日本音響学会誌 79 巻 1 号（2023），pp. 18–25

技 術 報 告

x-vector を用いた日本語電話音声に対するテキスト独立型話者照合システムの検討*

多 谷 邦 彦[*1]　サクティ サクリアニ[*2,*3]　藤 原 修 治[*1]　中 村　哲[*2]

［要旨］ 本論文では，電話を通して録音された日本語発話音声を用いたテキスト独立型話者照合実験の結果を報告する。法科学において，電話を通じて録音された音声による話者照合技術は有効なものであり，効果的に活用するためには，電話録音の影響，雑音による影響，年齢や性別等の話者特性，更に，近年の生活環境の変化により身近なものとなっているマスクの影響を分析することが重要である。近年，DNN を用いた話者照合手法が報告されていることから，この技術を用いた話者照合実験を行い，録音条件や話者特性が照合結果に及ぼす影響を分析した。電話録音した 113 人の音声の照合実験では EER=0.28%であった。また，テスト音声に付加する雑音が SNR=15 dB 以上であれば EER=2%以下，発話時間が 5 秒以上であれば EER=1.5%以下であった。更に，マスク着用や年齢及び性別の話者特性は話者照合に影響を与えないことが分かった。

キーワード　話者照合，x-vector，テキスト独立型，電話，マスク
Speaker verification, x-vector, Text independent, Telephon, Face mask

1. はじめに

警察庁が公開する資料 [1] によると，2020 年の特殊詐欺の認知件数は 13,550 件にのぼり，その多くは犯行に電話が使用されている。また，認知件数の数倍の予兆電話（個人情報などを探る電話）が架電されている。被害を未然に防ぐ対策を強化すると共に，電話機に設置する自動通話録音機などで得られた犯人の音声を手がかりに捜査が進められる。法科学分野において，電話を通じて録音された音声による話者照合技術は，事件音声と被疑者を結びつける有効な手段であり，人々の安心と安全を守るために重要で欠かせないものである [2]。

従来より音声による話者照合研究は行われており，近年では DNN（Deep Neural Network）を用いたアプローチにより高い照合性能を発揮することが明らかになっている。音声認識ツールキット Kaldi [3] が公開されており，多くの研究者が利用し，Kaldi を使用した話者照合実験についても研究結果が報告されている [4, 5]。また，研究に利用可能な英語話者の大規模音声データベースが公開されており，数千人規模の発話データを利用することができる [6]。我が国においても，ATR 音声言語データベース [7]，話し言葉コーパス CSJ [8] や JVS corpus [9] など，研究用音声データセットが数多く存在している。

一方で，新型コロナウイルス（COVID-19）の世界的大流行により私達の生活様式が大きく変化している。屋内外を問わず日常的にマスクを着用して生活し，会話をすることが多くなっている。話者照合研究を行う上では，マスク着用の有無が発話音声に及ぼす影響を考慮することが重要である。

本研究では，最新技術や有用なデータセットを利用し，特殊詐欺事件の特徴や近年の生活習慣の変化を考慮した上で，我が国における法科学的利用を念頭においたテキスト独立型話者照合技術について精度評価を詳細に行い，その有用性を検証したので，結果を報告する。

2. 関 連 研 究

法科学における話者認識を目的とする研究は科学警察研究所などの機関で行われている。電話を通した音声や雑音環境下の音声を対象とすることが特徴であり，

---

* Examination of text-independent speaker identification system for Japanese telephone speech using x-vector,
by Kunihiko Taya, Sakriani Sakti, Shuji Fujihara and Satoshi Nakamura.

*1 京都府警察本部刑事部科学捜査研究所
*2 奈良先端科学技術大学院大学先端科学技術研究科情報科学領域
*3 北陸先端科学技術大学院大学先端科学技術研究科情報科学系人間情報学研究領域
（問合先：多谷邦彦　〒602–8556 京都市上京区下長者町通新町西入藪之内町 85 番地 3 京都府警察本部刑事部科学捜査研究所）
（2022 年 7 月 11 日受付，2022 年 9 月 2 日採録決定）
[doi:10.20697/jasj.79.1_18]

照合精度を上げる研究が行われている [10–12]。

また，雑音環境下での音声認識では，中村や北岡らが日本語データセットを作成し精度評価を行っている [13, 14]。近年の携帯電話では通信速度確保のために帯域制限がかかり音質が低下していることから，中西らは帯域拡張を行った音声の話者照合評価を報告している [15]。他にも，塩田らはポップノイズに基づくアルゴリズムを spoofing 対策に用いたもの，宋らは短時間音声での精度向上について，話者照合研究を報告している [16, 17]。いずれの研究も日本語話者を対象とした音声認識や話者照合の研究が報告されている。

外国語音声を用いた話者照合研究も数多くされており，電話音声を用いた話者照合研究では [18–20]，雑音環境下の話者照合では [21, 22]，更には，法科学的な目的や近年の COVID-19 の世界的流行を踏まえ，ヘルメットやマスクの影響に注目した話者照合研究が報告されている [23, 24]。

