話者 V2S 攻撃：話者認証から構築される声質変換とその音声なりすまし可能性の評価

中村 泰貴[1]　齋藤 佑樹[2]　高道 慎之介[2]　井島 勇祐[3]　猿渡 洋[2]

概要：本稿では，話者認証を用いた声質変換（音声の話者性などを変換する技術）の構築法を提案し，声質変換を用いた音声なりすまし攻撃の可能性を評価する．話者認証は，音声入力により登録済みユーザを特定可能な，ユーザへの負担が小さい生体認証である．基本的に，話者認証システム自体には登録済みユーザの音声データは含まれないが，悪意のある攻撃者がそのシステムを暴露した場合，攻撃者は登録済みユーザの音声を復元できてしまう可能性を孕む．特に，攻撃者が声質変換を用いた場合，攻撃者の音声を登録済みユーザの音声に変換することで，登録済みユーザのあらゆる発話を再現できてしまう可能性がある．本稿では，この攻撃を話者 verification-to-synthesis (V2S) 攻撃と呼び，ホワイトボックスな話者認証システムへの話者 V2S 攻撃により声質変換モデルを学習する方法を提案する．提案法では，話者認証システムにおいて入力音声からの特徴抽出を行う話者認識モデルに由来する，話者の話者性を再現するための制約と，攻撃者によって事前に用意された音声認識モデルに由来する，音声の発話内容を保持するための制約に用いて声質変換モデルを学習する．実験的評価では，提案法により生成した音声の自然性・話者再現性（登録済みユーザの音声の話者性をどの程度再現できるか）を，登録済みユーザの音声データから学習される通常の声質変換と比較する．実験的評価の結果，提案法の性能は，ごく少量の音声データから学習される通常の声質変換と同程度であることを示す．

キーワード：話者認証，音声なりすまし攻撃，声質変換，音声認識

Taiki Nakamura[1]　Yuki Saito[2]　Shinnosuke Takamichi[2]　Yusuke Ijima[3]　Hiroshi Saruwatari[2]

Abstract: automatic speaker verification, voice impersonation attack, voice conversion, automatic speech recognition

Keywords: This paper presents a new voice impersonation attack using voice conversion (VC). Enrolling personal voices for automatic speaker verification (ASV) offers natural and flexible biometric authentication systems. Basically, the ASV systems do not include the users' voice data. However, if the ASV system is unexpectedly exposed and hacked by a malicious attacker, there is a risk that the attacker will reproduce the enrolled user's voices. Especially, voice conversion (VC), a method for transforming speaker individuality, has the potential to transform the attacker's any voice to the enrolled speaker's one. We name this the "speaker verification-to-synthesis (V2S) attack" and propose VC training with the ASV and pre-trained automatic speech recognition (ASR) models. The experimental evaluation compares converted voices between the proposed method that does not use the targeted speaker's voice data and the standard VC that uses the data. The experimental results demonstrate that the proposed method performs comparably to the existing VC methods that trained using a very small amount of parallel voice data.

1. はじめに

話者認証 (automatic speaker verification: ASV) [1], [2] は，音声入力により登録済みユーザを認証できる，ユーザ負担の小さい生体認証である．この技術は，音声以外の生体情報を用いずにユーザを識別できるため，指定された

[1] 東京大学 工学部
Faculty of Engineering, The University of Tokyo, Japan.
[2] 東京大学 大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo, Japan.
[3] NTT メディアインテリジェンス研究所
NTT Media Intelligence Laboratories, NTT Corporation, Japan.

© 2018 Information Processing Society of Japan

キーワードが発話に含まれるかを検出するキーワードスポッティング [3] や，スマートフォンの音声検索などに利用される．特に，テキスト非依存話者認証は，セリフを限定しない任意の発話から個人を認証できるため，ユーザにとってより自然な形での生体認証として期待される．

