Vol.2020-NL-246 No.33
Vol.2020-SLP-134 No.33
2020/12/3

深層学習に基づくなりすまし検出の言語依存性に関する調査

奥野 桜子[1,a)]　塩田 さやか[1,b)]　貴家 仁志[1,c)]

概要：本論文ではなりすまし検出システムの言語依存性について調査する．近年，人間の身体的特徴を利用して本人認証を行う生体認証システムが日常生活の中に普及してきている．このうち人間の声を生体情報として用いる生体認証を話者照合と呼ぶ．話者照合はマイクがあれば実現可能であり導入コストが低いという利点がある．一方で，録音再生音声や合成音声を用いたなりすまし攻撃も容易に可能であることからなりすまし攻撃に対する対策が急務とも報告されている．本研究では，深層学習を用いてなりすまし検出実験を行い，言語や収録環境などの異なるデータベースを用いてシステムの性能を評価した．さらに，システムに言語ラベルを追加した場合の実験も行った．これらの結果から，本来は言語に依存性のないと期待されるなりすまし検出においてモデル内に言語の依存性が残っていたこと，さらになりすまし検出に関係ない要因による依存性が残っていることを報告する．また，言語ラベルの追加によって言語依存性が緩和できることを報告する．

Investigation on language dependence of deep-learning based spoofing detection for speaker verification

Abstract: In this paper, we investigate language dependence of deep-learning based spoofing detection. Recently, biometric authentication systems that perform personal authentication using human physical features have become widespread. Automatic speaker verification (ASV) is one of the biometric authentication by using human voice. Since ASV systems can be installed with a microphone only, it is easy to introduce to many applications. However, it has been reported that ASV systems suffer from spoofing attacks, e.g., speech synthesis and replay. Therefore, it is increased importance to develop reliable ASV systems. In this paper, we perform some experiments using deep-learning based spoofing detection and evaluate the performance of the systems in terms of difference in databases or languages. Moreover, we also propose a language-aware modeling for spoofing detection. From the experimental results, language dependence remains in the conventional deep-learning based model in spoofing detection, which was originally expected to be language independent. It also shows that language dependency can be reduced by language-aware training.

1. はじめに

近年，人間の身体的特徴を利用して本人認証を行う生体認証システムが日常生活の中に普及してきている．実用化の例として，スマートフォンの指紋認証や，入出国管理に用いられる顔認証などがある．生体認証システムのうち，人間の声を生体情報として用いる技術を話者照合と呼ぶ．話者照合はマイクがあれば実現可能であり導入コストが低いという利点がある．また，スマートスピーカやスマートフォンとの親和性も高いことから実用化が期待されている技術である．一方で，録音再生音声や合成音声を用いたなりすまし攻撃も容易に実行可能であることからなりすまし攻撃に対する対策が急務とも報告されている [1]．なりすまし攻撃の例には，スマートスピーカーなどのシステムにおいて，登録話者の音声を録音した音声を再生する事で第 3 者がシステムを誤操作させることなどが挙げられる．そこで，話者照合に対するなりすまし攻撃の対策法について統一したデータで比較評価をするための ASVspoof (Automatic Speaker Verification spoof) Challenge が 2015 年から隔年で開催されるようになった．これまでに公開されてきたデータベースは徐々に大規模かつ複雑化してきてはいるが，ASVspoof Challnge 以外で公開されているデータベースがなく，多角的な評価や知見が十分に存在してい

[1] 現在，東京都立大学 システムデザイン研究科 情報科学域
Presently with Tokyo Metropolitan University, Faculty School of Systems Design, Department of Computer Science
a) okuno-sakurako@ed.tmu.ac.jp
b) sayaka@tmu.ac.jp
c) kiya@tmu.ac.jp

