

録音再生攻撃環境を考慮したマルチタスク学習における なりすまし音声検出の検証*

☆菅野滉大, 塩田さやか (東京都立大学)

1 はじめに

マイクさえあれば導入できる手軽さから音声を用いた生体認証の需要が高まる一方, そのセキュリティを脅かすなりすまし音声攻撃への対策が急務となっている. そのため, なりすまし音声攻撃を検出するためのなりすまし音声検出という技術が活発に研究されている. なりすまし音声攻撃のうち, 録音再生攻撃について, これまでにいくつかのデータベースが公開され [1–4], それらのデータベースをもとに様々なモデルが提案されてきている [5]. 録音再生攻撃の検出は, なりすまし音声の録音再生・攻撃収録の過程で含まれるリプレイノイズという微細な音響的歪みの検出が重要であることが報告されているが, 微細な特徴であることから検出が難しく, またモデルの頑健性も課題となっている.

これまでに, なりすまし音声検出にマルチタスク学習を用いた手法が提案されている [6]. この手法では, なりすまし音声検出タスクに加え, 録音再生攻撃の際の再生機器や再収録機器などのリプレイノイズを分類するサブタスクを同時に学習することで, ASVspooof 2017 評価セットで性能が大幅に向上したことが報告されている. しかし, [6] では, 各タスクの損失重みを均等に設定しており, その重み調整が性能に与える影響については検証がされていなかった. また, 他のデータベースでの頑健性についても評価されていないという課題もある.

なりすまし音声検出におけるマルチタスク学習は, リプレイノイズという録音再生攻撃によるなりすまし音声の微細な特徴をサブタスクとして学習させる手法であり, ネットワークをリプレイノイズに敏感な特徴量抽出器として学習できることが期待される. これにより, 学習データ以外のデータベースに対しても頑健な検出性能を発揮できると考えられる. 本研究では, この効果を検証するため, 最先端のなりすまし音声検出モデルにマルチタスク学習を適用し, 複数のデータセットを用いた評価を実施することでマルチタスク学習が汎化性能に与える影響を検証する. さらに, マルチタスク学習におけるタスクごとの損失重み調整が検出精度に与える影響についても調査する. 実験では, 最先端モデルの一つである AASIST を基盤とし, wav2vec 2.0 を特徴抽出器として組み合わせ

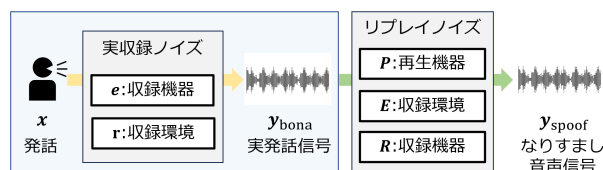


Fig. 1: 録音再生攻撃における音声入力過程のモデル
たモデル [7] に対してマルチタスク学習を行った. 実験結果より, 適切なタスクの損失重みを与えることで, 複数のデータセットに対して性能を向上させることを報告する.

2 関連研究

2.1 なりすまし音声検出

なりすまし音声攻撃とは, 攻撃者が他人の声になりすまして話者照合システムを突破しようとする攻撃である. なりすまし音声攻撃のうち, 本研究で焦点を当てている録音再生攻撃では, 事前に実発話を不正録音した音声を再生機器から再生し, その再生された音声を, システムの収録機器を通じて入力する方法を想定している. なりすまし音声検出は, 入力された音声人間が実際に話している実発話かなりすまし音声かを判定する技術である. 多くの研究では, なりすまし音声検出のための国際コンペティションである ASVspooof challenge [8] で公開されているデータセットを用いてモデルが学習され, 検出性能が評価されている. しかし, 既存の多くのアプローチで学習データと評価データのドメインが異なると性能が低下する課題を抱えている.

