論文 / 著書情報
Article / Book Information

論題(和文)	音声と耳介画像情報を用いたマルチモーダル話者照合の高精度化
Title(English)	
著者(和文)	宮崎 太郎, 浅見 太一, 岩野 公司, 古井 貞熙
Authors(English)	Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会 2004年秋季講演論文集, Vol. , No. 2-4-7, pp. 99-100
Citation(English)	, Vol. , No. 2-4-7, pp. 99-100
発行日 / Pub. date	2004, 9

音声と耳介画像情報を用いたマルチモーダル話者照合の高精度化 *

◎宮崎 太郎　浅見 太一　岩野 公司　古井 貞熙 (東工大)

1　はじめに

我々の研究室では，モバイル環境での手軽でかつ高精度な個人認証の手法として，音声と耳介画像を用いたマルチモーダル個人認証を提案している [1]. これは，音声の欠点である時期変動や周辺雑音による認証精度の劣化を軽減するために，音声と共に耳介画像情報を用いる手法である．耳介は時期差による変化が非常に小さく，安定した情報であり，また個人認証に必要な生理学的な特徴を持つことが報告されている [2]．しかし，これまでの結果では，特に耳介画像のみを用いた個人認証の精度がそれほど高くないために，システム全体として，高い性能が得られていなかった.

そこで本稿では，特に耳介画像を用いた個人認証の精度を向上させることで，システム全体の高精度化を目指す.

2　提案手法の概略

本研究で用いたシステムの概略図を図 1 に示す.

従来法 [1] においては，耳介画像からの特徴量抽出に主成分分析 (PCA) のみを用いていた．本稿ではそれに加え，近年，顔画像を用いた個人認証などに用いられその効果が報告されている，独立成分分析 (ICA)[3] を用いて特徴量を抽出し，PCA による特徴量と併せて使用する．画像特徴量は GMM によってモデル化し，PCA，ICA それぞれの特徴量ごとに特定話者モデル，不特定話者モデルを作成する.

音声特徴量は HMM を用いて，特定話者モデル，不特定話者モデルを作成する.

照合には，入力特徴量の，申告話者モデルに対する対数尤度を用いる．その際，不特定話者モデルに対する入力特徴量の対数尤度を，申告話者モデルに対する対数尤度から引くことで，正規化を行う [4]. この，正規化を行った対数尤度のことを，以下では「照合スコア」と呼ぶ．2 種類の画像特徴量による照合スコアに重み付けして足し合わせたものを融合画像スコアとし，さらに，得られた融合画像スコアと音声スコアを重み付けして足し合わせたものを融合スコアと呼ぶ．この融合スコアが閾値を越えたとき，入力特徴量が申告話者のものであると判断する.

こうして，音声による特徴量と，画像による特徴量 (2 種類) の，合計 3 つの特徴量を融合することで，話者照合の精度の向上を実現する.

3　実験条件

3.1　音声・耳介画像データ

音声・耳介画像の各データは時期差を考慮し，1ヵ月毎に 5 回に渡って収録した．収録話者数は 37 名で，全て男性話者である．各話者は 1 時期に 50 個の 4 桁連続数字を発声している．音声は 16kHz，16bit で標本化，量子化した．また，音声の収録と共に，各話者について，髪がかからないようにした右耳の正面からの画像を各時期に 1 枚撮影した．撮影時の解像度は 720×540 である．また，撮影条件を整えるためにフラッシュを利用した.

収録された 5 時期分のデータのうち，1〜3 時期目のものを学習データ，4，5 時期目のものを評価データとした．不特定話者モデルの学習データに含まれている詐称者と含まれていない詐称者を用意するために学習データを 19 名と 18 名の 2 グループに分ける．このようにすることで，各話者の評価データは「本人のデータ (1 人分)」,「不特定話者モデルの学習に含まれている詐称者のデータ (19 人分)」,「不特定話者モデルの学習に含まれていない詐称者のデータ (17 または 18 人分)」となる.

