一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2017-32(2017-08)

プロの物真似タレントの声真似が話者照合に与える影響と音響特徴の分析

岩野　公司†　　堀畑　拓斗†

†東京都市大学 メディア情報学部 〒224-8551 横浜市都筑区牛久保西 3-3-1

E-mail:　†iwano@tcu.ac.jp

あらまし　話者照合の実用化を考えると，最も手軽な成りすまし攻撃である「声真似（模倣）」に対するシステムの脆弱性を十分に把握しておく必要がある．我々の先行研究では，物真似に特別な技術を有さない一般人（素人）の声真似の攻撃力の分析を行っており，その成りすましが十分成功する可能性があることを明らかにしている．そこで本稿では，声質を他人に似せることに高い技術を有している人物が行った声真似が，素人の声真似に比べて，どの程度の攻撃力を有しているか，また，声真似によってどのように声の音響特徴を変化させているかについて分析を行う．1 名のプロの物真似タレントが事前に面識のない 6 名の物真似を行った音声を，HMM に基づく話者照合システムの詐称者発声として入力し性能を調査したところ，素人の声真似よりも高い攻撃力を有していることが明らかになった．また，プロの物真似音声のケプストラム特徴をカルバック・ライブラー情報量に基づく発声間距離を用いて分析したところ，プロの物真似タレントは素人よりも声質を大きく変化させていない一方で，対象者には確実に近づいている様子が明らかになった．

キーワード　話者照合，物真似音声，成りすまし攻撃，音響特徴分析，プロの物真似

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）」が挙げられる[2]. 我々はこれまでに，物真似に特別な技能を有さない一般人（声真似の素人）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[3]. 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．従来までに，プロの物真似タレントによる声真似が話者照合性能に与える影響を分析した研究として，Hautamäki らの研究[4]などが挙げられる．文献[4]では，プロの声真似によって，GMM-UBM 法に基づく話者照合[5]と i-vector とそのコサイン距離に基づく話者照合[6]のそれぞれのシステムで性能劣化が生じることが報告されているが，一般の素人の行う声真似との比較については言及されていない．Mariéthoz らは，GMM-UBM 法に基づく話者照合に対し，プロの物真似タレントの声真似と，一般の素人の声真似を与えた場合の影響を調査している[7]. しかし，声真似の対象者に物真似タレントが得意とする人物を設定していることから，プロと素人の詐称能力を公平に比較することができない．それに対し本研究では，プロ，素人ともに，これまでに物真似を行ったことがない（共通の）人物を対象とした声真似を対象にした分析を行うことから，両者の詐称の攻撃力を純粋に比較することが可能である．

以降では，まず 2 章において本研究で使用する音声データについて説明を行う．3 章でプロが行った声真似を詐称攻撃に利用した場合の話者照合性能の調査を行い，その攻撃力を素人の声真似と比較する．4 章では，プロが声真似を行うことによってどのように発声の音響特徴が変化するのかについて，ケプストラムの発声間距離を利用した分析により明らかにする．最後に 5 章で本稿の結論を述べる．

2. 使用する音声データ

我々の先行研究[3]では，一般人（素人）の声真似が話者照合システムに与える影響の分析を行っている．この研究では，3 週間，約 2 日ごと（計 9 日）に本学学生の地声（声真似を行っていないときの発声）と物真似音声の収録を行っている．発声内容は 4 桁の連続数字であり，発声する 4 桁数字は日ごとに変更しているが，全員が同じ内容の発声を行っている．

本研究では，このうちの男子学生（6 名）が行った 5 日分の地声の音声を用いて，性能分析のための話者照合システムを構築する．また，この 5 日とは別の 1 日に行われた，本人の地声と他の 5 人の物真似を行ったときの連続数字発声を，比較のための素人の声真似の分析に用いる．声真似については，対象者の声を聴取し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声真似を行った場合（訓練あり）の試行が存在するので，その双方を分析に用いることとした．地声，5 人の声真似それぞれについて 4 桁連続数字を 10 個ずつ発声しているため，各人について，地声 10 発声，訓練なしの声真似 50 発声（10 発声×5 名），訓練ありの声真似 50 発声（10 発声×5 名）のデータが素人の声真似の分析に用いられることになる．

