E-014

声質変換なりすまし攻撃への同一話者判定 DNN の頑健性
Robustness of Same Speaker Identification DNN against AI Spoofing Attacks

長橋 龍ノ介* 亀田 健太郎* 柘植 覚† 西田 昌史 ‡ 堀内 靖雄* 黒岩 眞吾*
Ryunosuke Nagahashi, Kentaro Kameda, Satoru Tsuge, Masafumi Nishida, Yasuo Horiuchi, Shingo Kuroiwa

1. はじめに

我々は，短い発声での話者照合精度向上を目的に，話者ベクトルを介さずに 2 つの音声を直接比較する同一話者判定 DNN[1]の検討を行っている．同手法は専門家の音声鑑定技術から発想を得ており，同一音素系列の局所的な差異を検出できる構造を持っている．この特長から，全体的には詐称対象話者の音声に似ているものの，局所的には違和感を感じさせる人工的な詐称音声の検出にも効果が高いと考えた．そこで本稿では，同一話者判定 DNN の声質変換なりすまし攻撃に対する頑健性を，話者照合分野でのデファクト標準となっている x-vector との比較により，その有効性を検証した．

2. 話者照合手法

2.1 同一話者判定 DNN

同一話者判定 DNN は，同一テキスト音声の 2 つの音響特徴量時系列を入力してそれらが同一の話者によるものか否かを判定する．図 1 に同一話者判定 DNN の構造を示す．入力部では，長さの等しい 2 つの音響特徴量時系列をチャネル方向に結合（Concatenate）し，2 チャネルの特徴マップとして中間部へ出力する．なお，入力特徴量のフレーム長 T は任意であるが，入力前処理により 2 つの入力のフレーム長は同一である．中間部は，CNN 1 層と CNN にショートカット接続を導入した 3 段の ResNet ブロックで構成され，活性化関数には ReLU 関数を用いる．その後，3 段目の ResNet ブロックからの出力特徴マップに対して，チャネルごとに平均プーリングを行いスカラーとした後，チャネル数次元のベクトルに集約する．本稿では，最後の出力特徴マップのチャネル数を 128 としたため，プーリング処理によって 128 次元のベクトルが得られる．出力部は，全結合層と Sigmoid 関数からなり，[0,1] のスコアを出力する．DNN は，2 つの入力音声が同じ話者によるものである場合には 1，異なる話者の場合には 0 を出力するように学習される．話者照合は，この出力値に対する閾値により行われる．

2.2 入力前処理

同一話者判定 DNN では入力部での結合処理のため，2 つの入力音響特徴量のフレーム長をそろえる必要がある．そこで特徴量の抽出後，DP マッチングにより 2 つの特徴量セグメントの DTW アライメントを求め，それに従ってフレーム長をそろえる．本稿では，フレーム間の距離の計算にはコサイン距離を用いた．DTW アライメント処理では，求まったアライメントに従い，対応するフレームを複製することで 2 つの特徴量セグメントのフレームの長さをそろえる．

＊千葉大学 Chiba University
† 大同大学 Daido University
‡ 静岡大学 Shizuoka University

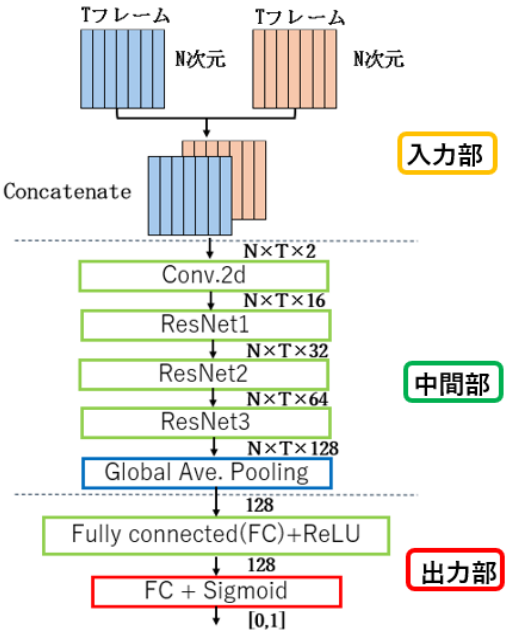

図 2　同一話者判定 DNN

また，DNN の学習はミニバッチにより行う．その際，本稿で使用する PyTorch では，ミニバッチ内でもフレーム長 T をそろえる必要がある．そのため，音声データの複製（繰り返し）を用いた複製パディング処理を採用した．複製パディングでは，ミニバッチ内のペアの最大フレーム長を求め，最大フレーム長に満たないペアは前半のデータを複製し，後尾に接続することで最大フレーム長に合わせる処理を行う．

3. 話者照合実験

3.1 実験条件

本稿では，科学警察研究所の「大規模話者骨導音声データベース」[2]に収録されている 66 種類の単語 （2~7 モーラ）の気導音声を用いた．収録は 2~3 カ月の時期差を設けて 2 度行われ，収録時期ごとに一連の発話内容の音声を小休憩を挟みながら 2 回収録している．録音はサンプリング周波数 44.1kHz，量子化精度 16bit で，防音室にて行われた．実験では，音声を 16kHz にダウンサンプリングして使用した．

同一話者判定 DNN の学習には男女 200 名ずつの計 400 名 60 単語 （2~5 モーラ），評価には学習データとは異なる男性 76 名，女性 82 名の計 158 名の学習に含まれない 6 単語を用いた．正例である同じ話者による音声のペアを「同一話者ペア」，負例である異なる話者の音声によるペアを「異なる話者ペア」と呼ぶ． また，本稿では同時期データ及び異時期データを全て使用して 1 つのモデルを学習した．用い