2.2 録音再生攻撃における音声入力過程

録音再生攻撃において, 実発話が収録される過程と, なりすまし攻撃音声収録される過程を Fig.1 に示す. 実発話信号 y_{bona} は, 話者の発話 x が, 収録環境の特性 e , ASV システムの収録機器の特性 r を介してシステムに入力される. これは畳み込みとして以下のように表される.

$$y_{\text{bona}} = x * e * r \quad (1)$$

ここで, 太字の記号は離散時間信号ベクトルを表し, * は畳み込み演算を表すものとする. 一方, なりすまし

*Investigation of Multi-Task Learning for Environment-Aware Replay Spoofing Detection,
by KANNO Kouta, and SHIOTA Sayaka (Tokyo Metropolitan University)

し音声信号 y_{spooof} は、実発話信号が録音され、その後再生されるという過程を経る。この過程で、再生機器の特性 P 、収録環境の特性 E 、収録機器の特性 R といった音響特性が再生された信号に追加で積み込まれる。したがって、なりすまし攻撃音声信号は以下のように表される。

$$y_{\text{spooof}} = y_{\text{bona}} * P * E * R \quad (2)$$

ここで大文字で書かれた、 P 、 E 、 R をまとめてリプレイノイズと呼ぶ。これらのリプレイノイズは、実発話信号に存在しないため、リプレイノイズの有無が、実発話となりすまし音声の主要な特徴となる。本稿では、以降実発話となりすまし音声はそれぞれ y_{bona} 、 y_{spooof} を指すこととする。

2.3 マルチタスク学習によるなりすまし音声検出

本節では、従来法である [6] について説明する。従来法では、2.2 節で述べたリプレイノイズに着目し、なりすまし音声検出とリプレイノイズ分類を同時に行うマルチタスク学習を提案している。リプレイノイズ分類タスクとして、再生機器分類、収録環境分類、収録機器分類の3つのタスクを追加しており、合計4つのタスクでモデルを学習している。マルチタスク学習におけるすべてのタスクの損失は等しく重み付けされており、識別モデルには LCNN [9] が使われている。各タスクの損失重みは均等に設定されており、評価も単一データセットに限られていた。また検出モデルも LCNN が採用されており、より高性能な最新アーキテクチャでは未検証であった。

3 提案法

本研究では、従来法の課題を踏まえ、マルチタスク学習の各タスクの損失重みを調整したときの性能の変化や、複数データセットで評価したときの頑健性について大規模事前学習モデルを用いた高性能モデルで調査した。

3.1 データセット

従来法では ASVspoof 2017 で公開されたデータベースを学習データに用いていたが、本研究では ASVspoof 2021 の評価計画に従い、ASVspoof 2019 データセットを学習に用いた。そのため、使用する学習データのメタ情報に即して追加タスクを設定する必要があり、モデル構造を従来法から変更する必要がある。ASVspoof 2017 では、リプレイノイズ情報として再生機器、収録環境、収録機器が提供されていたが、ASVspoof 2019 では、再生機器と再生時の再生機器と収録機器の距離に関する情報のみが提供されている。したがって、本研究におけるマルチタスク学