話者認証システムの更なる普及を見据え，話者認証を利用した音声なりすまし攻撃の可能性を議論すべきである．具体的には，悪意のある攻撃者が話者認証システムを暴露した場合，攻撃者は所望の登録済みユーザ（以降，攻撃対象話者）の音声を復元できてしまう可能性がある．我々はこの攻撃を話者 V2S (verification-to-synthesis) 攻撃と称し（図 1 の下半分），特に，声質変換 (voice conversion: VC) [4], [5] を用いた話者 V2S 攻撃について議論する．声質変換とは，入力音声の非言語情報（話者性など）を所望の情報に変換する技術である．2013 年頃からは，事前に収録された 2 話者（元話者・目標話者）の音声データと深層ニューラルネットワーク (deep neural network: DNN) のような統計モデル（声質変換モデル）に基づく声質変換法が提案され，元話者による任意の発話の話者性を目標話者の話者性に変換できるようになりつつある（図 1 の上半分）．本稿では，攻撃話者を元話者，攻撃対象話者を目標話者とみなし，攻撃者が声質変換の利用により攻撃対象話者の話者性を復元し，その話者性で任意の発話を再現できるかを議論する．通常，話者認証システム自体には登録済みユーザの音声データは含まれないため，システムが暴露された場合でも，通常の声質変換モデルの学習は不可能である．しかしながら，話者認証は音声の話者性を定量的に評価する機能を有するため，話者認証システムを詐称するように声質変換モデルを学習することで，攻撃対象話者の話者性を再現できてしまう可能性がある．

本稿では，ホワイトボックスの話者認証システムへの話者 V2S 攻撃による声質変換モデルの学習法を提案する．ここで，攻撃者は，話者認証システムにおいて，入力音声からその話者性を認識する話者認識モデルの DNN 構造と，攻撃対象話者のラベルを知っていると仮定する．この仮定は，やや非現実的だが，より現実的なブラックボックスの話者認証システム（すなわち，攻撃者にとって話者認証システムがどのように構築されているかが未知）への話者 V2S 攻撃の性能限界を調査する上で重要である．話者認識モデルは，音声に含まれる話者性のみを認識し，音韻性（発話内容）を認識しない．故に，話者認識モデルを詐称するように声質変換モデルを学習するだけでは，攻撃対象話者の話者性を再現できるものの，音韻性の失われた（すなわち，発話内容の分からない）音声しか生成できない．そこで提案法は，話者認識モデルを詐称するだけではなく，攻撃者によって事前に用意された音声認識（automatic speech recognition: ASR, もしくは speech-to-text）モデルにより予測される音素事後確率 [6] が声質変換前後で変わらない

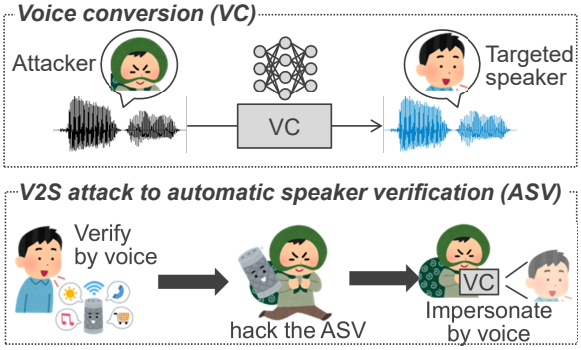

図 1　声質変換と話者 V2S 攻撃

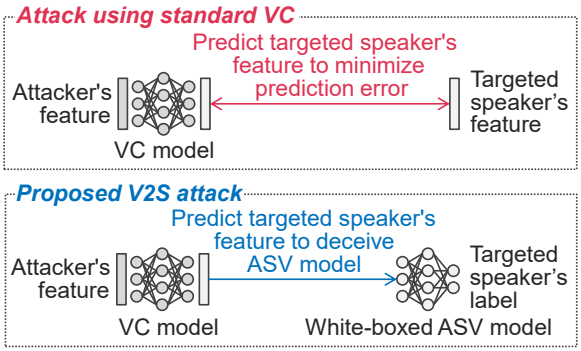

図 2　通常の DNN 声質変換（**2** 節）と提案する話者 V2S 攻撃（**3** 節）．VC と ASV はそれぞれ，声質変換と話者認証を表す．

ように，声質変換モデルを学習する．音素事後確率は，音声の発話内容を音素のソフトラベルで表現したものであるため，音素事後確率を一致させるように声質変換モデルを学習することで，入力音声の音韻性の保持が可能となる．実験的評価では，音声データから学習される通常の声質変換と，提案法により学習される声質変換の音声の品質を比較する．前者は，攻撃者が攻撃対象話者の音声データそのものを入手した場合を想定しており，提案法の音声品質の上限である．実験結果より，提案法の音質と話者再現性（攻撃対象話者の話者性をどの程度再現できるか）は，攻撃対象話者のごく少量の音声データから学習される通常の声質変換と同程度であることを示す．

