科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　２２　日現在

機関番号：　１２５０１
研究種目：　基盤研究(B)
研究期間：　2009～2013
課題番号：　２１３０００６０

研究課題名（和文）長期間収録音声コーパスに基づく話者内音声変動に頑健な話者認識手法の研究

研究課題名（英文）Robust Speaker Recognition with Intra-Speaker Variability Compensation based on Long-Term Recorded Speech Corpus

研究代表者
黒岩　眞吾（Kuroiwa, Shingo）

千葉大学・融合科学研究科（研究院）・教授

研究者番号：２０３３３５１０

交付決定額（研究期間全体）：（直接経費）　　13,800,000円、（間接経費）　　4,140,000円

研究成果の概要（和文）：本研究では、音声が長期間及び短期間にどのように変化するかを調査するための音声コーパスを構築すると共に、民生利用及び法科学の分野での利用を視野に、話者内音声変動に頑健で信頼性の高い話者認識手法の検討を行った。具体的には、10年間に渡り、毎週1回朝・昼・夕に同一話者が防音室で音素バランス文等を発声した音声データを国立情報学研究所・音声資源コンソーシアムを通じ『AWA長期間収録音声コーパス』として公開した。また、同コーパスを利用して話者内音声変動に頑健な話者認識手法を提案すると共に、法科学分野における話者認識で有用な特徴量、機械と人の話者認識特性の比較、話者モデル構築手法、照合手法を提案した。

研究成果の概要（英文）：This research project aimed to build a new speech corpus that enables many researchers to investigate changes in human voices during a day, a month or several years, and to develop accurate and robust speaker recognition methods for industrial and forensic uses. The speech corpus named "AWA Long-Term Recorded Speech Corpus (AWA-LTR), which is released by Speech Resources Consortium of National Institute of Informatics (NII-SRC), consists of 6 speaker's read speech data recorded at morning, noon, and evening every week for several years (2 to 10 years). Using this corpus, we have developed intra-speaker variability compensation methods that improve the robustness of speaker recognition techniques. We also studied effective speech features for forensic speaker recognition, a comparison between human and machine speaker recognition abilities, accurate and robust speaker modeling methods and speaker verification methods.

研究分野：　情報学

科研費の分科・細目：　知覚情報処理・知能ロボティクス

キーワード：　音声学　話者認識　話者照合　話者識別　音声データベース　法科学　話者内音声変動　『AWA長期間収録音声コーパス』

１．研究開始当初の背景

近年、音声による個人認証技術である話者認識（話者照合・話者識別）技術は、セキュリティ分野以外にも、映像データを対象とする情報検索、音声認識の個人適応、ロボットの個人識別等、様々な応用が期待されている。研究開始当初、音声の分野での最大の国際会議である、Interspeech2008 では、話者認識をタイトルとするセッションが６つあり，周辺技術を含めると 100 件以上の発表が行われた。また、アジア地区においても中国言語資源コンソーシアム（Chinese Corpus Consortium）が 2006 年に話者認識に関する国際的なコンテストを実施する等、国際的には話者認識技術に対する期待が大きく、また研究活動にも活性化が見られた。これに対し、日本国内では、音声に関する最大の会議である日本音響学会 2008 年秋季研究発表会においては発表件数がわずか４件と日本での話者認識研究は低迷した状況にあった。これは、日本において共通して利用可能な話者認識コーパスが存在しないことが最大の原因であると考えられた。話者認識は、言語依存性が少ないように思われがちであるが、中国語・英語の間で有効な手法が異なることが既に知られており、日本語の話者認識研究のための音声コーパスを構築・公開することが急務であると考えた。また、欧米においては比較的長時間の音声を用いた話者認識の研究が主流であったのに対し、日本では、民生利用や法科学分野での利用を想定した比較的短い音声での話者認識技術が必要とされていた。そのため、欧米で性能が高いと報告されている手法も、比較的短い音声の場合に必ずしも有効でない可能性があり、各手法の再評価が必要であった。さらに、比較する音声間に時期差がある場合に照合性能が大きく低下する問題や、機械と人間の話者認識特性の違い等、大量データに基づく機械学習が主流となる中で見過ごされている音声の個人性に関する本質的な分析を行う必要性も高かった。

