Vol.2021-SLP-135 No.4
2021/2/18

日仏共同CREST VoicePersonaeプロジェクトの紹介

山岸 順一[1]

概要：VoicePersonae プロジェクトは、国立情報学研究所、仏 Avignon 大学、仏 Eurecom 研究所を参加メンバーとする日仏共同 CREST プロジェクトである。JST がフランス 国立研究機構 (ANR) との協力枠組み合意に基づき、2018 年から実施している共同研究提案公募であり、本プロジェクトはその第一号として採択された。本プロジェクトでは、音声合成・声質変換・音声強調などの音声波形生成タスクに関する機械学習技術を高精度化すると同時に、話者認識システムの安全性と頑健性を高め、音声のプライバシー保護に関する新しい技術を実現することを目標としている。また同時に、音声変換と生体検知、音声の匿名化と再識別化と言った目的が相反する技術をどちらも加速させるチャレンジを積極的に開催している。更に、他の生体情報へ研究成果を適用することで、deepfake ビデオの自動検出、顔認証システムに対する StyleGAN によるウルフ顔生成攻撃なども提案するなど、分野にとらわれない幅広い研究も行っている。

Introduction of Japan-France joint CREST Voice Personae project

1. はじめに

VoicePersonae プロジェクトは、国立情報学研究所、仏 Avignon 大学、仏 Eurecom 研究所を参加メンバーとする CREST プロジェクト (2018-2023) である。JST がフランス 国立研究機構 (ANR) との協力枠組み合意に基づき、2018 年から実施している共同研究提案公募であり、本プロジェクトはその第一号として採択された。国立情報学研究所の音声合成や声質変換の知見、仏 Avignon 大学および Eurecom 研究所の話者認識とプライバシーに関する知見を活かしたプロジェクトを実施している。具体的には、音声合成・声質変換・音声強調などの音声波形生成タスクに関する機械学習技術を高精度化すると同時に、話者認識システムの安全性と頑健性を高め、音声のプライバシー保護に関する新しい技術を実現することを目標としている。また同時に、音声変換と生体検知、音声の匿名化と再識別化と言った目的が相反する技術をどちらも加速させるチャレンジを積極的に開催・運営している。更に、他の生体情報へ研究成果を適用し、deepfake ビデオの自動検出技術、顔認証システムに対する StyleGAN によるウルフ顔生成攻撃など分野を超えた幅広い研究も行っている。

本稿では、3 年度目までの成果を簡単に紹介する。各技術の詳細や実験結果については、各論文を参照されたい。

[1] 国立情報学研究所 コンテンツ科学研究系
National Institute of Informatics, Chiyoda, Tokyo 101–8430, Japan

2. 声の個人性に関するモデリング技術の高精度化と融合

2.1 Neural vocoder における成果

深層学習による音声合成や声質変換では、通常ボコーダを利用し、音声波形を生成する。本プロジェクトでは、ソースフィルター・ボコーダーのフィルター部をニューラルネットワーク化したモデル「ニューラルソースフィルター (NSF)」を提案し、ソースフィルターモデルの音質を大きく改善できる事を示した [1], [2]。さらにクリーキー音声などの発声にモデルを対応させる取り組みも行なった [3]。また、フィルター部ではなく、音源部をニューラルネットワーク化した「GAN excited linear prediction (GELP)」をフィンランド Aalto 大学と共同提案した [4]。また NSF が音声信号だけではなく、楽器音にも適用可能である事も示した [5]。この研究成果が一般に広く使われるためには、多くのユーザ・企業が手軽に利用可能なプログラムの公開も必要である。そこで、NSF を C++および Pytorch で実装したコードも無償で公開済みである。公開された NSF コードは研究用途に広く利用されていることは勿論、無料歌声合成ソフト「NEUTRINO」にも採用され、現在 Youtube 等に NSF 波形生成技術を利用した多くの楽曲が公開されている。なお本技術は、Google Brain に

より提案された「Differential DSP (DDSP)」[6] とも非常に関連が深い。

2.2 音声合成と話者認識の融合

話者認識技術を活用することで、End-to-end 音声合成の機能を向上できることも示している。例えば、[7] では話者認識モデルに利用される話者埋め込みベクトルを音声合成に導入する事により音声合成の話者性も適切に制御できる事を示している。これに加え、方言認識モデルの中間表現を更に導入することにより音声合成の方言も制御できることを示した [8]。これらは MIT との共同研究である。

2.3 音声合成と声質変換の融合

テキスト音声合成と声質変換は、それぞれ文字を音声に、入力話者の音声を目標話者の音声に変換する異なる音声生成タスクであるが、入力情報が文字であるが、変換元話者の音声であるかという点を除けば、共通点は多い。そこで、これらの異なる複数のタスクを理論的に融合することを目標に、音声合成と声質変換とを同時に実行可能な新たなニューラルネットワーク構造の提案を幾つか行った。音声合成ではテキストを入力し、声質変換では他人の音声を入力し、目標話者の音声を生成する。言い換えると、入力情報は異なるが、音声生成部は同一である。そこで、これらの異なる入力を柔軟に受け付け、どちらの入力からも高品質な音声を生成可能である新たなモデルを新たにデザインし発表を行なった [9], [10]。これにより、これまで個別に研究されてきたテキスト音声合成と声質変換を統一的な枠組みで考えることが可能になり、データベース等の学習データも相互利用することが可能になる。単なる手法の改善とは異なり、音声生成タスクの方法論自身を発展させることができた良い成果であると考える。これらはシンガポール国立大学 (NUS) との共同研究である。

また、異なる音声生成タスク間でモデルやデータベースを共有するだけでなく、異なる音声生成タスク間で転移学習も可能になった。そこで、テキスト音声合成から声質変換へニューラルネットワークを転移学習させる新たな枠組みも提案し [11]、従来の声質変換技術を大きく改善できる事を示した。この発表の後提案された Transfomer 音声合成システムを声質変換システムに転移学習させる手法とも関連が深い [12]。

2.4 音声強調・音声明瞭性強調

音声波形生成タスクは他にも、雑音を含む音声からクリーン音声のみを取り出す音声強調 (Far-end speech enhancement) や、音声を雑音に負けないように変換する音声明瞭性強調 (Near-end listening enhancement, speech intelligibility enhancement) にも密接に関係する。

そこで、これらのタスクにも取り組み、音声合成において発話様式を制御する Style Token モデルを音声強調に応用し、ノイズタイプを潜在変数として学習する「Noise token モデル」を提案した [13]。更に、音声合成の Stlye Transfer と呼ばれる、参照音声を元に合成音声の発話様式を変換する技術と Noise Token モデルとを組み合わせ、音声強調に応用することで、雑音や反響下の音声をクリーン音声へ変換する手法も提案した [14]。

また音声明瞭性タスクにおいて、SIIB[15] や ESTOI[16] といった音声明瞭性の指標を敵対的生成ネットワーク (Generative Adversarial Network) の Discriminator の出力値として獲得し、不明瞭な音声を聴きやすい音声に自動変換に利用する iMetricGAN [13] を提案し、Hurricane Challenge 2.0 [17] で優秀な成績を収めた。台湾のアカデミア・シニカとの共同研究成果である。

2.5 エンタメ応用

さらに音声合成技術、特に声の個人性に着目したモデリング技術の挑戦的応用例およびエンタメ応用例として、日本の伝統芸能である落語に着目し、真打噺家の話芸を音声合成により学習・再現するという検討も行なった [18]。情報伝達や質問回答を主たる目的とする従来の音声合成技術とは目的が全く異なり、聞き手を楽しませる、いわば「AI 噺家」の実現を目標としている。真打・2つ目・前座という異なる階級の落語家との比較を通し、通常の自然性改善だけでなく、役の区別や、話の流れや構成が理解できる様にモデリングを改善する必要があることが判明している [19]。

2.6 Voice Conversion Challenge 2020

前述の通り、本プロジェクトでは、チャレンジの開催・運営を重要視している。そこで、戸田 CREST および海外の数大学（シンガポール国立大学、中国科学技術大学、東フィンランド大）と協力し、音声のアイデンティティを自動変換する声質変換技術を相互比較する Voice Conversion Challenge 2020 を開催した [20]。2016 年から隔年で実施 [21], [22] しているチャレンジの 3 回目である。今回のチャレンジでは、言語内・異言語間での声質変換に着目し、新たなデータベースを構築・無償公開、そのデータベースを利用し海外 33 組織が構築した手法を統一評価した。チャレンジの結果からは、上位 7 システムによる変換音声は目標音声と話者類似性の観点で有意差が無いという興味深い結果を確認できた。また本チャレンジの国際ワークショップも 2021 年 9 月にオンライン開催した。

3. 生体認証の安全性と頑健性向上

3.1 ASVspoof challenge 2019

前述した音声の個人性再現技術は、エンターテインメント等にて新たな価値をもたらすと考えられるものの、悪用された場合には話者認識システム等においてセキュリ

ティー上の問題も発生する。話者認識の安全性と頑健性の向上のため、本プロジェクトの仏パートナーおよび東フィンランド大と協力し、話者照合に対するなりすまし攻撃を自動的に防御するライブネス検出を共通コーパスで比較する ASVspoof challenge 2019 を世界規模で開催した [23]。

音声合成や声質変換によるなりすまし攻撃を想定した「Logical access タスク」と、音声の単純な再生によるリプレイ攻撃を想定した」Physical access タスク」の 2 種類を想定し、それぞれ大規模データベースを構築した [24]。Logical access タスクにおいては、19 種類の異なるアルゴリズムによる合成音声や変換音声を、Google, iFlytek, NTT 等の複数企業の協力のもと用意した。Physical access タスクにおいては、様々な条件によるリプレイをシミュレーションにより大量に生成した。

大規模データベースはチャレンジ参加希望の 154 組織に配布され、そのうち 50 組織が実際にライブネス検知モデルを構築し、未知の攻撃手法が大量に含まれる評価データの真贋判定を行った。データベースは 2019 年秋に無償公開され、更に多くの機関に利用されている。チャレンジ参加者のライブネス検知モデルの精度評価およびランキングは、等価誤り率 (EER)、および、後段の個人認証との統合スコア (t-DCF) [25] により行った。様々な分析の結果、人間には聴覚上差がわからない様な詐称音声でも適切に識別可能である事が確認されている。

また、音声の生体検知に関する学術発表の場を研究コミュニティーに提供するため、国際会議 Interspeech 2019 および ASRU 2019 の両方において、スペシャルセッションを開催した。それに加え、国際ジャーナル誌 Computer Speech Language における特集号も企画した。その他、データベース構築に協力した Google およびチャレンジ参加企業 ID R&D Inc のプレスリリースがあったことから、米国における新聞報道も多数あり、社会的に大きな反響も得た。日仏プロジェクト全組織による取り組みである。

3.2 話者認識と生体検知システムの同時学習

話者認識システムと音声の生体検知システムとを同時に学習する枠組みについても新たに検討した。生体認証および生体検知の指標 (EER や DCF) は通常微分不可能である事から、深層モデルの学習には直接利用されず、単純な end-to-end 学習は不可能である。そこで、教師あり学習ではなく、強化学習を新たに導入することで、話者認識システムと生体検知システムの両方を同時に高精度化し、安全性と頑健性を高める研究を行なった [26]。東フィンランド大との共同研究成果である。

4. 音声のプライバシー保護

4.1 X ベクトルを用いた話者匿名化

本プロジェクトでは、音声の匿名化法の研究、とりわけ、音声に含まれる話者性の匿名化にも注力している。本テーマは非常に新しい研究トピックであり、現在の音声分野を幅広く見渡しても、どの様に話者匿名化を実現できるのか未だ明確に定義されていないのが現状である。しかし、Youtube 等の SNS 上の音声データ等から音声合成システムを作ることも現実可能になりつつあることから、喫緊の対策が必要なことは言うまでもない。

そこで、音声の自然性や音声から知覚可能な年代や性別といった話者の属性情報を保ったまま、音声の個人性を変えることを目的とする話者匿名化を提案した [27]。これは、音声を抑揚、音素情報、xベクトルという話者性を表すベクトルの３つの情報に分解し、xベクトルのみを近傍の K 人の話者と平均化することで匿名化する手法である。音声波形を再合成するモジュールにはニューラルソースフィルタを利用し、高品質な音声生成を可能にした。英語話者の音声データを利用した実験から、x ベクトルの空間において、k 匿名化を行うことで、話者認識システムおよび人間の聴覚上の話者識別性能が有意に下がることを確認した。さらに、単なる平均値による K 匿名化ではなく、話者空間における確率密度の混合分布を考慮した改良版も提案した [28]。