2. 攻撃対象話者の音声データから学習される通常の DNN 声質変換

本節では，攻撃対象話者の音声データを使用する通常の DNN 声質変換（図 2 の上半分）を解説する．DNN 声質変換では，事前に獲得した音声データから，音声特徴量を変換する声質変換モデルを構築する．構築後には，攻撃者（声質変換の元話者）の任意の発話を攻撃対象話者（声質変換の目標話者）の音声に変換することが可能となる．この技術は，攻撃者が攻撃対象話者の音声データを入手した場合に可能な音声なりすまし攻撃である．故に，攻撃対象話者の音声データを用いない話者 V2S 攻撃の品質上限とみなされる．ここでは，DNN 声質変換の代表的な手法とし

© 2018 Information Processing Society of Japan

て，攻撃者・攻撃対象話者によるパラレル音声データ[*1]から学習される声質変換（**2.1 節**）と，目標話者の音声データ（すなわち，ノンパラレル音声データ）のみから学習される声質変換（**2.2 節**）に触れる．

2.1 攻撃者・攻撃対象話者によるパラレル音声データを用いた DNN 声質変換 [7]

$\boldsymbol{x} = [\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_T^\top]^\top$ と $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_t^\top, \cdots, \boldsymbol{y}_T^\top]^\top$ をそれぞれ，攻撃者または攻撃対象話者によるパラレル音声データから抽出された音声特徴量の時系列とする．これらの音声特徴量は，音声波形に対するフレーム分析により得られる．$\boldsymbol{x}_t$ と $\boldsymbol{y}_t$ はそれぞれ，フレーム t における，攻撃者または攻撃対象話者の音声特徴量ベクトルである．通常，このベクトルの次元数は数十程度（後述する実験的評価では 78 次元）である．t と T はそれぞれ，フレームインデックスと総フレーム数である．声質変換モデル（例えば DNN）を $\boldsymbol{G}(\cdot)$ とすると，攻撃者の音声特徴量系列を変換して得られる変換音声特徴量系列は，$\hat{\boldsymbol{y}} = \boldsymbol{G}(\boldsymbol{x}) = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_t^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ として与えられる．この声質変換モデルのパラメータは，攻撃対象話者の音声特徴量時系列 $\boldsymbol{y}$ と変換音声特徴量系列 $\hat{\boldsymbol{y}}$ の間の二乗誤差 $L_{\mathrm{MSE}}(\boldsymbol{y}, \hat{\boldsymbol{y}})$（次式）を最小化することで推定される．

$$L_{\mathrm{MSE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{T}(\hat{\boldsymbol{y}} - \boldsymbol{y})^\top (\hat{\boldsymbol{y}} - \boldsymbol{y}) \qquad (1)$$

最終的な音声波形は，ボコーダ処理（音声特徴量から音声波形を生成する処理）により生成される．

2.2 攻撃対象話者によるノンパラレル音声データのみを用いた DNN 声質変換 [8]

事前に収録した多数話者の音声データを用いることで，攻撃対象話者によるノンパラレル音声データのみから声質変換モデル $\boldsymbol{G}(\cdot)$ を構築することが可能である．本稿では，d-vector を用いたモデル適応に基づく手法 [8] を採用する．この手法は，攻撃対象話者の音声特徴量系列から得られる d-vector（DNN に基づいて得られる話者表現ベクトル [2]）を声質変換モデルに入力することで，攻撃者の音声特徴量時系列を攻撃対象話者の音声特徴量時系列に変換できる．

3. 話者 V2S 攻撃：話者認証から学習される DNN 声質変換

本節では，新たな音声なりすまし攻撃として 話者 V2S 攻撃（図 2 の下半分）を提案する．**2 節**で概説した通常の声質変換とは異なり，攻撃者は，攻撃対象話者の音声データを一切用いずに，攻撃対象話者が登録されたホワイトボックスの話者認証システムから声質変換モデルを構築する．声質変換モデルの学習には，話者認証システムの構成要素として含まれる話者認識モデルだけではなく，攻撃者によって事前に用意された音声認識モデルも使用する．話者認識モデルを詐称することで攻撃対象話者の話者性を復元でき，音声認識モデルを使用することで元音声による音韻性を保持できる．