２．研究の目的

音声による個人認証技術である話者認識（話者照合・話者識別）技術は、セキュリティ分野以外にも、映像データを対象とする情報検索、音声認識の個人適応、ロボットの個人識別等、様々な応用が期待されている。本研究では、音声に含まれる情報のうち「話者性」の工学的解明を目標に見据え、実社会で利用し得る話者認識技術を確立することを目的に研究を行った。具体的には、話者認識および音声研究用大規模コーパスの構築と話者認識評価基盤の整備、話者内音声変動のモデル化とそれに基づく頑健な話者認識方式の研究、及び、法科学における音声による個人認証に有効な手法を明らかにするために、以下の研究項目を実施した。

（１）音声工学、法科学における話者認識研究で利用可能な大規模音声コーパスを構築する。

（２）話者認識技術を客観的に評価可能な評価基盤を構築し、各々の技術の比較評価を行う。

（３）話者内音声変動、音韻性、話者性を分離可能な音響特徴空間を構成する。

（４）計算量にとらわれない高精度かつ頑健な新しい話者認識手法を開発する。

（５）法科学の観点から、人間及び機械による音声からの個人認証の有効性を検証すると共に、それらに有効な特徴パラメータを検討する。

３．研究の方法

（１）音声コーパスの構築

同一話者が、毎週 1 回、朝・昼・夕に音素バランス文等を防音室で発声した音声を収録し、コーパスとしての整備（音声区間切り出しやファイリング）を進め公開する。また、協力者にマイクを配布し数か月おきに自宅等で収録した音声を収集し実環境評価用音声コーパスを作成する。

（２）話者認識評価環境の構築と比較実験

ベクトル量子化（VQ）、混合正規分布（GMM）、サポートベクターマシン（SVM）、i-vector 等のテキスト独立型話者認識分野での標準手法をインプリメントし、様々な条件で性能の比較を行う。

（３）話者内音声変動に頑健な話者認識手法

上述の（１）で整備した音声コーパスを利用し、話者内音声変動をモデル化すると共に、それを用いた頑健な話者認識手法を構築する。

（４）新しい話者照合手法とスコア正規化法

識別誤り最小化学習、深層学習など計算量は大きいが照合性能の向上が期待される様々な機械学習手法を話者照合に応用する。また、既存の話者から未知話者のモデルを統計的に大量生成し、スコア正規化を行う手法等、照合時における頑健性及び精度向上を目指す。

（５）法科学における話者認識手法

テキスト独立型話者認識に加え、テキスト依存型話者認識において機械による話者照合と人間（聴取及び目視）による話者照合の性能や特徴を調査する。また、人手を介さないと抽出しにくい特徴量も用いた話者認識（母語や方言の識別を含む）を試みる。

４．研究成果

（１）音声コーパスの構築

毎週１回、朝・昼・夕の時間帯に研究代表者が音素バランス文 50 文等の 10 分強の音声を 5 年間防音室で収録した。また、うち、2010 年度に収録した１年分のデータの切り出しを行い『AWA 長期間収録音声コーパス(AWA-LTR)』として 2012 年 6 月に国立情報学研究所・音声資源コンソーシアム(NII-SRC)を通じて公開した。さらに、過去に収録した

データを含め、NII-SRC の協力により切り出し等の整備を行い 2012 年 3 月までに収録した話者 6 名の 1 年半～9 年半分の音声データを 2014 年度中に公開予定である。本コーパスの話者毎の音声収録期間を表 1 に示す。また AWA-LTR に関する情報提供を行うホームページを開設した。

AWA-LTR は同一の話者が防音室内で同じ文章を定期的にかつ長期間にわたり発声した音声を収録したもの（体調の如何に関わらず収録を行っているため、風邪等の体調不良時の音声も含まれている）であり、国際的にも唯一無二なコーパスである。そのため、AWA-LTR は、話者認識だけでなく、広く音声学の分野で活用されるコーパスと考えられる。また、研究代表者が今後も収録を継続することでその価値はさらに高まると期待される。

表 1．話者と発声期間（開始は 2003 年 10 月）

話者	開始時年齢	終了年月
男性 1	39 歳	継続中
男性 2	30 歳	2008 年 3 月
男性 3	20 代前半	2008 年 3 月
男性 4	20 代前半	2005 年 3 月
女性 1	20 代前半	2007 年 3 月
女性 2	20 代前半	2005 年 3 月

