携帯端末への話者照合を用いたセキュリティロック
A security lock by speaker verification using a smart phone

山室 慶太
Keita Yamamuro
法政大学情報科学部ディジタルメディア学科
E-mail:keita.yamamuro.4m@cis.hosei.ac.jp

Abstract

Advance of information technology makes the handheld device has information of many kinds. These include important information. Conventional security Locke may not completely protect information. This article proposes Security Locke by the speaker verification to an android device. Speaker verification used to record voice sample in an android device. The sound samples were recorded a Japanese Newspaper Article Sentences of a month for 20 speakers. The acoustic models were updated by putting an old sound sample and a new sound sample together. The recognition results were got 6.0 % of false rejection rate and 12.6 % of false acceptance rate by result of performance experiment.

1 まえがき

情報技術の発達により，RFID 技術を用いた電子マネーや PIM による個人情報の管理などの様々な個人情報を携帯電話などの一つの携帯端末で管理・利用することが多くなってきている．これらの端末にはパスワードを認証キーとしたセキュリティが掛かっているが，パスワードによる認証では忘却や紛失によって本人でも認証できなくなることや，漏洩や盗難によって他人が認証される恐れがある．そこで生体認証の一つである話者認証に注目した．生体情報である音声を用いる話者認証を利用することで認証キーの忘却や紛失，他人への漏洩などの危険性を減らすことができる．また，パスワードとの併用によってより強固なセキュリティロックを実装することも考えられる．

話者認証技術は現在までに，雑音に強固な音声認証として文献 [1] や文献 [2] のような高精度な認証精度を持たせる研究や，文献 [3] のような多様な音響環境下における音声データ収集システムの構築などの性能を向上させるための研究がされている．しかし，生体認証のひとつである静脈認証はいくつか製品化されているが，話者認証が実用化された例はあまりないのが現状である．静脈の場合，静脈を読み取るための特別な機材が必要なため小型化が難しいという問題点を抱えている．それに対し，話者認証技術はマイクを利用するため特別な機材が必要なく，簡単な機材で認証を行えるため，携帯電話のような音声入力端子がある携帯端末と相性が良く，また誰でも簡単に利用できる音声を用いることで機械操作に不得意な人でも直感的な操作を可能にできるなどの利点が考えられる．

しかし，このような話者認識システムを利用するには照合精度があまり高くないという問題点がある．その原因の一つとして人の音声は変化，あるいは変動することが挙げられる．音声の変化とは声変わりのような発声器官の生理変化によるもので，音声の変動とは風邪などの体調変化のような調音法の変化である．そのため同じ音声データを用いた話者モデルを利用し続けると認識率が低下する恐れがある．このようなことを避けるため話者モデルを定期的に更新する必要がある．しかし，話者モデルを更新するためには新たに音声データを収集する必要があり，話者への負担が考えられる．

そこで，本論文では電話時の会話音声を学習データとして再利用することを提案する．モデルを更新するための学習データを新たに録音するかわりに，この方法を用いることで話者への負担を減らすことができる．また，会話音声を利用し話者モデルの更新を行うことで認識率の低下を防ぐ．

2 話者照合システム

話者照合システムとは，音声の特徴を用いた個人の照合システムである．先行研究として，話者照合は文献 [4] にある富士通の VoiceGATEII というシステムで利用されている．この VoiceGATEII は音声によるパスワードで話者照合を行うシステムで，電話やインターネット経由の話者認証によりユーザーが簡単にテレホンバンキングや通信販売が利用できる．このシステムの利点として，電話を使うため特別な機材が必要ない，暗証番号の発話によって番号と音声の 2 つの認証を行える，認証キーを発話するだけなので難しい操作を必要としない，といったものが挙げられる．

