様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月 ２４ 日現在

機関番号：８２５０５
研究種目：基盤研究(C)（一般）
研究期間：2013～2018
課題番号：２５３５０４８８

研究課題名（和文）標準化・正規化変換を利用した発話様式や時期変動に頑健な話者認識手法に関する研究

研究課題名（英文）Research on speaker recognition method that is robust to the differences in speaking styles and timing of recording speech using the Standardized-Normalization Transformation

研究代表者

長内　隆（Osanai, Takashi）

科学警察研究所・法科学第四部・部長

研究者番号：７０３９２２６４

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：話者認識では、比較する音声資料の収録環境、発話様式、時期変動の違いが認証性能の劣化、つまり誤認識を引き起こす要因の一つと言われている。そこで、このような音声資料の環境の違いにも頑健な話者認識手法について調べた。本課題では、我々がこれまでに構築した多様な音声データベースを用いた。また、我々が先に提案し、話者認識性能の向上に効果があることを証明した標準化・正規化変換を利用した。その結果、標準化・正規化変換は、このような音声資料の環境の違いにも頑健な話者認識手法として有効であることを示した。

研究成果の学術的意義や社会的意義
非協力的な話者を扱うことが多い法科学分野においては、多様な音声資料に適応できる話者認識に期待する声は大きい。例えば、振り込め詐欺事件において、同一犯による犯行の解明には、犯人の音声を比較して同一話者か否かを判断する必要があるが、それぞれの事件の会話はさまざまである上、関係者を装うなど、話し方も多様となるケースが多い。本研究の成果を利用することで、多様な音声資料であってもそれぞれの事件の犯人の同一性を示すことが期待できる。

研究成果の概要（英文）：In speaker recognition, it is said that the differences in recording environments, speaking styles, and timing of recording speech samples, are one of the factors that cause deterioration of authentication performance. Therefore, we investigated the speaker recognition method that is robust to such differences. In this research, we used various speech databases that we constructed so far. In addition, we used the Standardization-Normalization Transformation, which was proposed earlier in our research and proved to be effective in improving the speaker recognition performances. Results of the experiments showed that the Standardization-Normalization Transformation is an effective method for conquering the differences in the recorded speech data.

研究分野：情報工学

キーワード：話者認識　発話様式　時期変動　特徴量変換　犯罪捜査支援

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

テキスト依存型話者認識では、同一発話の２つの音声資料の音響特徴量間の距離を基にそれらの発話が同一話者によるものか否かを判断する。また、テキスト独立型話者認識では、登録音声資料から登録話者モデルを構築し、認識音声資料の登録話者モデルに対する尤度を基に認識音声の発話者が登録話者か否かを判断する。その際、比較する音声資料や登録時と認識時の音声資料の収録環境、発話長、発話内容、発話様式、時期変動などの条件の違い、つまり、音声資料のミスマッチによって、話者認識性能が低下すると言われている。

話者認識システムの構築において、このような音声資料のミスマッチを防ぐ最も効果的な方法は、(1)ミスマッチが生じないように指示した上で登録音声、認識音声を獲得すること、(2)ミスマッチが生じたときに、認識時の音声を登録時の音声に適合するように収録し直すことである。この考え方の下で運用する話者認識システムであれば、事前に話し方などを指示することでミスマッチを防ぐことができる。また、このようなシステムにあっては、利用者は自分自身を認識してもらいたいケースが多いので、誤認識に対する再収録にも快く応じてくれると考えられる。このように、いわゆる協力的な話者を対象としている場合、上記の対応は現実的な対応と言える。

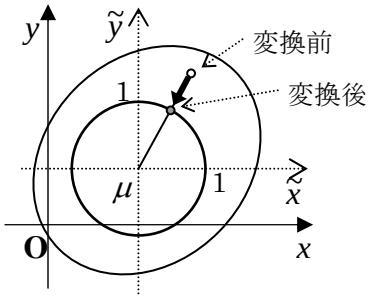
しかし、必ずしもそのような話者だけを対象にできるわけではない。例えば、法科学分野では、犯罪現場における音声がそれに該当する。また、会議の議事録作成やテレビ番組の字幕表示などでは、収録した音声が不十分であるからと言って再収録することはできない。このように、収録環境の指定や、再収録が困難場合でも、例えば、発話長、発話内容に関しては、登録話者モデルを構築する際の音声資料の音韻に偏りがなく、発話長の長い音声資料が別途利用可能であれば、発話長や発話内容の影響を抑えることが可能である。しかし法科学分野ではそれも困難なケースが多い。したがって、与えられた音声資料を用いるといった前提において、音声資料の収録環境、発話様式、時期変動などの違いが話者認識に与える影響を精査し、音声資料のミスマッチに頑健な話者認識手法の改善に取り組む必要がある。

