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
SP2014-134 (2015-1)

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

岩野　公司†‡　曽根　泰斗‡　坂本　香菜子‡

† 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
‡ 東京都市大学 環境情報学部 〒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　話者照合システムにおいて，物真似（模倣）攻撃に対する脆弱性を正しく把握し，その対応策を講じることは極めて重要である．本稿では，一般人（物真似の素人）を対象として模倣音声を収録し，それを用いて模倣行為で生じる音響特徴の変動の分析と，HMM に基づく話者照合システムの性能に与える影響の調査を行った．男女各 6 名の通常発声と模倣音声を用いて，模倣によるケプストラム特徴の変動分析を行ったところ，「模倣の努力によって本人とは異なる声が生成されているが，模倣対象者には近づかない」傾向が確認された．一方，話者照合性能に対する模倣攻撃の影響を調査した結果，全体的には模倣音声と対象者の音響特徴の間に隔たりがある状況下であっても，成りすましが成功する場合が存在し，結果として等誤り率が約 1.2〜2.7 倍に増加することがわかった．

キーワード　話者照合，物真似（模倣）音声，音声模倣攻撃，音響特徴分析

Analysis of effects of voice mimicry on speaker verification and acoustic features of the imitated voices

Koji IWANO†‡　Taito SONE‡　and　Kanako SAKAMOTO‡

† Faculty of Informatics, Tokyo City University　3-3-1 Ushikubo-nishi, Tsuzuki-ku, Yokohama, 224-8551 Japan
‡ Faculty of Environmental and Information Studies, Tokyo City University　3-3-1 Ushikubo-nishi, Tsuzuki-ku, Yokohama, 224-8551 Japan
E-mail: † iwano@tcu.ac.jp

Abstract　It is quite important to recognize the vulnerability of speaker verification (SV) systems against voice mimicry attacks and to propose countermeasure methods against them. This paper describes an acoustic feature analysis using imitated speech uttered by non-professional speakers and a performance degradation of an HMM-based SV system by the mimicry attacks using the imitated speech. We collected normal/imitated speech uttered by 6 males and 6 females, and analyzed the cepstral feature changes by the imitation. The analysis results show the tendency that the acoustic features of the speakers' voice significantly change by their efforts of the imitation, whereas the distance of the acoustic features between imitated and target speaker's speech is still large. The experimental results on the SV system performance show that the mimicry attacks increase equal error rates by 1.2 to 2.7 times. These facts indicate that non-professional mimicry attacks rarely succeed in spoofing the SV system and yield actual performance degradations in spite of the situation where there is a large gap of acoustic features between imitated and target speech.

Keywords　Speaker verification, Imitated speech, Voice mimicry attack, Acoustic feature analysis

1. はじめに

近年，音声による個人認証（話者照合）に対する期待が高まっており，様々な研究が進められ，実用製品の開発なども進んでいる[1]．セキュリティーシステムといった実用例を考えると，話者照合が晒される危険性は多様なものとなる．文献[2]では，想定される攻撃として，声真似（模倣）[3-6]，録音再生[7]，音声合成[8, 9]，声質変換[10, 11]を挙げている．これらの攻撃に対する話者照合の脆弱性を正しく把握し，その対応策を講じることは極めて重要となる．

これらのうち，本研究では，最も手軽な攻撃手段である「声真似（模倣）」に焦点をあてる．音声模倣攻撃に関するこれまでの研究では，「プロの物真似タレントによる声真似[4, 5]」や，「模倣対象者に声質が近い一般人（素人）による声真似[3, 6]」を仮定し，それによる照合性能の劣化の可能性を報告している．しかし，実際に攻撃を行う場面を考えると，「対象者の物真似ができるプロのタレントに依頼して模倣させる」ことや

「対象者の声質に近い一般人を選定し，その者に模倣させる」といった状況は特殊であると考える．

そこで我々は，「素人が，声質の近さとは無関係に決定された対象者（ただし，声は知っている者）に成りすますための模倣を行う」という，より一般的な状況を想定し，このような状況で発声された物真似音声の音響特徴の分析と，話者照合性能に与える影響について調査を行う[12]．日本語の物真似音声の分析については，文献[13]で，プロの物真似タレントの発声を対象にした分析が行われており，基本周波数（F_0）パターンやフォルマント周波数が目標対象者に近づく傾向があることが報告されている．それに対して，本研究の分析対象者は一般の素人となっている．なお，このような状況を想定して収録された日本語模倣音声データはこれまでに存在していないため，本研究では，音声データベースの構築から行っている．

