E-018

音声フレームの重要度推定による音声発生源判別の精度改善
Improvement of Speech Source Discrimination Accuracy by Estimating Importance of Speech Frames

高橋 陽南† 村松 駿‡ 吉田 孝博†
Hinata Takahashi Shun Muramatsu Takahiro Yoshida

1. はじめに

近年、音声認識技術の発展に伴い、スマートフォンやスマートスピーカーのみならず家電等においても音声対話型ユーザーインターフェース (VUI：Voice User Interface) が普及している。これにより、音声を使用してより快適な機器の操作が行えるようになったが、テレビやスマホの動画で再生される音声により、他の VUI 機器が誤作動してしまう事例が発生している。今後さらに、ユーザ周囲の VUI 機器が増えていくことで機器同士が干渉する危険性も考えられる。したがって、実際に人が機器へ直接話す音声と他の機器から発せられる音声を判別する技術が必要とされる。

そこで本研究室では、同一話者の肉声・再生音声・通話音声、ならびに合成音声の計 4 種類の音声発生源を MFCC（メル周波数ケプストラム係数）と CNN（畳み込みニューラルネットワーク）を用いて判別する手法を開発した[1]。類似技術として、ASVspoof[2]などで報告されている、セキュリティ目的でのなりすまし音声検知技術があるが、これは 2 値分類を行う技術である。一方で、音声発生源判別法は複数の VUI 環境下での機器の誤動作を低減させる目的で 4 値分類を行う点が異なる。その後の音声発生源判別法の研究[3]にて 6 つの特徴量が比較され、MFCC-D が最適であることが判明したが、SNR = 0 dB の雑音環境下における正解率が 78.3%であり不十分であった。この従来手法では、音声内の各フレームの判別スコアの単純な和により各音声の発生源を判別しており、各フレームで異なる雑音の影響、すなわち重要度を考慮していないことが一因と考えた。

そこで本研究では、雑音環境下での精度改善を目指し、音声発生源判別に重要なフレームを CNN により推定し、重要度の高いフレームの判別スコアの和を使用する提案手法を考案した。そして、音声発生源判別精度の評価実験にて有効性を確認した。

2. 音声発生源判別法と提案手法

2.1 提案手法を適用する音声発生源判別法

提案手法を適用する音声発生源判別法[3]の処理手順を図 1 中の橙色にて示す。まず、音声の開始時間をオーバーラップしながらシフトすることで、複数のフレームに分割する。その後、各フレームから特徴量を抽出し、CNN に入力してモデルの学習や発生源判別を行う。判別時には、音声内の全フレームの出力確率の単純な和が最大となるクラスを判別結果としている。

2.2 各フレームの重要度推定法（提案手法）

提案手法は、雑音環境下での判別精度改善を目指し、図 1 中の青色にて示すような処理で、各フレームの重要度を CNN で学習・推定させ、その結果を音声発生源判別の各クラスの出力確率の算出時に反映させる手法である。従来手法に対して、新たに重要度を推定する CNN モデルの学習と各フレームの重要度の推定を行う処理を追加した。

音声発生源判別 CNN から出力される各フレームの正解クラスの確率値に対して閾値を設けて、音声発生源判別に使用する重要フレームと使用しない不要フレームのラベル付けを行う。そして、重要度推定 CNN に各フレームの音声特徴量と教師ラベルを入力し学習することで、出力を 2 クラスとする重要度推定モデルを作成する。ラベル付けの際には 5 回分の各フレームに対する音声発生源判別モデルの出力確率を使用し、その出力確率の平均値が 0.9 以上となるフレームを重要なフレームとラベル付けした。音声発生源判別時には、重要度推定モデルによる各フレームの重要度（Softmax による出力確率）によって、重み付けを行う。具体的には、n 個のフレームに分割した元の音声の肉声、通話音声、再生音声、合成音声のそれぞれの確率について次式で求める。