我々が最も関心を持っていることは，話者照合において高い性能を発揮している x-vector を使用し，空調音や雑音がある一般的な室内で発せられた「日本語」が「電話」を通して録音された音声を対象として話者照合することであり，更に，近年の生活習慣を考慮し，話者のマスク着用が照合精度に影響を与えるかどうかを検証することである。しかしながら，このような研究報告は現時点ではなされていない。そこで，我々は日本語話者を対象とした録音実験を行い，その照合精度について詳細な評価を行った。第 3 章で実験方法について説明し，第 4，5 章で実験結果について報告する。

3. 実 験 方 法

日本人 113 人を対象に電話を通した音声の録音を行い，学習用音声データセットを使用して x-vector を用いた話者照合実験を行った。

3.1 x-vector を用いた話者照合システム

3.1.1 x-vector [4, 5]

話者表現を抽出するために**表–1** に示す DNN モデルを構築する。ネットワークの埋め込み層から得られるベクトルを x-vector と呼び，ここに話者情報が埋め込まれているとみなすものである。学習用音声から抽出した 30 次元の MFCC 特徴量に特徴量正規化（Cepsturm Mean and Variance Normalization, CMVN）を行ったものが DNN への入力となる。ただし，発話時間が短いものは除外される。本実験では，TDNN6 から抽出した 512 次元のベクトルを x-vector とする。ここで，T は DNN に入力するフレーム数であり，N は学習に用いた話者数である。

表–1 DNN の構成

Layer	Layer context	Input size	Output size
TDNN1	$[t-2, t+2]$	150	512
TDNN2	$\{t-2, t, t+2\}$	1,536	512
TDNN3	$\{t-3, t, t+3\}$	1,536	512
TDNN4	$\{t\}$	512	512
TDNN5	$\{t\}$	512	1,500
stats pooling	$[0, T)$	$1{,}500T$	3,000
TDNN6	$\{0\}$	3,000	512
TDNN7	$\{0\}$	512	512
softmax	$\{0\}$	512	N

3.1.2 PLDA [25]

話者照合においては，二つの音声データから抽出された x-vector の類似性を測り，確率モデル PLDA（Probabilistic linear discriminant analysis）を用いて照合結果を判断する。二つの音声が同一の話者モデルから生成される尤度 P_{s} と，異なる話者モデルから生成される尤度 P_{d} の対数尤度比

$$Score = \log \frac{P_{\mathrm{s}}}{P_{\mathrm{d}}}$$

を照合スコアとする。照合スコアが閾値よりも高い場合は同一人と判断する。DNN から抽出された 512 次元の x-vector は事前に LDA で 200 次元に縮約され，この 200 次元のベクトルを PLDA による照合に用いる。

3.2 評価用音声

日本法科学技術学会ヒト対象医学的研究等倫理審査委員会の承認（承認番号 R3M2）を得た上で，評価用音声として，20 代から 60 代までの日本人 113 人から電話を通した音声を録音した。性別は男性 73 人，女性 40 人である。また，年齢は 20 代 9 人，30 代 46 人，40 代 36 人，50 代 19 人，60 代 3 人である。録音場所は会議室内であり，話者は固定電話（アナログ）の受話器を手に持ち，原稿に書かれた内容を発話する（**図–1**）。原稿は 1 文当たり 10 語程度の 100 文で構成されており，特殊詐欺等を想定した内容である。話者はこれらを順次読み上げていくが，前半の 50 文と後半の 50 文で話者のマスク着用状況は変わり，前半にマスクを付けていた話者は後半はマスクを外し，逆に，前半にマスクを外していた話者は，後半はマスクを付けて読み上げる。話者の架電先の電話機に IC レコーダ（型番：PCM-A10）を有線接続し，話者の音声を録音する。会議室には什器類がなく声が反響し易かったため，反響を軽減させるために段ボール箱や緩衝材を設置した。

3.3 学習用音声データセット

学習用音声データには CSJ（1,417 人）及び JVS（90 人）を使用した。電話録音する評価用音声のサンプリング周波数が 8 kHz であることから，学習用音声デー

図–1　評価用音声の録音室

図–2　学習用音声データセットの録音方法

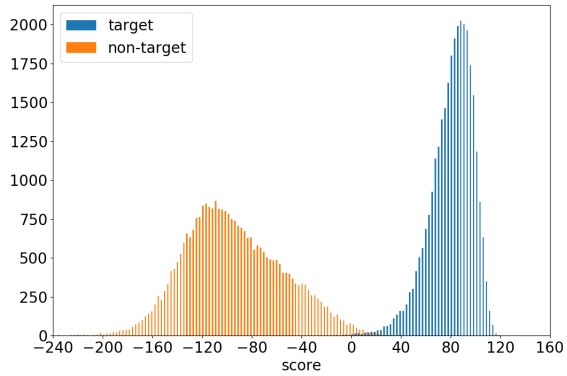
タをすべて 8 kHz にダウンサンプリングした。また，CSJ 及び JVS の音声データをスピーカ（型番：MSP5）で再生し，電話を通じて評価用音声と同様に録音したデータも利用した（図–2）。従って，一つの音声データについて，8 kHz にダウンサンプリングしたものと，電話経由で録音したものの 2 種類の音声データを学習に使用することとなる。