3.1 話者性を再現するための話者認識モデル

S 名の事前に登録された話者の中の 1 名を特定する話者認識モデルを $\boldsymbol{V}(\cdot)$ とする．ここで，s_y 番目の話者のラベルは，one-hot の話者コード $\boldsymbol{l}_y = [l_y(1), \cdots, l_y(s), \cdots, l_y(S)]^\top$ として次式で与えられる [9].

$$l_y(s) = \begin{cases} 1 & \text{if } s = s_y \\ 0 & \text{otherwise} \end{cases} \qquad (1 \le s \le S) \qquad (2)$$

ここで，s は話者のインデックスである．話者認識モデル $\boldsymbol{V}(\cdot)$ は，話者事後確率（入力音声の話者が S 名の中で誰なのかを定量的に表す指標）をフレーム毎に予測する．話者事後確率の時系列を $\boldsymbol{V}(\boldsymbol{y}) = [\boldsymbol{v}_1^\top, \cdots, \boldsymbol{v}_t^\top, \cdots, \boldsymbol{v}_T^\top]^\top$ とし，フレーム t における事後確率を $\boldsymbol{v}_t = [v_t(1), \cdots, v_t(s), \cdots, v_t(S)]^\top$ とする．ここで，$v_t(s)$ は入力音声 $\boldsymbol{y}_t$ が s 番目の話者により発話されたものである確率を表す．$\boldsymbol{V}(\cdot)$ のモデルパラメータは，話者ラベル $\boldsymbol{l}_y$ と話者事後確率 $\boldsymbol{V}(\boldsymbol{y})$ の間の softmax cross-entropy $L_{\mathrm{SCE}}(\boldsymbol{l}_y, \boldsymbol{V}(\boldsymbol{y}))$（次式）を最小化することで推定される．

$$L_{\mathrm{SCE}}(\boldsymbol{l}_y, \boldsymbol{V}(\boldsymbol{y})) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{S} l_y(s) \log v_t(s) \qquad (3)$$

1 節で述べたように，本稿では，話者認識モデル $\boldsymbol{V}(\cdot)$ の DNN 構造と，攻撃対象話者のラベルが既知であると仮定する．故に，学習済みの話者認識モデルを用いることで，変換音声 $\hat{\boldsymbol{y}}$ の話者性と攻撃対象話者の話者性の違いを損失 $L_{\mathrm{SCE}}(\boldsymbol{l}_y, \boldsymbol{V}(\hat{\boldsymbol{y}}))$ により定量的に評価でき，さらに，この損失を用いた backpropagation によって他の DNN を学習できる．

3.2 音韻性を保持するための音声認識モデル

話者認識モデル $\boldsymbol{V}(\cdot)$ を詐称する損失 $L_{\mathrm{SCE}}(\boldsymbol{l}_y, \boldsymbol{V}(\hat{\boldsymbol{y}}))$ を用いることで，攻撃対象話者の話者性を再現する声質変換モデルの学習が実現できる．しかしながら，この学習は，声質変換の前後で音韻性が不変であることを保証しない．そこで，本稿では，入力音声の発話内容を予測するように事前に構築された音声認識モデル $\boldsymbol{R}(\cdot)$ を利用する．$\boldsymbol{R}(\cdot)$ は入力音声の音素事後確率をフレーム毎に予測するため，攻撃者の音声と変換後の音声の音韻性の乖離度を，音素

[*1] 攻撃者と攻撃対象話者による同一内容の発話音声データであり，例えば，両話者が「こんにちは」「ありがとう」と発話した事前収録音声データの対を指す．前述したように，十分な音声データ量を用いて学習された DNN 声質変換は，事前収録音声データに含まれない発話においても，攻撃者の声質を攻撃対象話者の声質に変換することが可能である．

© 2018 Information Processing Society of Japan

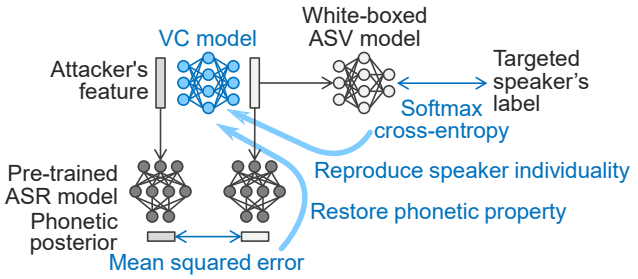

図 3　提案する話者 V2S 攻撃の学習

事後確率の二乗誤差 $L_{\mathrm{MSE}}(\boldsymbol{R}(\boldsymbol{x}), \boldsymbol{R}(\hat{\boldsymbol{y}}))$ として定量的に評価できる．さらに，$\boldsymbol{R}(\cdot)$ を DNN で表現することで，損失 $L_{\mathrm{MSE}}(\boldsymbol{R}(\boldsymbol{x}), \boldsymbol{R}(\hat{\boldsymbol{y}}))$ を用いた backpropagation による声質変換モデルの学習が実現できる．