以降では，まず 2 章で模倣音声データベースの構築について説明する．3 章では，収録された音声を用いて，素人の模倣行為によって生じる，音声の音響特徴（ケプストラム）の変動の様子を分析する．4 章では，今回想定した話者照合システムの説明と，評価データに模倣音声を含めたときの照合性能について調査する．最後に 5 章で本稿の結論を述べる．

2. 模倣音声データベースの構築

2.1. 音声データの収録

本学学生の男性 6 名，女性 6 名を被験者（発話者）とした音声収録を行った．これらの被験者は声質の近さを考慮せずに選ばれており，全員，物真似の素人である．異性間の模倣は現実的ではないと考え，男性 6 名で 1 グループ，女性 6 名で 1 グループを構成し，それぞれのグループ内でお互いに声を真似しあうことで模倣音声を収録する．全ての被験者同士は知り合いであり，普段からそれぞれの声を良く聞いている．

収録は約 2 日ごと，3 週間にわたり継続的に行った．一人あたり合計で 9 日収録を行っている．各話者は 1 日の収録で，

① 特に意図を持たず，本人の声として自然に行う発声（1 セッション）
② グループ内の他者（5 名）を模倣しようと努力して行った発声（5 セッション）
③ 過去に行った本人自身の発声を聴取した上で，本人として受理されようと努力して行った発声（1 セッション）

の計 7 セッションの発声を行う．各話者はそれぞれのセッションでランダムの 4 桁連続数字を 10 回発声する．初日は①のセッションのみ，2 回目以降は初日に収録された他者または本人の音声を聴いた上で②，③の収録を行う．

なお，マイクにはオーディオテクニカ社の AT-VD6 を，オーディオキャプチャにはローランド社の UA-101 を用いた．個々の発声は，前後の余分な無音区間を除去するため，音声認識エンジン Julius 付属の adintool[14]による音声区間検出（VAD）を行った上で，16kHz サンプリング・16bit 量子化で収録される．

2.2. 模倣支援機能を有する収録システムによる再収録

一般人が，自分の声質の近さとは無関係に選ばれた対象者の発声を模倣することには大きな困難が予想される．もし，発話者が模倣発声に対して照合システムで用いられる「照合スコア（申告者らしさ）」を参照することができれば，システムを欺くためにどのように声を変化させればよいかを把握することができ，模倣のスキルが向上する可能性がある．

そこで，照合スコアを発声ごとに参照することができる，「模倣支援の機能を有する音声収録システム」を構築し，それによって模倣発声の収録を行った．被験者は 2.1 節の男性 6 名であり，初回収録から数ヵ月後に本システムを用いて再収録を行った．再収録では，各日の収録について，2.1 節の①～③に続けて，

④ 模倣対象者に対する照合スコアができるだけ大きくなるような努力をして収録した，グループ内の他者（5 名）を模倣した発声（5 セッション）
⑤ 本人自身に対する照合スコアができるだけ大きくなるように努力して行った発声（1 セッション）

を行っている．なお，マイク，オーディオキャプチャは初回収録時と同じものを利用している．

2.3. 模倣音声の主観評価

収録データの模倣音声に対し，「模倣の上手さ」を示すスコアを主観評価によって付与した．このスコアリングは，収録した 3 週間分の音声のうちの 4 日間分に対して実施し，収録被験者と異なる 5 名の被験者の主観評価によって行った．各評価者は，模倣対象者の音声と模倣音声の両者の音声を聴いた後，7 段階 (1: 全く似ていない～7: 非常によく似ている) で評価を行い，その 5 名の平均値を各回のスコアとした．なお，男性 6 名の再収録分の主観評価については，5 名の評価者のうち，2 名は初回収録分の評価者と同じであるが，3 名は異なっている．なお，全ての評価者は，発話者（6 名）全員と知り合いであり，普段からその声をよく耳にしている．