２．研究の目的

多様な音声資料に頑健な話者認識手法を確立することを目的とする。話者認識システムにおいて、比較する音声資料や登録時と認識時の音声資料の録音条件（電話音声、マイクロホン音声など）、発話様式（話す速さ、声の大きさ、声の高さなど）、時期変動（短期変動、経年変化など）などの違いが認証性能の低下、つまり誤認識を引き起こす要因の一つと言われている。このようなミスマッチが生じると考えられる音声資料の多様性に頑健な音響特徴量の探索に関する検討を行い、ミスマッチの程度を図る指標の導出を試みる。

３．研究の方法

登録時と認識時のミスマッチが生じる要因として収録環境、発話様式、時期変動に着目し、各要因のミスマッチが話者認識性能に与える影響について調べる。ここでは、話者認識に有効な特徴量変換として先に提案した標準化・正規化変換(Standardized-Normalized Transformation: 以下 SNT と記す)[1,2]を利用する。SNT は、その音響特徴量の統計量である平均値、標準偏差を用いて標準化を行い、更にその特徴量ベクトルのノルムが１になるように正規化を行う変換である。この概念を図１に示す。この変換は、平均値を中心とする超球面上に特徴量を射影することである。この超球面上に射影した特徴量間の距離を利用することで話者認識性能の向上に効果があることを示した。これは話者性に由来する成分が超球面上に多く存在すること意味している。逆に、中心から放射状に存在する径方向には話者認識にはあまり有効でない成分が含まれており、それが発話様式など多様な音声成分と関連がある可能性がある。対象とする音響特徴量を話者認識で良く利用される LPC ケプストラム係数などが音声資料の多様性の影響をどの程度受けるかについて検証する。



特徴量ベクトル：$\mathbf{C}_t^{\mathrm{T}} = \left[C_{t1},\cdots,C_{ti},\cdots,C_{tp}\right]$を

平均値：$\boldsymbol{\mu}^{\mathrm{T}} = \left[\mu_1,\cdots,\mu_i,\cdots,\mu_p\right]$,

標準偏差：$\boldsymbol{\sigma}^{\mathrm{T}} = \left[\sigma_1,\cdots,\sigma_i,\cdots,\sigma_p\right]$で以下の変換を施す。

◆標準化変換：

$$\overline{\mathbf{C}}_t^{\mathrm{T}} = (\mathbf{C}_t - \boldsymbol{\mu})^{\mathrm{T}} = \left[C_{t1}-\mu_1,\cdots,C_{ti}-\mu_i,\cdots,C_{tp}-\mu_p\right]$$

◆正規化変換：

$$\tilde{\mathbf{C}}_t^{\mathrm{T}} = \left[\frac{C_{t1}-\mu_1}{\sigma_1},\cdots,\frac{C_{ti}-\mu_i}{\sigma_i},\cdots,\frac{C_{tp}-\mu_p}{\sigma_p}\right]$$

図１　SNT の概念

この SNT 変換を利用して、以下、(1)～(3)のミスマッチについての影響を検討する。なお、科学警察研究所でこれまでに様々な条件下で構築した多数話者音声データベースを保有しており、本研究でもこれらを利用する。音声データベースの概要を表１に示す。

