論文 / 著書情報
Article / Book Information

論題(和文)	音声と耳介画像を用いたマルチモーダル話者照合
Title(English)	
著者(和文)	岩野公司, 広瀬智治, 上林英悟, 古井 貞熙
Authors(English)	Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会 2003年春季講演論文集, Vol. , No. 3-3-3, pp. 109-110
Citation(English)	, Vol. , No. 3-3-3, pp. 109-110
発行日 / Pub. date	2003, 3

音声と耳介画像を用いたマルチモーダル話者照合 *

◎岩野 公司　広瀬 智治　上林 英悟　古井 貞熙 (東工大)

1　はじめに

近年，携帯電話をデバイスとしたインターネット上のオンラインショッピングや電子決済などの利用が増加しており，モバイル環境における高精度な個人認証の必要性が高まっている．特に，人間の生体に固有の情報を用いる「バイオメトリック個人認証」は，盗難や偽造，紛失の恐れが少ないことから注目され，研究が進められている．

バイオメトリクスとしては，指紋・虹彩・網膜・音声・顔などがあげられるが，モバイル環境での利用を考えると，入力の手軽さや，入力機器の簡便性から，音声による話者照合が不可欠な認証技術となる．しかし，音声は 1) 時期差による特徴変動，2) 周辺雑音などの影響によって認証性能が劣化することから，音声のみでは実用的な性能を得ることが困難である．

そこで，他のバイオメトリクスを併せて利用し，音声の不利を補い，高精度な個人認証を行う「マルチモーダル・バイオメトリック話者照合」について検討を進めている．有力なバイオメトリクスとしては「顔画像」があげられ，音声との組み合わせによる認証性能の向上が報告されている．しかし，顔画像についても，髪型の変化・ひげの有無・化粧の有無などの時期差の影響を受けやすく，それによる照合性能の劣化が予想される．

そこで，音声と併せて「耳介画像（耳の形状情報）」を利用することを考える．耳介を用いた個人識別の実現可能性は古くから指摘され，文献 [1] では，耳介情報に個人を識別する生理学的な特徴が含まれていることが報告されている．耳介情報のみを用いて自動的な個人識別を行った例には，文献 [2], [3] などがあるが，未だ研究例が少なく，十分な認証性能が得られていない．しかし，耳の形状は時期差による変化が極端に小さいことがわかっており [1] ，音声による話者照合に耳介画像情報を組み入れることで，時期差や雑音に対する頑健性が向上するものと期待される．

以上の観点から，音声と耳介画像を組み合わせた話者照合手法を提案する．本稿では，その雑音環境下における照合性能評価を行い，融合の有効性を示す．

2　音声と耳介画像を用いたマルチモーダル話者照合

2.1　音声と耳介画像情報の融合

特徴量 x が入力されたとき，申告話者 S^c である確率 $p(S^c|x)$ は以下のように定義される．ただし，音声特徴量を x_s，耳介画像特徴量を x_e とし，S^c_s は申告話者の音声，S^c_e は申告話者の耳介画像情報を表す．

$$\begin{aligned} p(S^c|x) &= p(S^c_s|x_s)\cdot p(S^c_e|x_e) \\ &= \frac{p(x_s|S^c_s)p(S^c_s)}{p(x_s)}\cdot\frac{p(x_e|S^c_e)p(S^c_e)}{p(x_e)} \end{aligned} \quad (1)$$

ここで，音声・耳介特徴量の生起確率 $p(x_s)$, $p(x_e)$ を，それぞれの不特定話者モデルからの特徴量の出現確率 $p(x_s|S^g_s)$, $p(x_e|S^g_e)$ を用いて表すと，

$$p(S^c|x) = \frac{p(x_s|S^c_s)p(S^c_s)}{p(x_s|S^g_s)p(S^g_s)}\cdot\frac{p(x_e|S^c_e)p(S^c_e)}{p(x_e|S^g_e)p(S^g_e)} \quad (2)$$