4. 評価用音声の照合結果

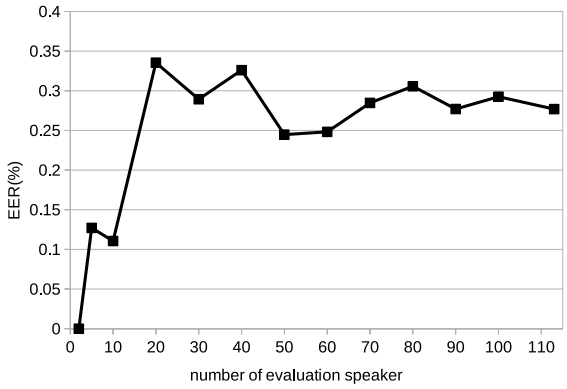
評価用音声のテスト条件を変えて，照合精度の変化を検証した。

4.1 評価用音声 113 人の照合結果

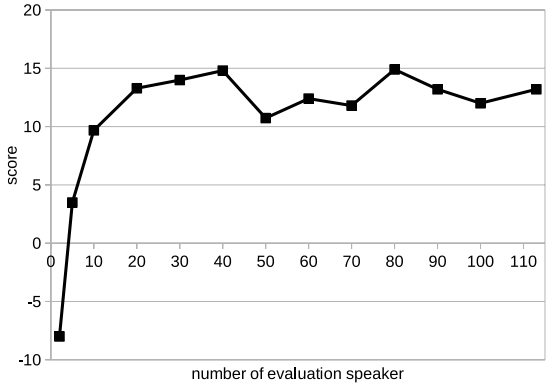
録音した音声から無音区間を削除し，1 発話データが 20 秒間となるようにトリミングしたものを照合実験に使用する。113 人から得られた発話データ数は 2,675 であり，同一人音声のテスト回数は 113 人で計 30,701 回となった。別人音声のテスト回数は 3,545,774 回であり，無作為に 30,701 回を抽出した。話者照合の評価尺度には，本人拒否率（FRR）と他人受入率（FAR）が一致する点である等価エラー率（EER）を用いた。テスト結果のヒストグラムを図–3 に示す。右側の山が同一人（target）の照合であり，左側の山が別人（non-target）の照合である。概ね同一人は正の範囲に，別人は負の範囲に分布していて EER=0.28%であり，閾値は score=13.2 であった。



図–3　評価用音声 113 人の照合スコアのヒストグラム



図–4　評価用音声の話者数の変化による EER の変動



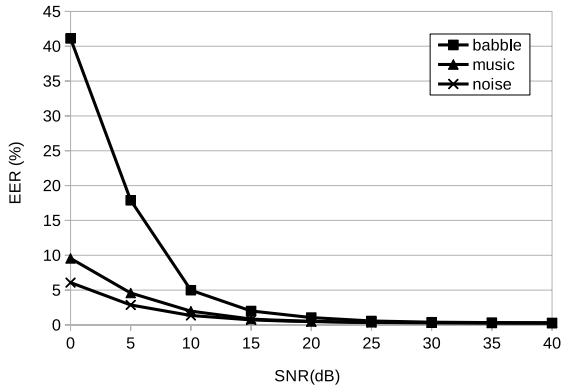
図–5　評価用音声の話者数の変化による閾値スコアの変動

4.2 評価用音声の話者数の妥当性

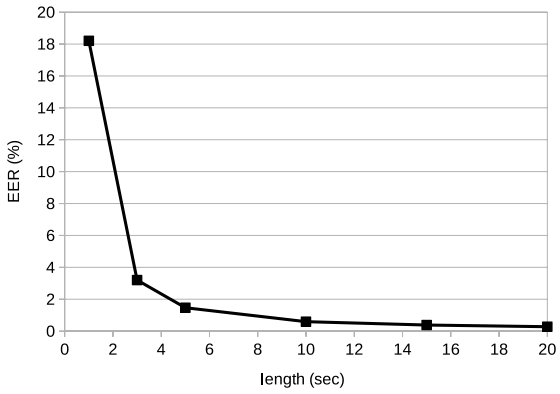
本システムの評価用音声の話者数 113 人が妥当であるかを検証する。テストで照合する話者数を 2 人から徐々に増やしていき，EER 及びそのときの閾値となる score の変動をみた。話者数が増加するにつれて EER も増加しているが，80 人以降は 0.30 付近で安定し，閾値となる score も 13 付近で安定している（図–4，図–5）。評価用音声の話者数は 113 人であるため，本システムの評価が適切に行えていると判断できる。

4.3 雑音の影響

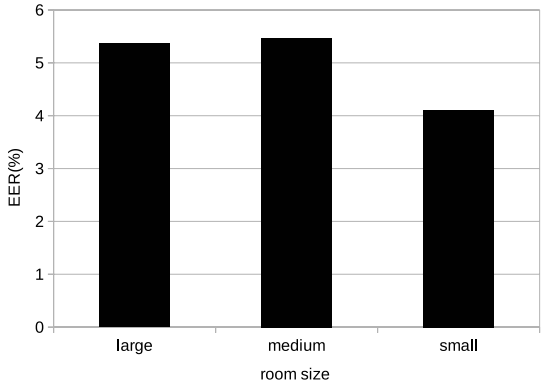
評価用音声に SNR=0 dB～40 dB の範囲で雑音を付加し，EER の変動をみる。加える雑音の種類は，bab-