評価の結果，初回収録時の模倣音声における平均スコアは男性で 1.81（標準偏差 0.35），女性で 3.41（標準偏差 0.39）となった．スコアが低いことから全体的

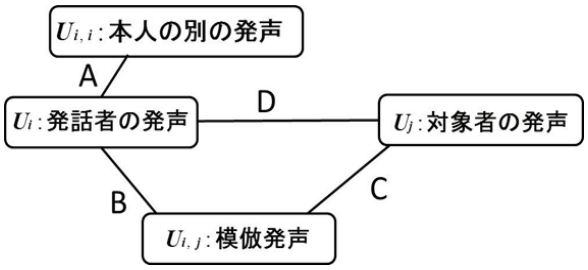

図 1　分析対象とする発声間距離

Fig. 1: Analysis target: distances among four kinds of utterances by imitators and target speakers.

に模倣はうまくいっておらず，一般の素人による模倣の難しさが読み取れる．また，再収録の模倣音声については，模倣支援使用時の収集音声のみに対してスコアリングを行い,その平均スコアは 1.43(標準偏差 0.25)であった．したがって，模倣支援を利用した状況であってもスコアは低く，人間が知覚できるような，はっきりとした模倣の上達は見られなかった．

3. 物真似音声の音響特徴分析

3.1. 発声間の距離

本研究では，以下の 4 つの発声に対して特徴量を抽出し，そこから「発声間の距離」を求めて物真似（模倣）の分析を行う．まず，それぞれの発声を以下のように定義する．

- U_i：①で収録された，発話者 i の自然な発声
- U_j：①で収録された，模倣対象者 j の自然な発声
- $U_{i,j}$: ②（あるいは④）で収録された，発話者 i が対象者 j を模倣しようと努力して行った発声
- $U_{i,i}$: ③（あるいは⑤）で収録された，発話者 i が本人として受理されようと努力して行った発声

図 1 に発声空間の模式図を示し，今回分析対象とした 4 つの発声間距離（A～D）を表す．

3.2. ケプストラムに基づく発声間距離の定義

文献[15]では，話し言葉音声と読み上げ音声のケプストラム特徴の分析のために，音素間のマハラノビス距離を利用している．本研究でも同様に，各発声間についてケプストラムに基づくマハラノビス距離を算出し，分析に利用する．

発声間距離の算出方法は以下の通りである．まず，分析対象音声をフレームごとに 12 次元 MFCC，その 1 次微分成分,対数パワーの 1 次微分成分の計 25 次元の特徴ベクトルに変換する．次に，各発声を 3 状態の HMM（混合数 1）でモデル化し，2 状態目のモデルパラメータ（25 次元の平均・分散ベクトル）をその発声の音響特徴として取り出す.発声 U_x, U_y 間の距離 $D(U_x, U_y)$ を式(1)のように計算する[14]．K はベクトルの次元数（25）であり，μ_{xk} と σ^2_{xk} は，発声 U_x の平均，分散ベクトルの k 次元目の要素である．

表 1　グループごとの発声間のケプストラム距離（A~D）の平均値（括弧内は標準偏差）

Table 1: Means of cepstral distances A~D for each speaker group (standard deviations in parentheses).

	A	B	C	D
女性①	0.43 (0.20)	0.53 (0.18)	0.73 (0.21)	0.76 (0.20)
男性①	0.88 (0.65)	0.77 (0.49)	1.56 (0.63)	1.51 (0.50)
男性②（模倣支援なし）	0.89 (0.49)	1.35 (1.46)	2.06 (0.95)	2.04 (0.78)
男性②（模倣支援あり）	1.26 (0.43)	1.34 (0.46)	1.43 (0.53)	2.04 (0.78)

$$D(U_x,U_y)=\sqrt{\frac{K\sum_{k=1}^{K}(\mu_{xk}-\mu_{yk})^2}{\sum_{k=1}^{K}\sigma_{xk}^2+\sum_{k=1}^{K}\sigma_{yk}^2}} \quad (1)$$

