様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 １８ 日現在

機関番号：２２６０４
研究種目：研究活動スタート支援
研究期間：2013～2014
課題番号：２５８８００２６

研究課題名（和文）カーネル法に基づいた歪みに頑健な話者照合システムの構築

研究課題名（英文）Constructing noise robust speaker verification system based on kernel method

研究代表者
塩田　さやか（Shiota, Sayaka）

首都大学東京・システムデザイン研究科・助教

研究者番号：９０７０５０３９

交付決定額（研究期間全体）：（直接経費）　　2,100,000円

研究成果の概要（和文）：統計的手法を用いた音声信号処理技術を用いた生体認証手法である話者照合の性能改善について研究を行った。本研究では、従来の話者照合システムとは異なり、登録話者のモデルを統計的手法を用いて学習するのではなく、カーネル法を用いることで登録音声と評価音声の距離を直接評価することを行った。研究の当初にはカーネル法で音声の距離を測る方法も問題があったが研究協力者との協議の結果、小規模な話者照合実験において高い照合性能を得ることができた。今後はさらに複雑な環境において高い照合性能が得られるように照合結果の求め方や、パラメータの設定についてもより詳細に検討する予定である。

研究成果の概要（英文）：Speaker verification system is one method of biometrics authentication system. The conventional speaker verification system based on statistical machine learning methods require enough number of training data for training speaker dependent models. In this research, a kernel based method is used for calculate two samples distance. The distance is called MMD (Maximum Mean Discrepancy) and the two sample test is based resampling method. However, I found that the MMD based resampling method is not suitable for time series data. Recently, a wild bootstrap technique based MMD method has been reported for time series data. Thus, the wild bootstrap based MMD method is carried out for speaker verification system. Since the proposed method uses a kernel framework, some parameters are affected sensitively but a small experiment achieved a high accuracy score. In future work, the some conditions should be tested.

研究分野：音声信号処理

キーワード：話者照合　音声信号処理　統計的機械学習　カーネル法

１．研究開始当初の背景

本研究の対象とする話者照合とは、音声を用いた生体認証のことを指す。指紋や静脈を用いた生体認証では特別は装置を必要とするが、対して話者照合はマイクが一つあれば実現可能なシステムなので導入が簡単という利点がある。しかしながら現在の話者照合システムはまだ認識性能が十分ではないという問題がある。この原因の一つとして，音声は環境雑音や発声変動により歪んでしまうためモデル化が難しいことが挙げられる。生体認証には金融取引や施設・携帯のセキュリティなど幅広い実用化が期待される一方で，認識誤りが致命的であることから高い認識性能が要求される。そのため，話者照合の頑健性，識別性能の向上が必要不可欠である。1996 年以降, 米国立標準技術研究所 (NIST) の主催する話者照合に関するコンペティション NIST SRE [2] が定期的に開催されるなど，世界各国の研究機関が活発に話者照合に関する研究を行っている。話者照合における代表的なアプローチには，混合ガウスモデル (GMM) を用いた話者のモデル化[1]，サポートベクターマシン(SVM) [2]を用いた識別などがある。特に近年では, 接合因子分析(JFA)という因子分析に基づいた手法が提案され[3], 従来手法よりも高い識別率を得ることができると NIST SRE [4] でも紹介されたことから JFA や i-vector[5]と呼ばれるモデル化手法が新たな基盤技術となりつつある。しかし、話者照合を実用化するためには実環境を想定する必要がある。そのため、想定されることとして、環境雑音やチャンネル雑音の影響に依存しないことやユーザの発話長が非常に短いことや体調等による声質の変動があることなどが挙げられる。これらの環境に置いて，JFA を用いたとしても十分な性能が得られていないという問題があり、話者照合の技術に対して、さらなる識別率の改善が求められている。

２．研究の目的

音声を用いた生体認証手法である話者照合では、環境雑音や発声変動により入力音声が歪むために、その性能は著しく低下する。JFA などの従来の話者照合では登録話者の音声から話者モデルを予め学習しておき、照合時には入力された音声との類似度を計算するという手段をとってきた。しかしながら、モデルを学習するということは、ある程度使用される環境を予め考慮しておく必要があり、環境雑音や発声変動に対する頑健性が十分ではないという問題点がある[6]。そこで、本研究では環境雑音や発声変動の歪みを吸収し、話者の安定した特徴を捉えることが出来る話者照合技術を確立することを目的とする。本研究で用いる手法の基となるカーネル法はカーネル回帰子による非常に表現力の高いモデルを内包している。本研究で用いる枠組みにおいては、照合時に高次元のカーネル空間で入力音声の分布と登録話者の音声の分布との距離を直接測るので詳細な差異を頑健に捉えることができる。本研究の成果により、セキュリティとして話者照合を普及させることや映像・音声のビッグデータの自動分類や音声による検索など、話者照合技術が真に活用されるようになることが期待できる。