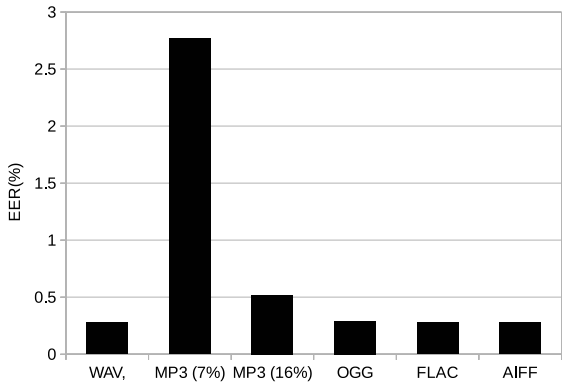
図–6 雑音（babble，music，noise）の付加程度の変化による EER の変動



図–8 テストに用いる 1 発話当たりの時間による EER の変動



図–7 残響（reverberation）による EER の変動



図–9 音声データのファイル形式別の EER

ble（多言語の会話），music（音楽），noise（ノイズ）の 3 項目であり，MUSAN データセットを利用した [26]。また，RIR データセットを利用して残響（reverberation）をシミュレーションした [27]。部屋の大きさは，large，medium，small の 3 種類である。テスト回数は同一人も別人もともに 30,701 回ずつであり，組み合わせは 4.1 節と同じである。各テストの EER を図–6，図–7，表–A.1，表–A.2 に示す。SNR=15 dB 以上であれば EER=2%以下に抑えられているが，雑音が大きいほど特に babble では照合精度の低下が顕著にみられた。残響については，3 種類で照合精度に大きな差はみられなかった。

4.4 発話時間の影響

テストに使用する評価用音声は無音区間を削除してから 20 秒間にトリミングしているが，本項では，このトリミング区間を短くしたときの EER の変動をみる。トリミング開始点は同じであり，終了点を短くしていくため，テストに使用するデータ数は変わらない。テストに用いる 1 発話の時間が 5 秒以下になると，照合精度の低下が顕著になった（図–8，表–A.3）。

4.5 圧縮の影響

評価用音声のテストには非圧縮音声を使用していたが，本項では，これらのファイル形式を変換したときの EER を比較する。ファイル形式変換により，データサイズが最大 7%まで圧縮されるが，ファイルサイズが 10%以下になるような極端な圧縮でない限り，話者照合への影響はほとんどみられなかった（図–9）。なお，各ファイル形式に変換したときのファイルサイズを WAV 形式のファイルサイズで除したものを圧縮率として表–A.4 に示す。

4.6 話者の録音時期差の影響

20 代から 50 代の 20 人を対象として，7 か月後に同じ実験環境で録音を行い，時期差がある音声の照合精度の評価を行った。時期差なし音声は，20 人の 1 回目と 2 回目の録音音声を使用するが，照合は同一時期の録音音声同士である。また，時期差ありでは，1 回目と 2 回目の音声を照合する。時期差なしでは EER=0.11%であったが，時期差ありでは EER=2.06%であった。この 20 人のうち 1 人の照合精度が特に低く，これを除いた 19 人では EER=0.43%であった。時期差のある音声の照合においては精度が低下し，また，個人間で低下の程度にばらつきがあることを考慮しなければならない。

5. 話者の特性による照合精度の評価

話者の特性が照合精度に及ぼす影響を検証するため，

表–2 マスク着用における照合テストの分析結果

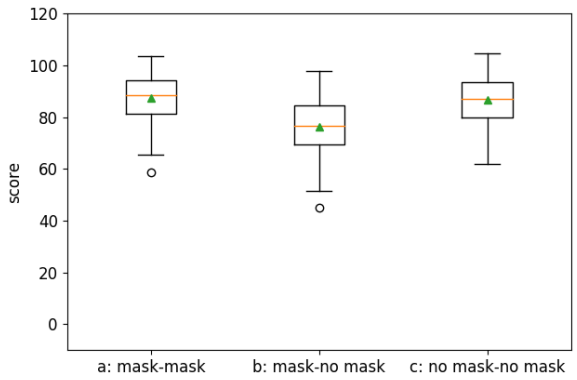
	マスク着用の有無	データ数	標準偏差	等分散検定	分散分析	多重比較	効果量
同一人	a: mask-mask	113	8.9	等分散と仮定可	一元配置分散分析	Tukey-Kramer 法	Cohen's d
	b: mask-no mask	113	10.8	Bartlett 検定	有意差あり	a-b: 有意差あり	a-b: 1.12
	c: no mask-no mask	113	9.6	($p = 0.10$)		b-c: 有意差あり	b-c: 1.00
別人	a: mask-mask	6,328	37.1	等分散と仮定不可	Kruskal-Wallis 検定	Steel Dwass 法	Cliff's delta
	b: mask-no mask	6,328	36.0	Bartlett 検定	有意差あり	a-b: 有意差あり	a-b: 0.062
	c: no mask-no mask	6,328	36.6	($p = 0.048$)		a-c: 有意差あり	a-c: 0.075