本研究では新たに，テレビなどに良く出演している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．本人の地声による連続数字発声を 10 発声収録した上で，上述の男子学生 6 名の物真似による連続数字発声（それぞれ 10 発声ずつ）を依頼し，それをプロの声真似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，声真似を行うことも初めてである．収録前に各学生の発声を数回聴取してもらい，数分間の練習を行った後で声真似を行っており，この条件は素人の「訓練なし」と同様である．

3. プロと素人の声真似の攻撃力の分析

3.1. 分析用話者照合システムの構築

2 章で説明した男子学生 6 名 5 日分のデータを利用して，隠れマルコフモデル（HMM）に基づく話者照合システムを構築する．この照合手法では，申告話者モデルと不特定話者モデル（UBM）を 3 状態の HMM でモデル化しているが，GMM-UBM 法[5]と同じ照合の枠組みを利用している．なお，UBM は 6 名の 5 日分の 300 発声で学習する．

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．

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

物真似音声を詐称者音声として入力しときの性能の変化を正確に把握するため，物真似音声の入力を想定しない環境でシステムの性能を最大化させておく．そこで，申告話者モデルと不特定話者モデルの混合数を 1, 2, 4, 8, 16, 32, 64, 128, 256 と変化させながら，素人の地声の発声とプロの地声の発声を詐称者音声としたときの照合性能を調査した．その結果，混合数が 64 のときに最も等誤り率が小さくなり，素人に対しては 1.7%，プロに対しては 0%の等誤り率となった．以降の実験では，この混合数 64 の照合システムを性能分析に使用する．

なお，i-vector とコサイン距離に基づく話者照合手法についても同様に性能の調査を行ったが，4 桁連続数字という短い発声を対象とする今回の実験では，HMM に基づく手法の方が高い性能となったため，分析には採用しなかった．

3.2. 声真似が照合性能に与える影響の分析

まず，男子学生（素人）の発声について，しきい値を変化させたときの詐称者受理率と本人棄却率の変化の様子を図 1 に示す．詐称者受理率については，詐称者の音声として「地声による発声（声真似なし）」「訓練を行わずに物真似を行ったときの発声（声真似あり・訓練なし）」「訓練を行って物真似を行ったときの発声（声真似あり・訓練あり）」を用いた場合の 3 通りを示している．しきい値の小さい領域をみると，詐称者受理率が，「声真似なし＜声真似あり・訓練なし＜声真似あり・訓練あり」の順に有意に増加しており，この結果は先行研究[3]で得られた傾向と一致する．ただし，等誤り率付近では「訓練あり」と「訓練なし」の詐称者受理率に逆転がみられ，先行研究の結果と異なっている．これは，先行研究では 3 日分の素人の発声データを評価（分析）に使用しているのに対し，今回の実験では（プロが行った物真似発声の量と合わせるために）1 日分に減らしているため，誤差が大きくなっていることが影響していると考えられる

図 2 には，詐称者の発声として，プロの地声の発声を使用した場合（声真似なし）と，プロの物真似音声を使用した場合（声真似あり）の詐称者受理率の変化の様子を示す．地声を利用した場合には，詐称者受理率が（素人に比べても）小さいのに対して，声真似によって詐称者受理率が大きく増加しており，成りすましに成功する確率が上昇している様子がわかる．

図 3 に，素人とプロの物真似音声を詐称者の発声として使用した場合の詐称者受理率の比較を示す．全体の傾向をみると，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くなっていることがわかる．なお，等誤り率は，素人（訓