$$a=\sum_{i=0}^{n} x_i w_i \quad (1)$$

ここで、x_i はフレーム i の確率値、w_i はフレーム i の重要度である。

3. 音声発生源判別精度の評価方法

3.1 音声データセット

本研究のモデルの学習と精度比較には、先行研究[1][3]で収集された肉声、再生音声、通話音声、合成音声の 4 種類の音声データセットを用いた。肉声は人が PCM レコーダに対して直接話した音声、再生音声は録音した肉声をスピーカから再生して録音した音声、通話音声は録音した肉声を Skype に通して録音した音声、合成音声は PC 上で音声合成エンジンを使用して作成した音声である。「今日の天気は晴れです」、「明日のスケジュールを教えて」、「リビングの電気をつけて」の 3 種類の発話内容で 10 人分の音声を録音している。また、合成音声は 10 種類の音声合成エンジンを使用している。

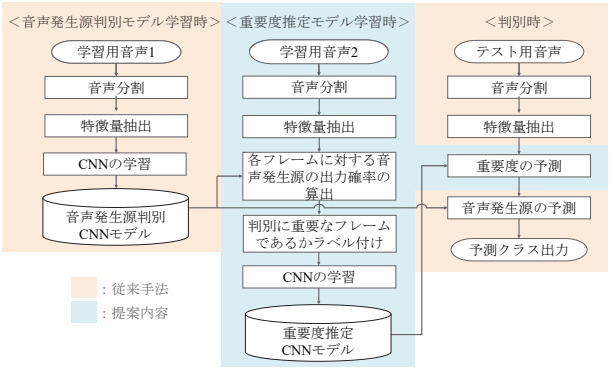


図 1 提案手法のフローチャート

† 東京理科大学 Tokyo University of Science
‡ 東京大学 The University of Tokyo

Input
(conv2D 3×3@32 ReLU)×2
Pooling 2×2
conv2D 3×3@32 ReLU
Pooling 2×2
Flatten
FC 1024, ReLU
Dropout 0.5
FC 4, Softmax
Output

(a) 音声発生源判別

Input
conv2D 3×3@8 ReLU
Pooling 2×2
Dropout 0.3
conv2D 3×3@16 ReLU
Pooling 2×2
Flatten
FC 64
Dropout 0.5
FC 2, Softmax
Output

(b) 重要度推定

図 2　提案手法のモデル構造

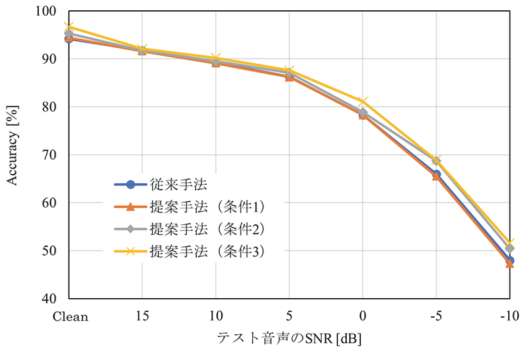
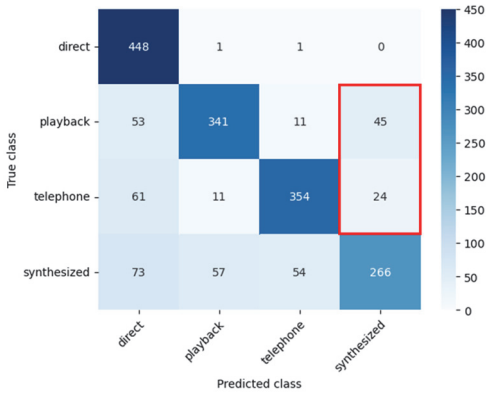
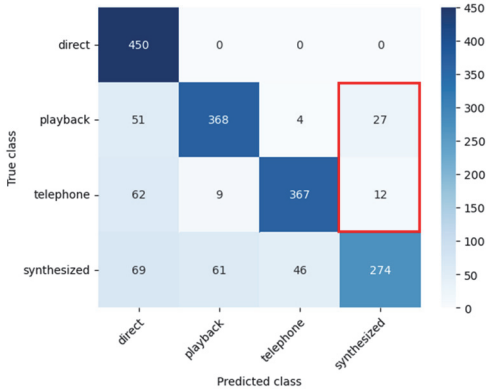