音声データについては，学習データには SN 比で 30dB の白色雑音を付加させ，評価用データに関しては SN 比で 5，10，15，20，30dB の白色雑音を付加させたデータを用意した.

画像データについては，特徴量を抽出する前に，人手により位置・角度を調整し，輪郭強調の処理を行い，80×80 の領域で切り出す．さらに，色情報を落として 8bits グレイスケールに変換し，円形領域に切り出したものを用いた.

3.2　音声・耳介画像特徴量

音声特徴量には，MFCC12 次元，ΔMFCC12 次元，Δ 対数パワー 1 次元の計 25 次元のベクトルを用いた．特徴量はフレーム長 25ms，フレーム周期 10ms で抽出し，入力音声毎に CMS を行っている.

画像に関してはまず，円形に切り出された画像を用いて，PCA と ICA により基底を作成した．それぞれの基底の作成には不特定話者モデル作成用のデータの 1 時期目のものだけを用い，前述の 2 つのグループに対しそれぞれ別の基底空間を作成した．得られた基底空間に耳介画像を投影したベクトルを画像特徴量として用いる.

3.3　音声・耳介画像のモデル化

音声特徴量は数字 HMM でモデル化を行い，画像特徴量は GMM でモデル化する．耳介画像については学習データが 1 話者あたり 3 時期分の 3 枚しか存在しないが，撮影の際の位置のずれを考慮し，上下左右に 2，4，6 画素，耳介の中心を画像の中心から移動した画像を作成する．また，それぞれに対し，うなずきなどによる耳介画像の回転を考慮して-30 °〜30 °まで 1 °ずつ回転させた画像を作成し，全てを学習に利用する．なお，評価用データに関しては平行移動の操作は行わなず，回転のみを行う.

∗ Improvement of a multi-modal speaker verification method using speech and ear images

By Taro Miyazaki, Taichi Asami, Koji Iwano and Sadaoki Furui (Tokyo Institute of Technology)

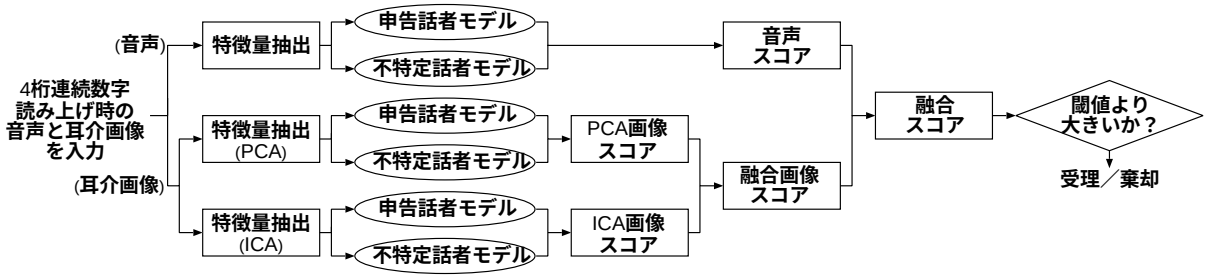

図 1. 本研究におけるシステム

表 1. 耳介画像のみを用いた場合の各手法の等誤り率 (%)