4.1 節の実験結果について統計分析ソフト R [28] を用いて統計検定を行った [29–32]。分析する特性は，マスク着用の有無（有り：mask，無し：no mask），性別（男：M，女：F），年齢（23～39 歳：グループ A（55 人），40～64 歳：グループ B（58 人））である。はじめに正規性及び等分散性の検定を行い，次に分散分析及び多重比較を行い，最後に効果量（Cohen's d もしくは Cliff's delta）を算出した。いずれの検定においても，有意水準を 5%とした。一人の話者に対して発話データが複数あるため，同一話者組み合わせで複数回の照合テストを行うが，統計検定には平均値を用いる。例えば，話者 S（性別 M，年齢 A）が 21 発話あり，うち mask が 10 発話，no mask が 11 発話とすると，話者 S の同一人の照合テストの回数は mask 同士で 45 回，no mask 同士で 55 回，mask と no mask で 110 回行うことになるが，話者 S のマスク有無に関する変数はそれぞれの平均値の三つだけである。別人の照合テストにおいても同様である。

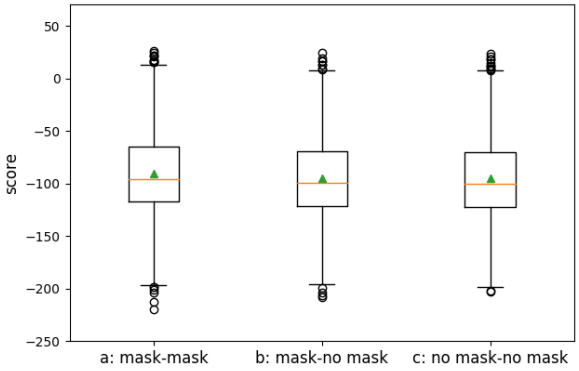
5.1 マスクの着用の影響

マスク着用の有無が照合に影響を与えるかどうかを検定する。照合する 2 音声の話者が，ともにマスク有りの場合（a: mask-mask），一方がマスク有りで他方がマスク無しの場合（b: mask-no mask），ともにマスク無しの場合（c: no mask-no mask）の 3 水準で検定を行い，結果を表–2 に示す。

初めに，同一人の照合テストについて各水準のデータ分布を箱ひげ図で示す（図–10）。各水準の score をヒストグラム及び正規確率プロット（Q-Q Plot）で図示することにより，各水準の正規性を確認した。また，等分散性の検定に Bartlett 検定を用いた。正規性と等分散性が仮定できたため一元配置分散分析を行った結果，マスク有無に差異があることが認められた。次に，Tukey-Kramer 法による多重比較検定を行ったところ，ab 間及び bc 間に有意差が認められた。ab 間及び bc 間の効果量（Cohen's d）は大きかった。以上のことから，照合する 2 音声のマスク着用状態が異なるとき (b)，両方の音声の着用状態が同じ場合 (a, c) と比べて score が低下する傾向があることが認められた。



図–10 同一人照合テストのマスク有無によるスコア分布



図–11 別人照合テストのマスク有無によるスコア分布

別人の照合テストについても，同様に各水準のデータ分布を箱ひげ図で示す（図–11）。各水準の score をヒストグラム及び正規確率プロットで図示することにより，各水準の正規性を確認し，等分散性の検定に Bartlett 検定を用いたところ，正規性は仮定できたが等分散性が仮定できなかった。そこで，Kruskal-Wallis 検定を行った結果，マスク有無に差異があることが認められたため，Steel Dwass 法による多重比較検定を行った。ab 間及び ac 間に有意差が認められたが，ともに効果量（Cliff's delta）は小さかった。以上のことから，別人の照合においては，マスク着用状態による差異はあるものの効果量が小さいため考慮する必要はないといえる。

5.2 性別の影響

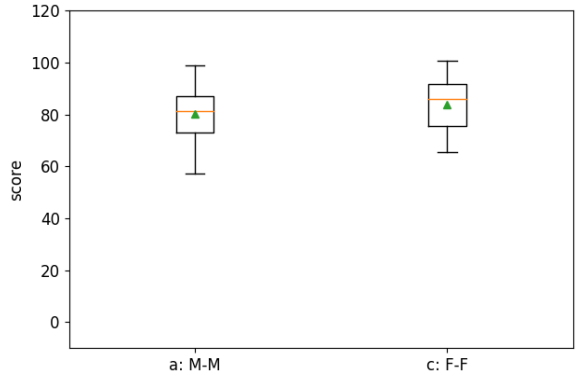
性別が照合に影響を与えるかどうかを検定する。照合する 2 音声の話者が，男同士の場合（a: M-M），一