3.3. 発声間距離による模倣音声の分析結果

分析対象データには主観評価によるスコアリングを行った 4 日間分の音声を用いた．得られたグループごとの発声距離（A~D）の平均値を表 1 に示す．初回収録の男性・女性音声に対する結果を「男性①」「女性①」,再収録時に模倣支模を使わずに収録された音声の分析結果を「男性②（模倣支援なし)」，支援を使って収録された音声の分析結果を「男性②(模倣支援あり)」としている．なお,「男性①」「女性①」の分析結果が先行発表[12]の数値と若干異なっているのは，今回の分析対象音声には新たに VAD による冒頭・末尾の無音区間の除去を行っていることに起因する．

表中の話者間の距離（D）を見ると，男性の方が女性よりも値が大きく，話者間の声質が離れている（個性が大きい）傾向が見られる．今回の実験で，男性同士の模倣が難しく，女性よりも主観評価値が低くなったのは，このことが一因であると考えられる．全体的な傾向としては，

- 模倣によるケプストラムの移動距離（B）が，本人内の変動距離(A)よりも大きいことから，「模倣による努力によって本人とは異なる声を出そうとしている」こと
- 模倣による移動距離（B）は，話者間距離（D）に比べるとまだ小さく，また，模倣発声と対象者発声の間の距離（C）も依然として本来の話

者間距離（D）程度に大きいことから，「模倣対象者にうまく近づけていない」こと

が読み取れる．また，模倣支援を使用した場合としなかった場合を比較すると，模倣による移動距離（B）には差が見られないが，模倣発声と対象者発声の間の距離（C）が，支援の使用によって小さくなっており，対象者にやや近づいている傾向が読み取れる．

表 2, 3 に，話者ごとの「模倣による移動距離（B)」を「本人内の変動距離（A)」と「話者間距離（D)」で割った比を示す．女性グループ（女性①）の結果を表 2 に，男性グループ（男性①②）の結果を表 3 に示す．男性では話者 M05，女性では F01, F04 などは，本人内の変動に比べ模倣による特徴変動が比較的大きく，その距離は話者間距離の 7 割程度に達していることがわかる．これらの話者は，自分の声質を大きく変動させる能力を持っていると考えられ，(対象者に高確率で近づけないまでも）模倣による成りすましが成功する可能性を秘めていると考えられる．

4. 話者照合性能への影響の調査

4.1. 使用する話者照合システム

収録された模倣音声が話者照合性能に及ぼす影響を評価するため，本研究では，HMM でモデル化された申告者モデルと不特定話者モデル（UBM）を利用する話者照合を利用する．この手法では，各話者を（3 章における分析と同様に）3 状態の HMM でモデル化しているが，基本的には GMM-UBM 法[16]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，入力音声はフレームごとに 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される．この特徴量も 3 章の分析で用いたものと同じである．得られた特徴量系列 X を申告者モデル（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 $S(X)$ は尤度比（対数尤度の差）で定義される．

$$S(X) = \log P(X \mid C) - \log P(X \mid U) \qquad (2)$$

この照合スコアが設定したしきい値θよりも大きければ申告者として受理し，小さければ詐称者とみなされ棄却する．2.2 節で解説した，模倣支援時に発話者に提示するスコアには，この $S(X)$を用いている．

今回の実験では，3 週間のうちの前半 2 週間（6 日分）で収録されたデータを学習に，後半 1 週間（3 日分）で収録されたデータを評価に用いる．申告者モデルの学習には，2.1 節の①で発声された音声を使用する．したがって，一つの申告者モデルは 60 個の 4 桁連続数字発声で学習される．UBM はグループ内の全話者の発声データで学習する．この学習にも，2.1 節の①で発声された音声を用いており，UBM は 360 個の 4 桁連続数字発声で学習される．

表 2　話者ごとの模倣によるケプストラム距離変動の分析結果（女性）

Table 2: Analysis results of cepstral distance changes caused by imitations (Female).

ID	女性①	
	B/A	B/D
F01	2.30	0.65
F02	1.30	0.58
F03	0.82	0.71
F04	3.51	0.65
F05	0.89	0.80
F06	1.08	0.84

表 3　話者ごとの模倣によるケプストラム距離変動の分析結果（男性）

Table 3: Analysis results of cepstral distance changes caused by imitations (Male).