また，本研究では話者照合に繰り返し更新を行った話者モデル用いることを想定している．これまでに話者照合の話者モデルの作成方法について文献 [10] のような研究が行われている．人の音声は時期により特徴が変化してしまい，同じ話者モデルを使い続けていると認識率に影響が出てしまう．この研究では複数の時期に録音した 10 文章を使い周期的にモデルの更新を行った．その更新したモデルと未更新のモデルを本人棄却率と他人受理率の二つの誤り率を用いて性能を比較している．その結果，更新したモデルの誤り率は未更新のモデルの誤り率よりも約 4 割減少し，この手法の有効性を示していた．

これらの研究から，話者認証を用いたシステムは誰でも簡単に利用できる音声を用いることで，機械操作に不得意な人でも直感的な操作を可能にできることや，自分の声を認証キーとするためキーを持ち歩く必要もなく，忘れることもないという利点が考えられる．その他にも，特殊な取り込み装置が不要な点や会話音声から学習データを取得できる点から携帯電話などの端末との相性が良いことが考えられる．

3 android 端末へ実装する話者照合システムの概要

話者照合を行うシステムとしては，音声データの収集を行う図 1 と話者モデルを用いて話者照合を行う図 2 の二種類の工程で行われるものを想定している．

図 1 の工程では，話者モデルを構築するための音声データの収集を行う．まず，携帯端末上で会話音声の録音を行う．録音した音声は一時的に携帯端末内でまとめて保管しておく．この音声データはすべて定期的に TCP 通信によってサーバー側の PC へ転送される．音声データを携帯端末で保管せず，サーバーで保管することで多くの音声データを収録日時など話者モデルに利用する情報と一緒に管理する．このようにすることで多くの音声データによる携帯端末の保存領域への負荷を軽減する．また「各年の 12 月の音声データ」といったより詳細な情報

Supervisor: Prof. Katunobu Itou

を持つ話者モデルを構築することができると考えられる．

図 2 の工程では，話者モデルを用いて話者の照合を行う．まず，収集した音声データを利用して話者モデルの構築を行う．このとき，すでに話者モデルを構築している場合はその際に利用した音声データと新しく登録された音声データを使い，話者モデルの更新を行う．構築した話者モデルは TCP 通信によって携帯端末に転送される．この話者モデルとマイク入力された音声データを用いることで話者照合を行う．

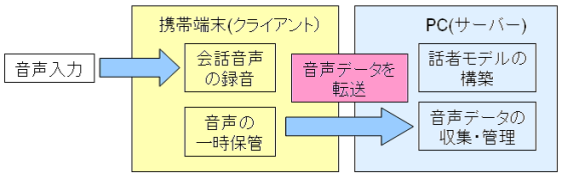

図 1. 音声データ収集の工程

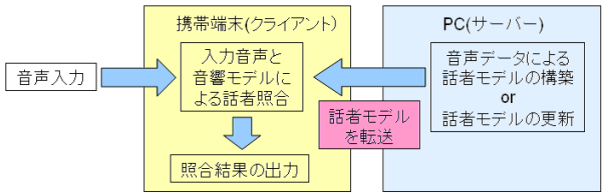

図 2. 話者照合の工程

3.1 音声特徴量の抽出

話者照合に用いる話者モデルを構築するために各話者の音声情報を音声特徴量として音声データから抽出する必要がある．今回の研究では MFCC，ΔMFCC，基本周波数（F0）情報，ΔF0，Δ 対数パワーをそれぞれ特徴量として用いている．

3.2 ケプストラム分析

音声を振幅伝達特性と振幅スペクトルに分離する信号処理をケプストラム分析と呼ぶ．手順としては音声波形のパワースペクトルを対数変換し逆フーリエ変換を行う．またその結果として計算した値をケプストラム係数と呼ぶ．このケプストラム係数の低次には振幅伝達特性が高次には音源特性が反映されている．そのため，低次のケプストラム係数を用いることでスペクトルの包絡を図 3 のように表現できる．この話者照合ではこのケプストラム係数を音声特徴量として用いるために音声データから抽出する．[6]