3.3　声質変換モデルの学習方法

提案法における声質変換モデル学習の損失関数 $L(\cdot)$ は，次式で与えられる．

$$L\left(\boldsymbol{x}, \hat{\boldsymbol{y}}, \boldsymbol{l}_y\right) = L_{\mathrm{SCE}}\left(\boldsymbol{l}_y, \boldsymbol{V}\left(\hat{\boldsymbol{y}}\right)\right) + \omega L_{\mathrm{MSE}}\left(\boldsymbol{R}\left(\boldsymbol{x}\right), \boldsymbol{R}\left(\hat{\boldsymbol{y}}\right)\right) \tag{4}$$

ここで，ω は第二項の重みを調整するハイパーパラメータである．式 (4) の最小化により，攻撃対象話者の音声 $\boldsymbol{y}$ を明示的に用いない声質変換モデルの学習が実現できる．図 3 に話者 V2S 攻撃の概念図を示す．

3.4　提案法に関する議論と考察

話者 V2S 攻撃は，音声なりすまし攻撃 [10] の一種であり，その目的は，攻撃者が攻撃対象話者の声になりすますことである．これまでの音声なりすまし攻撃の研究では，攻撃対象話者の音声データを用いた合成音声 [10]（本稿では **2** 節が該当）や録音再生音声 [11] を用いた攻撃方法が提案されてきた．話者 V2S 攻撃は，前者に類似した攻撃だが，攻撃対象話者の音声データを使用せずに話者認証システムから攻撃対象話者の声質を復元する方法である．関連研究として我々は，音声なりすまし攻撃を検知する anti-spoofing を攻撃（詐称）して声質変換モデルを改善する方法を提案 [12] している．一方で本稿の話者 V2S 攻撃は，話者認証システムを攻撃して声質変換モデルを構築するものである．

上記の観点から，話者 V2S 攻撃は，パブリックドメインの音声コーパスから話者認証システムを詐称できる音声データを選択する手法 [13] に関連する．これらの手法は，話者認証システムを詐称する点で共通するが，話者 V2S 攻撃は更に，攻撃対象話者のあらゆる発話を人工的に生成することを目標とする．

敵対的攻撃 (adversarial attack) は，学習済み認識モデル（例えば，音声認識 [14] や画像認識 [15]）に対する，良く知られた攻撃方法である．話者 V2S 攻撃の目的は，敵対的攻撃の目的と全く異なる．敵対的攻撃は，認識モデルを誤認識させるようなデータサンプルを生成するのに対し，話者 V2S 攻撃は，認識モデルから攻撃対象の属性（本稿では話者性）を復元することを目的とする．

上記の観点から，話者 V2S 攻撃は，classifier-to-generator 攻撃 [16] や membership inference 攻撃 [17] に強く関連する．これらの攻撃は，学習済み認識モデルの学習データもしくはその確率分布を推定するものである．これらの攻撃を利用することで，話者 V2S 攻撃による変換音声の品質を改善できる可能性がある．すなわち，話者認識モデルの学習に利用した発話を推定することで，声質変換の性能改善が期待できる．

提案法は，**2.2** 節の手法やデータ選択手法 [13] のように多人数話者の事前収録音声データを使用しない．話者 V2S 攻撃の将来的な展望として，この事前収録音声データと転移学習 [18] を使用した手法が考えうる．また，より現実的なタスクにおける攻撃を見据え，本稿では無視した以降の要素も考慮する必要がある．