表 1　これまで構築した多数話者音声データベースの概要

	第 1 世代(G1)	第 2 世代(G2)	第 3 世代(G3)	科研費 DB(H17～18)
話者	男性 740 名	男性 2,849 名	男 313 名 女 319 名	男 27 名 女　4 名
年令	19～59 歳	18～59 歳	18～76 歳	21～40 歳
収録方法	1 チャンネル： 黒電話	1 チャンネル： ファッション電話	4 チャンネル： 携帯電話、 マイクロホン （それぞれ気導音、骨導音）	2 チャンネル： ファッション電話、 マイクロホン
収録回数・時期差	2 時期 2 回 (約 4 か月)	3 時期 3 回 (約 3 か月)	2 時期 2 回 (2～3 か月)	2 時期 2 回 (約 2 か月)
発話様式	速さ：普通 大きさ：普通 高さ：普通	速さ：普通 大きさ：普通 高さ：普通	速さ：普通 大きさ：普通 高さ：普通	速さ：速、普通、遅 大きさ：大、普通、小 高さ：高、普通、低
発話内容	母音、単語、短文			

(1) 収録環境の影響

収録環境の異なる音声を比較する場合、照合性能の低下を避けるために、特徴量の平均値を減算するケプストラム平均値正規化（Cepsutral Mean Normalization：以下 CMN と記す）が利用される。SNT にも、特徴量の平均値を減算する処理が含まれており、CMN と同様、収録環境の影響を軽減する効果が期待できるので、両者の効果を検証した。また、SNT に必要な統計量算出に用いる音声資料の違いや連続音声への適用についても合わせて検証した。本検討では、多チャンネル同時収録をした第 3 世代の音声データベースを利用した。マイクロホンや携帯電話を介して 4 チャンネル同時収録した音声をフレーム長 23.2 ミリ秒、フレームシフト 11.6 ミリ秒で分析し、12 次の LPC ケプストラム係数を音響特徴量とした。収録環境に対する頑健性を調べるため、男女別に異なるチャンネル間で照合実験を行い、得られた照合率で評価を行った。なお、CMN に必要な平均値は、当該資料と同時期に発話した ATR 音素バランス文から求めた。ここでは、以下の実験を行った。

（実験 1）単音を対象とした実験：SNT で用いる統計量算出の音声資料の影響

男女各 200 名が発話した 100 単音に対して、平均特徴量間のユークリッド距離から照合率を求めた。SNT で用いる統計量は照合に用いる話者とは異なる 100 名の音声から求めた。照合に用いる当該単音、同一母音、ATR 音素バランス文のそれぞれから算出した統計量を用いて照合実験を行い、得られた照合率を比較した。

（実験 2）連続音声を対象とした実験

男女各 200 名が発話した 66 単語に対して､動的計画法を用いて求めた距離から照合率を求めた。SNT で用いる統計量は、照合に用いた話者とは異なる男女各 100 名の音声のうち、CMN を施した ATR 音素バランス文から求めた。

(2) 発話様式の影響

協力的話者ならば､登録時と照合時の発話様式に大きな違いを生じさせないことができるが、法科学分野において、それを期待することは難しい。発話様式が極端に異なるケースでは話者の異同識別鑑定が困難となることも少なくない。そこで、発話様式の違いが照合性能に与える影響と SNT を利用することによる影響の軽減について調べた。本検討では、さまざまな発話様式の音声を収録した科研費 DB を利用した｡成人男性 18 名が､無響室内で通常の発話に加えて、話す速さ（速い・遅い）、声の高さ（高い・低い）、声の大きさ（大きい・小さい）を変えて発声した 5 母音、12 短文を用いた。照合は時期変動を除外するために同時期間の比較とした。任意の発話様式で発話された平均特徴量間のユークリッド距離から照合率を求めた。

(3) 時期変動の影響

時期差のある発話に対する SNT の有効性について検討した。本検討では、3 時期にわたって収録した第 2 世代の音声データベースを利用した。男性話者 300 名が固定電話を介して、3 時期（時期差 2～3 か月、各時期 3 回発話）にわたって発話した 25 単語の音声資料を用いた。音声資料の比較は、時期差を同時期、2～3 か月、4～6 か月の 3 通りとし、単語別に動的計画法を用いて話者内、話者間距離を計算し、照合率等を求めた。音響特徴量に対して、無変換、CMN、SNT、CMN+SNT の 4 通りの処理を行い、それぞれにおける照合性能を比較した。