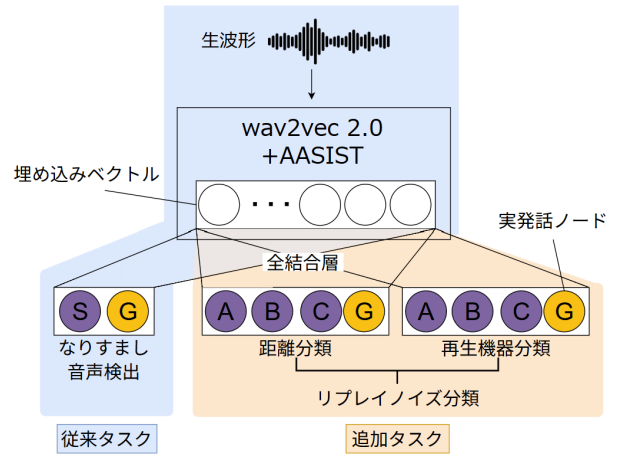


Fig. 2: マルチタスク学習を行うモデル構造

習は、なりすまし音声検出に加え、再生機器分類と再生時の再生機器と収録機器の距離分類の3つのタスクで学習を行う。

3.2 提案モデル

3.1 節に基づき、マルチタスク学習を行う。基にするモデルには、ASVspoof 2021 評価セットで高い性能が得られることが報告されている wav2vec 2.0 と AASIST を組み合わせたモデル (w2v+AASIST) [7] を採用した。これは自己教師あり学習を用いて事前学習された wav2vec 2.0 フロントエンド [10] と、なりすまし音声検出に特化した AASIST バックエンドを組み合わせたモデルである。Figure 2 に提案するモデル構造を示す。従来のなりすまし音声検出用の全結合層はそのまま残し、距離分類および再生機器分類のための全結合層を新たに追加している。全結合層より前の層はすべてのタスクで重みを共有しており、この構造は従来法と同様である。学習時の損失関数についてはタスクごとの損失を考慮し、以下の式のように、重みを付けて加算し、逆伝搬を行っている。

$$L_{\text{sum}} = w_{\text{spooof}} \cdot L_{\text{spooof}} + w_{\text{distance}} \cdot L_{\text{distance}} + w_{\text{device}} \cdot L_{\text{device}} \quad (3)$$

ここで、 L_{spooof} はなりすまし音声検出タスク、 L_{distance} は距離分類タスク、 L_{device} は再生機器分類タスクの損失をそれぞれ表している。また、 w_{spooof} 、 w_{distance} 、 w_{device} は各タスクの損失重みである。

4 なりすまし音声検出実験

4.1 実験条件

本実験では、モデルの学習データに ASVspoof 2019 を使用した。評価データには ASVspoof 2019, ASVspoof 2017, ASVspoof 2019 real, ASVspoof 2021, JSpAW の各評価セットを用いた。これらのデータセットは、いずれも実発話と録音再生されたなりす

Table 1: 評価用データセットの音声収録時間（時間）

Dataset	Bonafide	Spoof
ASVspoof2019 [2]	19.7	114.8
ASVspoof2019real [2]	0.6	2.2
ASVspoof2017 [1]	1.2	10.8
ASVspoof2021 [3]	83.2	547.0
J-SpAW [4]	0.5	4.4

まじ音声で構成されている。各評価セットにおける発話数については、Table 1 にまとめている。

マルチタスク学習においては、タスクごとの損失重みを調整した際の検出精度への影響を調査するため、複数の損失重み設定で学習を行った。学習の最大エポック数は100とした。再生機器 (Device) のカテゴリは perfect (A), high (B), low(C), 再生機器と収録機器の距離 (Distance) のカテゴリは 10-50cm (A), 50-100cm (B), 100cm 以上 (C) となっており、Fig.2 の紫の数字 (A, B, C) に該当する。実発話の場合には G がラベル付されている。データ数の不均衡を補正するため、損失計算時には、正規発話となりすまし音声、および各リプレイノイズカテゴリそれぞれのデータ数の逆数の比を重みとして乗算した。フロントエンドである wav2vec 2.0 は XLS-R 300M を用いて事前に学習された重みからファインチューニングを行い、バックエンドの AASIST 部分は特定のシード値に基づいてパラメータを初期化して学習した。モデルの選択基準としては、ASVspoof 2019 の開発セットにおいて、なりすまし音声検出タスクで最も高い正解率 (Accuracy; Acc) を示したエポックを採用した。評価基準はなりすまし音声誤検出率と実発話誤棄却率が等しくなる等価エラー率 (Equal error rate; EER) である。

4.2 結果

Table 2 に、w2v+AASIST の各損失重みにおける開発セットでの各タスクの正解率 (Development accuracy; Dev Acc) と各評価セットでの EER を示す。なお、表中の損失重みおよび Dev Acc における Spoof, Dist., Devi. は、それぞれなりすまし音声検出、距離分類、再生機器分類を表す。ASVspoof 2019 を用いた評価では、マルチタスク学習を導入した No.2 から No.7 までのほぼすべての組み合わせにおいて、なりすまし検出タスクのみを学習したシングルタスクモデルの性能を上回った。特に Spoof, Dist., Devi. の損失重みが 8:1:1 の No.7 では、EER が約 21%削減された。この結果は、w2v+AASIST のような高性能なモデルにおいてもマルチタスク学習がなりすまし音声検出の性能向上に有効であることを示唆している。学習時とは異なるデータセットに対する評価で