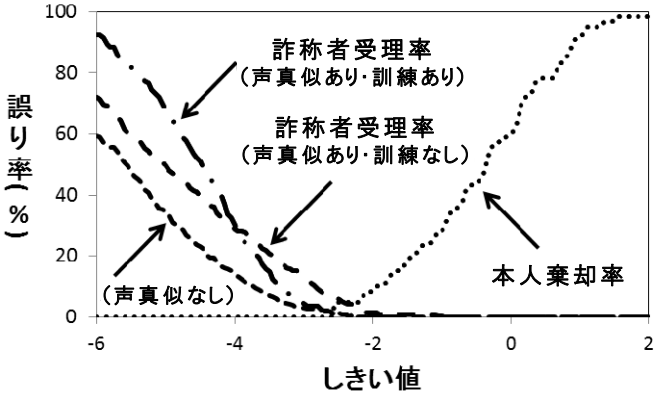

図 1　素人の声真似に対する詐称者受理率

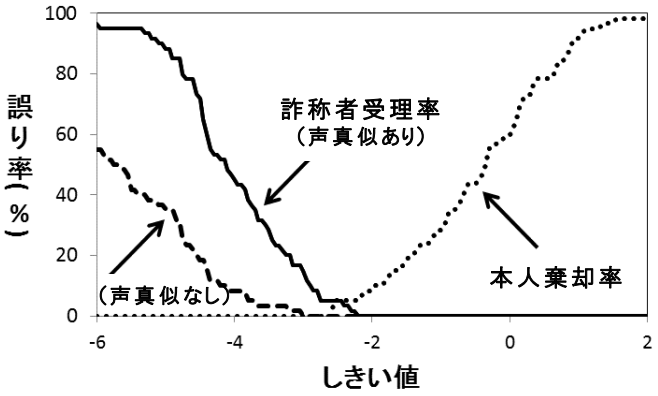

図 2　プロの声真似に対する詐称者受理率

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．

4. 発声間距離を用いた物真似音声の音響特徴の分析

4.1. ケプストラムに基づく発声間距離の定義

声真似という行為によって発声の音響特徴，特に，話者照合で重要であるケプストラム特徴がどのように変化するかについて，プロと素人の傾向の違いを明らかにする．この分析には，先行研究[3]で提案した「ケプストラムに基づく発声間距離」の概念を利用する．この手法では，以下の 3 つの発声に対して音響モデルを構築し，それぞれの発声間の距離を調べる．

- U^i：発話者 i の自然な発声（地声）
- U^j：物真似の対象者 j の自然な発声（地声）
- $U^{i\sim j}$: 発話者 i が対象者 j を真似て行った発声（物真似音声）

図 4 に分析対象となる 3 つの距離（A, B, C）を示す．A は「物真似によって生じる音響特徴量の変化（移動距離）」，B は「物真似音声と対象者の地声との間の音響的な隔たり」，C は「発声者と対象者の間の地声の音響的な隔たり」を表している．

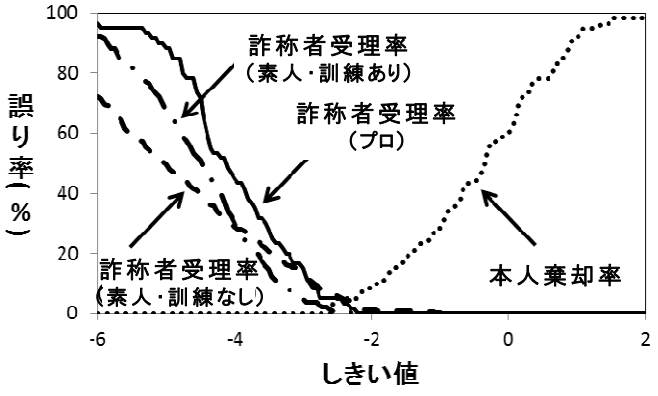

図 3　プロと素人の声真似に対する照合性能の比較

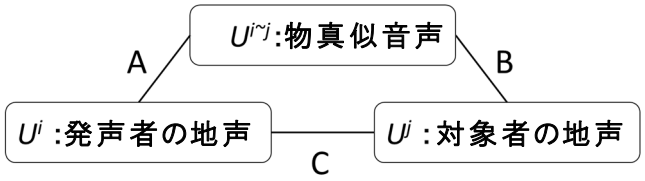

図 4　分析対象とする発声間距離（A～C）

これらの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（話者照合で利用する特徴量と同じ，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）に変換し，各発声を 3 状態の HMM（混合数 K）でモデル化する．先行研究[3]では，モデルの混合数を K=1 とし，単一正規分布間のマハラノビス距離に基づいて発声間距離を定義していたが，今回の研究では分析精度の向上を目指して，混合正規分布への対応と，カルバック・ライブラー（KL: Kullback-Leibler）情報量[8]の導入を検討する．この KL 情報量に基づいて，発声 U^a, U^b 間の発声間距離 $D(U^a, U^b)$を 2 種類提案する．