- **波形生成処理**: 本稿の声質変換モデルは音声特徴量を生成し，音声波形を別途のボコーダ処理で生成する．より現実的なタスクでは，ボコーダ処理（例えば，neural vocoder [19]）も含めた学習・生成が必要である．
- **空間伝達**: 本稿では声質変換システムから生成された音声（特徴量）が直接的に話者認証システムに入力されることを想定している．より現実的なタスクでは，音声波形が声質変換システムから放射されて話者認証システムに収音されるまでの，音響的な空間伝達特性を考慮する必要がある．
- **特徴量分析処理**: 本稿では話者認証システムに入力された音声（特徴量）が直接的にそのシステムで使用されることを想定している．より現実的なタスクでは，音声の統計量 [1] を使用する場合，end-to-end 型など [20], [21] も考慮する必要がある．

本稿では，ホワイトボックスの話者認証システムに対する攻撃法を提案した．より現実的なタスクでは，話者認証システムはブラックボックスであり，特徴量分析法・認証システムの構成・話者ラベルは攻撃者にとって未知である．これらの課題を扱うために，敵対的生成ネットワーク [22], [23] や強化学習 [24] に基づくアプローチが必要になると考えられる．

4.　実験的評価

4.1　実験条件

声質変換・話者認証・音声認識に使用する音声特徴量は 39 次元（1 次から 39 次）のメルケプストラム係数（スペクトル特徴量に相当）とその動的特徴量である．音声分析合成には STRAIGHT [25] を使用する．音声分析時のフ

© 2018 Information Processing Society of Japan

レームシフトは 5 ms である．攻撃者は男性 1 名，攻撃対象話者は男女各 2 名とし，声質変換モデルの構築は攻撃者と攻撃対象話者のペア毎に行う（すなわち，同性間と異性間の声質変換がある）．評価では，各ペアの 2 話者が発話した 25 発話のパラレル音声データを用いる．

話者 V2S 攻撃の声質変換モデルは，78 ユニットの入力層，$\{256, 128\}$ ユニットの ReLU [26] 隠れ層，78 ユニットの線形出力層から成る Feed-forward DNN である．話者認識モデルは，78 ユニットの入力層，$\{1024 \times 3, 8 \times 1\}$ ユニットの sigmoid 隠れ層，260 ユニット（登録話者数に相当）の線形出力層から成る Feed-forward DNN である．音声認識モデルは，78 ユニットの入力層，1024×4 ユニットの sigmoid 隠れ層，56 ユニット（音素数に相当）の線形出力層から成る Feed-forward DNN である．話者認識モデル・音声認識モデルの学習には，260 名（男性 130 名・女性 130 名）の日本語母語話者音声データを使用する．この 260 名に攻撃者は含まれず，攻撃対象話者は含まれる．声質変換モデルのパラメータ最適化には，学習係数を 0.1 とした AdaGrad [27] を使用する．声質変換モデルの学習の反復回数は 25 回とし，学習に使用する攻撃者の音声データ量は 200 発話である．式 (4) の重み ω は 0.01 とする．話者 V2S 攻撃では，声質変換モデルを用いて攻撃者のスペクトル特徴量を攻撃対象話者のスペクトル特徴量に変換する．0 次のメルケプストラム係数と帯域平均化非周期性指標 [28], [29] は変換しない．音高（F_0）の変換に関して，話者 V2S 攻撃では攻撃対象話者の音高は観測できず，本稿で提案する枠組みでは音高の変換は困難である．そこで本稿では，攻撃対象話者の音高の統計量（平均と分散）は攻撃者にとって既知であると仮定し，攻撃者の音高を線形変換することで，目標話者の音高に変換する [5]．

パラレル音声データを用いた DNN 声質変換（**2.1 節**）における声質変換モデルのアーキテクチャと学習時の反復回数は，話者 V2S 攻撃で使用するものと同じである．この学習データ量については，後の節で議論する．音高，0 次のメルケプストラム係数，帯域平均化非周期性指標の変換処理についても，話者 V2S 攻撃と同様である．ノンパラレル音声データを用いた DNN 声質変換（**2.2 節**）における声質変換モデルのアーキテクチャ，学習データ，それ以外のハイパーパラメータについては，変分オートエンコーダ [30]・音素事後確率 [6]・d-vector [2] を使用する既存研究 [8] と同様である．

4.2 実験結果

通常の声質変換と，提案する話者 V2S 攻撃の変換音声を評価した．ここでは，変換音声の自然性に関するプリファレンス AB テストを実施した．評価した手法は以下の通りである．

- **ParaVC**: { 5, 10, 30 } 発話のパラレル音声データを用いた DNN 声質変換（**2.1 節**）
- **NonparaVC**: ノンパラレル音声データを用いた DNN 声質変換（**2.2 節**）
- **V2S**: 提案する 話者 V2S 攻撃（**3 節**）