４．研究成果

(1) 収録環境の影響

（実験 1）当該単音から求めた統計量を用いて SNT を施したときの 100 単音各々の照合率の平均値を図 2 に示す。ここでは、SNT で使用する統計量算出用の音声資料として、当該音韻、同一母音、音素バランス文から算出した統計量を用いて SNT を施した。これより、当該単語から

算出した照合率が高いものの、これらの間には有意な差がないことから、音素バランス文から算出した統計量を用いても照合性能改善の効果があることが示された。これより、連続音声を対象とした話者照合に本手法が適用可能であることが示唆された。

（実験 2）男性話者が発話した単語「警察」で、マイクロホン収録音声と携帯電話収録音声を比較した照合率は、変換せずにそのまま照合したときの照合率（ベースライン：BL と表記）が 67%であったのに対し、CMN を施すことで 77.8%と 10.8point 向上し、さらに SNT を施すことで 81.9%と 14.9point 向上した。66 単語について、男女別、異なるチャンネル間、計 792 条件で同様の照合実験を行い、CMN を施した時、CMN 後に SNT を施した時のそれぞれの照合率と BL の照合率との差の頻度分布を図 3 に示す。CMN を施すことで、全条件のうち約 92%のケースで照合性能が改善し、平均 4point の改善が、さらに SNT を施すことですべてのケースで照合性能が改善し、平均 9point の改善がなされた。

(2) 発話様式の影響

通常発話と各発話様式における発話間で照合実験を行った。速さ、高さ、大きさに着目し、それぞれ通常発話との比較並びに速い vs. 遅いといったように条件が顕著に異なるケースの比較における話者照合率を図 4 に示す。これより、発話様式の違いは照合性能の低下を招き、大きさの違いによる影響が強いことが示された。SNT の効果について検討した結果、数%程度の照合率の向上が認められた。

(3) 時期変動の影響

各処理における照合率を図 5 に示す。時期差によって照合率が低下するが、2〜3 か月と 4〜6 か月の差は小さい。時期毎にみると、CMN を施すことでやや照合率の低下が見られたが、これは同一環境下で収録した音声資料を対象としたためと考えられた。CMN の有無にかかわらず、SNT を施すことで照合率はわずかに改善された。

(4) 得られた成果の国内外における位置づけ並びに今後の展望

本課題の研究と並行して、オーストラリアの法音声学者と、法科学分野で積極的に利用されるフォルマント情報を効果的に利用するための比較方法について共同研究を行なった。ここでは、一般にフォルマント周波数を抽出することは困難であるので、フォルマント帯域を指定した距離尺度[3]を利用することで、陽にフォルマント周波数を検出せずに比較する手法である。この手法に SNT を組み入れることで性能の改善が図られる可能性が期待できる。今後、共同研究を継続して取り組むこととする。

これまで提案した SNT による特徴量変換を利用することで、収録環境の違い、発話様式の違い、時期変動に対する影響が軽減できることが示された。商用利用を想定した話者認識システムの開発にあっては、このような知見はそれほど必要とされていないと思われるが、法科学分野では非常に有効な成果といえる。今回、ミスマッチの程度を図る指標の導出には至らなかったので、今後は、その点の解明を進める予定である。