ID	男性①		男性②（模倣支援なし）		男性②（模倣支援あり）	
	B/A	B/D	B/A	B/D	B/A	B/D
M01	2.57	0.45	0.57	0.44	0.96	0.63
M02	0.68	0.41	1.80	0.52	1.07	0.63
M03	1.23	0.76	1.53	0.39	1.25	0.65
M04	1.01	0.54	1.11	0.37	0.96	0.52
M05	2.09	0.50	2.82	1.40	1.07	0.83
M06	0.31	0.46	1.60	0.54	1.16	0.62

模倣発声を含まない評価（「模倣なし」）を行う場合には，評価データとして 2.1 節の③で発声された音声のみを利用する．このとき，申告者以外の全ての話者の音声が詐称者データとして用いられ，詐称者受理率の算出に用いられる．模倣発声を含む評価(「模倣あり」)行う場合には，2.1 節の③で発声された音声を申告者受理の評価に利用し，詐称者データには，②（もしくは④）で発声された「他 5 名による対象申告者への模倣音声」を使用する．

4.2. 実験結果

まず，申告者モデルと不特定話者モデル（UBM）の混合数混合数を増やしながら，それぞれのグループにおける「模倣あり」と「模倣なし」の場合の照合性能の検証を行う．

図 2 に，モデルの混合数を変化させたときの，「女性①」「男性①」「男性②」それぞれのグループに対する「模倣あり」「模倣なし」の場合の等誤り率（Equal

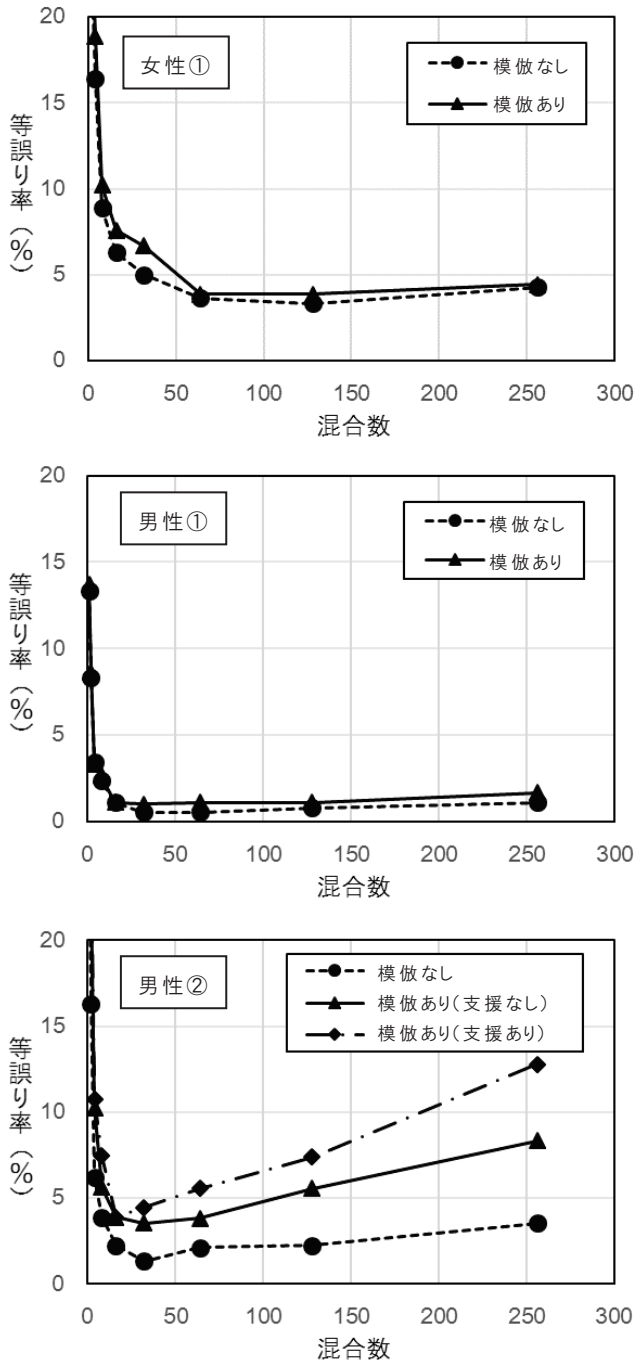

図 2: 模倣が等誤り率に与える影響の検証
（上：女性①，中：男性①，下：男性②）

Fig. 2: Effect of voice imitations on equal error rates. Top: Female (first recording session), Middle: Male (first recording session), Bottom: Male (second recording session).