表–3 性別の差による照合テストの分析結果

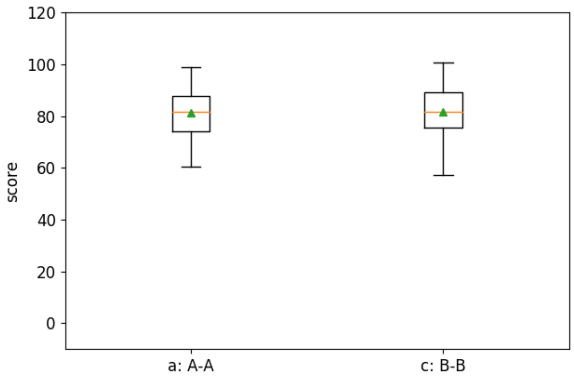
	性別	データ数	標準偏差	等分散検定	分散分析	多重比較	効果量
同一人	a: M-M	73	9.3	等分散と仮定可	一元配置分散分析	—	Cohen's d
	c: F-F	40	9.2	F 検定 ($p = 0.94$)	有意差あり		a-c: 0.41
別人	a: M-M	2,628	33.2	等分散と仮定不可		Steel Dwass 法	Cliff's delta
	b: M-F	2,920	23.3	Bartlett 検定	Kruskal-Wallis 検定	a-b: 有意差あり	a-b: 0.59
	c: F-F	780	26.5	($p = 2.2e-16$)	有意差あり	a-c: 有意差あり	a-c: 0.55
						b-c: 有意差あり	b-c: 0.93

表–4 年齢の差による照合テストの分析結果

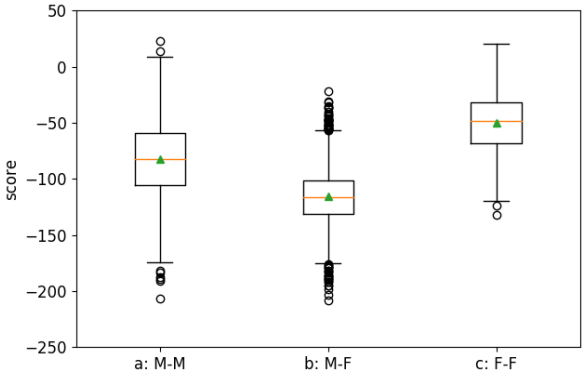
	年齢	データ数	標準偏差	等分散検定	分散分析	多重比較	効果量
同一人	a: A-A	55	9.3	等分散と仮定可	一元配置分散分析	—	—
	c: B-B	58	9.6	F 検定 ($p = 0.81$)	有意差なし		
別人	a: A-A	1,485	34.8	等分散と仮定可	一元配置分散分析	Tukey-Kramer 法	Cohen's d
	b: A-B	3,190	36.0	Bartlett 検定	有意差あり	a-b: 有意差あり	a-b: 0.18
	c: B-B	1,653	36.9	($p = 0.06$)		a-c: 有意差あり	a-c: 0.27
						b-c: 有意差あり	b-c: 0.08



図–12 同一人照合テストの性別によるスコア分布



図–14 同一人照合テストの年齢によるスコア分布



図–13 別人照合テストの性別によるスコア分布

方が男で他方が女の場合（b: M-F），ともに女の場合（c: F-F）の 3 水準で検定を行い，結果を**表–3** に示す。また，各水準のデータ分布を**図–12**，**図–13** に示す。

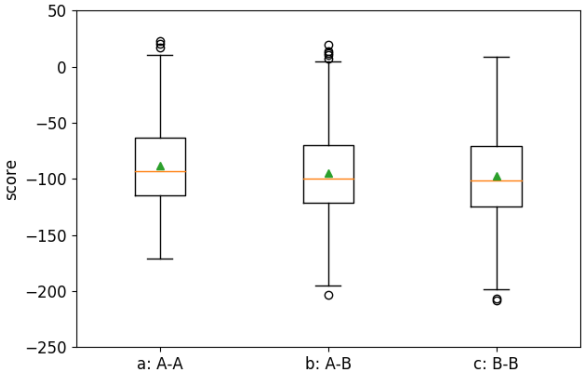
5.1 節と同様に正規性について確認し，等分散性は同一人の場合は F 検定，別人の場合は Bartlett 検定を行った。同一人では，正規性と等分散性が仮定できたため，一元配置分散分析を行った。別人では，正規性は仮定できたが等分散性が仮定できなかったため，Kruskal-Wallis 検定を行い，Steel Dwass 法による多重比較検定を行った。

男性同士と比較して，女性同士の照合テストではスコアがやや高い傾向がみられ，話者の性別は照合テストのスコアに影響を与えると考えられる。

5.3 年齢の影響

年齢が照合に影響を与えるかどうかを検定する。年齢を A，B の 2 区分に分け，照合する 2 音声の話者が A 同士の場合（a: A-A），一方が A で他方が B の場合（b: A-B），ともに B の場合（c: B-B）の 3 水準で検定を行い，結果を**表–4** に示す。また，各水準のデータ分布を**図–14**，**図–15** に示す。

5.1 節と同様に正規性について確認し，等分散性は同一人の場合は F 検定，別人の場合は Bartlett 検定を行った。ともに正規性と等分散性が仮定できたため，一元配置分散分析を行った。同一人では年齢による差がないことが棄却されなかったが，別人では年齢に有意差が認められたため Tukey-Kramer 法による多重比較検定を行った。