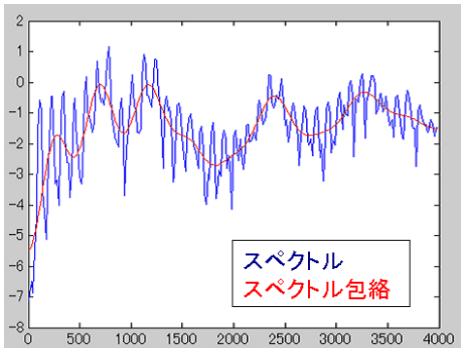

図 3. 「あ」のスペクトルとスペクトル包絡

3.3 MFCC パラメータ

ケプストラムのパラメータにはいくつか計算方法がある．その中でも今回は MFCC(メル周波数ケプストラム係数) をスペクトルから計算することで音声特徴量の抽出を行う．MFCC の計算ではメル周波数軸上で等間隔に配置した三角窓によりフィルタバンク分析を行う．フィルタバンク分析とは周波数軸上に配置された複数のフィルタ群の出力に基づき行うスペクトル分析のことである．またメル周波数軸上でフィルタバンク分析を行う利点は，メル周波数が人の低い周波数は細かく聞き取り，高い周波数は粗く聞き取るという感覚尺度に近く，音声のような人が聞いて違いがわかる音の分析に向いているとおもわれるためである．最終的に，フィルタバンク分析により得られた各帯域におけるパワーを離散コサイン変換することで MFCC を求めることができる．この MFCC は逆離散コサイン変換をすることでスペクトルに戻すことができる．そのため MFCC を 0 次を除去することで直流成分の除去などの音声処理も行うことができる．また，ΔMFCC は時間軸方向へ差分をとることで MFCC の時間的変化量を特徴として用いている．

図 4 は男性話者，女性話者，多数話者の 50 音すべての MFCC を平均化したものである．

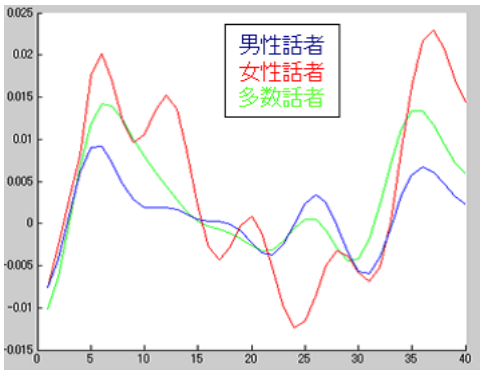

図 4. 各話者ごとの MFCC の平均

3.4 基本周波数（F0）の情報

F0 とは人の声の高さの情報で声の最も低い周波数である．MFCC は声道の特性を表現しているため，音韻的な情報を特徴としている．それに対し F0 の情報は韻律の情報を特徴としている．文献 [8] などでは F0 情報は個人性を含むとされ，これまでに韻律情報を利用した話者認識の研究がいくつかされている．また，F0 情報を用いることで認識性能の対雑音性が向上することが報告されている [2]．そのため今回音声特徴量として MFCC のほかに F0 とその差分である ΔF0 の情報を利用した．

F0 情報の抽出には STRAIGHT-TEMPO を用いている．複数の手法を使い雑音の影響下で F0 の情報を求めている文献 [9] において STRAIGHT-TEMPO は高精度な推定を行うことができるという結果が出ている．本研究で扱う音声データには雑音が混ざっているものも多くなるため STRAIGHT-TEMPO を使い特徴量を抽出した．また，ΔF0 は F0 情報の前後 10ms ごとの値から最小 2 乗法によって得られる傾きを ΔF0 としている．