（2）話者認識評価環境の構築と比較実験

代表的なテキスト独立型話者認識手法である VQ、最尤推定（ML）および事後確率最大化（MAP）による GMM, GMM-SVM、i-vector の実験環境を構築した。登録用音声として音素バランス文 5 文、照合用音声して音素バランス文 1 文を用いた話者照合実験では、これらのうち GMM－SVM が最も照合率が高く、登録音声数が増えるに従いさらに性能の向上が確認された。ただし、発声に含まれる音素の出現頻度の偏りが大きい場合は GMM(MAP)の照合率が GMM-SVM を越えることもあり、登録音声が 1～3 文の場合には GMM(MAP)の照合率の方が高い。また、これらの手法では話者モデルの種となる背景話者モデル(UBM)が、照合率に大きく影響することも明らかとなった。一方で、米国立標準技術研究所(National Institute of Standards and Technology; NIST)主催の話者認識コンテスト等で高い性能を示したことからテキスト独立型話者照合で現在デファクトスタンダードとなっている i-vector は、上述の実験条件において GMM-SVM よりも照合率が低かった（勿論、登録・照合共に数分程度の音声を用いることが可能な条件下では i-vector の照合率が最も高くなる。つまり、現状の i-vector は上述の条件と類似の条件下で用いられる話者照合システムには適さないと考えられる）。以上の知見には、イノベーティブな要素は少ないが、話者照合を民生利用する上で有益な情報となっている。また、構築した実験環境の一部は共同研究等を通じて社会に還元されており、今後も同様の活動を続けていく。

短い音声での話者照合では、登録時と照合時で出現が排他的となっている音素が問題となる。GMM や GMM-SVM ではその問題は局所的であるため Missing Feature Theory 等の手法が適用できるが、i-vector 上では全ての次元にその問題が波及してしまう。今後は、短い発声でも i-vector の優位性を活かせる手法や、音声認識と話者認識を統合した手法の検討を行っていく。

（3）話者内音声変動に頑健な話者認識手法

AWA-LTR を用いて、長期的及び短期的な話者内音声変動の調査を行った。その結果、韻律情報も含めた分析では、朝と昼夕間で顕著な違いが現れた。一方で、話者認識に用いられているパワースペクトルを基本とする各種特徴空間(MFCC 等)上では、長期的な音声変動が大きく、また、音素によってその変動の主成分軸が異なることが明らかとなった。また、AWA-LTR の 1 名の話者の音声データより抽出した話者内音声変動の大きい部分空間の補空間で、科学警察研究所が構築した話者認識研究用音声コーパス(『大規模話者骨導音声データベース』)を用いて GMM-SVM による話者照合実験を行ったところ、時期差のあるデータに対して照合率の向上を確認した（2 つのコーパスは、話者および収録環境オープンである）。さらに、音素毎に同様の射影を行いさらなる照合率の向上を確認した。以上の結果は、パワースペクトルを基とする特徴空間上で話者内音声変動が音素毎に異なる一方、話者間では共通部分があることを示唆している。

以上の研究は機械学習的アプローチでは得にくい話者内音声変動を物理的にとらえることにつながる研究であり、今後、音声の生成機構に基づく研究に発展させていく予定である。また、直近の課題として、AWA-LTR の 6 名の音声データを用い話者間で共通する話者内音声変動や季節による音声の周期的変動等より詳細な調査を行っていく。

（4）新しい話者照合手法とスコア正規化法

話者照合手法として、識別学習のひとつである Soft Margin Estimation(SME)を用いた話者モデル構成手法、深層学習を用いた話者照合手法、複数話者の音声が重なっている場合にも適用可能な話者照合手法を提案した。また、実環境での頑健性を向上させるために、深層学習を用いた残響除去手法や背景音声除去手法を提案した。さらに、本研究が対象としてきた短い発声に対する頑健性向上手法として Missing Feature Theory を利用した話者照合手法も提案した。スコア正規化手法としては、ロバスト統計の一つである順位統計量に基づく閾値設定手法や疑似話者モ

デルを用いた正規化手法を提案した。これらのうち、SME や深層学習による話者照合では若干の改善は見られたものの有意な照合率向上は現在までに確認できていない。一方で、大量のデータをシミュレーションにより生成可能な残響除去では深層学習により優位な改善が確認できた。複数話者の音声の重なりに対しては話者インデキシングをタスクとして、照合とモデル更新を繰り返すことで構成した混合音声モデルによりインデキシング性能の改善を確認した。スコア正規化では、従来の T-norm 等の正規化法に順位統計量を組み合わせることで照合スコア（等誤り率、及び minimum detection cost function: minDCF）の改善及び閾値の安定性を確認した。