図–15　別人照合テストの年齢によるスコア分布

同一人の場合は年齢による有意差が見られなかった。別人の場合では有意差が見られるが効果量は小さいため，考慮する必要はないといえる。

6. ま と め

本研究では，電話を通して録音した日本語発話音声について，x-vector を用いた話者照合実験を行い精度を検証した。室内環境下で録音した 113 人の音声では EER=0.28%であり，極めて良好な照合結果が得られた。雑音や発話時間について評価し，テスト音声に付加する雑音が SNR=15 dB 以上であれば EER=2%以下，発話時間が 5 秒以上であれば EER=1.5%以下であった。20 人について 7 か月の録音時期差がある音声で照合した結果，精度の低下がみられ，個人間に差があることを考慮する必要があることが分かった。発話時のマスク着用，話者の性別及び年齢の話者特性は，統計検定により照合スコアに影響を与えることが分かった。しかしがながら，図–3 及び図–10～図–15 に示すとおり，同一人と別人の照合スコアの分布は話者特性による差異よりも十分大きく分かれているため，話者特性は照合結果を左右するほどの影響を与えないといえる。

以上より，照合対象となる音声（話者）の特性を把握した上での本話者照合システムの利用は，法科学における話者識別技術として有用であることを検証することができた。

文　献

[1] 警察庁ホームページ, “特殊詐欺認知・検挙状況等について,” https://www.npa.go.jp/publications/statistics/sousa/sagi.html（参照 2022-05-09）.
[2] 長内　隆, 石原俊一, “法科学分野における話者認識の動向,” 音響学会誌, 69, 365–370 (2013).
[3] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer and K. Veselý, “The Kaldi speech recognition toolkit,” *Proc. Autom. Speech Recognit. Underst.* (*ASRU*) *Workshop* (2011).
[4] D. Snyder, D. Garcia-Romero, D. Povey and S. Khudanpur, “Deep neural network embeddings for text-independent speaker verification,” *Proc. Interspeech 2017*, pp. 999–1003 (2017).
[5] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, “X-vectors: Robust DNN embeddings for speaker recognition,” *Proc. Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*) *2018*, pp. 5329–5333 (2018).
[6] N. Arsha, S. C. Joon and Z. Andrew, “VoxCeleb: A large-scale speaker identification dataset,” *Proc. Interspeech 2017*, pp. 2616–2620 (2017).
[7] 吉田芳郎, 袋谷丈夫, 竹沢寿幸, “ATR 音声データベース,” *Proc. Annu. Conf. JSAI*, 16, 124–125 (2002).
[8] K. Maekawa, H. Kikuchi and W. Tsukahara, “Corpus of spontaneous Japanese: Design, annotation and XML representation,” *Proc. Int. Symp. Large-scale Knowledge Resource* (*LKR 2004*), Tokyo Institute of Technology, pp. 19–24 (2004).
[9] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji and H. Saruwatari, “JVS corpus: Free Japanese multi-speaker voice corpus,” *arXiv preprint*, 1908.06248 (2019).
[10] 長内　隆, 蒔苗久則, 網野加苗, “音と法科学,” 音響学会誌, 72, 74–80 (2016).
[11] 鎌田敏明, 峯松信明, 長内　隆, 蒔苗久則, 谷本益巳, “雑音環境下における話者照合,” 信学技報, 音声 106(614), pp. 55–60 (2007).
[12] T. Osanai, Y. Kinoshita and F. Clermont, “Exploring sub-band cepstral distances for more robust speaker classification,” *Proc. 17th Speech Sci. Technol. Conf.*, pp. 41–44 (2018).
[13] S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa, N. Kitaoka, T. Nishiura, A. Sasou, M. Mizumachi, C. Miyajima, M. Fujimoto and T. Endo, “AURORA-2J: An evaluation framework for Japanese noisy speech recognition,” *IEICE Trans. Inf. Syst.*, **E88-D**, 535–544 (2005).
[14] N. Kitaoka, T. Yamada, S. Tsuge, C. Miyajima, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, T. Takiguchi, S. Tamura, S. Matsuda, T. Ogawa, S. Kuroiwa, K. Takeda and S. Nakamura, “CENSREC-1-C: An evaluation framework for voice activity detection under noisy environments,” *Acoust. Sci. & Tech.*, 30, 363–371 (2009).
[15] 中西亮介, 塩田さやか, 貴家仁志, “非線形帯域拡張法に基づく話者照合の検討,” 情処研報, Vol. 2017-SLP-115, No. 4 (2017).
[16] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen and T. Matsui, “Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification,” *Proc. Interspeech 2015*, pp. 239–243 (2015).
[17] 宋　裕進, 塩田さやか, 高道慎之介, 村上大輔, 松井知子, 猿渡　洋, “短時間発話を用いた話者照合のための音声加工の効果に関する検討,” 情処研報, Vol. 2021-SLP-136, No. 29 (2021).
[18] R. Li, D. Chen and W. Zhang, “Voiceai systems to NIST SRE19 evaluation: Robust speaker recognition on conversational telephone speech,” *Proc. Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*) *2020*, pp. 6459–6463 (2020).
[19] M. Senoussaoui, P. Kenny, N. Dehak and P. Dumouchel, “An i-vector extractor suitable for speaker recognition with both microphone and telephone speech,” *Proc. The Speaker and Language Recognition Workshop*, Odyssey, pp. 28–33 (2010).
[20] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, F.

