# 話者照合のためのなりすまし音声検出に対するノイズ重畳の影響調査

渡部 そら† 塩田 さやか†
† 東京都立大学システムデザイン学部情報科学科

## 1. はじめに

人の音声を用いた生体認証技術である話者照合技術は近年活発に研究されている．重要課題の一つに，なりすまし攻撃である再生音声を検出するなりすまし音声検出がある．これまでに様々ななりすまし音声検出の手法が提案されているが，静音環境下でのデータを用いた評価が主流となっており，ノイズに対する影響について調査がなされていなかった．そこで本論文では，ノイズ重畳したデータを学習に加えることがなりすまし音声検出に与える影響について調査した．

## 2. なりすまし音声検出

録音再生によるなりすまし音声検出タスクとは，直接話者の音声を収録した実発話と，一度録音された実発話音声をスピーカー再生し再収録したなりすまし音声と識別するタスクである．なりすまし音声検出においては再生時のノイズ情報などを特徴として捉えることが多く，シミュレーションによるデータ拡張などがしづらいタスクとなっている．そこで本研究では実際にノイズ重畳がシステムに与える影響について調査する．

## 3. 実験条件

本実験では，なりすまし音声検出のコンペティション ASVspoof [1] でベースラインの一つとして公開されている LFCC-GMM と呼ばれるモデルを用いて評価を行った．入力特徴量には 20 次元の LFCC を用い，GMM の混合数は 512 となっている．モデル学習には 2 つのデータセットを使用した．1 つは ASVspoof から公開された訓練用データセット ASVspoof2019（実発話 4,307 件，なりすまし音声 38,893 件）であり，もう 1 つは ASVspoof2019 のデータセットに SNR が 30dB のホワイトノイズを重畳してデータ拡張したデータセット ASVspoof2019+WN である．評価には次の 2 つのデータセットを使用した．1 つは ASVspoof2021 で公開された評価用データセット（実発話 126,630 件，なりすまし音声 816,480 件）であり，もう 1 つは，独自に収録された音声コーパスである．この独自コーパスは実発話 2100 件となりすまし音声 2100 件で構成されており，4 つの環境下（室外，静かな室内，空調が効いている室内，背景音楽がある室内）で収録されている（1つの環境につき実発話 525 件，なりすまし音声 525 件）．評価指標には等価エラー率（Equal Error Rate; EER）を用いた．

## 4. 実験結果

表 1, 2 に ASVspoof2019 及び ASVspoof2019+WN それぞれで学習したモデルに対して用意した 2 つのデータセットで評価した際の EER を学習回数別に示す．表 1 と表 2 を比較すると，ほとんどの場合でノイズ重畳したデータを学習に加えた方が高い EER となっており，精度が低下していることがわかる．このことから，従来の音声認識や話者認識などで広く用いられるデータ拡張であるノイズ重畳はなりすまし音声検出に必要な特徴を抽出しにくくしてしまうことを示している．つまりモデルの頑健性向上としてのノイズ重畳によるデータ拡張はモデルの汎化性能を逆に下げてしまう可能性が高いことがわかる．次に，学習回数について見てみると，ほとんどの場合で学習回数が 10～20 回の時に最も低い EER となっている．対数尤度の収束状況から確認できる学習の進み具合としては 100 回程度であることが確認できていたが，実験結果からは少ない学習回数の方が識別性能が高いため学習が難しいタスクであることがわかる．このことから汎化性能を向上させるためには複雑なモデルを用い，かつデータ拡張についても慣例的に広く用いられるものではなくよりタスクに適したものを探す必要があることがわかった．

表1. ASVspoof2019 で訓練時の EER(↓)

| 学習回数(回) | | 10 | 20 | 30 | 40 | 60 | 100 |
|---|---|---|---|---|---|---|---|
| ASVspoof2021 | | 35.93 | **34.38** | 34.71 | 34.97 | 35.37 | 35.19 |
| 独自コーパス | 室外 | **1.52** | 2.42 | 2.67 | 1.81 | 4.25 | 3.24 |
| | 室内(静か) | **1.76** | 8.64 | 7.26 | 9.63 | 24.76 | 14.83 |
| | 室内(空調) | **0.38** | 4.50 | 5.00 | 4.15 | 15.05 | 4.23 |
| | 室内(音楽) | **7.69** | 48.97 | 43.40 | 59.13 | 81.87 | 76.12 |

表 2. ASVspoof2019+WN で訓練時の EER(↓)

| 学習回数(回) | | 10 | 20 | 30 | 40 | 60 | 100 |
|---|---|---|---|---|---|---|---|
| ASVspoof2021 | | 45.17 | **45.15** | 46.11 | 45.10 | 45.30 | 46.32 |
| 独自コーパス | 室外 | **5.92** | 10.70 | 7.37 | 6.93 | 7.24 | 10.86 |
| | 室内(静か) | **8.18** | 17.38 | 17.14 | 13.14 | 12.65 | 24.54 |
| | 室内(空調) | **5.43** | 21.97 | 9.54 | 14.67 | 15.43 | 22.64 |
| | 室内(音楽) | 23.61 | 39.60 | 25.70 | **20.11** | 35.58 | 40.71 |

## 5. まとめ

本論文では，なりすまし音声検出におけるノイズ重畳の影響を確認した．ノイズ重畳によるデータ拡張は頑健性向上に貢献しないことから，今後の課題としてより適切なデータ拡張手法の検討や深層学習モデルでの評価などが挙げられる．

## 参考文献

[1] ASVspoof “ASVspoof 2021:Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan”