社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

ブラックボックス型敵対的攻撃に対する話者照合システムの脆弱性に関する調査

甲斐　優人†　塩田さやか††　貴家　仁志††

† 首都大学東京 システムデザイン学部
E-mail: †kai-hiroto@ed.tmu.ac.jp，††{sayaka, hitoshi}@tmu.ac.jp

あらまし　近年，機械学習に基づくシステムへの敵対的攻撃に対する脆弱性が危惧されている．敵対的攻撃とは，あるデータに人間には知覚ができない程度の微妙な摂動を敵対的構造になっているニューラルネットワークを用いることで推定し，元の音声に付与させることでシステムの認識結果を故意に変えるものである．ネットバンキングやネットショッピングなどの利用時に，本人確認の主な手段として使用されている生体認証システムにおいても，敵対的攻撃に対する脆弱性が指摘されている．生体認証の 1 つである話者照合においても同様であるが実験的にはその頑健性が示されていなかった．そこで本研究では，攻撃対象となるシステムの内部を想定しないブラックボックス型敵対的攻撃に着目し，照合精度の変化を調査した．実験では，生成された敵対的サンプルとランダムなホワイトノイズを重畳した音声を複数の話者照合システムに入力して評価し，敵対的攻撃に対する話者照合の脆弱性について分析した．それぞれの照合精度の変化の結果から，ブラックボックス型の敵対的攻撃であっても敵対的攻撃特有の摂動を付与した音声の方が，ランダムに生成されたホワイトノイズを重畳した音声よりも照合性能を大幅に低下させることがわかった．
キーワード　敵対的攻撃，話者照合，ブラックボックス型攻撃

Vulnerability investigation of speaker verification against black-box adversarial attacks

Hiroto KAI†, Sayaka SHIOTA††, and Hitoshi KIYA††

† Faculty of System Design, Tokyo Metropolitan University
E-mail: †kai-hiroto@ed.tmu.ac.jp，††{sayaka, hitoshi}@tmu.ac.jp

Abstract　Recently，vulnerability against adversarial attacks is being feared for machine learning-based systems. Adversarial attacks are aimed to intentionally change results of systems based on machine learning algorithms. Data which a small noise is added to an original speech are called perturbation. The perturbation is estimated by using adversarial structured neural networks, and the generated data are almost indistinguishable by a human when compared to the original data. Biometric authentication systems which are used as a primarily means for user authentication in applications like internet banking and online shopping are also pointed out in terms of vulnerability against such adversarial attacks. The same can be said for speaker verification (SV), one of the biometric authentication, however its robustness has not been indicated experimentally. Therefore, we investigated the change in accuracy when attacking a system where the internal design is unexpected using adversarial attacks. In our experiments, we have evaluated the results when inputting generated adversarial examples and audios mixed with white noise to SV systems, and analyzed the vulnerability of speaker verification against adversarial attacks. From the results, the performances of SV systems became worse when the generated adversarial examples were inputted, compared with audio with white noise were inputted.
Key words　adversarial attack，speaker verification，black-box attack

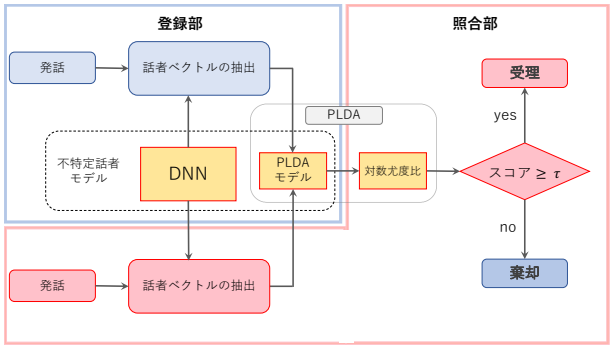


図 1　x-vector に基づく話者照合システムのフロー図

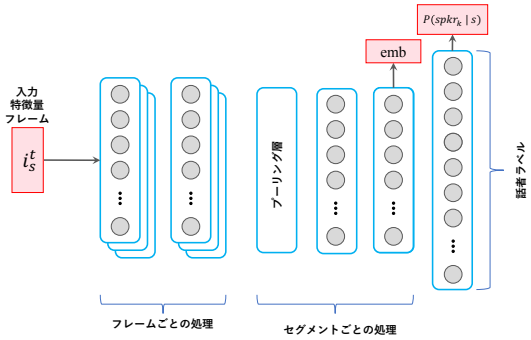


図 2　x-vector の抽出に用いる DNN の構造

1. まえがき

話者照合とは，システムに入力された音声があらかじめ登録されている話者のものであるか否かの判定を行う生体認証技術である．インターネットの普及により，オンライン上，電話越しでの話者照合による本人確認の需要が考えられるため，ネットバンキングやネットショッピングなどの e コマースへの導入が期待されている．話者照合は i-vector に基づく手法 [1] や，深層学習 (Deep Neural Network; DNN) に基づく手法 [2-4]，また，probabilistic linear discriminant analysis (PLDA) に基づく手法 [5] などによりその認証精度が非常に向上してきている．特に最新のシステムとして x-vector に基づく話者照合に関する研究が活発に行われている [6-8]．

DNN に基づく手法が画像認識や音声認識などに大きく貢献することが報告されている．一方で，ニューラルネットワークを用いた識別システムへの攻撃についても研究がされてきた [9]．その 1 つに敵対的攻撃と呼ばれるものがある．これは人間には知覚できない程度の微小な摂動を敵対的構造になっているニューラルネットワークを用いることで推定し，元の音声に付与することで故意にシステムの結果を歪めることができるものである．近年，様々なシステムへの敵対的攻撃に対する脆弱性が懸念されており，話者照合システムにおいてもその脆弱性が危惧されている．先行研究では，話者照合において，話者モデルの学習と敵対的サンプルの生成に使うデータベースや特徴量が一致する時と，異なる時の照合精度の変化を調査しているものがある [10]．しかし，より現実的な攻撃手法であるブラックボックス型に対する脆弱性は実験的に調査されていない．そこで本研究では，攻撃対象となるシステムの内部を想定しないブラックボックス型敵対的攻撃に着目し，x-vector に基づく話者照合における照合精度の変化を調査した．

2. 話者照合

話者照合とはユーザーの入力音声を用いて，入力音声が本人であるか否かを判定するシステムである．一般的に話者照合システムは登録部と照合部の二つに分けられており，登録部において照合したい話者の音声の声を用いて登録話者モデルを作成する．照合部では入力された音声の特徴量と登録部で作成された特定話者モデルとのスコアを計算し，閾値以上であれば受理，未満であれば棄却するという流れになっている．

2.1 x-vector に基づく話者照合

近年，話者照合における state-of-the-art な手法の一つとして x-vector に基づく手法 [8] が広く用いられている．これは，可変長の発話から固定次元の話者ベクトルにマッピングする DNN を構築し，埋め込み層を用いて話者表現を抽出するものである．図 1 に x-vector に基づく話者照合システムのフロー図を示す．図 1 より，登録部においてまず不特定話者モデルと呼ばれる DNN の構築を行う．これは，x-vector と呼ばれる話者表現を抽出するためのものである．この DNN の構造を図 2 に示す．i_s^t は入力音声の特徴ベクトルであり，フレーム数 $t = (1, ..., T)$ の発話 s から抽出されたものである．プーリング層よりも前の層ではフレームごとに入力された特徴量が処理される．プーリング層では，前の層で計算された出力の平均と分散を計算する．埋め込み層 (emb) で話者表現をマッピングし，x-vector を抽出する．各層へ入力されるフレーム数や入力サイズを表 1 に示す．話者を表す x-vector は図 2 の埋め込み層で表現される．そのため登録話者の音声が入力されると，その音声を用いた x-vector が抽出され登録される．照合時にも入力されたテスト音声から x-vector が抽出される．照合部では，この登録されていた x-vector とテスト部の x-vector を比較することで本人か否かを判定する．次節では識別手法として広く用いられている PLDA について述べる．

2.2 PLDA

PLDA は抽出された話者ベクトルから話者性に寄与しない情報を低減する手法でありチャネル変動等を軽減することが知られている [5]．また，i-vector や x-vector に基づく手法の back-end としても有効であることが報告されている．2.1 節で紹介した x-vector に基づく手法においても PLDA のモデルは不特定話者データから次のように求められる．まず発話 u から抽出された x-vector $\boldsymbol{\omega_u}$ をその生成過程を無視して式 (1) のように生成されたと考える．

$$\boldsymbol{\omega_u} = \bar{\boldsymbol{\omega}} + \boldsymbol{\Phi\delta} + \boldsymbol{\Gamma\zeta_u} + \boldsymbol{\epsilon_u} \tag{1}$$

ここで，$\boldsymbol{\Phi}$ と $\boldsymbol{\Gamma}$ は話者とチャネルの部分空間を張る基底行列であり，$\boldsymbol{\delta}$ と $\boldsymbol{\zeta_u}$ は話者及びチャネル因子を表しており，それぞれ標準正規分布に従う．$\boldsymbol{\epsilon_u}$ は残差成分を表し，平均ベクトル $\mathbf{0} \in \boldsymbol{R}^{CD_F}$，対角共分散行列 $\boldsymbol{\Sigma} \in \boldsymbol{R}^{CD_F \times CD_F}$ のガウス

分布に従う．$\bar{\boldsymbol{\omega}}$ は x-vector 空間におけるオフセットである．式 (1) から確率生成モデルを考える．

$$p(\boldsymbol{\omega_u}|\boldsymbol{\delta}, \boldsymbol{\zeta_u}) = N(\bar{\boldsymbol{\omega}} + \boldsymbol{\Phi\delta} + \boldsymbol{\Gamma\zeta_u}, \boldsymbol{\Sigma}) \tag{2}$$

式 (2) より登録話者の x-vector $\boldsymbol{\omega_1}$ と照合話者の x-vector $\boldsymbol{\omega_2}$ を用いて $\boldsymbol{\omega_1}$，$\boldsymbol{\omega_2}$ が同一話者モデルから生成されたか ($\boldsymbol{H_1}$) 否か ($\boldsymbol{H_0}$) に関する仮説に対して次式の対数尤度比

$$\log \frac{p(\boldsymbol{\omega_1}, \boldsymbol{\omega_2}|\boldsymbol{H_1})}{p(\boldsymbol{\omega_1}|\boldsymbol{H_0})p(\boldsymbol{\omega_2}|\boldsymbol{H_0})} \tag{3}$$

を計算し，照合時のスコアとして用いて評価する．

3. 敵対的攻撃

話者照合システムを攻撃する際に使う敵対的サンプルを生成する必要がある．敵対的サンプルに含まれる摂動の生成にはいくつか手法があり，本研究で使用した手法を以下に述べる．

3.1 敵対的サンプルの生成手法

入力 $\boldsymbol{x}$ が与えられた時，$\boldsymbol{x}$ の敵対的サンプル $\hat{\boldsymbol{x}}$ を次のように表せる．

$$\hat{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\delta} \tag{4}$$

ここで $\boldsymbol{\delta}$ は 原音声に付与される摂動を表す．また，$\boldsymbol{\delta}$ の特徴は $\boldsymbol{\delta} \in \mathbb{R}^{\boldsymbol{d}}$ で人間が聞いても知覚できないが，$\boldsymbol{\delta}$ を付与した音声はネットワークの予測結果を変えることができる．

文献 [11] で，音声認識結果を歪ませる敵対的攻撃について報告されている．この攻撃では，式 (5) を最適化することで，$\boldsymbol{\delta}$ を求めている．

$$\begin{aligned} &\text{minimize} \quad |\boldsymbol{\delta}|_2^2 + cL(\boldsymbol{x} + \boldsymbol{\delta}, t) \\ &\quad \text{such} \quad \text{that} \quad dB_x(\boldsymbol{\delta}) \leq \tau \end{aligned} \tag{5}$$

ここで，c は敵対損失の重みを表す定数，L はロス関数，t はターゲットラベル，$dB_x(\boldsymbol{\delta})$ は $\boldsymbol{\delta}$ のデシベル値，τ は十分大きい閾値を表す．式 (5) の最適化の計算過程で最適な解 $\boldsymbol{\delta^*}$ を得られた時，τ を小さくするとともに式 (5) の最適化を再開する．同時に，定式の解決にあたって，音声，複数の特徴量に基づく識別器，ネットワーク，そして最終的なロスのすべてを微分することで敵対的サンプルを生成している．