一つ目は，発声 U^a, U^b それぞれのモデルの第 2 状態の正規分布のうち，混合重みが最大となる正規分布のインデックスを k^{a_max}, k^{b_max} とし，該当するそれぞれの正規分布 $N^a_{k^{a\_max}}$, $N^b_{k^{b\_max}}$ を用いて式(2)のように発声間距離を定義する．

$$D_{KL\_max}(U^a, U^b) = SKL(N^a_{k^{a\_max}}, N^b_{k^{b\_max}}) \tag{2}$$

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

$$SKL(N^a, N^b) = KL(N^a||N^b) + KL(N^b||N^a) \tag{3}$$

$$KL(N^a||N^b) = \int_{-\infty}^{\infty} N^a(\boldsymbol{x}) \log \frac{N^a(\boldsymbol{x})}{N^b(\boldsymbol{x})} d\boldsymbol{x} \tag{4}$$

表 1　カルバック・ライブラー情報量に基づく距離定義による発声間距離の分析結果

距離	話者	A	B	C	A/C	B/C
D_{KL_max}	素人（訓練なし）	37.1	45.9	31.0	1.19	1.48
	素人（訓練あり）	24.8	25.9	31.0	0.80	0.83
	プロ	17.2	20.9	27.0	0.64	0.77
D_{KL_ave}	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	20.9	0.65	0.69

表 2　マハラノビス距離に基づく距離定義による発声間距離の分析結果

距離	話者	A	B	C	A/C	B/C
D_{MD_max}	素人（訓練なし）	3.01	3.11	3.13	0.96	0.99
	素人（訓練あり）	3.04	3.08	3.13	0.97	0.98
	プロ	1.83	2.42	2.45	0.75	0.99
D_{MD_ave}	素人（訓練なし）	1.91	2.44	2.39	0.80	1.02
	素人（訓練あり）	2.18	2.29	2.39	0.91	0.96
	プロ	1.79	2.28	2.12	0.84	1.08

二つ目は，発声 U^a, U^b それぞれのモデルの第 2 状態の正規分布に対し，正規分布間の対象化 KL 情報量が最も小さくなるペアを選びながら，その値を平均化して発声間距離を求めるものである．定義を式(5)に示す．

$$D_{KL\_ave}(U^a,U^b)=\frac{1}{2}\left\{\sum_k^K w_k^a\cdot\min_l SKL\left(N_k^a,N_l^b\right)+\sum_k^K w_k^b\cdot\min_l SKL\left(N_l^a,N_k^b\right)\right\} \quad (5)$$

従来の距離定義を利用した分析と比較するため，文献[3]で利用したマハラノビス距離に基づく距離定義も用意する．マハラノビス距離に基づく音響距離の定義は，自然発話の音響特徴の分析などにも利用され，その有効性が確認されている[9]．2 つの多次元正規分布 N^a,N^b が与えられたとき，両者の間の距離 MD は以下の式(6)で計算される．

$$MD(N^a,N^b)=\sqrt{\frac{M\sum_m\left(\mu_m^a-\mu_m^b\right)^2}{\sum_m(\sigma_m^a)^2+\sum_m\left(\sigma_m^b\right)^2}} \quad (6)$$

ここで，M はベクトルの次元数（25）であり，μ_m^x と $(\sigma_m^x)^2$ は正規分布 N^x の平均・分散ベクトルの m 次元目の要素を表している．この MD を式 (2), (5) における SKL の代わりに使用したものを，それぞれ D_{MD_max}，D_{MD_ave} と定義する．