Error Rate: EER）を示す. なお, 混合数は申告者と UBM で共通とし，1~256 まで変化させている．これらの図より，全てのグループにおいて，模倣により照合性能の劣化が見られることがわかる.「模倣なし」で最も性能が良好であった混合数における，模倣による等誤り率の上昇を見ると，女性①で 3.33%→3.89%，男性①で 0.56%→1.02%，男性②で 1.33%→3.54%（模倣支援なし）となり，誤り率が 1.2～2.7 倍に増加していることが分かる．したがって，今回のような「声質の近さを考慮せずに構成された話者グループ」の内の模倣であっても，成りすましが成功している状況が発生し，それによる照合性能の劣化が確認された．

「男性②」の結果からは，再収録時に利用した模倣支援の効果を読み取ることができる．照合スコアによる支援を行った模倣の方が，支援を行わない場合に比べて照合性能が劣化している様子がわかる．また,「男性①」の結果と比べると，支援を行っていない場合であっても模倣による照合性能の劣化が顕著になっている．この要因の一つには，照合スコアを用いた日々の訓練によって発声者の模倣のスキルが上がり，成りすましの成功率が上がった可能性が考えられる．

図 3 に,「女性①」「男性①」「男性②」それぞれのグループにおける，模倣による詐称者受理率の変化の様子を示す．モデルの混合数は，等誤り率が最良となるものを選択した(女性 128, 男性 32)．この図からも，全てのグループにおいて，模倣による詐称者受理率の上昇が読み取れる.「男性②」の結果からは，図 2 の結果と同様に模倣支援を行うことによる更なる詐称者受理率の上昇が読み取れる．

5. まとめ

本稿では,「一般の素人が，その者の声質とは無関係に決定された対象者に成りすますために物真似（模倣）を行う」という状況を想定し，その模倣による音響特徴の変動の分析と，話者照合性能に及ぼす影響の調査を行った．各話者の発声間のケプストラム距離による分析からは,「模倣による努力によって本人とは異なる声を出そうとしているが，模倣対象者には中々近づかない」ことが示され，素人による模倣の難しさが確認された．一方，模倣による特徴変動が大きい話者も存在することもわかった．模倣が話者照合性能に与える影響を検証したところ,（上述のように）必ずしも模倣音声が対象者の特徴に近づいていない状況であっても，成りすましが成功してしまっているケースが存在し，結果として等誤り率が 1.2～2.7 倍に増加し，照合性能に悪影響を及ぼすことが確認された．また，照合スコアを参考に「模倣の訓練・支援」を行いながら収録した模倣音声については，支援を受けずに発声した模倣音声よりも照合性能の劣化に大きな影響を及ぼすこともわかった．

今後の課題として，話者や試行ごとの詳細な分析を行うことや, UBM の強化によるベースライン性能向上時の模倣の影響評価, i-vector に基づく話者照合システムに対する模倣の影響評価，また，このような模倣攻