Vol.2020-NL-246 No.33
Vol.2020-SLP-134 No.33
2020/12/3

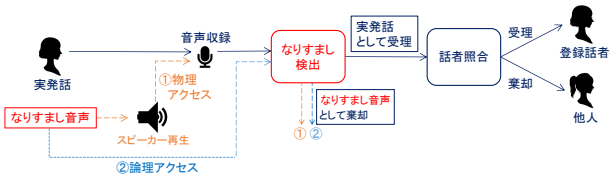

図 1 話者照合システムとなりすまし検出システムおよびなりすまし攻撃のアクセスフロー

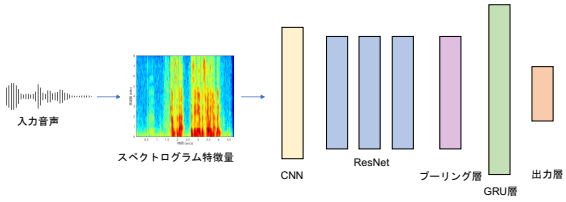

図 2 CNN-GRU モデル

なかった．

なりすまし検出の手法として，ASVspoof Challenge に則った枠組みだけでなく声の生体検知 [2,3] やマルチチャンネル信号を用いたなりすまし検出法 [4,5] など様々な対策法がこれまでに報告されてきている．それらの評価に用いられているデータベースは言語や収録環境などが ASVspoof とは異なる．これらのデータを学習や評価に用いることでこれまでに提案されてきた手法の多角的な評価が可能になると考えられる．そこで，本研究では，ASVspoof Challenge 2019 で高い性能を得ている CNN-GRU (Convolutional neural network - Gated recurrent units) を基に，データベース，言語，収録環境への依存性を評価する．文献 [2,5] で紹介されているデータベースは日本語話者による日本語の発話が収録されており，言語の依存性については用意に評価ができることから本研究ではさらに CNN-GRU モデルに言語ラベルを追加して実験を行うことで言語に対する依存性がどの程度緩和できるかについても調査を行った．実験においては，英語だけ，日本語だけおよび両方のデータを学習データとして用いた場合のなりすまし検出実験を行い，性能の改善及びそれ以外の要因についても考察を行った．これらの結果から，本来は言語に依存性のないと期待されるなりすまし検出においてモデル内に言語の依存性が残っていたこと，さらになりすまし検出に関係ない要因による依存性が残っていることを報告する．

2. 関連研究

2.1 なりすまし検出

話者照合はマイクがあれば実現可能であり導入コストが低いという利点があるが，他の生体認証技術と同様にスマートフォンや IC レコーダーなどを使うことで簡単になりすましを行えることが問題視されている．そこで，図 1 に示すようになりすまし検出を話者照合の前段で行うことを考える．近年，なりすまし検出手法について比較評価するコンペティション ASVspoof Challenge が隔年で開催されているが，その際に想定されているなりすまし攻撃のフローが 2 系統ある．その 1 つが登録者の録音音声や合成音声などのなりすまし音声を認証時にスピーカーで再生する物理アクセス（図 1①) であり，もう 1 つがシステムの入力系統に直接合成音声等を割り込ませる論理アクセス (図 1②) である．物理アクセスと論理アクセスの根本的な違いはシステムに入力される音声がスピーカーで再生されてマイクで再収録されるというプロセスが存在するか否かということである．録音再生音声による物理アクセスは，専門的な知識を必要とせず容易に行えてしまうため，特に認証のフレーズが固定の場合現実的かつ検出が難しいなりすまし攻撃となっている．

2.2 ASVspoof Challenge

ASVspoof Challenge では，2015 年に論理アクセス攻撃，2017 年に物理アクセス攻撃に対する対処法についてのコンペティションを開催してきた [6,7]．それらのコンペティションを開催した結果，それぞれのなりすまし攻撃に対して，先行研究として様々な音声特徴量およびモデルを用いた対策手法が提案された [8–11]．過去 2 回のコンペティションで公開されたデータベースはデータ量が中規模であり，深層学習が十分にできないことが指摘されたため，2019 年に開催された ASVspoof 2019 [12] では論理アクセスと物理アクセスそれぞれを想定した大規模なデータが公開された．これらのデータベースを用いて様々な深層学習に基づくなりすまし検出法が提案されてきている [13–15]．

2.3 CNN-GRU

本研究では，ASVspoof 2019 において高い検出性能を得ており，実装されたコードも公開されている CNN-GRU モデルをベースラインとなるシステムとして用いる．図 2 に示した CNN-GRU モデルの概要のように，入力音声からスペクトログラムを抽出し特徴量としており，畳み込み層を用いて入力特徴を処理し，フレームレベルの埋め込みを抽出する．畳み込み層は，学習を容易にするために，同一性マッピングを持つ残差ブロックで構成されている．次に抽出されたフレームレベルの特徴を発話レベルの特徴に集約するため，GRU 層を採用している．出力層では入力された音声が実音声かなりすまし音声かのどちらであるかを示している．GRU を含むネットワークは様々なタスクで性能を示しており，なりすまし検出においても有効であることが報告されている [16]．

