1R-03

d-vectorを用いた話者モデルの選択に基づく咽喉マイクの特徴マッピング

村手　涼雅　西田　昌史　綱川　隆司　西村　雅史
静岡大学大学院　総合科学技術研究科

1. はじめに

多人数会話環境や騒音下環境では，周囲の雑音や発話重畳が原因で音声認識が困難になることがある．そこで，我々は周囲雑音に頑健な接触型マイク（咽喉マイク）を用いる．咽喉マイクを用いた音声認識は，Paul ら[1]による咽喉マイクの音響特徴分析に関する研究など，今もなお活発に議論されている研究テーマの一つである．一方，これまで用いられてきた空気伝導型マイク（接話マイク）から取得した音声との音響的差異が発生するため，咽喉マイクのみでは十分な音声認識精度が保証できない．さらには，咽喉マイク独自のデータセットが存在せず，音声認識に必要な音響モデル学習のための大規模なデータセットの構築には膨大なコストがかかる．

そこで，林ら[2]は接話マイク音声と咽喉マイク音声のパラレルデータを学習に用いた Long short-term memory (LSTM) による特徴マッピング手法を提案した．また，鈴木ら[3]は，上記の手法を利用し，擬似的に咽喉マイクのデータセットを構築している．

本研究では，特徴マッピングの精度改善を目的とするために，特徴マッピングの話者依存性に注目した．話者ごとに独立したマッピングモデルを学習した上で，入力データの話者と音響的に近い話者のモデルを，d-vector[4]によって選択する手法を提案する．特徴マッピングには林ら[2]が提案した LSTM による手法を利用する．d-vector は，話者認証技術として提案された話者埋め込み手法の一つであるが，近年では音声合成のための話者適応[5]に使われることもある．

2. DNN 特徴マッピング

林ら[2]が提案した特徴マッピング手法は，咽喉マイクと接話マイクのパラレルデータを LSTM で学習し，咽喉マイク音声から接話マイクの振幅スペクトルを推定する．そして，推定した振幅スペクトルと咽喉マイク音声から取得した位相スペクトルを逆短時間フーリエ変換し，接話マイクの音声を復元させる．　本研究では，この手法をベースに提案手法に関する実験を行う．

Throat microphone feature mapping based on speaker model selection using d-vector
Ryoga Murate, Masafumi Nishida, Takashi Tsunakawa, Masafumi Nishimura (Shizuoka University)

3. 特徴マッピングの話者依存性評価

最初に，男性話者 10 名分のデータセットを用いて特徴マッピングの話者依存性評価を行う．テスト話者と同一話者のデータを学習した同一話者モデルと，テスト話者を除いた 9 名の話者のデータを学習した複数話者モデルのマッピング精度を，メルケプストラム歪み（Mel-cepstral distortion, MCD）[6]によって評価した．MCD は変換後の音声と目標とする音声の誤差を示す指標であり，その値が小さいほど誤差が小さいと評価される．本研究では変換後の咽喉マイク音声と接話マイクから収録した同一発話の MCD を求める．

図 1 に一話者あたり 103 発話の MCD の平均値を算出した結果を箱ひげ図で示す．同一話者モデルでマッピングした音声の方が，複数話者モデルでマッピングした音声よりも全体的に MCD が低く，特徴マッピングの話者依存性を確認できた．すなわち，学習データに含まれない話者であっても，類似度の高い話者のマッピングモデルを選択できれば，特徴マッピングの精度を向上できると予想する．

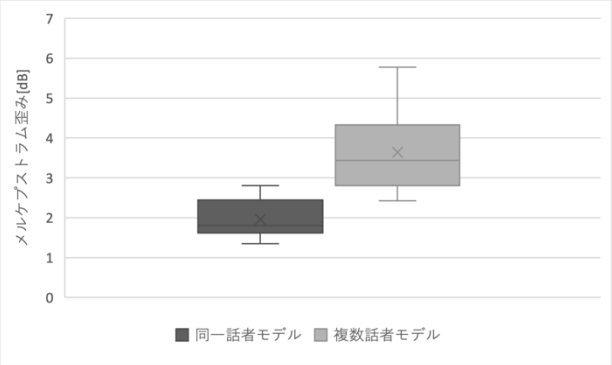

図 1. MCD の比較による特徴マッピングの話者依存性評価

4. 提案手法

本研究では，d-vector を用いることで，対象話者と類似している話者のマッピングモデルを選択する手法を提案する．d-vector は，予め話者認識モデルを学習し，その中間層出力を話者特徴ベクトルとして用いる手法である．類似度計算のために一話者あたり複数の発話の d-vector を抽出しておく．この処理を話者登録と呼ぶ．次に，マッピングモデル選択までの流れを説明する．はじめに，テストデータの d-vector と話者登録によ