3.2 話者照合に対するブラックボックス型敵対的攻撃

ホワイトボックス型敵対的攻撃とは攻撃者が対象となるシステムの内部の設計や使われている深層学習用のネットワーク構造やパラメータを知り得ることができる状態で敵対的攻撃を行う場合を指す．一方，ブラックボックス型敵対的攻撃とは攻撃者は対象となるシステムの入力および出力結果のみ知り得ることができる状態で敵対的攻撃を行う場合を指す．ブラックボックス型敵対的攻撃の方が現実的ではあるが敵対的攻撃に関する研究ではホワイトボックス型敵対的攻撃を想定した研究が多い．本研究では，報告の少ないブラックボックス型の敵対的攻撃を前提として話者照合システムへの影響について考える．

2 章で述べた x-vector に基づく話者照合システムにおいて，入力は音声信号，出力が照合スコアとなる．ブラックボックス型の敵対的攻撃を行うには他の深層学習に基づくシステムを用意し，そのシステムを騙すような勾配を計算することで微小なノイズを付与した入力音声を生成し，その音声を話者照合システムへ入力する．敵対的攻撃は対象となるシステムだけでなく他のシステムにも影響がある攻撃であると報告されているため [12]，実際の検証が必要だと考えられる．

表 1　DNN の各層の構築

層	層のコンテキスト	全コンテキスト	入力 x 出力
フレーム 1	{t - 2, t + 2}	5	120x512
フレーム 2	{t - 2, t, t + 2}	9	1536x512
フレーム 3	{t - 3, t, t + 3}	15	1536x512
フレーム 4	{t}	15	512x512
フレーム 5	{t}	15	512x1500
プーリング層	[0, T)	T	1500Tx3000
セグメント 6	{0}	T	3000x512
セグメント 7	{0}	T	512x512
Softmax	{0}	T	512xS

4. 実　　験

4.1 データベース

本実験では，Kaldi-toolkit [13] を用いて x-vector に基づく話者照合システムを構築，評価した．データベースには VoxCeleb [14, 15] および Speaker In The Wild (SITW [16]) を用いた．VoxCeleb は二つのデータセット VoxCeleb1 [14], VoxCeleb2 [15] で構成されており，どちらのデータセットも Youtube にアップロードされた著名人のインタビュービデオから収集されている．VoxCeleb1 は学習用のデータベースが話者数 1211，発話数は 148,642，照合用のデータベースが話者数 40，発話数が 4874，VoxCeleb2 は話者数 5994，発話数は 1,092,009 含んでいる．SITW は登録データが話者数 119，発話数 1,958 となっており，テストデータは話者数 180，発話数 2,883 がそれぞれ含まれている．SITW は収録状況を制御したデータベースではなく，本来の背景ノイズ等を含み，より実環境に近いデータベースとなっている．SITW と VoxCeleb は別々の環境で収録または収集されているが，2 つのデータベースには話者 60 名が重複している．そのため，重複している話者を学習前に VoxCeleb のデータベースから削除した．また，データ拡張の一種として重畳するノイズのデータベースには MUSAN [17] と RIRNOISE [18] を用いた．MUSAN データベースは 900 以上のノイズと 42 時間の様々なジャンルの音楽，12 言語の 60 時間にわたる会話が含まれている．RIRNOISE は部屋の残響ノイズである．ノイズデータベース以外の全てのデータベースの言語は英語であり，16 kHz でサンプリングされている．

4.2 実 験 条 件

本実験ではデータベースに対する依存性の調査も含めるために 2 つの x-vector に基づく話者照合システムを構築した．それぞれのシステムの構築条件を以下に示す．