Vol.2020-NL-246 No.33
Vol.2020-SLP-134 No.33
2020/12/3

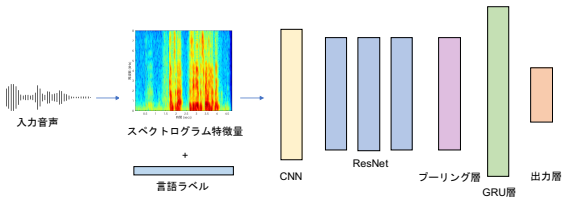

図 3 提案モデル

3. 提案法

3.1 データベースの依存性

本研究グループで使用可能なデータベースの詳細は表 1 に示すように 4 種類のみとなっている．ASV2017 および ASV2019 はそれぞれ 2017 年，2019 年に開催された ASVspoof Challenge において公開されたデータベースを示している．ASV2019 データベースの音声数が最も多く，DNN (deep neural network) を学習するのに十分な数となっている．これは，深層学習を使った手法が話者認識や話者照合などの先行研究で高い精度を出しているためである．ASVspoof 2019 で公開されたデータセットは物理アクセス，論理アクセス両方であったが，本研究で ASV2019 というのは物理アクセスのためのデータセットのみを指すこととする．ASVspoof 2019 で用意された物理アクセスデータは，録音再生音声を用意するために再収録を実際に行うのではなく，録音再生による音声の歪みをコンピュータでシミュレーションして生成している．これは，再収録のコストが高く，深層学習が可能な程度の規模のデータベースを用意することが難しかったために取られた手段である．一方，2017 年に開催された ASVspoof 2017 では，実際に再収録されたデータが物理アクセスのデータベースとして公開されている．これは，データ量としては中規模と言える量ではあったが多くのコンペティション参加者が深層学習の実装が難しいと指摘したために 2019 年の録音再生音声のシミュレーションにつながったデータベースである．VLD (voice liveness detection)，IRT は文献 [2, 5] において用いられているなりすまし攻撃を想定して実際に録音再生を行ったデータベースとなっている．これらのデータはステレオチャンネルで収録されているため実際に有用となるデータ量は記載されているデータ数の半分以下になっている．また，ASV2019 及び ASV2017 データベースは発話内容が全て英語となっているが，VLD 及び IRT は発話内容が日本語のデータベースとなっており，使用言語に違いがある．

3.2 言語依存性

前述のとおり，なりすまし検出のための大規模な公開されているデータは種類が少なく，また収録環境や言語なども統一された要素が少なくなっている．本来，なりすまし検出は発話内容に依存せず，実発話か再生音声かを検出するシステムである．しかしながらデータベースが存在していないことから言語に関しては依存性が確認されていなかった．文献 [5] において，ASVspoof Challenge のデータを用いたシステムとの比較が行われているが ASVspoof Charenge のデータで学習されたモデルでの評価値が著しく低いことからもデータベースの違い，つまりドメインの違いが性能に大きく影響を与えていることが確認できる．また，ASVspoof 2019 で開発されたシステムが ASV2017 を用いて検証すると精度があまり出ないという報告もされている [17]．さらに文献 [5] で用いられているデータベースは日本語話者による発話が収録されているため，言語の依存性について評価することが可能となる．また，言語の依存性を除去するモデルを用いることで性能評価の結果がどのように変化するかについても検討する必要がある．

表 1 なりすまし検出のためのデータベース

		ASV2019 (Sim)	ASV2017 (Real)	VLD (Real)	IRT (Real)
言語		英語	英語	日本語	日本語
データ数	学習	83700	4726	18000	
	テスト	134730	13306		5100

表 2 実験条件

特徴量	スペクトログラム
攻撃音声	録音再生音声 (物理アクセス)
学習データ	ASV2017, ASV2019, VLD
テストデータ	ASV2017, ASV2019, IRT
スペクトル抽出	
サンプリング周波数	16kHz
窓	ハミング窓
窓長	800 ms
窓シフト	480 ms
FFTbin	2048
学習条件	
モデル	CNN-GRU
学習回数	200
正規化	Batch Normarization
バッチサイズ	32, 256
Learning Rate	0.0005