F0 と ΔF0 をそれぞれ 10ms ごとに求めた結果は次の図 5 のようになる．

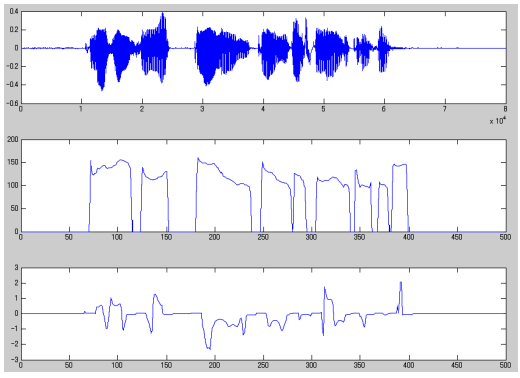

図 5. 上：元の音声波形，中：F0 情報 下：ΔF0 情報
発話内容「この文章を丁寧に声に出して読んで下さい」

3.5 GMM の基本構造

話者認識に使われる話者モデルは HMM（隠れマルコフモデル）や GMM（混合ガウス分布モデル）と呼ばれるモデルが多く用いられている．HMM は音声認識などでよく使われている

話者モデルであるが，HMM は話者モデルを構築するには学習データを多く用意する必要があるという欠点がある．そのため少ない学習データでモデルを構築できる GMM（混合ガウス分布モデル）が用いられる．

確率分布関数を用いる際に，ガウス分布は単一のピークを持つ分布関数であり，複雑な分布形状を表現することはできない．そこで，複数のガウス分布の重みつき和を用いて複数のピークを持つような分布を表現するため混合ガウス分布が用いられる．多くの HMM では各状態ごとにこの混合ガウス分布を用いて出力確率分布を構成している．それに対し，各ガウス分布に重み付けをし，一つの状態にまとめたものを GMM と呼ぶ．そのため状態数に違いがあり，HMM では多くの場合に状態数は３であるが，GMM での状態数は１となる．

混合ガウス分布は平均 μ，分散 σ^2 を持つガウス分布を $N(o;\mu,\sigma^2)$ とすると，

$$f_o(o) = \sum_{w=1}^{W} \lambda_w N(o;\mu_w,\sigma_w^2) \tag{1}$$

$$\sum_{w=1}^{W} \lambda_w = 1 \tag{2}$$

によって表され，分布を規定するパラメータは状態ごとに定まる $\Theta = \{\lambda_w, \mu_w, \sigma_w^2\}$(混合の重み，平均，分散) である．混合分布では，次のようにパラメータ推定を行うことができる．

$$\hat{\lambda}_m w = \frac{\sum_{n=1}^{N} \varphi(n,m) \frac{\lambda_w N(o(n);\mu_w,\sigma_w^2)}{f_o(o(n))}}{\sum_{n=1}^{N} \varphi(n,m)} \tag{3}$$

$$\hat{\mu}_m w = \frac{\sum_{n=1}^{N} \varphi(m,n) \frac{\lambda_w N(o(n);\mu_w,\sigma_w^2)\cdot o(n)}{f_o(o(n))}}{\sum_{n=1}^{N} \varphi(m,n)} \tag{4}$$

3.6 話者モデルの学習

音声データから抽出した音声特徴量を用いて GMM の学習を行う．GMM の学習には EM アルゴリズムを用いる．

EM アルゴリズムは，観測できない隠れたパラメータを最尤推定を行い求める手法である．EM アルゴリズムは反復法によって局所最適解を求めるアルゴリズムで，モデルの尤度の期待値を計算する式（5）とそこで求めた尤度の期待値を最大化するようなパラメータを求める式（6）を交互に繰り返す．

$$Q(\theta|\theta^t) = E_Z|x,\theta^t[logL(\theta;x,Z)] \tag{5}$$

$$\theta^{(t+1)} = argmaxQ(\theta|\theta^t) \tag{6}$$

4 音響モデルの性能実験

話者照合の識別精度を確認するために話者モデルの性能実験を行う必要がある．累積した様々な音声データにあわせて，音声の変化を表現しやすいようにモデルの構造を複雑化しながら話者モデルを再構築することが理想的である．そのため今回は二つの話者モデルを用意し，それぞれデータを学習させたモデルの性能実験を行い比較した．