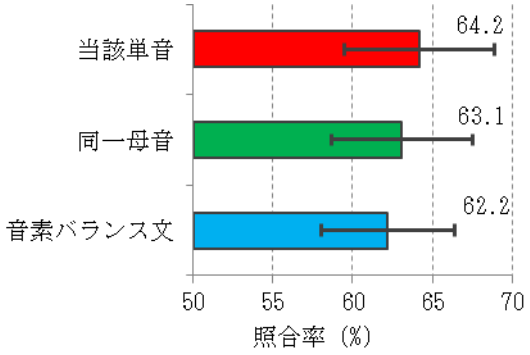


図 2　統計量算出資料の違いによる照合率

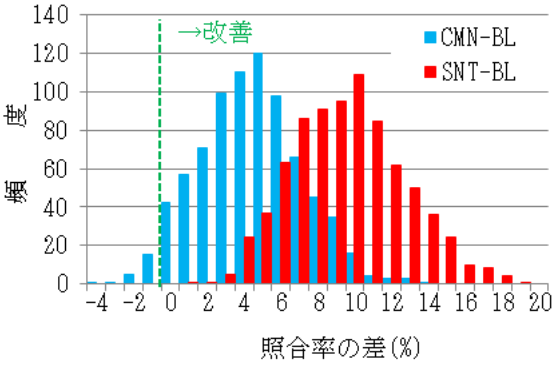


図 3　SNT による照合率改善の程度

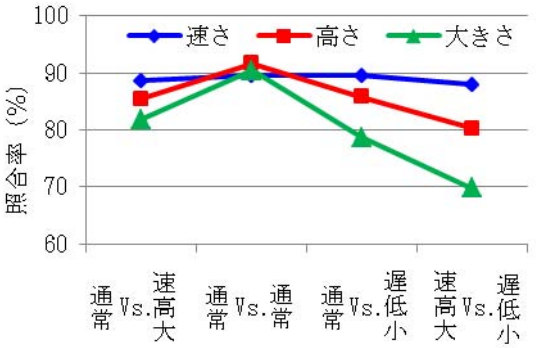


図 4　発話様式の違いによる照合率

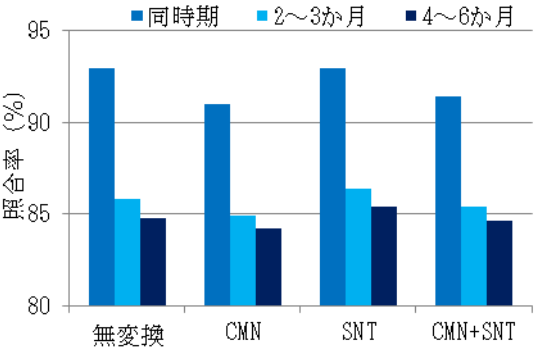


図 5　比較する資料の時期差と照合率

＜引用文献＞
[1] 単独発声母音を用いた話者照合における特徴量変換，長内隆，尾関和彦，谷本益巳，音響誌，62(12)，pp.848-855，2006.
[2] 法科学的話者照合のための標準化･正規化クロス VQ 歪みの利用，長内隆，尾関和彦，谷本益巳，音響誌，63(12)，pp.708-715，2007.
[3] Frequency-band specification in cepstral distance computation, F. Clermont, P. Mokhtari, 5th Australian International Conference on Speech Science & Technology, pp.354-359, 1994.

５．主な発表論文等
〔雑誌論文〕（計　7 件）
① Reference data on Japanese vowel devoicing: Effects of speakers' and parents' places of origin and within-speaker reproducibility, Kanae Amino, Hisanori Makinae, Toshiaki Kamada, Takashi Osanai, Acoustical Science and Technology, 39(3), pp.207-214, 2018, Doi:10.1250/ast.39.207, 有.
② 音と法科学，長内隆，蒔苗久則，網野加苗，日本音響学会誌，72(2)，pp.74-80，2016，有.
③ Cross-language differences of articulation rate and its transfer into Japanese as a second language, Kanae Amino, Takashi Osanai, Forensic Science International, 249, pp.116-122, 2015, Doi:10.1016/j.forsciint.2015.01.029, 有.
④ Native vs. non-native speaker identification using Japanese spoken telephone numbers, Kanae Amino, Takashi Osanai, Speech Communications, 56, pp.70-81, 2014, 有.
⑤ 多数話者による単独発話母音から抽出したフォルマント周波数の特性，鎌田敏明，蒔苗久則，網野加苗，長内隆，科学警察研究所報告，63(1)，pp.19-23，2014，有.
⑥ 法科学分野における話者認識の動向，長内隆，石原俊一，日本音響学会誌，69(7)，pp.365-370，2013，無.
⑦ 法科学分野における話者認識のための大規模音声データベースの構築，蒔苗久則，鎌田敏明，長内隆，科学警察研究所報告，62(1-2)，pp.53-57，2013，有.