3.3 提案モデル

なりすまし検出システムの言語依存性を評価するために，既存の CNN-GRU に基づくシステムに対して言語ラベルを付与したモデルを検討する．言語ラベルを付与する方法としては様々なものが考えられるが，本研究では各フレームに 1 次元のラベルを付与することでモデル構造には影響を与えない方法を用いることとした．図 3 に示す通り入力特徴量であるスペクトログラムに言語ラベルを付与してデータの入力を行っている．

表 3 実験結果 (%)

				言語ラベル無し			言語ラベル有り		
	英語		日本語	英語		日本語	英語		日本語
	ASV2019 (Sim)	ASV2017 (Real)	VLD (Real)	ASV2019 (Simulation)	ASV2017 (Real)	IRT (Real)	ASV2019 (Simulation)	ASV2017 (Real)	IRT (Real)
(A)			✓	44.43	43.94	45.82	**43.29**	44.95	47.33
(B)		✓		39.15	18.85	57.35	**37.81**	23.34	**54.25**
(C)	✓			8.65	49.77	32.94	**7.50**	50.54	39.33
(D)		✓	✓	39.59	24.36	55.59	**37.72**	27.5	**54.67**
(E)	✓		✓	9.76	42.06	28.17	**7.92**	43.17	29.09
(F)	✓	✓		6.74	25.52	53.19	**6.21**	26.64	**41.35**
(G)	✓	✓	✓	6.68	23.96	36.62	7.28	35.35	47.33

4. 実験

4.1 実験条件

本実験では表 1 に示すデータベースを用いてなりすまし検出性能の比較を行った．ASV2019, ASV2017, VLD の 3 種類を学習データとして用い，実験においてはデータベース単体で学習したもの，2 種類を混ぜて用いたもの，3 種類全て用いたものの合計 7 種類の実験条件でシステムを評価している．それぞれの等価エラー率 (Equal Error Rate; EER) を計算し，シミュレーションデータとリアルデータ，英語と日本語の 2 つの観点から考察を行った．システムの学習条件を表 2 にまとめる．CNN-GRU に関する実験設定はすべて文献 [16] に準じている．比較手法としては言語ラベルを考慮したものを提案法、言語ラベルを考慮していないものを従来法としている．

4.2 実験結果

実験結果を表 3 に示す．結果は全て等価エラー率 (EER) で示している．最初に，言語ラベルなしの従来法である CNN-GRU の結果について述べる．学習データに ASV2019 を含まない場合 ((A),(B),(D))，どのテストデータに対しても EER が非常に高いことがわかる．これは，CNN-GRU が ASV2019 のために開発されたモデルあること，また，ASV2017 および VLD のデータ量が深層学習に基づくモデルの学習に不十分であることを示している．一方で ASV2019 のみ，ASV2019 と ASV2017 もしくは VLD をあわせて学習データとして用いてシステムを構築している ((C),(E),(F),(G)) に関してもテストデータが ASV2017, IRT の場合には EER が非常に高い．また，ASV2017 がテストデータの際の結果から，(B) の ASV2017 で学習し ASV2017 で評価した場合が一番 EER が低く，ASV2019 を追加しデータ量を補填した効果がないことがわかる．テストデータが IRT の場合にも全体的に EER が高くなっているが，学習データ量がある程度確保され、かつ日本語のデータを含んだ学習を行った (E) が一番低い EER となっていることが確認できた．これより，ASV2019 で学習したシステムはデータベースの依存性が高く，他のデータベースに対する頑健性が低いといえる．

次に，言語ラベルを付与したモデルの結果について比較する．テストデータが ASV2019 の場合には言語ラベルを付与した場合ほぼ全てのシステムで EER の改善が見られ，言語の依存性が緩和されていることがわかる．特に ASV2019 と ASV2017 を学習データとして用いた (E) では ASV2019 をテストデータとしたときに EER が 6.74% から 6.21%と改善しており，ASV2019 と ASV2017 で共通するラベルが存在することで性能の改善につながっていると確認できる．これらの結果からデータベース内全体で共通，またデータベースをまたいで共通な要素を陽に示すことで性能が改善したと考えられる．一方で，IRT をテストデータとして用いた場合の EER は依然として高く，言語以外にも性能が改善しない要因が含まれていると考えられる．ASV2019 と ASV2017 は同じ言語ではあるが収録環境の違いや，なりすまし攻撃の生成過程がシミュレーションによるか実収録かなど様々な違いが存在している．本実験で ASV2019 と ASV2017 に同じラベルを付与した結果同一条件である言語の要素を除去することができたと考えられる一方，全部のデータベースを合わせて学習データとして用いた (G) での EER が高くなることから，なりすまし検出として本質的に必要な要素以外のものがまだ多く残っていると考えられる．今後はそれらの要因についても除去することでより頑健ななりすまし検出法が実現可能だといえる．