は、適切な重みが与えられたときにはシングルタスクのモデルよりも性能が改善する傾向にあったが、適切な重みでない場合には性能が悪化する場合もあった。また、ASVspoof 2021 の評価セットでは改善が得られなかった。この要因として、データセットごとに含まれるリプレイノイズの特性が異なるため、学習時に重視すべきタスクのバランスもまた異なってくる事が考えられる。そのため、マルチタスク学習で期待していた頑健性あまり得られていないということがわかる。次に、損失重みの違いに関する詳細な考察を述べる。ASVspoof 2019 では、なりすまし音声検出の重みを 0.8 と高くし、各リプレイノイズ分類の重みを 0.1 ずつに抑えた構成で、最も低い EER となった。このことからリプレイノイズ分類は、損失重みを低く設定し、なりすまし音声検出タスクの学習を補助する正則化として機能させることが有効であることが分かった。また、リプレイノイズ分類タスクの中では、特に再生機器分類が EER 改善に大きく寄与する傾向が見られた。具体的には、リプレイノイズ分類タスクを単一に制限した表中 No.5 (再生機器分類のみ) と No.6 (距離分類のみ) の比較では、再生機器分類を学習している前者が全ての評価セットで低い EER となった。同様に、なりすまし音声検出の重みを固定し、距離分類と再生機器分類の重みを入れ替えた表中 No.3 と No.4 のモデルの比較においても、再生機器分類の重みを高く設定した No.3 のモデルが全ての評価セットで低い EER となった。この結果から、リプレイノイズの中でも特に再生機器に由来する音響特徴が、なりすまし音声検出において重要な手がかりとなる事が考えられる。一方で、補助タスクの選択と重み付けには注意が必要であることも明らかになった。開発セットでの正解率を見ると、距離分類タスクは他のタスクよりも正解率が低く、学習が困難であったことが分かる。実際に、距離分類の重みを最も大きくした表中 No.6 のモデルでは、シングルタスクモデルより検出精度が悪化する結果となった。このことから、マルチタスク学習において、学習が困難な補助タスクの損失重みを不適切に高く設定すると、主タスクの精度向上を阻害する可能性があるという知見が得られた。

Figure 3 に、w2v+AASIST の全結合層より前の層から得られた埋め込みベクトルを t-SNE を用いて可視化した結果を示す。シングルタスク学習モデルと、本研究の評価で最も低い EER を達成した表中 No.7 のマルチタスク学習モデルの埋め込みベクトルを可視化した。この図から、マルチタスク学習モデルの埋め込みベクトルが、各タスクのラベルに応じて明確なクラスを形成していることがわかる。また興味

Table 2: 各損失重みにおける開発セット正解率および各評価セットでの EER (%)

No.	損失重み			Dev Acc (%)			EER (%)				
	Spoof	Dist.	Devi.	Spoof	Dist.	Devi.	2019	2019r	J-SPAW	2017	2021
1	1	-	-	96.24	-	-	3.86	26.94	29.27	32.58	40.09
2	1/3	1/3	1/3	96.52	68.95	89.14	3.81	34.21	22.75	24.34	42.60
3	10/18	3/18	5/18	96.35	71.47	89.67	3.65	22.18	28.51	25.74	41.16
4	10/18	5/18	3/18	96.35	70.84	87.79	3.57	34.63	32.89	30.59	43.58
5	7/10	-	3/10	96.77	-	88.87	3.23	21.71	25.25	28.82	40.68
6	7/10	3/10	-	96.05	70.05	-	4.21	31.20	29.13	34.75	40.88
7	8/10	1/10	1/10	96.19	70.48	87.78	3.06	33.38	33.06	27.43	41.86