評価者には，変換音声をランダムな順で提示し，より自然に聴こえる方を選択させた．同様に，話者再現度に関するプリファレンス XAB テストを実施した．リファレンス X は，攻撃対象話者の自然音声とした．全ての主観評価は，我々のクラウドソーシング評価システム上で実施した．各評価で 40 名が参加し，各評価者は評価データのうち 10 個の音声対を評価した．最終的な評価者数の総計は，2（AB，XAB テスト）× 2（同性間，異性間変換）× 4（手法の組み合わせ数）= 640 であった．

表 1 と 表 2 にそれぞれ，同性間・異性間の声質変換におけるプリファレンス AB テストの結果を示す．同様に，表 3 と 表 4 にそれぞれ，同性間・異性間の声質変換におけるプリファレンス XAB テストの結果を示す．表 1 と 表 2 の結果から，提案する話者 V2S 攻撃の音声の自然性は，NonparaVC の自然性を下回ることが分かる．一方で，同性間・異性間変換の両方において，話者 V2S 攻撃の音声の自然性は，ParaVC (5 utts) の音声の自然性と同程度であることが分かる．表 3 と 表 4 から，話者 V2S 攻撃の話者再現度は，通常の声質変換の話者再現度に劣ることが分かる．ただし，同性間変換の ParaVC (5 utts) に対しては，同程度の話者再現度であることが分かる．以上より，提案する話者 V2S 攻撃は，ごく少量（5 発話程度）のパラレル音声データから学習される通常の声質変換と同程度の品質を達成できることがわかる．

表 1 音声の自然性に関するプリファレンス AB テストの結果（同性間変換）．太字は p 値が 0.05 より小さい手法である．

A	スコア	p 値	B
ParaVC (5 utts)	0.388 vs. **0.612**	1.221×10^{-10}	V2S
ParaVC (10 utts)	0.475 vs. 0.525	0.158	V2S
ParaVC (30 utts)	0.458 vs. **0.542**	0.016	V2S
NonparaVC	**0.598** vs. 0.402	2.694×10^{-8}	V2S

表 2 音声の自然性に関するプリファレンス AB テストの結果（異性間変換）． 太字は p 値が 0.05 より小さい手法である．

A	スコア	p 値	B
ParaVC (5 utts)	0.490 vs. 0.510	0.572	V2S
ParaVC (10 utts)	**0.593** vs. 0.407	1.365×10^{-7}	V2S
ParaVC (30 utts)	**0.610** vs. 0.390	3.174×10^{-10}	V2S
NonparaVC	**0.538** vs. 0.462	0.034	V2S

5. まとめ

本稿では，声質変換を用いた音声なりすまし攻撃として，話者 verification-to-synthesis (V2S) 攻撃を提案した．話者 V2S 攻撃において声質変換モデルは，攻撃対象話者の話者性を復元するためにホワイトボックスの話者認証シス

© 2018 Information Processing Society of Japan

表 3　話者再現度に関するプリファレンス XAB テストの結果（同性間変換）．太字は p 値が 0.05 より小さい手法である．

A	スコア	p 値	B
ParaVC (5 utts)	0.530 vs. 0.470	0.090	V2S
ParaVC (10 utts)	**0.615** vs. 0.385	$< 10^{-10}$	V2S
ParaVC (30 utts)	**0.675** vs. 0.325	$< 10^{-10}$	V2S
NonparaVC	**0.660** vs. 0.340	$< 10^{-10}$	V2S

表 4　話者再現度に関するプリファレンス XAB テストの結果（同性間変換）．太字は p 値が 0.05 より小さい手法である．

A	スコア	p 値	B
ParaVC (5 utts)	**0.585** vs. 0.415	1.324×10^{-6}	V2S
ParaVC (10 utts)	**0.713** vs. 0.287	$< 10^{-10}$	V2S
ParaVC (30 utts)	**0.705** vs. 0.295	$< 10^{-10}$	V2S
NonparaVC	**0.588** vs. 0.412	$< 10^{-10}$	V2S

テムを詐称し，かつ，発話の音韻性を保持するために音声認識モデルを利用して構築される．実験的評価から，この枠組みにより学習・生成された音声の自然性・話者再現度は，攻撃対象話者によるごく少量（5 発話程度）の音声データを用いて学習される通常の声質変換と同程度であることを明らかにした．今後は，多数話者の事前収録音声データを用いた 話者 V2S 攻撃を扱う．

謝辞：本研究の一部は，セコム科学技術支援財団の助成を受け実施した．

参考文献

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 788–798, 2011.

