4S-01

発話スタイル依存型話者照合

高山　響† 西田　昌史† 柘植　覚‡ 黒岩　眞吾§ 西村　雅史†
静岡大学† 大同大学‡ 千葉大学§

1. はじめに

話者照合とは、入力された音声が本人か否かを判別する生体認証技術として、セキュリティシステムに応用されている。話者照合システムの多くは、学習時と照合時に同じ発話内容を発話するテキスト依存型、照合時にシステムが指定した発話内容を発話するテキスト指定型、照合時にユーザが自由な発話内容を発話するテキスト独立型の 3 種類に分類される。これらの方式に対してより頑健な話者照合法として、森山ら[1]の歌声を入力とした話者照合システムがある。この手法では、話者の照合と音程の照合を組み合わせ、本人の音声と本人だけが知っている音程の情報の 2 つの鍵を用いることで、詐称者に歌声の音程が漏洩していない場合、詐称者受理率を大幅に削減できることを示した。

この研究を踏まえ、本研究では、頑健な話者照合法として、ユーザが詐称者に知られないようにあらかじめ感情や方言、歌声などの発話スタイルを指定し、指定した発話スタイルで発話しなければ照合されない発話スタイル依存型話者照合法を新たに提案する。

2. 提案手法

提案するシステムの概略図を図 1 に示す。

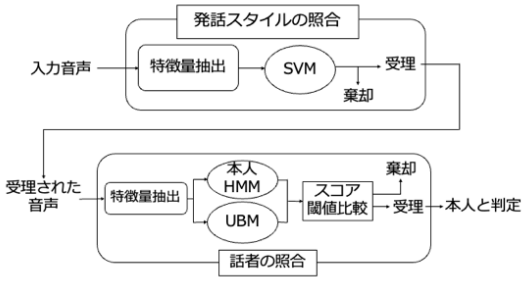

図 1 提案手法概略図

提案手法である発話スタイル依存型話者照合システムは、まず、指定された発話スタイルで正しく発話できているかの照合を行い、受理された音声に対して話者の照合を行う。

本研究では、最初の試みとして、発話スタイルを感情に設定する。ユーザはシステムに対してあらかじめある感情を一つ指定し、システムは感情の照合と話者の照合の 2 段階の照合を行う。感情の照合の詳細は 2.1 節で、話者の照合の詳細は 2.2 節で説明する。

Utterance Style-dependent Speaker Verification
Hibiki Takayama†, Masafumi Nishida†, Satoru Tsuge‡, Shingo Kuroiwa§, Masafumi Nishimura†
Shizuoka University†, Daido University‡, Chiba University§

2.1. 感情の照合

感情認識では、特徴量として基本周波数やパワー、スペクトルなどに関する様々な音響特徴量とその統計量を用いるのが主流である。本研究では特徴量として、openSMILE という特徴抽出ツールで抽出した eGeMAPSv02[2]という 88 次元の特徴セットに対して、ランダムフォレストを用いて 10 次元まで次元削減したものを使用した。

感情の照合では、ユーザが指定した感情で発話したか否かを判別する。したがって本研究では、SVM を用いて、指定した感情か、それ以外の感情かの 2 クラス分類により感情の照合を行う。

2.2. 話者の照合

現在の話者照合には、深層学習に基づく手法と深層学習以外に基づく手法が存在する。Ali ら[3]は、3 つの感情音声コーパスを対象として、深層学習(LSTM、CNN、GRU)を用いた手法と深層学習以外の手法(GMM、SVM、KNN、ANN)で照合精度の比較を行った。その結果、3 つのコーパスの内、2 つで GMM が全モデルの中で最高の性能を達成した。この結果から、感情音声を対象とした話者照合では、深層学習に基づく手法が必ずしも優れているわけではないことがわかる。したがって本研究では、従来広く用いられている GMM-HMM を用いた。

本研究では、まず、本人登録音声を用いて各登録話者の本人モデルを、本人以外の音声を用いて UBM(Universal Background Model)を学習する。

照合時は、入力音声に対する本人モデル、UBM から得られる対数尤度の差を尤度スコアとして求め、尤度比検定を用いて照合した。

特徴量の抽出、話者モデルの学習、尤度計算には HTK(Hidden Markov Model Toolkit) [4]を用いた。

3. 評価実験

まず、2.1 節、2.2 節で述べた手法で実装した提案手法による照合性能の評価を行った。次に、本人、詐称者の感情を完全に正しく照合できるという仮定のもとに、後段の話者照合性能の評価を行い、理想的に感情照合器を学習できた際の提案手法の有効性を検証した。

3.1. 実験条件

実験には JTES コーパスに含まれる日本人話者 100

名(男女各 50 名)が、4 つの感情(怒り、喜び、悲しみ、平静)で各 50 文発話した音声を使用した。

本人 1 名あたりの学習データには本人の全感情 160 発話、評価データには、本人は必ず指定した感情で発話すると仮定し、本人の指定した感情 10 発話、詐称者 99 名の全感情 200 発話を用いた。