5. おわりに

本研究では，CNN-GRU を基に，データベース，言語，収録環境への依存性を評価した．また，言語ラベルを追加して実験を行うことで言語に対する依存性がどの程度緩和できるかについても調査を行った．実験結果より，本来は言語に依存性のないと期待されるなりすまし検出においてモデル内に言語の依存性が残っていたこと，さらになりすまし検出に関係ない要因による依存性が残っていることが確認された．また，言語ラベルを用いることで言語依存性

を緩和できることが確認できた．

今後の課題として，他のモデルを用いた場合との性能の比較や，言語以外のラベルを付与したモデル学習，ラベル付与以外の学習方法の提案などが挙げられる．

謝辞 本研究は，JSPS 科研費若手研究 JP19K20271 と ROIS-DS-JOINT(023RP2020)，セコム財団挑戦的研究助成の助成を受けたものである．

参考文献

[1] Z. Wu, *et al.*,“Spoofing and countermeasures for speaker verification: A survey” in Proc. Speech Communication, pp.130–153, 2015.
[2] S. Shiota, *et al.*,“Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification” in Proc. Interspeech, pp.239–243, 2015.
[3] L. Zhang, *et al.*,“VoiceLive:A phoneme localization based liveness detection for voice authentication on smartphones” in Proc. the 2016 ACM SIGSAC Conference on Computer and Communications SecurityACM MobiCom, pp.1080–1091, 2016.
[4] S. Shiota, *et al.*,“Voice Liveness Detection for Speaker Verification based on a Tandem Single/Double-channel Pop Noise Detector’ in Odyssey, pp.259–263, 2016.
[5] R. Yaguchi, *et al.*,“Improving replay attack detection by combination of spatial and spectral features” in Proc. APSIPA Annual Summit and Conference, 2019.
[6] ASVspoof 2015, http://www.asvspoof.org/index2015.html
[7] ASVspoof 2017, http://www.asvspoof.org/index2017.html
[8] M.S. Saranya, *et al.*,“Decision-level feature switching as a paradigm for replay attack detection” in Proc. Interspeech, pp.686–690, 2018.
[9] G. Suthokumar, *et al.*,“Modulation Dynamic Features for the Detection of Replay Attacks” in Proc. Interspeech, pp.691–695, 2018.
[10] F. Tom, *et al.*,“End-To-End Audio Replay Attack Detection Using Deep Convolutional Networks with Attention” in Proc. Interspeech, pp.681–685, 2018.
[11] K. Sriskandaraja, *et al.*,“Deep Siamese Architecture Based Replay Detection for Secure Voice Biometric” in Proc. Interspeech, pp.671–675, 2018.
[12] ASVspoof 2019, http://www.asvspoof.org/index2019.html
[13] X. Cheng, *et al.*,“Replay detection using CQT-based modified group delay feature and ResNeWt network in ASVspoof 2019” in Proc. APSIPA Annual Summit and Conference, IEEE, pp.540–545, 2019.
[14] Y. Yang, *et al.*,“The SJTU Robust Anti-Spoofing System for the ASVspoof 2019 Challenge” in Proc. Interspeech, pp.1038–1042, 2019.
[15] M. Todisco,*et al.*,“ASVspoof 2019: Future horizons in spoofed and fake audio detection” in Interspeech 2019, 2019.
[16] J. weon Jung,*et al.*,“Replay attack detection with complementary high-resolution information using end-to-end DNN for the ASVspoof 2019 challenge” in Proc. Interspeech 2019, pp.1083–1087, 2019.
[17] A. Gomez-Alanis,*et al.*,“A Light Convolutional GRU-RNN Deep Feature Extractor for ASV Spoofing Detection” in Proc. Interspeech 2019, pp.1068–1072, 2019.