Richardson, S. Shon, F. Grondin, R. Dehak, L. P. García-Perera, D. Povey, P. A. Torres-Carrasquillo, S. Khudanpur and N. Dehak, "State-of-the-art speaker recognition for telephone and video speech: The JHU-MIT submission for NIST SRE18," *Proc. Interspeech 2019*, pp. 1488–1492 (2019).
[21] J. Zhou, T. Jiang, Q. Hong and L. Li, "Extraction of noise-robust speaker embedding based on generative adversarial networks," *Proc. APSIPA Annu. Summit Conf. 2019*, pp. 1641–1645 (2019).
[22] H. Taherian, Z. Wang and D. Wang, "Deep learning based multi-channel speaker recognition in noisy and reverberant environments," *Proc. Interspeech 2019*, pp. 4070–4074 (2019).
[23] R. Saeidi, I. Huhtakallio and P. Alku, "Analysis of face mask effect on speaker recognition," *Proc. Interspeech 2016*, pp. 1800–1804 (2016).
[24] R. K. Das and H. Li, "Classiflcation of speech with and without face mask using acoustic features" *Proc. APSIPA Annu. Summit Conf. 2020*, pp. 747–752 (2020).
[25] S. Ioffe, "Probabilistic linear discriminant analysis," *Proc. Computer Vision-ECCV 2006*, pp. 531–542 (2006).
[26] D. Synder, G. Chen and D. Povey, "MUSAN: A music, speech, and noise corpus," arXive:1510.08484v1 (2015).
[27] T. Ko, V. Peddinti, D. Povey, M. Seltzer and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," *Proc. Acoust. Speech Signal Process.* (*ICASSP*) *2017*, pp. 5220–5224 (2017).
[28] R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2015). URL https://www.R-project.org/(参照 2021-09-24).
[29] 池田郁男, "統計検定を理解せずに使っている人のために I," 化学と生物, **51**, 318–325 (2013).
[30] 池田郁男, "統計検定を理解せずに使っている人のために II," 化学と生物, **51**, 408–417 (2013).
[31] 池田郁男, "統計検定を理解せずに使っている人のために III," 化学と生物, **51**, 483–495 (2013).
[32] 森 敏昭, 吉田寿夫, 心理学のためのデータ解析テクニカルブック (北大路書房, 京都, 1990).

付 録

A.1 各照合実験結果一覧

表–A.1 雑音 (babble, music, noise) の付加程度の変化による EER の変動

	EER (%)		
SNR (dB)	babble	music	noise
0	41.1	9.6	6.1
5	17.9	4.6	2.9
10	5.0	2.0	1.4
15	2.0	0.85	0.72
20	1.1	0.50	0.46
25	0.56	0.38	0.36
30	0.39	0.32	0.29
35	0.33	0.28	0.28
40	0.29	0.26	0.28

表–A.2 残響 (reverberation) による EER の変動

ROOM	large	medium	small
EER (%)	5.4	5.5	4.1

表–A.3 1 発話当たりの時間による EER の変動

Length (sec)	1	3	5	10	15	20
EER (%)	18.2	3.2	1.5	0.59	0.38	0.28

表–A.4 音声データのファイル形式別の EER

type	WAV	MP3	MP3	OGG	FLAC	AIFF
rate (%)	100	7	16	23	60	100
EER (%)	0.28	2.8	0.52	0.29	0.28	0.28

多谷 邦彦

2007 年東北大学理学部物理学科卒業。2020 年神戸大学博士 (工学)。2009 年より京都府警察科学捜査研究所に勤務し鑑定及び研究に従事。日本音響学会, 日本法科学技術学会, 電子情報通信学会各会員。

サクティ サクリアニ

2002 年ドイツウルム大学修士課程修了。2003–2011 年 ATR, NICT で音声言語処理の研究に従事。2008 年ドイツウルム大学博士課程修了。2011 年 NAIST 助教。2018 年 NAIST 特任准教授, RIKEN AIP 研究員。現在, JAIST 准教授, NAIST 客員准教授, RIKEN 客員研究員。電子情報通信学会, JNS, SFN, ASJ, ISCA, IEICE, IEEE 各会員。

藤原 修治

2005 年同志社大学文学研究科博士前期課程修了。2005 年より京都府警察科学捜査研究所に勤務し鑑定及び研究に従事。日本心理学会, 日本生理心理学会各会員。

中村 哲

1981 年京都工芸繊維大学電子工学科卒業。1992 年京都大学博士 (工学)。シャープ (株) 研究所, ATR, NICT 等を経て 2011 年より奈良先端大 教授, 音声翻訳, 音声対話等の音声言語処理の研究に従事。ISCA 理事・フェロー, IPSJ フェロー, IEEE フェロー。