機械学習は話者認識においても強力なツールになると期待されている。しかし、本人モデルの学習データや照合に用いるデータが極端に少ない話者照合では、現在までにそれらの学習手法を直接適用するだけで照合率を向上できたとの報告はない。大量に収集可能な不特定話者の音声から得られる知見や統計情報をモデル化し、何らかの方法で個々の話者に適応していく手法が今後の研究の鍵になると考えられる。そのため、今後も i-vector や深層学習の技術を少量のデータでも活かすことを可能とする手法の探求を継続する。

（５）法科学分野における話者認識手法

法科学分野における話者照合の有効性を調査するため、VQ に基づくテキスト独立型話者照合において、発声内容の共通性の指標を提案し、共通性が照合性能に与える影響を検討した。その結果、共通性が高いほど照合性能が高くなること、及び、音声サンプルが 30 秒程度以上あれば内容の影響を受けにくくなることを確認した。一方で、法科学分野ではテキスト依存型話者認識が利用できる場面が多いことから、携帯電話音声を対象に発話内容と話者照合性能の関係を明らかにした。さらに、人間の聴取による話者照合と DP（Dynamic Programing）による話者照合を比較する実験を行い、同程度の性能であること、人間が聴取に加えスペクトログラムの目視も行うことで照合誤りを半減できること、DP が誤ったデータの 90%を人間が正しく照合できるとの実験結果が得られた。これらの結果は、法科学において人間と機械が協力することで、より高い個人認証が可能となることを示唆している。また、グローバル犯罪に対応していくため、日本語非母語話者の特徴を分析すると共に出身地識別に有効なパラメータを検討した。その結果、F0 パタン、モーラを基準とする調音速度等のプロソディ情報や、母音の無性化頻度が識別に有効であることが判明した。

今後、これらの特徴量を機械による話者照合へ適用することで性能向上が期待できるため、特徴量の自動抽出法を含め今後も検討を続ける。

（６）その他

話者認識分野で活躍する国内の研究者に呼びかけ音響学会誌 2013 年 7 号において小特集「話者認識に関する研究の動向」をまとめた。

５．主な発表論文等

〔雑誌論文〕（計１１件）

① Kanae Amino, Takashi Osanai, “Native vs. non-native accent identification using Japanese spoken telephone numbers,” Speech Communication, 査読有, Vol.56, pp.70-81, 2014. DOI: 10.1016/j.specom.2013.07.010

② 鎌田敏明, 蒔苗久則, 網野加苗, 長内隆, “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,” 科学警察研究所報告, 査読有, Vol. 63, No. 1, pp.19-23, 2014.

③ 内田正洋, 篠崎隆宏, 堀内靖雄, 黒岩眞吾, “発話中の一部区間を用いた感情認識,” 電子情報通信学会論文誌, 査読有 Vol. J97-D, No.1, pp.236-238, 2014.

④ 黒岩眞吾, “小特集「話者認識に関する研究の動向」にあたって,” 音響学会誌, 査読無, vol.69, No.7, p.340, 2013.

⑤ 王龍標, 西田昌史, 柘植覚, 網野加苗, “話者認識におけるロバストネス,” 音響学会誌, 査読無, vol.69, No.7, pp.357-364, 2013.

⑥ 長内隆, 石原俊一, “法科学分野における話者認識の動向,” 日本音響学会誌, 査読無, vol. 69, No.7, pp.365-370, 2013.

⑦ Kanae AMINO, and Takashi OSANAI, “Speaker characteristics that appear in vowel nasalisation and their change over time, Acoustical Science and Technology, 査読有 Vol.33, No.2, pp.96-105, 2012.

⑧ 岡本悠, 柘植覚, 堀内靖雄, 黒岩眞吾, “順位統計量を用いたテキスト独立型話者照合手法,” 電子情報通信学会論文誌, 査読有, Vol.J94-D, No.9, pp.1551-1560, 2011.

⑨ Wenbin Zhang, Haoze Lu, Yasuo Horiuchi, Satoru Tsuge, Kenji Kita, Shingo Kuroiwa, “Text-Independent Speaker Identification Based on Reducing Inter-Session Variability of Speech Feature Using PCA Transformation,” Journal of Signal Processing, 査読有, Vol.15, No.4, pp.275-278, July 2011.

⑩ Haoze Lu, Masafumi Nishida, Yasuo Horiuchi, Shingo Kuroiwa, “Text-Independent speaker identification in phoneme-independent subspace using PCA transformation,” International

Journal of Biometrics, 査読有, Vol.2, No.4, pp.379-390, 2010.

