6M-08

話者照合におけるプロの物真似タレントの声真似攻撃の影響の分析

堀畑　拓斗†　　岩野　公司†
東京都市大学†

1. はじめに

近年，様々な情報システムに対する高いセキュリティの確保を目的として，人間の生体情報を利用した個人認証技術が注目されている．その中でも「音声による認証（話者照合）」は，その手軽さなどから利用の期待が高まっており，様々な研究が進められている[1].

話者照合の実用化を考えると，成りすましによる攻撃に対する脆弱性を十分に把握しておくことが重要である．最も簡単な成りすまし攻撃として「声真似（模倣）」が考えられる．我々はこれまでに，声真似に特別な技能を有さない「素人」を対象にその攻撃力を分析し，成りすましが成功する可能性が十分にあることを明らかにしている[2]．一方，声質を他人に似せることを日ごろから訓練し，高い技術を有している人物が成りすましを行う可能性も十分に考えられ，その攻撃力を把握することも重要となる．

そこで本研究では，声真似に高い技術を有する人物として「プロの物真似タレント」を対象とし，素人と比較してどれだけの攻撃力を有し，どのような物真似音声の特徴を有しているのかを明らかにする．

2. 使用音声データ

プロ以外の（素人の）発声として，先行研究[2]で使用された6名の男子学生の音声を利用する．このうち，分析に用いる話者照合システムの構築用データとして，約2週間にわたる5セッション分のデータを使用する．セッション間は1日以上空いており，被験者1人あたり，それぞれのセッションで 4 桁連続数字を 10 回ずつ発声している．この5セッションとは別の日に行われた1セッションについて，本人の地声にと他の5人の物真似を行ったときの連続数字発声を収録しており，それらを照合性能の評価と分析に用いる．このうち物真似については，対象者の声を聴取したただけで模倣を行った場合（訓練なし）と，照合システムから出力される照合スコアを参考に，できるだけその値が大きくなるように訓練を行ってから模倣を行った場合（訓練あり）が存在し，その双方を分析に用いる．各被験者について，地声 10 発声，訓練なしの物真似 50 発声（10発声×5名），訓練ありの物真似50発声（10発声×5名）のデータとなる．

本研究では新たにプロの物真似タレント 1 名（40 代男性，キャリア約 20 年）による発声の収録を行った．本人の地声と，上記の男子学生 6 名の物真似による連続数字発声（各 10 回）を依頼し，それをプロの模倣の評価と分析に用いる．なお，プロは事前にこの 6 名とは面識がなく，物真似を行うことも初めてである．

Analysis of effects of voice mimicry by professional impersonators on speaker verification
Takuto Horihata†, Koji Iwano†, †Tokyo City University

3. プロ／素人による模倣攻撃力の分析

3.1 話者照合システムの構築

2 章のデータを利用し，隠れマルコフモデル（HMM）でモデル化された申告話者モデルと不特定話者モデル（UBM）を利用する話者照合システムを構築する．各話者を 3 状態の HMM でモデル化するが，基本的な原理は GMM-UBM 法[3]と同じ照合の枠組みを利用している．

照合の流れは以下のようになる．まず，入力音声をフレームごとに 12 次元 MFCC とその 1 次微分成分，対数パワーの 1 次微分成分の計 25 次元のベクトルに変換する．得られた特徴量系列 X を申告話者モデル（C）と不特定話者モデル（UBM: U）に入力し，それぞれのモデルから対数尤度 $\log P(X|C), \log P(X|U)$ を算出する．照合スコア $S(X)$ は式(1)で定義され，この照合スコアが設定したしきい値よりも大きければ申告話者として受理し，小さければ詐称者とみなされ棄却する．

$$S(X) = \log P(X|C) - \log P(X|U) \qquad (1)$$

3.2 プロ／素人の模倣に対する照合性能の比較

図 1 に物真似発声を詐称者の発声として用いたときの，しきい値に対する詐称者受理率と本人

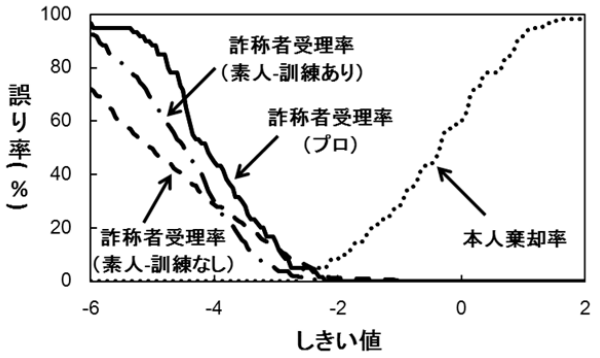

図 1　プロと素人の物真似に対する照合誤りの比較

棄却率の変化の様子を示す．詐称者受理曲線をみると，誤り率が「素人（訓練なし）＜素人（訓練あり）＜プロ」の順に有意に増加しており，プロの方が他人へのなりすましに成功する可能性が高いことがわかる．等誤り率は，素人（訓練なし）で 5.0%，素人（訓練あり）で 1.2%，プロでは 5.0%となった．なお，話者モデルの混合数は事後的な最適化により 64 と設定した．