一つは混合数 32 の 1 状態で構築した 1 状態モデルである．1 状態で構築したモデルとは，学習を行う際に話者の音声特徴を 1 状態で学習させた話者モデルとなっている．

二つ目は混合数 4 の 5 状態で音節の構造を持たせ構築した 5 状態モデル（音節構造モデル）である．音節の構造を持たせ構築したモデルとは，学習を行う際に母音 (V) や子音＋母音 (CV) といった音節ごとに話者の音声特徴を学習させた話者モデルとなっている．

それぞれの話者モデルは各話者ごとの音声情報を学習させた話者本人情報を持つモデルと男女 19 名分の話者情報を学習させた多数話者の情報を持つモデルを 1 セットとして考えて構築されている．この話者モデルとテストデータから求めた尤度を閾値と比較することで話者照合を行い，本人棄却率と他人受理率を求めることで性能を評価した．これらの話者モデルを話者照合に用いた流れは図 6 のようになっている．

この実験は通話時などの会話音声を随時録音し，学習データとして蓄積させていくことを想定している．そのため，学習データには日程を変えて録音した音声を複数用意して，それらを時期別に分けた 3 つの GMM に学習させた．各日程ごとの学習データを使い話者モデルの更新を行うことでモデルの本人棄却率と他人受理率の変化を調査し，更新を行わなかった話者モデルとの比較を行った．

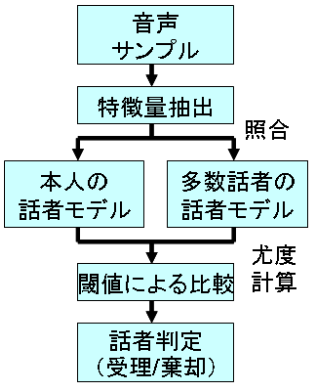

図 6. 照合性能評価の流れ

5 閾値の自動設定

閾値の設定は各話者のモデルごとに最適となる値が変わってくる．また，モデルを更新することによって，適切な閾値が変化するため新たに設定する必要がある．そのため，今回この閾値は話者モデルと同じく学習データを用いて毎回閾値の再設定を行っている．閾値の設定には更新した話者モデルと学習データを用いて一度認識を行い，その結果から尤度を平均した値を用いている．

閾値の設定は更新した話者モデルと学習データの結果から本人と認識したモデルの尤度と，他人と認識したモデルの尤度の平均をそれぞれ求め，その割合を使って各話者ごとに設定している．この閾値は更新したモデルとテストデータから求めた尤度を用いて話者の判別を行っている．

5.1 音声データ

実験に用いた音声データには話者 20 名に発話してもらった音素バランス文を約一ヶ月間にわたって録音したものを用いた．まず，初期学習用の音声データとして各話者ごとに 20 文の音素バランス文を収録した．その後，複数の日程で録音を行い，それぞれの日程で話者モデルのアップデート用学習データとして音素バランス文から 10 文，テストデータとして音素バランス文から 20 文収録した．録音したデータのパラメータはサンプリング周波数 16kHz，量子化ビット数 16bit となっている．この音声データは 1 から 12 次元までの MFCC12 次元とその 1 次差分，F0 情報を 1 次元，ΔF0 情報を 1 次元，Δ 対数パワー 1 次元の計 27 次元の音声特徴量をフレーム長 25ms，フレーム周期 10ms で抽出し，話者モデルの学習に用いた．音声の録音には android 端末 (docomo の HT-03A) の内蔵マイクを使って行った．