〔学会発表〕（計６２件）

① Yoko Takahashi, Shingo Kuroiwa, Yasuo Horiuchi, Satoru Tsuge, “Missing feature theory for speaker verification with short utterances,” International Workshop on Nonlinear Circuits, Communication and Signal Processing, pp.121-124, Honolulu, Mar. 1, 2014.
② Takaaki Ishii, Hiroki Komiyama, Takahiro Shinozaki, Yasuo Horiuchi, Shingo Kuroiwa, “Reverberant Speech Recognition Based on Denoising Autoencoder,” Interspeech2013, pp.3512- 3516, Lyon, Aug. 29, 2013.
③ Satoru Tsuge, Shingo Kuroiwa, “AWA Long-Term Recording Speech Corpus (AWA-LTR),” International Workshop on Nonlinear Circuits, Communication and Signal Processing, pp.17-20, Kailua-Kona, Mar. 5, 2013.
④ Kanae Amino, Takashi Osanai, “Foreign accent identification using articulation rate of Japanese read speech,” 14th Australasian International Conference on Speech Science and Technology, Sydney, Dec. 6,2012.
⑤ Haoze Lu, Wenbin Zhang, Takahiro Shinozaki, Yasuo Horiuchi and Shingo Kuroiwa, “PCA Transformation Based Inter-session Variability Suppression,” 8th International Conference on Natural Language Processing and Knowledge Engineering, CD-ROM, Hefei, Sep. 21, 2012.
⑥ 黒岩眞吾, 柘植覚, 張文彬, 篠崎隆宏, 堀内靖雄, “AWA 長期間収録音声コーパスと時期差の分析,” 音響学会春季研究発表会, pp.83-86, 横浜, Mar. 15, 2012.
⑦ Shiori Takenaka, Takahiro Shinozaki, Yasuo Horiuchi, Shingo Kuroiwa, “Pseudo Speaker Models for Text-Independent Speaker Verification Using Rank Threshold,” 7th IEEE Conference on Natural Language Processing and Knowledge Engineering, pp.265-268, Tokushima, Nov.29, 2011.
⑧ Kanae Amino, Takashi Osanai, “Realisation of the prosodic structure of spoken telephone numbers by native and non-native speakers of Japanese,” 17th International Congress of Phonetic Sciences, pp.17-21, Hong Kong, Aug. 18, 2011.
⑨ Masafumi Nakao, Satoru Tsuge, Minoru Fukumi and Shingo Kuroiwa, “Speaker vector combination method of air- and bone-conduction speech for speaker identification,” International Workshop on Nonlinear Circuits, Communication and Signal Processing, pp.417-420, Honolulu, Mar. 4, 2010.
⑩ Satoru Tsuge, Daichi Koizumi, Minoru Fukumi and Shingo Kuroiwa, “Speaker verification method using bone-conduction and air-conduction speech,” 2009 International Symposium on Intelligent Signal Processing and Communication Systems, pp.449-452, Kanazawa, Dec. 8, 2009.
⑪ Haruka Okamoto, Satoru Tsuge, A. Abdelwahab, Masafumi Nishida, Yasuo Horiuchi, Shingo Kuroiwa, “Text-Independent Speaker Verification Using Rank Threshold in Large Number of Speaker Models,” Iinterspeech2009, pp.2367-2370, Brighton, Sep. 9, 2009.

〔その他〕
ホームページ等
http://awa-ltr.xii.jp/
http://research.nii.ac.jp/src/AWA-LTR.html

６．研究組織
(1)研究代表者
黒岩　眞吾（KUROIWA, Shingo）
千葉大学・大学院融合科学研究科・教授
研究者番号：２０３３３５１０

(2)研究分担者
柘植　覚（TSUGE, Satoru）
大同大学・情報学部・准教授
研究者番号：００３２５２５０

長内　隆（OSANAI Takashi）
科学警察研究所・法科学第四部・部付主任研究官
研究者番号：７０３９２２６４

篠崎　隆宏（SHINOZAKI Takahiro）
東京工業大学・総合理工学研究科・准教授
研究者番号：８０３６１４４２

(3)連携研究者
堀内　靖雄（HORIUCHI Yasuo）
千葉大学・大学院融合科学研究科・准教授
研究者番号：３０２７２３４７

西田　昌史（NISHIDA Masafumi）
同志社大学・理工学部・准教授
研究者番号：８０３６１４４２