4.2 Voice Privacy Challenge

音声のプライバシー保護に関しても、分野を牽引し、研究を加速させるため、国際的なチャレンジを運営した。仏 Avignon 大が中心となり実施した”Voice Privacy Challenge”であり、仏 Eurecom 研究所、仏 Inria 研究所、および NII の協力の元、話者匿名化手法を相互に比較できる様、利用する音声データベース、評価セット、評価手順を規定した [29]。10 数の大学・企業・研究組織が提案した話者匿名化手法を相互評価する事を行い、国際会議 Interspeech 2020 および Speaker Odyssey 2020 におけるスペシャルセッションを開催した。また国際ジャーナル誌 Computer Speech & Language において音声プライバシーの特集号も企画した。

4.3 匿名化指標

この様な話者匿名化技術の適切な評価には、単なる変換音声の品質や話者認識精度による比較だけでなく、より適切な指標に基づく評価・分析が必要である。そこで話者匿名化後の音声が再識別化される最悪リスクに基づいて評価を行う指標 [30] や複数話者の匿名化後の音声がどれだけ相互に類似しているかを考慮した指標 [31] も提案した。

5. 他の生体情報への適用

5.1 映像への応用

ライブネス検出の知見を他の生体情報への適用する研究も開始した。具体的には、現在欧米を中心に社会問題とも

なっている deepfake や face2face という技術により自動生成され、見た目は非常に自然だが偽の顔映像を自動検出するモデルを構築した。具体的には、カプセルネットワークというニューラルネットワーク技術を利用することで、高精度に識別できる可能性が高いことを実験から示した [32]。本ネットワークは、下位カプセルネットワークと上位カプセルネットワークで構成される。下位カプセルネットワークは、顔のある部位に着目しそこに、真の画像にはないアーティファクトが存在するかどうかを判断する。上位カプセルネットワークは、下位カプセルネットワークの出力に基づき、最終的に、映像もしくは画像がフェイクかどうかを総合的に判定する。なお、下位カプセルネットワークが着目する特徴や部位はデータにより決まる。このネットワークの利用により、deepfake や face2face のどちらの偽の顔映像に対しても通常の CNN より少ないパラメータで頑健に検出できることを確認した [33]。

さらに、本技術を発展させ、単に映像に対して、真贋判定を行うだけでなく、改ざんされたピクセル領域を特定する事を同時に予測する新たなネットワークも提案した [34]。これは、真贋判定を行うタスクと改ざんされたピクセル領域を特定するタスクのマルチタスク学習に基づいている。ピクセル領域の特定には、画像のセグメンテーション技術を利用した。改竄されている領域を示すことにより、deepfake であるとの根拠を示せることにつながり、説明可能性が向上する。また、実験結果からは、改ざんされたピクセル領域の特定は、未知のフェイクビデオ生成手法に対しても有効に働き、様々な観点でメリットがある事を示した。

5.2 顔認識システムへの応用

更に、顔認証システムの深層生成モデルに対する脆弱性に関する研究も行なった。具体的には、複数の登録ユーザに特徴が一致するマスター顔（顔認証に対するマスター鍵）を、一般に公開されている深層生成モデル (StyleGAN [35]) および潜在変数のヒルクライミング手法による自動更新により、生成可能である事を示した [36]。スイス IDIAP 研究所との共同研究成果である。

5.3 自然言語生成への応用

自然言語生成への応用も行っている。具体的には、巨大なニューラル言語モデルにより非常に自然で流暢なクチコミを自動生成し、人間に識別可能であるかという点と、複数の識別モデルを組み合わせる事で、生成された口コミと人間が書いたクチコミと自動識別出来るかという点について調査を行っている [37]。この論文は米 Open AI が GPT ２というニューラル言語モデルを一般公開する際に、社会的に安全であり問題がないという根拠の一つになった [38]。さらに、単なる流暢な文章生成だけでなく、左派・右派向けの英語新聞記事風の文章も自動生成可能である事を示している [39]。