5.2 認識性能の評価

話者照合のシステム性能を評価する時，本人が棄却される誤り率（本人棄却率）と他人が受理される誤り率（他人受理率）が用いられる．これら 2 種類の誤り率と判定の閾値との関係はトレードオフとなっている．閾値を大きくすれば他人受理率の誤りが大きくなり，小さくすれば本人棄却率の誤りが大きくなる．一般に閾値は 2 種類の誤りの相対的な重要性に従って設定

表 1. 実験データ

収録期間	一ヶ月間 (2009 年 12 月)
収録人数	男性話者 10 名
サンプリング周波数	16kHz
量子化ビット数	16bit
パラメータ	MFCC12 次元、Δ MFCC12 次元 F0 1 次元、ΔF0 1 次元 Δ対数パワー 1 次元 （計 27 次元）

表 2. 音素バランス文

1. あらゆる現実をすべて自分のほうへねじ曲げたのだ。
2. 一週間ばかりニューヨークを取材した。
3. テレビゲームやパソコンでゲームをして遊ぶ。
4. 物価の変動を考慮して給付水準を決める必要がある。
5. 救急車が十分に動けず救助作業が遅れている。
(この他にあと 45 文を使用)

される．

5.3 実験結果

更新を行ったモデルと未更新のモデルから本人棄却率と他人受理率を求めた結果を図 7 と図 8 に示す．この結果は 3 回モデルの更新を行い，それぞれの認識結果の平均となっている．

1 状態モデルの認識率を比較した結果，本人棄却率の平均はそれぞれ未更新モデルが 8.1 %，更新したモデルが 6.0 %となり，モデルの更新を行うことで本人棄却率が未更新のものより約 25.0 %減少した．また他人受理率の平均はそれぞれ未更新モデルが 18.6 %，更新したモデルが 12.6 %となり，モデルの更新を行うことで他人受理率が約 30.0 %減少した．

次に 5 状態モデルの認識率だが，本人棄却率の平均はそれぞれ未更新モデルが 19.3 %，更新したモデルが 12.1 %となり，モデルの更新を行うことで本人棄却率が未更新のものより約 35.0 %減少した．また他人受理率の平均はそれぞれ未更新モデルが 26.1 %，更新したモデルが 18.0 %となり，モデルの更新を行うことで他人受理率が約 30.0 %減少した．こちらはモデル更新によって性能は向上したが，元々の認識率が悪かったため 1 状態モデルよりも性能はやや低い結果となった．その原因としては今回用いた話者モデルの構造に対して，学習データ数が少なかった可能性などが考えられる．

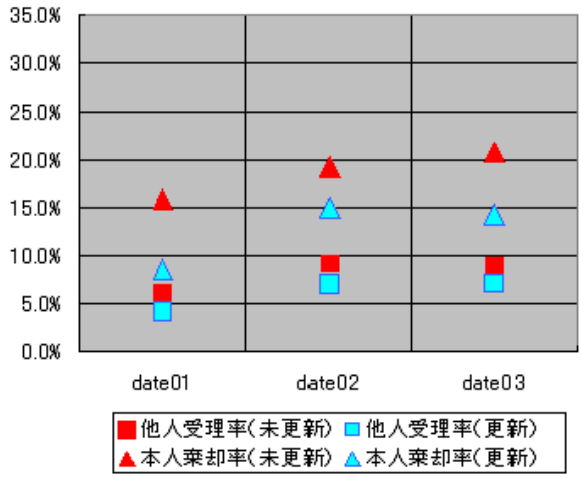

図 7. 本人棄却率と他人受理率の比較 (1 状態モデル)

6 あとがき

本研究では携帯端末向けの話者照合によるセキュリティロックを行うため，android 端末で録音した音声を用いて話者照合実験を行った．その結果，モデルを更新することで本人棄却率は未更新のモデルより約 25.0 %減少した 6.0 %，他人受理率は