３．研究の方法

本研究では、モデルの学習を行わないでサンプル間の距離をカーネル空間上で測る最大平均差異（MMD）[7]という指標を用い話者照合を行う。MMD 統計的機械学習の分野ですでにシミュレーション実験においてモデル学習を行わなくても分布間の識別を高い性能で行えることが証明されている手法である。すでに提案されている手法では互いに独立したサンプルに対しての高い性能が証明されており、音声のような高次元のデータに対しても識別を行うことが可能であることも証明されていた。そこで、MMD を提案した著者であり研究協力者でもある Arthur Gretton 氏とも議論を交わし、話者照合実験に適用することを行った。カーネル法に基づく手法では、カーネル関数の設定に微調整を行う必要があり、実験としては登録データと評価データおよびテストデータの 3 部に分けて適切な関数を求めつつ高い識別性能を得られるかどうかの検討を行った。話者照合の実験においては、NIST SRE で配布されている大規模データが標準タスクとして広く用いられているが、本手法ではまず簡単なタスクで適切に動くことを確認するために、防音室で収録された同時期の女性話者の音声データを用いた話者照合実験を行った。NIST SRE データは雑音およびチャンネルノイズが大きく、また、サンプリング周波数も低いというかなり厳しい条件での収録と成っている一方、学習に用いるデータは十分に用意されている。しかしながら実用性を考えるなら、学習や登録に用いるデータ自体もそれほど充実して集めることは難しいと考えられる。そこで本研究では、高い識別能力とともに短い発話だけで対応可能な手法ということに着目して実験を行う。具体的には適切なカーネル関数の他、入力に用いる特徴量の抽出方法や次元数、最終的な識別の決定方法および収録時期の変動などを考慮して実験を行っている。

４．研究成果

研究の当初は、MMD[8]で音声データであっても高い識別性能があることが期待されたが、実際に MMD という距離は時系列に依存したデータでは適切に動かないことがわかった。従来の話者照合手法である GMM などを用いた手法でも音声の時間変動については考慮していないため、MMD を用いた分布間の距離を時系列を無視して測るために bootstrap 法を用いることを行った。しかしながら、入力する特徴量として MFCC の静的特徴量と動的特徴量を用いることや、静的

特徴量を用いることなど時系列性を取り除くことを行った結果も、十分と言える性能を得られなかった。そこで、議論を交わした結果、Arthur 氏が時系列性を仮定した場合の MMD を用いた検定方法についての論文を発表し、本研究ではその手法を用いることを行った。具体的には wild-bootstrap 法[9]を用いた検定手法となる。これは MMD を用いて計算した分の距離を V-統計量として捉え、V-統計量から bootstrap 法を用いて再標本化する方法である。 ここで wild-bootstrap 法で再標本化される MMD の値は以下のように定義される。

$$\begin{aligned} M\hat{M}D_{k,b} =& \frac{1}{n_x^2}\sum_{i=1}^{n_x}\sum_{j=1}^{n_x}\bar{W}_i^{(x)}\bar{W}_j^{(x)}k(x_i,x_j) \\ & - \frac{1}{n_y^2}\sum_{i=1}^{n_y}\sum_{j=1}^{n_y}\bar{W}_i^{(y)}\bar{W}_j^{(y)}k(y_i,y_j) \\ & - \frac{2}{n_x n_y}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\bar{W}_i^{(x)}\bar{W}_j^{(y)}k(x_i,y_j) \end{aligned}$$

ここで、x, y はそれぞれ入力サンプル、nx,ny はサンプル x,y それぞれのサンプル長、W は時系列性を考慮した係数、k はカーネル関数を表す。文献[5]で紹介されている MMD とは異なり再標本化の際に時系列性を考慮するため、wild-bootstrap 法ではさらに時系列性に関する関数も設定をあらかじめ適切に設定する必要がある。そこで、本実験ではまず、同時期の複数の女性話者によって収録された音声データを用い、話者照合実験をおこなった。話者数は 5 名，各話者の発話数は 15 文章。1 文章につき約 3，4 秒の発話長となっている。特徴量としては MFCC（19 次元＋パワー）および 1，2 次動的特徴量を付与した合計 60 次元のものを用いた。実験結果を表 1 に示す。表 1 より，一文章単位の識別結果は FAR（False Acceptance Rate; 他人受理率）と FRR（False Rejection Rate; 本人拒否率）がそれぞれ 14.7%, 9.9%となった。本手法では、試行ごとに文章を受理または拒否するための閾値が自動的に設定される。本実験では再標本化された分布の有意水準 5%を閾値として算出している。従来の話者照合では、FAR と FRR の値が等価となる点を EER (Equal Error Rate)と呼びその値を性能比較に用いる。これは、従来の手法では閾値を全話者で事前に一定に定める必要があり、閾値の設定によって FAR と FRR が同時に設定されるためである。しかし、本手法では閾値が試行ごとに変化するため、EER を求めることが困難であるため、FAR と FRR の値で比較している。一文章ごとの結果は FAR・FRR 共に 10%程度となっている。ここでさらに一文章ごとではなく、登録文章全てにおける識別結果を元に多数決を取って最終的に受理または拒否を決定する方法を用いた結果が表 1 の多数決の結果となる。本