って得た d-vector との類似度を求め，その平均値を話者同士の類似度スコアとする．この処理を話者登録した話者数分行う．テストに用いる個々の発話で最も類似度スコアの高い話者を選択していき，その中で選択された頻度が最も高かった話者をマッピングモデルとして適用する．マッピングモデルを選択するまでの流れを図 2 に示す．

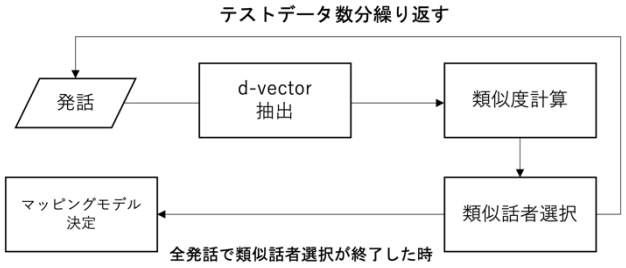

図 2. d-vector による類似話者選択

5. 評価実験

前述の複数話者モデルと，話者ごとに独立して学習した単一話者モデルから，d-vector に基づいてマッピングモデルを選択する提案手法のマッピング精度を比較する．マッピング精度の比較には話者依存性の評価同様 MCD を用いる．特徴マッピング及び話者認識モデルの学習とテストには 15 名の男性話者によるパラレル音声を用いる．なお，テストでは leave-one-out 法による交差検証を行なった．

はじめに，d-vector を抽出するためにテスト話者 1 名を除いた話者認識モデルを学習する．話者認識モデルは，中間層にユニット数が 256 の全結合層を 4 層積み重ね，出力層をユニット数が 14 の全結合層とする．学習には 26 次元のメルフィルタバンクを用いる．次に，テスト話者 1 名あたりの学習・テストデータ数について説明する．話者認識モデルの学習には，咽喉マイクで収録した音素バランス文 350 文×14 名分を用いる．話者登録には，学習データとは異なる音素バランス文 50 文×14 名分を用いる．テストデータには学習・話者登録とは異なる音素バランス文 103 文を用いる．

特徴マッピングモデルの学習には，257 次元の振幅スペクトルを入力とし，ユニット数が 512 の LSTM 層を 2 つ重ねた stacked-LSTM にユニット数が 257 の全結合層を出力層とした構造である．

表 1 に各手法における特徴マッピング精度の比較を示す．比較対象として，各テスト話者に対してマッピングモデルをランダムで 10 回選択した際の MCD の平均値及び最善のマッピングモデルを選択した際の結果を示す．テスト話者 15 名分の MCD の平均値を比較したところ，提案手法では複数話者モデルよりも僅かに低い MCD が得られた．また，複数話者モデルと比較して，提案手法では 15 名中 11 名の話者で MCD の改善が見られた．ランダム選択手法は他の 3 手法と比較して総じて高い MCD が確認できた．これは，提案手法で d-vector がある程度類似話者を選択できていることを示唆する．

表 1. 各手法における特徴マッピング精度の比較

	MCD[dB]
提案手法	3.49
複数話者モデル	3.69
ランダム選択	4.81
最善選択	3.11

6. おわりに

本研究では，d-vector を用いて事前に類似話者の特徴マッピングモデルを選択する手法を提案した．実験の結果，ベースラインである複数話者モデルと比較して僅かに MCD が改善し，d-vector によるマッピングモデル選択の有効性を確認した．今後の課題として学習・テストデータを更に増やして実験を行い，マッピングした音声を認識精度という観点でも評価を行う．また，x-vector[7]などの異なる手法の適用や，話者クラスタリングを行なった上で類似した複数の話者でマッピングモデルを学習する手法について検討を行う．

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す．

参考文献

[1] Subrata Kumer Paul et al, “Speech Recognition of Throat Microphone using MFCC Approach”, IRJET, 2020.
[2] Shengke Lin et al, ”DNN-based feature transformation for speech recognition using throat microphone”, APSIPA, 2017.
[3] 鈴木 貴仁他, "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.
[4] Ehsan Variani et al, “Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification”, ICASSP, pp.4080-4084, 2014.
[5] Rama Doddipatla et al, “Speaker Adaptation in DNN-Based Speech Synthesis Using d-Vectors.”, Interspeech, pp.3404-3408, 2017.
[6] R.Kubichek, “Mel-cepstral distance measure for objective speech quality assessment”, IEEE Pacific Rim Conference on Communications Computers and Signal Processing, 1993.
[7] David Snyder et al,” X-vectors: Robust DNN Embedding for Speaker Recognition”, ICASSP, pp.5329-5333, 2018.