6. 参加メンバー

これらの成果はプロジェクトメンバーの努力により実現している。現在のプロジェクトメンバーは以下の通りである。

- 山岸順一 (NII)
- 越前 功 (NII)
- Wang Xin (NII)
- Cooper Erica (NII)
- Kruengkrai Canasai (NII)
- Zhao Yi (NII)
- Luong Hieu-Thi (NII)
- Le Trung-Nghia (NII)
- 加藤 集平 (株式会社 RevComm)
- 安田 裕介 (総研大)
- Nguyen Hong Huy (総研大)
- Tieu Dung (総研大)
- Li Haoyu (総研大)
- Zeng Chang (総研大)
- Zhang Lin (総研大)
- Ji Yi (総研大)
- Williams Jennifer (University of Edinburgh)
- Jean-François Bonastre (Université d'Avignon)
- Driss Matrouf (Université d'Avignon)
- Natalia Tomashenko (Université d'Avignon)
- Paul-Gauthier Noé (Université d'Avignon)
- Nicolas Evans (Eurecom)
- Massimiliano Todisco (Eurecom)
- Andreas Nautsch (Eurecom)
- Jose Patino (Eurecom)

過去に本プロジェクトに参加したメンバーは以下の通りである。

- Fang Fuming (Alibaba)
- 高木 信二 (名工大)

7. おわりに

本稿では VoicePersonae プロジェクトのこれまでの成果を紹介した。本プロジェクトは 2023 年まで実施される予定である。

謝辞 本研究は，JST CREST (JPMJCR18A6, VoicePersonae project) の支援を受けたものである．

参考文献

[1] Wang, X., Takaki, S. and Yamagishi, J.: Neural Source-filter-based Waveform Model for Statistical Parametric Speech Synthesis, *ICASSP 2019 - 2019 IEEE In-*

ternational Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5916–5920 (online), DOI: 10.1109/ICASSP.2019.8682298 (2019).

[2] Wang, X., Takaki, S. and Yamagishi, J.: Neural Source-Filter Waveform Models for Statistical Parametric Speech Synthesis, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 402–415 (online), DOI: 10.1109/TASLP.2019.2956145 (2020).

[3] Wang, X. and Yamagishi, J.: Using Cyclic Noise as the Source Signal for Neural Source-Filter-Based Speech Waveform Model, *Proc. Interspeech 2020*, pp. 1992–1996 (online), DOI: 10.21437/Interspeech.2020-1018 (2020).

[4] Juvela, L., Bollepalli, B., Yamagishi, J. and Alku, P.: GELP: GAN-Excited Linear Prediction for Speech Synthesis from Mel-Spectrogram, *Proc. Interspeech 2019*, pp. 694–698 (online), DOI: 10.21437/Interspeech.2019-2008 (2019).

[5] Zhao, Y., Wang, X., Juvela, L. and Yamagishi, J.: Transferring Neural Speech Waveform Synthesizers to Musical Instrument Sounds Generation, *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6269–6273 (online), DOI: 10.1109/ICASSP40776.2020.9053047 (2020).

[6] Engel, J., Hantrakul, L. H., Gu, C. and Roberts, A.: DDSP: Differentiable Digital Signal Processing, *International Conference on Learning Representations*, (online), available from ⟨https://openreview.net/forum?id=B1x1ma4tDr⟩ (2020).

[7] Cooper, E., Lai, C., Yasuda, Y., Fang, F., Wang, X., Chen, N. and Yamagishi, J.: Zero-Shot Multi-Speaker Text-To-Speech with State-Of-The-Art Neural Speaker Embeddings, *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6184–6188 (online), DOI: 10.1109/ICASSP40776.2020.9054535 (2020).

[8] Cooper, E., Lai, C.-I., Yasuda, Y. and Yamagishi, J.: Can Speaker Augmentation Improve Multi-Speaker End-to-End TTS?, *Proc. Interspeech 2020*, pp. 3979–3983 (online), DOI: 10.21437/Interspeech.2020-1229 (2020).

[9] Zhang, M., Wang, X., Fang, F., Li, H. and Yamagishi, J.: Joint Training Framework for Text-to-Speech and Voice Conversion Using Multi-Source Tacotron and WaveNet, *Proc. Interspeech 2019*, pp. 1298–1302 (online), DOI: 10.21437/Interspeech.2019-1357 (2019).

[10] Luong, H. T. and Yamagishi, J.: NAUTILUS: A Versatile Voice Cloning System, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 2967–2981 (online), DOI: 10.1109/TASLP.2020.3034994 (2020).