〔学会発表〕（計　20 件）
① Exploring sub-band cepstral distances for more robust speaker classification, Takashi Osanai, Yuko Kinoshita, Frantz Clermont, 17th Australasian International Conference on Speech Science & Technology, 2018.
② Forensic voice comparison using sub-band cepstral distances as features: A first attempt with vowels from 306 Japanese speakers under channel mismatch conditions, Yuko Kinoshita, Takashi Osanai, Frantz Clermont, 17th Australasian International Conference on Speech Science & Technology, 2018.
③ 時期差のある単語発話を用いた話者照合における標準化・正規化変換の効果，長内隆，網野加苗，蒔苗久則，鎌田敏明，日本法科学技術学会第 24 回学術集会，2018.
④ 声道共鳴特性を用いた地域性情報と話者分類，鎌田敏明，蒔首久則，網野加苗，長内隆，日本法科学技術学会第 24 回学術集会，2018.
⑤ 言語形態を用いた地域性推定における共通語形の影響，網野加苗，蒔苗久則，鎌田敏明，長内隆，日本法科学技術学会第 24 回学術集会，2018.
⑥ 話者照合における発話様式の影響に関する予備的検討，長内隆，網野加苗，蒔苗久則，鎌田敏明，日本法科学技術学会第 23 回学術集会，2017.
⑦ Sub-band cepstral variability within and between speakers under microphone and mobile conditions: A preliminary investigation, Frantz Clermont, Yuko Kinoshita, Takashi Osanai, 16th Australasian International Conference on Speech Science & Technology, 2016.
⑧ 異なる環境下の単語発話を用いた話者照合における標準化・ 正規化変換の効果，長内隆，網野加苗，蒔苗久則，鎌田敏明，日本法科学技術学会第 22 回学術集会，2016.
⑨ 話者認識における静的特徴量と動的特徴量の比較，鎌田敏明，蒔苗久則，網野加苗，長内隆，日本法科学技術学会第 22 回学術集会，2016.
⑩ 正弦波モデルを用いたブラインド雑音抑圧，蒔苗久則，網野加苗，鎌田敏明，長内隆，日本法科学技術学会第 22 回学術集会，2016.
⑪ 非定常雑音の抑圧性能の評価に関する研究，蒔苗久則，網野加苗，鎌田敏明，長内隆，日本法科学技術学会第 21 回学術集会，2015.
⑫ 聴取による合成音声と 自然音声の識別，網野加苗，蒔苗久則，鎌田敏明，長内隆，日本法科学技術学会第 21 回学術集会，2015.
⑬ 収録環境の異なる音声を用いた話者照合における標準化・正規化変換の効果，長内隆，網野加苗，蒔苗久則，鎌田敏明，日本法科学技術学会第 21 回学術集会，2015.
⑭ 音声データベースの違いによる話者照合性能の比較，長内隆，網野加苗，鎌田敏明，蒔苗久則，日本法科学技術学会第 20 回学術集会，2014.

⑮ 母音問距離と時期差における話者の地域性，鎌田敏明，長内隆，蒔苗久則，網野加苗，日本法科学技術学会第 20 回学術集会，2014.
⑯ 言語形態を用いた出身地推定法の提案，網野加苗，蒔苗久則，鎌田敏明，長内隆，日本法科学技術学会第 20 回学術集会，2014.
⑰ 母音間距離と時期差における話者の地域性，網野加苗，蒔苗久則，鎌田敏明，長内隆，日本音響学会 2014 年春季研究発表会，2014.
⑱ 連続音声を対象とした音響特徴量聞の性別識別性能の比較，長内隆，網野加苗，鎌田敏明，蒔苗久則，日本法科学技術学会第 19 回学術集会，2013.
⑲ フオルマント周波数を用いた話者照合法の統計的評価，四宮康治，網野加苗，蒔苗久則，鎌田敏明，長内隆，日本法科学技術学会第 19 回学術集会，2013.
⑳ 本人および両親の出身地が母音の無声化率に与える影響，網野加苗，蒔苗久則，鎌田敏明，長内隆，日本音響学会 2013 年秋季研究発表会講演論文集，2013.

〔図書〕（計　0 件）
〔産業財産権〕
○出願状況（計　0 件）
○取得状況（計　0 件）
〔その他〕
なし

６．研究組織
(1) 研究分担者
なし
(2) 研究協力者
研究協力者氏名：鎌田　敏明
ローマ字氏名：(KAMADA toshiaki)
研究協力者氏名：蒔苗　久則
ローマ字氏名：(MAKINAE, hisanori)
研究協力者氏名：網野　加苗
ローマ字氏名：(AMINO, kanae)