4.2. 分析結果

表 1 に，D_{KL_max}，D_{KL_ave} の 2 つの距離を用いて,「素人（訓練なし)」「素人（訓練あり)」「プロ」の物真似音声を用いて距離の分析を行った結果を示す．素人の距離については全 6 話者の平均値であり，ここで示す分析結果は混合数 8 のモデルで得られたものである．「A/C」は「発声者と対象者の間の地声の距離（C）に対し,声真似によってどの程度音声特徴が変化したか」を示しており,「B/C」は「話者間距離（C）に対して，声真似によってどの程度対象者の音響特徴に近づいたか」を示している．この結果をみると，

- 素人は，声真似による特徴の変化量（A）が話者間距離（C）の 8 割以上の値となっており，声真似により特徴量が大きく変動している（努力によって地声とは異なる声を出そうとしている）．ただし，物真似発声と対象者の地声との距離（B）も話者間距離（C）の 8 割以上の値となっていることから，依然として対象者との間の隔たりが大きいことから，声真似が中々上手くいっていない．
- プロの物真似タレントは，声真似による特徴の変化量（A）が話者間距離（C）の 6〜7 割程度と，素人のそれよりも小さくなっており，声真似による特徴の変動自体は大きくない．一方で，物真似発声と対象者の地声との距離（B）は話者間距離（C）の 7〜8 割程度と，素人のそれに比べ小さくなっており，物真似音声と対象者の地声との差を確実に小さくしている．

という傾向が読み取れる．後者の結果は，3 章で示された話者照合システムに対する攻撃力の分析結果とも矛盾がなく，この発声間距離による分析の妥当性が伺

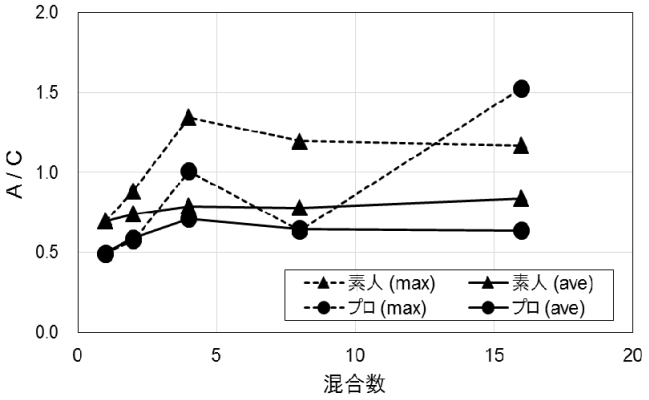

図 5　KL 情報量に基づく発声間距離で得られた物真似音声の分析結果（A/C）の混合数の変化に対する変動

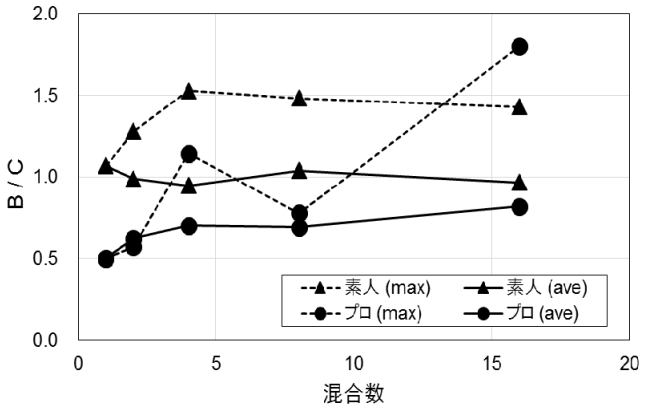

図 6　KL 情報量に基づく発声間距離で得られた物真似音声の分析結果（B/C）の混合数の変化に対する変動

える．

表 2 にマハラノビス距離に基づく距離定義を用いた場合の話者間距離の分析結果を示す．こちらの結果では,「話者間距離に対する物真似発声と対象者の地声の距離（B/C)」が素人とプロで同じ程度に見積もられているなど, 3 章の攻撃力の分析結果と一致していない．したがって，今回の分析データに対しては従来までのマハラノビス距離に基づく定義ではなく,提案する KL 情報量を用いた定義の方が有効であることがわかった．