[2] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, “Deep neural networks for small footprint text-dependent speaker verification,” in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 4080–4084.

[3] R. Prabhavalkar, R. Alvarez, C. Parada, P. Nakkiran, and T. Sainath, “Automatic gain control and multi-style training for robust small-footprint keyword spotting with deep neural networks,” in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4704–4708.

[4] Y. Stylianou, O. Cappé, and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[5] T. Toda, A. W. Black, and K. Tokuda, “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[6] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, “Phonetic posteriorgrams for many-to-one voice conversion without parallel data training,” in *Proc. ICME*, Seattle, U.S.A., Jul. 2016.

[7] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, “Voice conversion using artificial neural networks,” in *Proc. ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 3893–3896.

[8] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, “Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors,” in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 5274–5278.

[9] N. Hojo, Y. Ijima, and H. Mizuno, “Dnn-based speech synthesis using speaker codes,” *IEICE Transactions on Information and Systems*, vol. E101-D, no. 2, pp. 462–472, Feb. 2018.

[10] Z. Wu, X. Xiao, E. S. Chng, and H. Li, “Synthetic speech detection using temporal modulation feature,” in *Proc. ICASSP*, Vancouver, Canada, May. 2013, pp. 7234–7238.

[11] J. Lindberg and M. Blomberg, “Vulnerability in speaker verification - a study of technical impostor techniques,” in *Proc. EUROSPEECH*, Budapest, Hungary, Mar. 1999, pp. 1211–1214.

[12] Y. Saito, S. Takamichi, and H. Saruwatari, “Statistical parametric speech synthesis incorporating generative adversarial networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 755–767, Jun. 2018.

[13] T. Kinnunen, R. G. Hautamki, V. Vestman, and M. Sahidullah, “Can we use speaker recognition technology to attack itself? enhancing mimicry attacks using automatic target speaker selection,” in *Proc. ICASSP*, Brighton, United Kingdom, May 2019.

[14] H. Yakura and J. Sakuma, “Robust audio adversarial example for a physical attack,” in *Proc. IJCAI*, Macao, China, Aug. 2019, pp. 5334–5341.

[15] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv:1412.6572*, 2014.

[16] K. Kusano and J. Sakuma, “Classifier-to-generator attack: Estimation of training data distribution from classifier,” 2018. [Online]. Available: https://openreview.net/forum?id=SJOl4DlCZ

[17] S. Hisamoto, M. Post, and K. Duh, “Membership inference attacks on sequence-to-sequence models,” *arXiv:1904.05506*, 2019.

[18] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, “Transfer learning from speaker verification to multi-speaker text-to-speech synthesis,” vol. abs/1806.04558, 2018.

[19] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, “WaveNet: A generative model for raw audio,” vol. abs/1609.03499, 2016.

[20] C. Zhang and K. Koishida, “End-to-end text-independent speaker verification with flexibility in utterance duration,” in *Proc. ASRU*, Okinawa, Japan, Dec. 2017, pp. 584–590.

[21] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. L. Y. Bengio, and A. Courville, “Towards end-to-end speech recognition with deep convolutional neural networks,” *arXiv*, vol. abs/1701.02720, 2017.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Proc. NIPS*, Montreal, Canada, Dec. 2014, pp. 2672–2680.

[23] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, “MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement,” vol. abs/1905.04874, 2019.

[24] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, “DNN-based source enhancement self-

© 2018 Information Processing Society of Japan

optimized by reinforcement learning using sound quality measurements," in *Proc. ICASSP*, Orleans, U.S.A., Mar. 2017, pp. 81–85.

[25] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.

[26] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. ICML*, Haifa, Israel, June 2010, pp. 807–814.

[27] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, Jul. 2011.

[28] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *MAVEBA*, Firentze, Italy, Sep. 2001, pp. 1–6.

[29] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," in *Proc. INTERSPEECH*, Pittsburgh, U.S.A., Sep. 2006, pp. 2266–2269.

[30] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv*, vol. abs/1312.6114, 2013.

© 2018 Information Processing Society of Japan