[11] Luong, H. and Yamagishi, J.: Bootstrapping Non-Parallel Voice Conversion from Speaker-Adaptive Text-to-Speech, *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 200–207 (online), DOI: 10.1109/ASRU46091.2019.9004008 (2019).

[12] Huang, W.-C., Hayashi, T., Wu, Y.-C., Kameoka, H. and Toda, T.: Voice Transformer Network: Sequence-to-Sequence Voice Conversion Using Transformer with Text-to-Speech Pretraining, *Proc. Interspeech 2020*, pp. 4676–4680 (online), DOI: 10.21437/Interspeech.2020-1066 (2020).

[13] Li, H. and Yamagishi, J.: Noise Tokens: Learning Neural Noise Templates for Environment-Aware Speech Enhancement, *Proc. Interspeech 2020*, pp. 2452–2456 (online), DOI: 10.21437/Interspeech.2020-1030 (2020).

[14] Li, H., Ai, Y. and Yamagishi, J.: Enhancing Low-Quality Voice Recordings Using Disentangled Channel Factor and Neural Waveform Model, *Proc. SLT 2021*, pp. 734–741 (2021).

[15] Van Kuyk, S., Kleijn, W. B. and Hendriks, R. C.: An instrumental intelligibility metric based on information theory, *IEEE Signal Processing Letters*, Vol. 25, No. 1, pp. 115–119 (2017).

[16] Jensen, J. and Taal, C. H.: An algorithm for predicting the intelligibility of speech masked by modulated noise maskers, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 24, No. 11, pp. 2009–2022 (2016).

[17] Rennies, J., Schepker, H., Valentini-Botinhao, C. and Cooke, M.: Intelligibility-Enhancing Speech Modifications — The Hurricane Challenge 2.0, *Proc. Interspeech 2020*, pp. 1341–1345（オンライン）, DOI: 10.21437/Interspeech.2020-1641 (2020).

[18] Kato, S., Yasuda, Y., Wang, X., Cooper, E., Takaki, S. and Yamagishi, J.: Modeling of Rakugo Speech and Its Limitations: Toward Speech Synthesis That Entertains Audiences, *IEEE Access*, Vol. 8, pp. 138149–138161 (online), DOI: 10.1109/ACCESS.2020.3011975 (2020).

[19] Kato, S., Yasuda, Y., Wang, X., Cooper, E. and Yamagishi, J.: How Similar or Different Is Rakugo Speech Synthesizer to Professional Performers? (2020).

[20] Yi, Z., Huang, W.-C., Tian, X., Yamagishi, J., Das, R. K., Kinnunen, T., Ling, Z.-H. and Toda, T.: Voice Conversion Challenge 2020 — Intra-lingual semi-parallel and cross-lingual voice conversion —, *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, pp. 80–98 (2020).

[21] Toda, T., Chen, L.-H., Saito, D., Villavicencio, F., Wester, M., Wu, Z. and Yamagishi, J.: The Voice Conversion Challenge 2016, *Interspeech 2016*, pp. 1632–1636 (online), DOI: 10.21437/Interspeech.2016-1066 (2016).

[22] Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T. and Ling, Z.: The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods, *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, pp. 195–202 (online), DOI: 10.21437/Odyssey.2018-28 (2018).

[23] Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., Yamagishi, J., Evans, N., Kinnunen, T. H. and Lee, K. A.: ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection, *Proc. Interspeech 2019*, pp. 1008–1012 (online), DOI: 10.21437/Interspeech.2019-2249 (2019).

[24] Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., Sahidullah, M., Vestman, V., Kinnunen, T., Lee, K. A., Juvela, L., Alku, P., Peng, Y.-H., Hwang, H.-T., Tsao, Y., Wang, H.-M., Maguer, S. L., Becker, M., Henderson, F., Clark, R., Zhang, Y., Wang, Q., Jia, Y., Onuma, K., Mushika, K., Kaneda, T., Jiang, Y., Liu, L.-J., Wu, Y.-C., Huang, W.-C., Toda, T., Tanaka, K., Kameoka, H., Steiner, I., Matrouf, D., Bonastre, J.-F., Govender, A., Ronanki, S., Zhang, J.-X. and Ling, Z.-H.: ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech, *Computer Speech & Language*, Vol. 64, p. 101114 (online), DOI: https://doi.org/10.1016/j.csl.2020.101114 (2020).