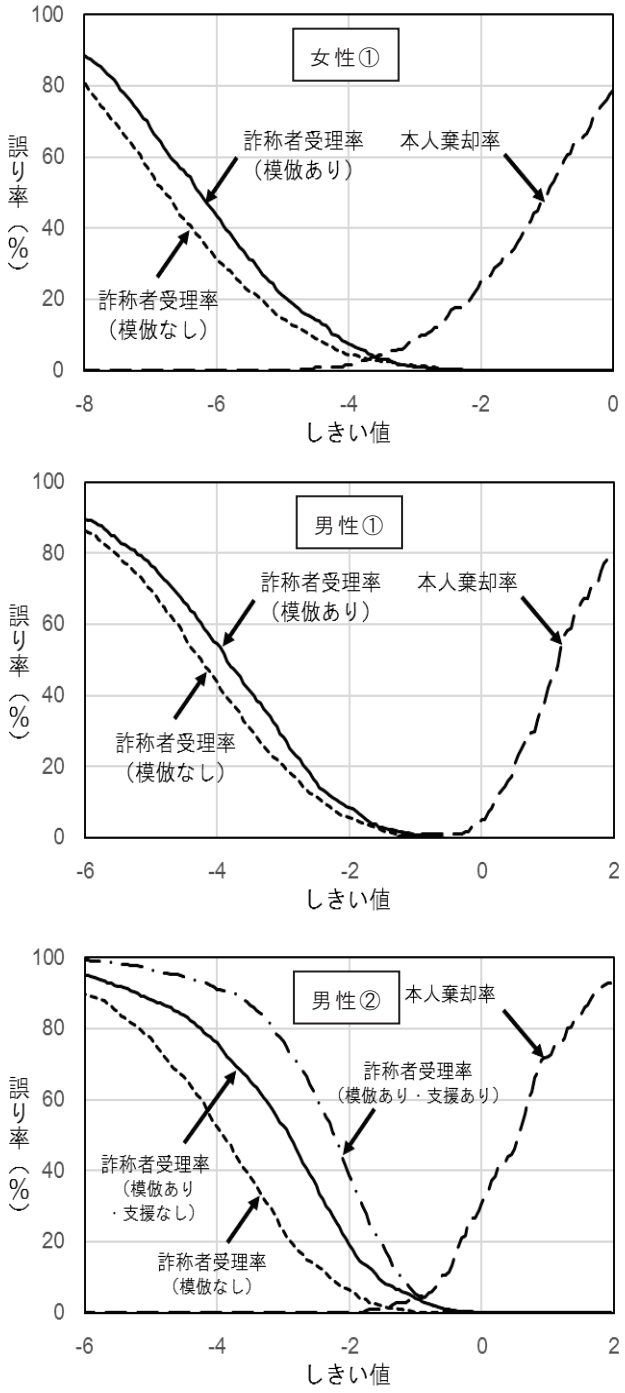

図 3: 模倣による詐称者受理率の変化の様子
（上：女性①，中：男性①，下：男性②）

Fig 3: The change of false acceptance rates by voice imitations. Top: Female (first recording session), Middle: Male (first recording session), Bottom: Male (second recording session).

撃への対策手法の提案などがあげられる．

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] N. Evans, et al., “Spoofing and countermeasures for automatic speaker verification,” Proc. INTERSPEECH, pp. 925-929, 2013.

[3] Y. W. Lau, et al., “Vulnerability of speaker verification to voice mimicking,” Proc. International Symposium on Intelligent Multimedia, Video and Speech Processing, pp. 145-148, 2004.

[4] J. Mariéthoz and S. Bengio, “Can a professional imitator fool a GMM-based speaker verification system?” IDIAP Research Report, no. Idiap-RR-61-2005, 2006.

[5] R. G. Hautamäki, et al., “I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry” Proc. INTERSPEECH, pp. 930-934, 2013.

[6] S. Panjwani and A. Prakash, “Crowdsourcing attacks on biometric systems,” Proc. Synposium on Usable Privacy and Security, pp. 257-269, 2014.

[7] J. Lindberg and M. Blomberg, “Vulnerability in speaker verification - A study of technical impostor techniques,” Proc. EUROSPEECH, vol. 3, pp. 1211-1214, 1999.

[8] T. Masuko, et al, “On the security of HMM-based speaker verification systems against imposture using synthetics speeech,” Proc. EUROSPEECH, vol. 3, pp. 1223-1226, 1999.

[9] P. L. De Leon, et al., “Evaluation of the vulnerability of speaker verification to synthetic speech,” Proc. Odyssey, pp. 151-158, 2012.

[10] T. Kinnunen, et al., “Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech,” Proc. ICASSP, pp. 4401-4404, 2012.

[11] E. Khoury, et al., “Introducing I-vector for joint anti-spoofing and speaker verification,” Proc. INTERSPEECH, pp. 61-65, 2014.

[12] 坂本，岩野，“話者照合への影響を考慮した模倣音声の音響分析，” 情報処理学会第 76 回全国大会講演論文集，no. 2, pp. 473-474, 2013.

[13] 北村，“物真似タレントによる物真似音声の分析，” 電子情報通信学会技術研究報告, vol. 107, no. 282, pp. 49-54, 2007.

[14] http://julius.sourceforge.jp/

[15] 中村他，“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.

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