4.3. 混合数の変化に対する検証

次に, 混合数の変化に対し, KL 情報量に基づく距離定義を用いた分析結果がどのように変化するかについて検証する．図 5, 6 はモデルの混合数を 1, 2, 4, 8, 16 と変化させたときの A/C, B/C の値の変化をそれぞれ示したものである．点線は D_{KL_max} を利用した場合，実線は D_{KL_ave} を利用した場合を示しており，プロの声真似の分析結果を丸印（●）で，素人（訓練なし）の声真似の分析結果を三角印（▲）で示している．この結果を見ると，最大の重みを有する正規分布のみを利用した D_{KL_max} では不安定な結果となっており，特に混合数が 16 の場合にプロの分析結果が 3 章の攻撃力の分析結果と大きく離れた結果となってしまっている．一方，全ての正規分布を考慮している D_{KL_ave} では混合数の変化に対して安定した結果が得られている．

以上より，D_{KL_ave} で定義される発声間距離を用いることで,実際の話者照合システムを構築することなく，話者照合に対する物真似音声の攻撃力を把握できる可能性があることがわかった．

5. まとめ

本稿では，プロの物真似タレントの声真似による攻撃が HMM に基づく話者照合システムの性能にどれだけの影響を及ぼすかの調査と，その物真似発声の音響特徴量の変化に対する分析を行った．6 名の男子大学生を登録した話者照合システムを構築し，1 名のプロの物真似タレントの物真似発声を詐称者発声として入力したところ，素人の声真似攻撃に比べて詐称者受理率が有意に増加するということが確認された．また，発声間距離に基づく物真似音声のケプストラム特徴の分析では，カルバック・ライブラー情報量に基づく距離定義を提案し，それを用いることで「素人は自分と異なった声質を出すことはできるが物真似の対象者には近づかない一方で，プロの物真似タレントは声の変化は比較的小さいが対象者には確実に近づく」という分析結果を得た．この結果は話者認識システムに対する攻撃力の分析結果と矛盾がないことから，この距離定義を利用した音響特徴の分析によって実際の話者照合システムを構築することなく，物真似音声の攻撃力を把握できる可能性も示された．

今回の実験では，使用したデータがプロ 1 名分の発声ということもあり，データ量が少ない．今度の課題としては，データ量の増加により分析結果の信頼性の向上を図る必要がある．また, 話者照合に DNN を利用するなど，GMM-UBM 法よりも高精度な話者照合システムの利用などを検討し，模倣音声の影響の調査を行う必要がある．最終的には，今回得られた分析結果を考慮することで，物真似に特別なスキルを有する詐称者が声真似攻撃を行った場合でも頑健となる話者照合手法の提案と実システムの開発が強く望まれる．

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．

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] Z. Wu, et al., “Spoofing and countermeasures for speaker verification: A survey,” Speech

Communication, vol. 66, pp.130-153, 2015.

[3] 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.

[4] R. G. Hautamäki, “I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry,” Proc. INTERSPEECH, 2013.

[5] D. A. Reynolds, et al., “Speaker verification using adapted Gaussian Mixture Models,” Digital Signal Processing, vol. 10, pp. 19-41, 2000.

[6] N. Dehak, et. al., “Front-end factor analysis for speaker verification,” ITASLP, vol. 19, no. 4, pp. 788–798, 2011.

[7] J. Mariéthoz and S. Bengio, “Can a professional imitator fool a GMM-based speaker verification system?” IDIAP Research Report, no. Idiap-RR 05-61, 2006.

[8] S. Kullback and R. A. Leibler, “On information and sufficiency,” Ann. Math. Statist., vol.22, no.1, pp.79–86, 1951.

[9] 中村ら，“話し言葉音声の音響的・言語的特徴の分析，” 電子情報通信学会技術研究報告, vol. 106, no. 78, pp. 19-24, 2006.