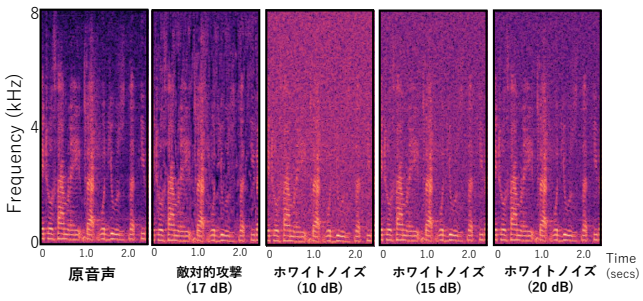
4.2.1 VoxCeleb1 を用いたシステム

DNN および PLDA のモデルの学習には VoxCeleb2 を用い，話者の登録および照合には VoxCeleb1 の照合用データを用い

表 2　VoxCeleb および SITW のシステムにおける各データセットの EER(%) と minDCF

	VoxCeleb			SITW		
	EER	minDCF (p=0.01)	minDCF (p=0.001)	EER	minDCF (p=0.01)	minDCF (p=0.001)
原音声 (ベースライン)	2.24	0.1454	0.1454	6.31	0.5291	0.7955
敵対的攻撃	**9.18**	0.6017	0.8297	**10.8**	**0.8236**	**0.9919**
ホワイトノイズ重畳 10 dB	8.22	**0.7273**	**0.8600**	9.64	0.8032	0.9775
ホワイトノイズ重畳 15 dB	6.17	0.5178	0.7042	7.99	0.6965	0.9491
ホワイトノイズ重畳 20 dB	5.01	0.4521	0.5276	7.14	0.6191	0.8886

* p は minDCF の重み係数



(発話内容 : “**very often I am and sometimes**”)

図 3　音声のスペクトログラム

てシステムを構築した．テストデータおよび敵対的サンプルの生成には VoxCeleb1 のテスト用データセットに含まれる話者 40 人から発話数 2955 を選択したものを用いた．VoxCeleb1 は各話者の様々な収録環境における発話で構成されていることから，すべての環境から発話を選ぶようにした．本実験では，選んだ原音声のみをテストに用いた実験をベースラインとした．

4.2.2　SITW を用いたシステム

DNN および PLDA のも出る学習には VoxCeleb1，2 を用い，話者の登録および照合には SITW の照合用データを用てシステムを構築した．テスト用データとして話者 180 人の中から 46 人，189 発話を選択した．SITW のテスト用データは発話長が長いため，各発話を 6 秒間隔に分割し，話者数 46，発話数 1370 を含む新たなデータセットをテストデータとし，このテストデータを用いた実験を SITW のベースラインとした．

4.2.3　評　　価

話者照合システムの評価には登録話者本人を誤って他人と認証してしまうエラー率 False Rejection Rate (FRR) と他人を登録者本人と認証してしまうエラー率 False Acceptance Rate (FAR) の 2 つの指標が存在する．本実験では，話者照合システムの評価には FRR と FAR が一致する等価エラー率 (Equal Error Rate; EER) と最小検出コスト関数 (minimum detection cost function; minDCF) を用いた．話者照合において，minDCF は一般的に，低い FRR を達成するよりも低い FAR を達成することが重要であるという考えに基づきシステムの性能を評価する．なお，minDCF の重み係数である p は 0.01，0.001 と定義する．

4.3　敵対的攻撃

敵対的攻撃には 3.1 節で述べた手法を用いる．本実験では，ブラックボックス型攻撃を想定しているため，攻撃者の攻撃対象となるシステムを深層学習に基づく音声認識システム Deep Speech [19] とし，入力音声の認識結果を常に “this is an adversarial example with a completely new transcription” となるように摂動を計算した．摂動を計算するための c には 1，ロス関数には CTCLoss [20]，最適化手法 には Adam [21] を使い，反復回数を 1000 回と指定し敵対的サンプルを生成した．これは 3.1 章で述べた手法および文献 [11] を準じている．本来の敵対的攻撃では人間が知覚できない程度のノイズの付与が想定されるが，本実験で生成した敵対的サンプルには多少のノイズが知覚できたため，その Signal-to-Noise Ratio (SNR) を測ったところ 20 dB 程度であった．

図 3 に本実験で用いたテスト音声のスペクトログラムの一部を示す．左から原音声，敵対的サンプル，ホワイトノイズ (10 dB)，ホワイトノイズ (15 dB)，ホワイトノイズ (20 dB) のスペクトログラムである．原音声と比較して，敵対的サンプルには少し歪みが生じている．一方，ホワイトノイズを重畳したサンプルは全周波数帯にまんべんなくノイズがのっており全体的に信号が歪んでいる．本実験で生成した敵対的サンプルは SNR が平均 20 dB 程度であったが同じ SNR のホワイトノイズを重畳したサンプルよりも歪みが明らかに少ないことがわかる．