感情の照合に用いた各感情の SVM は話者ごとに指定された感情 40 発話、指定以外の感情 120 発話の 2 クラスで学習を行った。

話者の照合で用いる本人モデル、UBM ともに全ての感情の発話を用いて学習させた。本人 1 名に対して UBM は 2 つ作成し、本人および詐称者として使用されない 50 名もしくは 49 名の音声で学習を行い、一つずつ組み合わせを変えて用いた。また、GMM-HMM の混合数を 8、状態数を 8 とし、特徴量には MFCC12+Δ+ΔΔを用いた。

提案手法による話者照合の評価には、設定した閾値に対して以下の FRR、FAR を算出した。

FRR：本人の指定した感情の発話の誤棄却率

FAR：詐称者の発話の誤受理率

3.2. 実験結果

まず、ユーザが各感情を指定した場合の感情照合精度の平均を表 1 に示す。

表 1 感情の照合精度[%]

指定された感情における本人の受理率	58.6
指定された感情における詐称者の受理率	34.9
指定以外の感情における詐称者の棄却率	78.7

表 1 から、もし詐称者がユーザに指定された感情で発話しても約 35%しか受理されず、詐称者の感情音声に頑健であることがわかった。一方で、本人が指定した感情の照合精度は低く、今後の課題といえる。

次に、提案手法による話者照合性能を表 2 に示す。各閾値での FRR と FAR の差が大きく、EER を求めることができなかったため、FRR と FAR の差が最小となる閾値における各エラー率で評価した。

表 2 提案手法による FAR、FRR[%]

	ユーザが指定した感情				平均
	怒り	喜び	悲しみ	平静	
FRR	34.8	50.2	31.0	49.5	41.4
FAR	32.2	24.5	20.1	22.0	24.7

表 2 から、現状の手法では感情照合の本人誤棄却率、詐称者誤受理率が高いため、話者照合のエラー率も高くなってしまうことがわかった。

次に、感情照合なしの従来の話者照合、提案手法における各閾値でのエラー率を図 2 に示す。全ての感情の結果を記載することは困難であるため、ユーザが悲しみを指定した場合の結果を記載した。また、照合精度は他の感情でも悲しみと同じ傾向であった。図 2 において FRR、FAR は従来手法の各エラー率、FRR_p、FAR_p は提案手法の各エラー率を意味する。

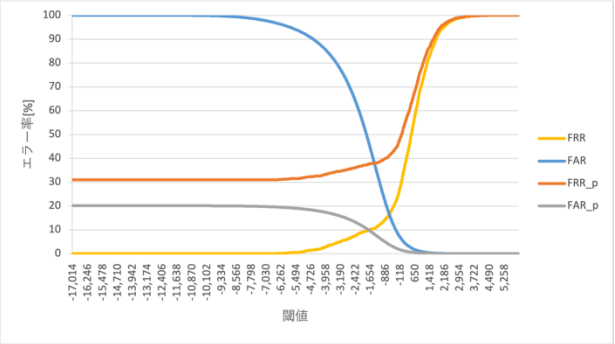

図 2 ユーザが悲しみを指定した場合のエラー率

図 2 中の FAR、FAR_p に注目すると、提案手法で詐称者の誤受理率を大幅に削減できることがわかった。セキュリティにおいては詐称者の誤受理を抑制することが重要であることから、提案手法は有効であると考えられる。

最後に、ユーザが指定された感情ごとに、従来の話者照合法の EER、感情を完全に正しく照合できると仮定した場合の提案手法の EER とその平均を表 3 に示す。

表 3 従来手法、提案手法の EER[%]

	ユーザが指定した感情				平均
	怒り	喜び	悲しみ	平静	
従来手法	18.5	17.5	16.6	16.9	17.4
提案手法	12.2	11.3	10.5	10.3	11.1

表 3 から、感情の照合器を理想的な精度になるよう学習できれば、従来手法よりも EER を改善できることが確認できた。ただし、表 3 の結果が提案手法の照合精度の上限値となるため、より高い精度を得るには話者の照合精度の改善が必要である。

4. おわりに

本研究では、発話スタイル依存型の話者照合法を提案し、発話スタイルとして感情を対象に評価を実施した。評価実験から、提案手法の有効性を確認できたが、現状の手法では感情照合精度が低く、実用的とはいえない。今後の課題として、感情照合、話者照合の精度を改善するため、深層学習を用いた手法の検討を行う。

参考文献

[1] 森山結衣, 堀内靖雄, 黒岩眞吾, “歌声を用いた話者照合システムの検討”, 信学技報, HCGSYMPO, pp.1-5, 2020.

[2] Florian Eyben, et al. “The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing,” IEEE Transactions on Affective Computing, 7(2), pp.190-202, 2016.

[3] Ali Bou Nassif, et al. “Empirical Comparison between Deep and Classical Classifiers for Speaker Verification in Emotional Talking Environments,” Information, 13(10), 2022.

[4] S.Young, et al. “The HTK book(Version3.4),” Cambridge University, 2006.