手法	等誤り率
PCA	7.0
ICA	16.2
PCA + ICA	5.6

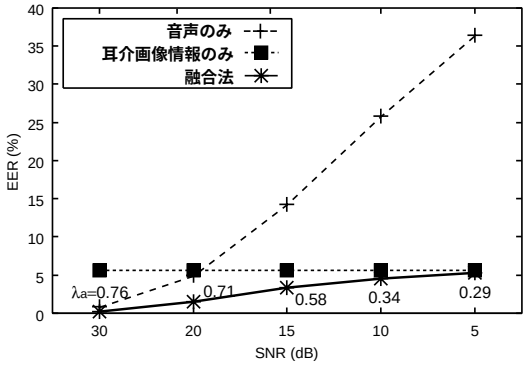

図 2. 3 種類の特徴量を融合した際の照合結果

4 照合実験結果

まず，耳介画像情報のみを用いて個人認証を行った際の，PCA，ICA のそれぞれについて単独で使用した場合と，PCA と ICA の結果を融合した場合の比較を表 1 に示す．それぞれの場合において，基底の次元数は実験的に最適なものを用いた．また，PCA と ICA の結果を融合する際の重みについても実験的に最適なものを利用した．ICA を単独で使用した場合では期待していた高い認証精度は得られなかったが，PCA，ICA を単独で用いる場合と比べ，この二つの手法を融合した場合に等誤り率が削減された．

さらに，この融合画像スコアを，音声スコアと融合した場合の評価を行った．音声のみを用いた場合，画像のみを用いた場合と融合法を用いた場合の比較を図 2 に示す．なお，図中の λ_a は融合する際の音声スコアの重みであり，各条件で事後的に最適化されている．全ての SN 比条件において，それぞれの手法を単独で用いた場合と比べ，融合法では照合性能の向上が確認できる．特にその傾向は SN 比の小さい場合に顕著で，SN 比が 15dB のときであれば，音声特徴量を単独で用いた場合と比べて 76.5%，耳介画像特徴量を単独で用いたときと比べて 40.4%の等誤り率を削減することができた．このことから，融合法は雑音環境に頑健な個人認証の手法であるといえる．

5 韻律特徴量の導入

韻律情報を用いた話者照合が雑音に対して頑健であることが報告されている [5]．そこで，文献 [5] の方法を，これまでに述べてきたマルチモーダル個人認証に適用することで，さらなる耐雑音性の向上を考える．

韻律特徴量には，$\log F_0$ と $\Delta \log F_0$ を用いる．これを 3.2 節で述べた音声特徴量と結合し，27 次元の新たな音声特徴量を作成する．特定話者モデル，不特定話者モデルは，音韻・韻律ストリームからなるマルチストリーム HMM によって構築される．

表 2. 韻律情報の有無による等誤り率の比較 (%)

SN 比	韻律情報なし	韻律情報あり
30dB	0.2	0.3
20dB	1.6	1.2
15dB	3.4	2.4
10dB	4.5	3.3
5dB	5.3	3.7

表 2 に，韻律情報を使う場合，使わない場合のそれぞれにおける，音声と耳介画像情報を融合したマルチモーダル話者照合の結果を示す．多くの SN 比条件において，等誤り率の改善が見られる．このことから，このマルチモーダル話者照合のシステムにおいても韻律情報が有効であることが確認できた．本システムにより，SN 比 5dB の条件で 3.7%の等誤り率を達成した．

6 まとめ

本稿では，音声と耳介画像を用いたマルチモーダル個人認証の高精度化を，2 種類の画像特徴量の融合することで実現した．また，韻律情報がこのシステムにおいても有効であることが確認できた．今後の課題としては基底の次元数，融合の際の各重みの自動設定などがあげられる．

参考文献

[1] 岩野公司, 広瀬智治, 上林英悟, 古井貞熙, “音声と耳介画像を用いたマルチモーダル話者照合,” 2003 年春季音講論, 3-3-3, pp.109-110 (2003-3).

[2] A. Iannarelli, *Ear Identification*. Forensic Identification series. Paramont Publishing Company, Fremont, California (1989).

[3] M. S. Bartlett, H. M. Lades, and T. J. Sejnowski, “Independent component representation for face recognition,” Proc. SPIE, Conf. on Human Vision and Electronic Imaging, pp.528-539.

[4] 松井知子, 古井貞熙, “テキスト指定形話者認識のための事後確率に基づく尤度正規化法,” 1993 年秋季音講論, 1-7-20, pp.639-640 (1993-10).

[5] 浅見太一, 岩野公司, 古井貞熙, “雑音に頑強な話者照合のための基本周波数情報の利用,” 信学技報, vol.104, no.87, pp.1-6 (2004-5).