM_FT	TM音声	7.71	67.90
CSJ_CM2TM_FT		8.83	63.24
+LM score		8.74	63.61
CSJ_CM_CM2TM_FT		6.61	72.48
+LM score		**6.30**	**73.77**

4. おわりに

本研究では咽喉マイク音声認識性能を向上させるため,声質変換モデルを用いたデータ拡張に基づくアプローチを行った結果,提案手法が有効であることを示した.ただし,接話マイク音声の認識率と比較すると改善の余地がある.今後は雑音環境下での音声認識実験に取り組む予定である.

謝辞

この成果は, 国立研究開発法人新エネルギー・産業技術総合開発機構(NEDO)の委託業務(JPNP20006)の結果得られたものである.

参考文献

[1] 鈴木貴仁, 緒方淳, 綱川隆司, 西田昌史, 西村雅史, "咽喉マイクを用いた大語彙音声認識のための特徴マッピングによるデータ拡張と知識蒸留", 情報処理学会論文誌,Vol.62,No.6, pp.1373-1381, 2021.

[2] J. Li, W. Tu, and L. Xiao, "FreeVC: Towards high-quality textfree one-shot voice conversion," arXiv preprint arXiv:2210.15418, 2022.

[3] Isletennos,"MMVC(RealTime-Many to Many Voice Conversion)", https://github.com/isletennos/MMVC_Trainer.git, 2022

[4] Baevski,et al."wav2vec2.0: A Framework for SelfSupervised Learning of Speech Representations", arXiv preprint arXiv:2006.11477, 2020.

[5] Kenneth Heafield. KenLM: faster and smaller language model queries. In Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation, pp. 187–197, 2011.