図 3　各 SNR における判別精度



(a) 従来手法



(b) 提案手法（条件 3）

図 4　音声発生源判別結果の混同行列

3.2 各種パラメータ設定値

音声のサンプリング周波数は 44.1 kHz である。各音声は、フレーム長を 100 ms、シフト幅を 50 ms として分割した。

図 2 に音声発生源判別と重要度推定で使用した CNN の構造を示す。学習の最適化手法は Adam、音声発生源判別モデルの学習率は 0.0001、重要度推定モデルの学習率は 0.00004 とした。バッチサイズは 32、学習回数は 30 回とした。 特徴量には次元数 128、FFT 窓長 1024 点、シフト幅 64 点の MFCC-D を使用した。

3.3 評価条件

音声発生源判別評価は、音声発生源判別エンジンの学習用音声、重要度推定エンジンの学習用音声、提案手法のテスト用音声の間で、話者と雑音区間は異なるようにデータセットを使用した。発話内容は、両エンジンの学習用には同一内容の音声を使用し、テスト用の音声は学習用音声と異なる発話内容の音声を使用した。この条件で 5 分割交差検証を 9 回試行し、その平均値を判別精度とした。音声発生源判別の学習には Clean, SNR = 0 dB の音声を使用し、テスト用には Clean, SNR = 15, 10, 5, 0, −5, −10 dB の音声を使用した。また、重要度推定における雑音学習の最適な SNR を検討するため、学習用音声を、条件 1：Clean のみ、条件 2：Clean, SNR = 0 dB、 条件 3：Clean, SNR = 10, 0, −10 dB、の 3 条件で比較した。

4. 精度評価結果

図 3 に従来手法と提案手法について、各 SNR における音声発生源判別正解率を示す。Clean, SNR = 0 dB 時の従来手法の判別正解率はそれぞれ 94.2%, 78.3%であったが、特に提案手法の条件 3 は 96.7%, 81.1%となり、2.5 ポイント、2.8 ポイントの精度向上が得られた。故に、音声の各フレームの重要度を考慮する提案手法が、雑音環境下の音声発生源判別に有効であることを確認した。また、重要度推定モデルの学習用音声に複数の SNR の音声を使用することが有効であることも確認した。

図 4 に、従来手法と提案手法（条件 3）の SNR = 0 dB における混同行列を示す。従来手法と比較すると、提案手法では合成音声として誤認識される音声数が 30 個減少しており、これが精度向上に大きく寄与していることがわかる。

5. まとめ

本研究では、VUI 機器の誤作動防止に有効な音声発生源判別法の雑音環境下での精度改善を目的とし、音声の各フレームの重要度を判別時に考慮する手法を提案した。

精度評価実験により、 音声発生源判別正解率が最大 2.8 ポイント向上し、提案手法が有効であることを確認した。

今後の課題として、重要度推定モデル構造の検討や重要なフレームと重要でないフレームの判別閾値の検討が挙げられる。

参考文献

[1] T. Imaizumi and T. Yoshida, “Speech Source Discrimination Method for Plural Voice User Interfaces Environment,” 2019 IEEE 8th GCCE, pp.1081-1083, 2019, doi: 10.1109/GCCE46687.2019.9015607.
[2] Junichi Yamagishi, et. al, “ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection”. ASVspoof 2021 Workshop, pp. 47–54, 2021.
[3] 前田 健吾, 吉田 孝博, “音声対話型 UI 間の協調動作のための音声発生源判別法に適した特徴量と深層学習モデル”, 信学技報, Vol.122, No.223, pp.29-34, 2022.