[25] Kinnunen, T., Delgado, H., Evans, N., Lee, K. A., Vestman, V., Nautsch, A., Todisco, M., Wang, X., Sahidullah, M., Yamagishi, J. and Reynolds,

D. A.: Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification: Fundamentals, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 2195–2210 (online), DOI: 10.1109/TASLP.2020.3009494 (2020).

[26] Kanervisto, A., Hautamäki, V., Kinnunen, T. and Yamagishi, J.: An Initial Investigation on Optimizing Tandem Speaker Verification and Countermeasure Systems Using Reinforcement Learning, *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, pp. 151–158 (online), DOI: 10.21437/Odyssey.2020-22 (2020).

[27] Fang, F., Wang, X., Yamagishi, J., Echizen, I., Todisco, M., Evans, N. and Bonastre, J.-F.: Speaker Anonymization Using X-vector and Neural Waveform Models, *Proc. 10th ISCA Speech Synthesis Workshop*, pp. 155–160 (online), DOI: 10.21437/SSW.2019-28 (2019).

[28] Srivastava, B. M. L., Tomashenko, N., Wang, X., Vincent, E., Yamagishi, J., Maouche, M., Bellet, A. and Tommasi, M.: Design Choices for X-Vector Based Speaker Anonymization, *Proc. Interspeech 2020*, pp. 1713–1717 (online), DOI: 10.21437/Interspeech.2020-2692 (2020).

[29] Tomashenko, N., Srivastava, B. M. L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N., Patino, J., Bonastre, J.-F., Noé, P.-G. and Todisco, M.: Introducing the VoicePrivacy Initiative, *Proc. Interspeech 2020*, pp. 1693–1697 (online), DOI: 10.21437/Interspeech.2020-1333 (2020).

[30] Nautsch, A., Patino, J., Tomashenko, N., Yamagishi, J., Noé, P.-G., Bonastre, J.-F., Todisco, M. and Evans, N.: The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment, *Proc. Interspeech 2020*, pp. 1698–1702 (online), DOI: 10.21437/Interspeech.2020-1815 (2020).

[31] Noe, P.-G., Bonastre, J.-F., Matrouf, D., Tomashenko, N., Nautsch, A. and Evans, N.: Speech Pseudonymisation Assessment Using Voice Similarity Matrices, *Proc. Interspeech 2020*, pp. 1718–1722 (online), DOI: 10.21437/Interspeech.2020-2720 (2020).

[32] Nguyen, H. H., Yamagishi, J. and Echizen, I.: Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos, *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2307–2311 (online), DOI: 10.1109/ICASSP.2019.8682602 (2019).

[33] Nguyen, H. H., Yamagishi, J. and Echizen, I.: Use of a Capsule Network to Detect Fake Images and Videos (2019).

[34] Nguyen, H. H., Fang, F., Yamagishi, J. and Echizen, I.: Multi-task learning for detecting and segmenting manipulated facial images and videos, *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, IEEE, pp. 1–8 (2019).

[35] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J. and Aila, T.: Analyzing and improving the image quality of stylegan, *arXiv preprint arXiv:1912.04958* (2019).

[36] Nguyen, H. H., Yamagishi, J., Echizen, I. and Marcel, S.: Generating Master Faces for Use in PerformingWolf Attacks on Face Recognition Systems, *International Join Conference on Biometrics* (2020).

[37] Ifeoluwa Adelani, D., Mai, H., Fang, F., Nguyen, H. H., Yamagishi, J. and Echizen, I.: Generating Sentiment-Preserving Fake Online Reviews Using Neural Language Models and Their Human- and Machine-based Detection, *The 34-th International Conference on Advanced Information Networking and Applications (AINA-2020)* (2020).

[38] Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K. and Wang, J.: Release Strategies and the Social Impacts of Language Models (2019).

[39] Gupta, S., Nguyen, H. H., Yamagishi, J. and Echizen, I.: Viable Threat on News Reading: Generating Biased News Using Natural Language Models, *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, Online, Association for Computational Linguistics, pp. 55–65 (online), DOI: 10.18653/v1/2020.nlpcss-1.7 (2020).