4.4　結　　果

VoxCeleb と SITW それぞれのシステムにおける EER と minDCF を表 2 に示す．原音声および敵対的サンプルの EER を比べると，VoxCeleb と SITW のどちらの結果においても敵対的攻撃を照合時に用いた時の方が EER の大幅な悪化が確認できる．また，ホワイトノイズを重畳した音声の場合，敵対的攻撃と近い SNR (20 dB) であっても，敵対的攻撃ほど影響を与えていないことがわかる．この傾向は SITW のシステムの方が顕著に確認できる．その理由として，SITW は VoxCeleb に比べてより実環境に近いデータベースとなっているため，ノイズに対する頑健性が高く，ホワイトノイズによる影響が少なかったためだと考えられる．両データベースの 10 dB，15 dB，20 dB の EER をそれぞれ比較すると，SITW のほうが照合精度の変化が少ないことからもノイズに対する頑健性が確認できる．

次に，minDCF について比較すると特に VoxCeleb を用いたシステムではベースラインと比較して minDCF の悪化が大き

い．SITW を用いたシステムでは，ベースラインの minDCF もあまり低くないが，敵対的攻撃を行うと非常に悪化していることがわかる．それぞれの悪化はホワイトノイズを重畳した場合よりも顕著である．

以上のことからスペクトログラムでも比較した通り，ホワイトノイズを重畳した音声より敵対的攻撃で生成した音声のほうが音の劣化は少ないにもかかわらず最先端の話者照合システムの照合性能を大幅に悪化させてしまうことがわかった．つまりブラックボックス型の敵対的攻撃も考慮したより頑健な話者照合システムを構築する必要があるといえる．

5. ま と め

本論文では，ブラックボックス型敵対的攻撃に対する話者照合システムの脆弱性について調査した．2 つの話者照合システムに対して敵対的サンプルとホワイトノイズを重畳した音声のデータセットを作り，EER の変化を分析したところ，敵対的サンプル特有の摂動を付与した音声の方がランダムに生成されたホワイトノイズよりも照合精度を低下させたことを確認した．この結果，攻撃者にとって攻撃対象となるシステムの内部が不明な場合においても敵対的攻撃により，話者照合に大きく影響を与えられることが示された．

今後の課題として，様々な敵対的サンプルを用いた話者照合システムの敵対的学習があげられる．

謝　　辞

本研究の一部は JSPS 科研費若手研究 JP19K20271 と ROIS-DS-JOINT (021KP2019) の助成を受けたものである．

文　　献

[1] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2010.

[2] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur. Deep neural network embeddings for text-independent speaker verification. In *Interspeech*, pages 999–1003, 2017.

[3] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer. End-to-end text-dependent speaker verification. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5115–5119. IEEE, 2016.

[4] Wei-Ning Hsu, Yu Zhang, Ron J Weiss, Yu-An Chung, Yuxuan Wang, Yonghui Wu, and James Glass. Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5901–5905. IEEE, 2019.

[5] Simon JD Prince and James H Elder. Probabilistic linear discriminant analysis for inferences about identity. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[6] Daniel Garcia-Romero, David Snyder, Gregory Sell, Alan McCree, Daniel Povey, and Sanjeev Khudanpur. X-vector dnn refinement with full-length recordings for speaker recognition. In *Proc. Interspeech*, pages 1493–1496, 2019.

[7] Ahilan Kanagasundaram, Sridha Sridharan, Sriram Ganapathy, Prachi Singh, and Clinton B Fookes. A study of x-vector based speaker recognition on short utterances. 2019.

[8] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333. IEEE, 2018.

[9] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9):2805–2824, 2019.

[10] Felix Kreuk, Yossi Adi, Moustapha Cisse, and Joseph Keshet. Fooling end-to-end speaker verification with adversarial examples. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1962–1966. IEEE, 2018.

[11] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 1–7. IEEE, 2018.

[12] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

[13] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society, 2011.

[14] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.

[15] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

[16] Mitchell McLaren, Luciana Ferrer, Diego Castan, and Aaron Lawson. The speakers in the wild (sitw) speaker recognition database. In *Interspeech*, pages 818–822, 2016.

[17] David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015.

[18] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5220–5224. IEEE, 2017.

[19] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[20] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.