ISS-SP-222

深層学習を用いた出現音素の偏りに頑健な話者照合手法

佐藤　洋輔† 　小川　哲司†† 　堀内 靖雄† 　黒岩　眞吾†
† 千葉大学大学院融合科学研究科 　†† 早稲田大学大学院基幹理工学研究科

1. はじめに

話者照合とは，音声を用いて個人を認証する技術である．特に，発話内容に制約を課さないテキスト独立型話者照合は利便性の点で有望であるものの，登録時と照合時の発話に含まれる音素の分布の違いが照合性能に悪影響を及ぼすことが知られている．そこで本研究では，音声が持つ情報のうち音素の決定に寄与する情報（以下，音素情報）を除去することで，話者の決定に寄与する情報（以下，話者情報）を正確に抽出し，テキスト独立型話者照合システムの特徴量として用いることを試みた．

2. 出現音素の偏りに頑健な話者特徴抽出

深層学習を用いて音響特徴量から音素情報を除去する手法を提案する．提案手法では，1）音響特徴量から音素情報を表すベクトルの抽出，2）音素情報のベクトル表現の元の音響特徴空間へのマッピング，3）音響特徴と音素情報のみを持つ音響特徴の差を計算，という3つの処理により，音素変動の影響を受けにくい話者照合に適した特徴量を得る．

1）では，音素認識を行う Deep Neural Network (DNN)を構築する．この DNN の入力は音響特徴量 $\boldsymbol{x}$ であり，出力は音素事後確率分布である．このネットワークの出力層の一つ前の層はコード層と呼ばれ，そこから得られる情報は音素を識別するための情報が支配的である[1]．つまり，コード層には音素情報が集約されており，理想的には話者情報を含んでいない．そこで，このコード層の出力を音素情報のベクトル表現（音素情報ベクトル）として使用する．

2）では，抽出した音素情報ベクトルを元の音響特徴量の空間にマッピングする．このマッピングを行う DNN は，前段で抽出した音素情報ベクトルを入力とし，対応する元の音響特徴量 $\boldsymbol{x}$ を教師データとして学習する．したがって，このネットワークの出力 $\tilde{\boldsymbol{x}}$ は音素情報のみを含んだ音響特徴量となることが期待される．

3）では，得られた特徴量 $\tilde{\boldsymbol{x}}$ を用いて元の音響特徴量 $\boldsymbol{x}$ から話者照合に寄与しない音素情報を除去する．本研究では，音響特徴量は音素情報とその他の話者性などの因子の線形結合で表せると仮定する．したがって，以下のように音素情報の影響を受けない特徴量 $\hat{\boldsymbol{x}}$ が得られる．

$$\hat{\boldsymbol{x}} = \boldsymbol{x} - \tilde{\boldsymbol{x}} \qquad (1)$$

表 1 GMM/SVM システムによる等誤り率(%)

従来特徴	0.2564
提案特徴	0.1545

3. 評価実験

GMM-SVM フレームワーク[2]を用いて話者照合実験を行った．提案手法により音素情報の除去を行った音響特徴量と従来の音響特徴量を用いた場合を比較することで，提案手法の有効性を評価した．

本実験では，科学警察研究所により作成された大規模話者骨導音声データベース[3]に収録されているデータのうち，通常のマイクロホンで録音された男性157名の音声を用いた．平均発話時間は約4秒である．評価には，本人音声として78名の各5発声を，詐称者音声として登録話者78名の各390発声を用いた．

音響特徴量は MFCC 12 次元，対数パワー，各Δ項の計 26 次元を用いた．また，UBM，DNN の学習には，評価に用いていない 157 名の各 5 文，約 3000 秒の音声を用いた．音素情報ベクトルを抽出する DNN は 5 層（ニューロン数は 182,1000,500,200,36）の全結合型ネットワークであり，入力は前後 3 フレーム（計 7 フレーム）の音響特徴量を連結して用いた．音素情報ベクトルを復元する DNN も 5 層（ニューロン数は 200,1000,2000,4000,26）の全結合型である．このとき，長母音と短母音を同一音素として扱った．これは，入力される音響特徴量（7 フレーム）が 70ms 程度と短い音声の特徴しか扱えないため，長短の判別が困難であると考えたためである．

実験結果を表 1 に示す．音素情報を除去することで，従来手法の等誤り率を 40%削減することができた．

4. 今後の予定

今後は，DNN が抽出した音素情報の分析を行う予定である．

参考文献

[1] Y. Liu et al., ``Speaker verification with deep features.” Proc. IJCNN 2014, pp. 747-753, Jul. 2014.
[2] W. M. Campbell et al., ``Support vector machines using GMM supervectors for speaker verification,” IEEE Signal Processing Letters, vol. 13, no. 5, pp.308-311, May 2006.
[3] 蒔苗久則ほか, “大規模話者骨導音声データベースの構築と予備的な解析,” 信学技報. SP, 音声, vol.107, no.165, pp.97-102, 2007.