表 1　実験データの内訳

	話者数	同一話者ペア	異なる話者ペア
学習データ	400名	1600ペア(1598ペア)	3200ペア(3196ペア)
評価データ	158名	316ペア	632ペア

たペア数は単語ごとにそろえた．表 1 に 1 単語あたりのペア数の内訳を示す（ただし，学習に用いた単語の内，2 つは収録ミスによりペア数は少なくなっており，その数は表中の () 内 に記載した）．

評価実験では，表 1 に示した評価データを使用した「人の声」と，評価データの内，異なる話者ペアにおける一方の話者の音声を，もう一方の話者の音声に変換し，新たに作成したペアを負例とした「声質変換音声」で評価を行った．声質変換には RVC[3]を用いた．RVC による各話者への変換モデルの学習には同データベースの内，評価話者の ATR 音素バランス文 A セット 50 文（2 回× 2 時期の延べ 200 発声）を使用し，エポック数は 300 とした．

同一話者判定 DNN における特徴量にはフレームサイズ 25ms，フレームシフト 10ms で求めた 30 次元対数 MFB を使用した．また，同一話者判定 DNN との比較として，事前学習済みの xvector-jtubespeech[4]を x-vector の抽出器として用い，x-vector 間のコサイン類似度の閾値処理により話者照合を行う手法を用いた．なお，同一話者判定 DNN と x-vector のどちらの場合も前処理として，振幅に基づく音声区間検出を行い，前後の無音区間を除去，振幅の標準化 （平均 0, 標準偏差 1 にする） を行った．

追加実験として，同一話者判定 DNN の学習に，学習話者の中から 200 名を選択し，評価データと同様にして作成した声質変換音声ペア音声を負例として加えた場合での評価も行った．

3.2 実験結果

評価は全て同時期データを対象とした．まず表 2 に「人の声」の評価結果を示す．表より，全ての評価単語において x-vector に比べ同一話者判定 DNN の EER が大幅に低くなった．単語ごとの結果に注目すると，モーラ数が少なく発声時間が短い「おい」「はい」は全体的に他の単語よりも EER が高くなる傾向にあり，逆に「もしもし」や「サービスエリア」などでは低い傾向にあった．

続いて表 3 に「声質変換音声」の評価結果を示す．同一話者判定 DNN は，学習に声質変換音声ペアを追加する前と後の結果を各々示している． x-vector と同一話者判定 DNN の追加前を比較すると表 2 の結果と同じく，いずれの評価単語でも x-vector に比べ同一話者判定 DNN の EER は低くなった．さらに同一話者判定 DNN の追加後では，追加前の結果と比べて大幅に低い EER を達成することが出来た．また表 3 の結果でも，モーラ数が多く発声時間も長い単語の方が，全体的に EER は低くなる傾向になることが分かる．ここで同一話者判定 DNN の追加前の結果では「もしもし」が最も低い EER を達成している．これは「もしもし」には鼻腔の長さに関連性がある鼻音/m/が 2 つ含まれていることが理由の 1 つだと考えられる．一方で，表 2 の結果と比較して EER は高くなっており，同一話者判定 DNN も声質変換音声に対しては改良の余地が大きい．さらに，表には無いが，声質音声を学習に追加することで，「人の声」に対するエラー率が増大してしまうことも観測され，この点での改良も必要である．

表 2　人の声を対象とした話者照合結果(EER[%])

評価単語	x-vector		同一話者判定DNN	
	EER	閾値	EER	閾値
おい	11.7	0.88	2.5	0.40
はい	11.6	0.88	2.3	0.44
おまえ	9.3	0.88	1.0	0.51
くるま	6.6	0.88	1.3	0.86
もしもし	6.3	0.92	0.6	0.62
サービスエリア	6.6	0.86	1.2	0.62

表 3　変換詐称音声を対象とした話者照合結果(EER[%])

評価単語	x-vector		同一話者判定DNN			
			追加前		追加後	
	EER	閾値	EER	閾値	EER	閾値
おい	35.1	0.93	27.1	0.99	8.9	0.09
はい	34.5	0.92	28.9	0.99	10.8	0.48
おまえ	37.3	0.93	24.4	0.99	6.7	0.41
くるま	32.3	0.92	19.5	0.99	6.3	0.60
もしもし	27.1	0.95	15.4	0.99	6.3	0.68
サービスエリア	30.6	0.96	17.6	0.99	4.4	0.29

4. おわりに

本稿では声質変換なりすまし攻撃も想定した話者照合において同一話者判定 DNN の頑健性を検証した．その結果，声質変換音声を対象とした場合でも同分野でのデファクト標準となっている x-vector よりは高い精度となった．しかし，その性能は十分とは言えず，マルチタスク学習の導入等，さらなる精度向上に向けた検討が必要である．また，法科学分野での利用を想定した，判定理由の説明や可視化手法の検討も行う．

謝辞

本研究は，JSPS 科研費 JP19K11975, JP23K11165, JP24K07957, JP24K14988 の助成を受けたものです．

参考文献

[1] Manaka Takamizawa et al, “Same Speaker Identification with Deep Learning and Application to Text-Dependent Speaker Verification,” KES Human Centred Intelligent Systems 2022, pp.149-158, June 2022.

[2] 蒔苗ら，“大規模話者骨導音声データベース”, IEICE Technical Report SP2007-40,2007.

[3] RVC-Project,https://github.com/RVCProject/Retrieval-based-Voice-Conversion-WebUI, 引用日 2023/4/30.

[4] T.Hamada et al, “https://github.com/sarulab-speech/xvector jtubespeech”, 2022.