深いことに、シングルタスク学習モデルにおいても、明示的にラベルを与えていないにもかかわらず、距離や再生機器の種類といったリプレイノイズの特性ごとに、埋め込みベクトルがある程度まとまる傾向が見られた。これは、高性能ななりすまし音声検出モデルが、その学習過程でリプレイノイズの物理的特性を暗黙的に学習していることを示している。つまり、これらの結果からリプレイノイズの特性がなりすまし音声検出に本質的に関連しているという従来法の仮説を裏付け、リプレイノイズの特性をマルチタスク学習によって明示的にモデルへ学習させるアプローチの有用性を示している。

5 結論

本研究では、録音再生攻撃に対するなりすまし音声検出において、マルチタスク学習を適用した際の、タスクごとの損失重みの影響を調査した。また、マルチタスク学習が頑健性の向上につながると考え、複数のデータセットを用いた評価を実施した。結果として、適した重みが与えられたときにはマルチタスク学習がなりすまし音声検出の性能を効果的に向上させることを示した。しかしながら、マルチタスク学習によって必ずしも頑健性が上がるわけではないことも確認された。今後の課題として、学習時のタスクごとの損失重みと評価時のデータセットの特性について調査していく必要がある。

謝辞 本研究の一部は JSPS 科研費 JP24K14993, SCAT および ROIS データサイエンス共同利用共同研究拠点 (DS-JOINT) の助成 (課題番号: 026RP2025) の助成を受けたものである。

参考文献

[1] T. Kinnunen *et al.*, “The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection,” in *Proc. Interspeech*, pp. 2–6, 2017.

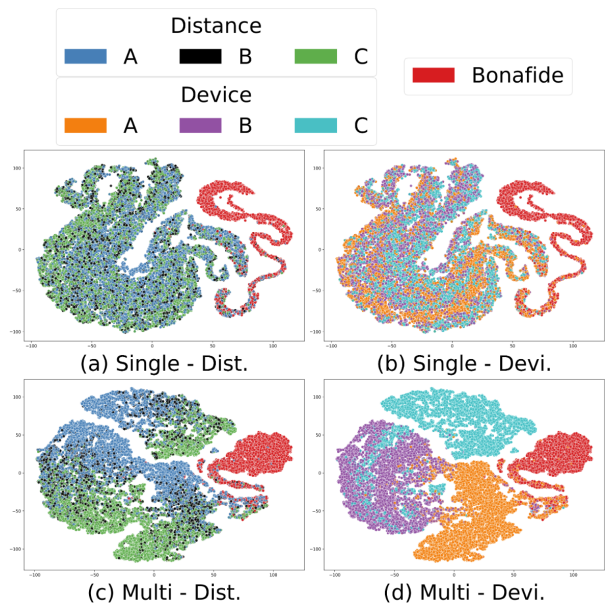


Fig. 3: t-SNE による埋め込みベクトルの可視化結果

[2] X. Wang *et al.*, “Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech,” *Trans. on Comput. Speech Lang.*, vol. 64, 2020.

[3] H. Delgado *et al.*, “ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan,” *arXiv preprint arXiv:2109.00535*, 2021.

[4] S. Shiota *et al.*, “J-SPAW: Japanese speaker verification and spoofing attacks recorded in-the-wild dataset,” in *Proc. Interspeech*, 2025(accepted).

[5] A. Khan *et al.*, “Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures,” *Artificial Intelligence Review*, vol. 56, pp. 513–566, 2023.

[6] H.-J. Shim *et al.*, “Replay spoofing detection system for automatic speaker verification using multi-task learning of noise classes,” in *Proc. TAAI*, pp. 172–176, 2018.

[7] H. Tak *et al.*, “Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation,” in *Proc. Odyssey*, pp. 112–119, 2022.

[8] <https://www.asvspoof.org/>.

[9] X. Wu *et al.*, “A Light CNN for Deep Face Representation With Noisy Labels,” *Trans. on Comput. Speech Lang.*, vol. 13, no. 11, pp. 2884–2896, 2015.

[10] A. Babu *et al.*, “XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale,” *arXiv*, vol. abs/2111.09296, 2021.