4. 発声間距離による物真似音声の音響分析

4.1 ケプストラムに基づく発声間距離の定義

プロと素人の物真似音声の特徴の違いを明らかにするため，以下の 3 つの発声に対して音響モデルを構築し，発声間の距離を調べる[2]．図 2 に分析対象となる 3 つの距離（A, B, C）を示す．

- U^i : 発話者 i の自然な発声（地声）
- U^j : 模倣対象者 j の自然な発声（地声）
- $U^{i\sim j}$: 発話者 i が対象者 j を模倣して行った発声

発声間距離の算出のため，まず，分析対象音声をケプストラムに基づく音響特徴量（12 次元 MFCC+12 次元ΔMFCC+Δ対数パワー）に変換し，各発声を 3 状態の HMM（混合数 K）でモデル化する．発声 U^x のモデルの 2 状態目の k 番目の正規分布を N_k^x と表すとき，発声 U^a，U^b 間の距離 $D(U^a,U^b)$ は式(2)のように定義される．

$$D(U^a,U^b)=\frac{1}{2}\left\{\sum_k^K w_k^a\cdot\min_l SKL\left(N_k^a,N_l^b\right)+\sum_k^K w_k^b\cdot\min_l SKL\left(N_l^a,N_k^b\right)\right\} \quad (2)$$

ここで，SKL は対称化 Kullback-Leibler 情報量であり，KL 情報量（KL）[4]を用いて式(3)のように定義される．

$$SKL(N^a,N^b)=KL(N^a||N^b)+KL(N^b||N^a) \quad (3)$$

4.2 分析結果

表 1 に素人とプロの発声に対する発声間距離の

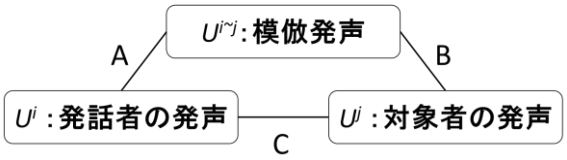

図 2　分析対象とする発声間距離

表 1　プロと素人の発声間距離の比較

	A	B	C	A/C	B/C
素人（訓練なし）	15.4	20.5	19.8	0.78	1.04
素人（訓練あり）	17.6	18.7	19.8	0.89	0.95
プロ	13.6	14.5	21.0	0.65	0.69

比較を示す．混合数は 8 である．「A/C」は「話者間距離に対し，物真似時にどの程度地声から特徴量を変化させているか」を表しており，「B/C」は「話者間距離に対し，模倣によってどれだけ対象者に近づいたか」を表している．この数値を見ると，素人は物真似時に地声から大きく特徴を変化させることはできても，訓練の有無にかかわらず対象者の特徴に近づくことはできないが，プロは比較的小さい変化であっても，確実に対象者の特徴に近づいていることがわかる．

5. まとめ

本研究では，話者照合に対するプロの物真似タレント物真似攻撃の分析を行った．その結果，素人に比べてプロの物真似タレントは対象者の声質に効率的に近づくことができ，成りすましに成功する可能性が高くなることがわかった．今後は，データの増加による分析結果の信頼性の向上や，今回の知見を活用した声真似攻撃対策手法の提案などを検討する必要がある．

謝辞　本研究は JSPS 科研費 基盤研究（C）25330206 の助成を受けたものです．

参考文献

[1] 越仲，篠田，“話者認識の国際動向，”日本音響学会誌，vol.69, no.7, pp.342-348, 2013.
[2] 岩野ら，“声真似が話者照合に与える影響と物真似音声の音響特徴の分析，”電子情報通信学会技術報告，vol.114, no.411, pp.43-48, 2015.
[3] D. A. Reynolds, et al., “Speaker verification using adapted Gaussian Mixture Models,” Digital Signal Processing, vol. 10, pp. 19-41, 2000.
[4] S. Kullback and R. A. Leibler, “On information and sufficiency,” Ann. Math. Statist., vol.22, no.1, pp.79–86, 1951.