表 1 実験結果

		FAR	FRR
Memory -0.96	一文章	14.	9.
	多数決	0	0
Memory -0.95	一文章	12.	12.
	多数決	0	0

実験では各話者の登録文章数は 15 文章なので、この場合は 8 文章以上が受理なら入力文章を登録話者として受理、7 文章以下なら拒否するという手順になる。結果より、多数決で決定する場合には Memory 変数に関わらずエラー率が 0%という非常に高い識別性能を得られることが確認できた。この結果より、Wild Bootstrap 法を用いた MMD 基準を用いた話者照合が高い識別性能を持つことがわかった。本実験は初期実験であるため、収録環境は防音室内、テスト文章と登録文章の収録時期が同時期という話者照合としては非常に簡単なものではあったが、従来法とは異なり、モデルを事前に学習することなく高い識別性能が得られたことは非常に有意義な結果であると言える。次に、データベースを別もので実験を行った。収録環境は先ほどの実験と同様、防音室であるが、テスト文章と登録文章の発話時期が異なり、また、発話内容も異なる。このときの結果は FAR が 35.7%、FRR が 19.8%となった。これはデータが変わったため、適切なカーネル関数の設定及び、時期が異なることから用いる時系列性に関する関数の設定などが適切でないことが考えられる。また、あわせて入力特徴量の次元数についても MFCC（12 次元＋パワー）にした合計 39 次元のものを用いた。それぞれの値を設定し、多数決で結果を求めたところ FAR26.0%、FRR13.3%となった。これは、使用したカーネル関数が時期差に関して敏感になりすぎており、時期が異なるときの変動まで識別してしまうためだと考えられる。本実験ではさらに特徴量の正規化についても検討を行ったが改善は得られなかったと考えられる。

今後の課題としては、まず、時期差がある場合にはどのように高い性能を得ることができるかを調べた上で、当初の目的であった雑音環境下など、より複雑な環境において提案法が高い識別性能を発揮できるよう条件を整えていくことが挙げられる。また、提案法は発話長に依存しないため、i-vector などの従来手法で問題であった少量のデータ量に関しても十分な性能が得られることが期待できる。そのため、短時間発話に対する性能に関しても調査し、発表する予定である。

＜引用文献＞

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” Digital Signal Processing, 10, 19-41 (2000).
[2] W. M. Campbell, D. E. Sturim, and D. A.

Reynolds, “Support vector machines using GMM supervectors for speaker verification,” IEEE Signal Processing Letters, 13, 308-311 (2006).
[3] Joint Factor Analysis Matlab Demo: http://speech.fit.vutbr.cz/software/joint-factor-analysis-matlab-demo
[4] NIST Speaker Recognition Evaluation (SRE) http://www.itl.nist.gov/iad/mig/tests/sre/
[5] N. Dehak, et. al., “Front-End Factor Analysis for Speaker Verification,” IEEE Trans. Vol. 19, no.4, 2011.
[6] Q. Jin, et. al., “Overview of Front-end Features for Robust Speaker Recognition,” APSIPA ASC, 2011.
[7] A. Gretton, et. al., “Optimal kernel choice for large-scale two-sample tests,” NIPS 2012.
[8] A. Gretton, et. al., “A kernel two-sample test,” J. March. Learn. Res., 13:723-773, 2012.
[9] K. Chwialkowski, et. al., “A Wild Bootstrap for Degenerate Kernel Tests,” ArXiv e-prints, 2014.

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　1　件）

①小川 哲司,塩田 さやか,“ i-vector を用いた話者認識，”日本音響学会誌 ISSN-03694232，一般社団法人日本音響学会 70 巻 6 号，pp.332-339，2014-06-01.http://ci.nii.ac.jp/naid/110009823396/，査読なし.

〔学会発表〕（計　1　件）

①塩田さやか，松井知子，貴家仁志，“日音響ノイズを用いた話者照合の検討”日本音響学会秋季大会，2014-09-05，北海道学園大学（北海道）.

〔図書〕（計　0　件）

〔産業財産権〕

○出願状況（計　1　件）

“生体検知装置、生体検知法およびプログラム，”発明者および権利者：山岸順一，塩田さやか，小野貴順，越前功，松井知子，特許 P150011538，2014-08-19，国内.

○取得状況（計　0　件）

〔その他〕

ホームページ等

http://www-isys.sd.tmu.ac.jp/Members/sayaka

６．研究組織

(1) **研究代表者**

塩田　さやか（SHIOTA, Sayaka）
首都大学東京・システムデザイン学部・助教
研究者番号：90705039

(2) **研究協力者**

松井 知子（MATSUI, Tomoko）
小川 哲司（OGAWA, Tetsuji）
貴家 仁志（KIYA, Hitoshi）
Konstantin Markov
Arthur Gretton
Gareth Peters