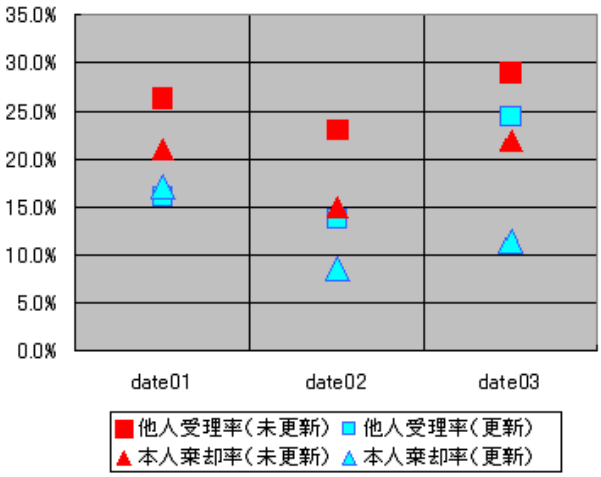

図 8. 本人棄却率と他人受理率の比較 (5 状態モデル)

未更新モデルより約 30.0 %減少した 12.6 %となり，どちらも未更新のモデルの認識結果よりも良い結果となった．

今後，携帯端末向けに話者照合の実装を目指すためには，より照合精度を上げる必要があると考えられる．そのためには学習データを多く蓄積できる利点を生かし，より音声の変化を表現しやすいようにモデルの構造を複雑化しながら話者モデルの更新を行う方法などが挙げられる．今回，話者モデルは 1 状態のものと 5 状態のものを構築したが，5 状態モデルは認識率があまり良い結果にはならなかった．そのため，学習データ数を増やすなど，様々な構造の話者モデルを構築することで照合精度を向上させる方法について検討していく．また，今回のデータ収録期間は一ヶ月間であったが，人の音声特徴は風邪などの個人の体調によって変化する．そのため，より長期的なデータの収集を行うことでこのような変化にどの程度対応することができるか検討していく必要がある．

参考文献

[1] 浅見太一, 岩野公司, 古井貞熙, “ ハフ変換による基本周波数情報を用いた雑音に頑健な話者照合 ” 日本音響学会 2004 年春季講演論文集, 3-Q-17, pp.177-178 (2004/03)

[2] 浅見太一, 岩野公司, 古井貞熙, “ 雑音に頑健な話者照合のための基本周波数情報の利用 ” 信学技報, SP2004-15, pp.1–6, （2004/03）.

[3] 原直, 宮島千代美, 伊藤克亘, 武田一哉, “ 多様な音響環境下における音声認識システム利用時のデータ収集システム ” 電子情報通信学会 2007 年論文誌

[4] 井上美明, 熊倉敏, “ 音声による話者照合システム「Voice-GATEII」, および話者識別システム「VoiceSync」” 日本機械学会 2000 年論文集, (2000/03)

[5] 板橋 秀一, “ 音声工学 ” 出版社: 森北出版 (2005/02)

[6] 鹿野 清宏, 河原 達也, 山本 幹雄, 伊藤 克亘, 武田 一哉, “ 音声認識システム ” 出版社: オーム社 (2001/05)

[7] 古井貞熙, “ 新音響・音声工学 ” 出版社: 近代科学社 (2006/09) 日本音響学会 2006 年春季講演論文集, 1-11-20 (2006/03)

[8] 松井知子, 古井貞熙, “ 音源・声道特徴を用いたテキスト独立形話者認識, ” 信学論, vol.J75-A, no.4, pp.703- 709 (1992-4).

[9] 石本 祐一, “ 時間情報と周波数情報を用いた実環境雑音下における基本周波数推定 ” 電子情報通信学会技術研究報告, vol.103, No.750, pp.49-54, Mar. 2004.

[10] 松井知子, 古井貞熙, “ 話者照合におけるモデルとしきい値の更新法 ” 電子情報通信学会技術研究報告, IEICE technical report. Speech 95(468) (1996/01/19)