となる．各話者について，申告話者の出現確率 $p(s^c_s)$, $p(S^c_e)$ は共通であると仮定し，不特定話者モデルの生起確率 $p(S^g_s)$, $p(S^g_e)$ は定数となるため，

$$p(S^c|x) \propto \frac{p(x_s|S^c_s)}{p(x_s|S^g_s)}\cdot\frac{p(x_e|S^c_e)}{p(x_e|S^g_e)} \quad (3)$$

となる．これは，特定話者モデルから得られた尤度を不特定話者モデルから得られた尤度で正規化することを意味している．音声・耳介画像それぞれの照合スコアを，

$$p_m = \log p(x_m|S^c_m) - \log p(x_m|S^g_m)\ \ (m = s, e) \quad (4)$$

と定義し，最終的な融合スコア p_{se} を，

$$p_{se} = \lambda_s p_s + \lambda_e p_e\ \ (\lambda_s + \lambda_e = 1) \quad (5)$$

と定義する．λ_s, λ_e は音声・耳介画像それぞれの照合スコアの重み係数である．この融合スコアが閾値を越えた時に，申告者本人であると判断する．

2.2　音声・耳介画像データ

音声・耳介画像データは時期差を考慮し，1 ヶ月毎に 5 時期に渡って収録を行った．収録話者数は 38 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おり，音声は 16kHz, 16bit で標本化・量子化した．また，各話者について，髪がかからないようにした右耳正面からの画像を各時期 1 枚ずつ撮影した．解像度は 720×540 である．なお，撮影の際にはフラッシュを付けた．

学習データは 1 ～ 3 時期目のデータ，評価データは 4, 5 時期目のデータとする．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，学習データを 19 名ずつの 2 グループに分ける．例えば，第 1 グループに属する話者を申告話者として照合実験を行う場合には，第 2 グループに属する全ての話者のデータで学習した不特定話者モデルを利用し，尤度の正規化を行う．このようにすることで，話者ごとの評価データは,「本人のデータ（1 名分)」「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 名分)」「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（18 名分)」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，評価データには 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，特徴量抽出に先立って以下の処理を順に行った．

(1) 耳孔を中心として，解像度が 80×80，8 bit グレイスケールの画像に変換する．その際，耳介の最も長い部分がおおよそ垂直となるように傾き補正する．位置と傾きの設定は手動で行った (図 1 (a)) .

∗ Multi-modal speaker verification using speech and ear images

By Koji Iwano, Tomoharu Hirose, Eigo Kamibayashi and Sadaoki Furui (Tokyo Institute of Technology)

図 1. 耳介画像の例

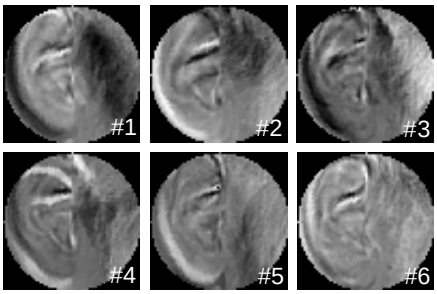

図 2. 固有耳画像の例

(2) ラプラシアンガウシアンフィルタによって輪郭強調処理を行う（図 1 (b)）.

(3) さらに，周辺の髪の影響の低減や，うなづく方向の耳介の回転を考慮し，円形の領域に切り出す（図 1 (c)）.

2.3 音声・耳介画像特徴量

音声特徴量には， MFCC 12 次元，ΔMFCC 12 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．分析周期は 10 ms，分析窓長は 25 ms とした．また，発声ごとに CMS を行った．

画像に関しては，まず，円形に切り出された画像を用いて主成分分析を行い，固有耳（eigen ear）を作成する．主成分分析用のデータは，不特定話者モデル学習用のデータの，1 時期目のみの画像を使用し，得られた固有空間における主成分得点を各画像の特徴量とした．本実験では，第 1 ～ 17 固有耳まで使用しているため，17 次元のベクトルとなる．第 1 ～ 6 固有耳画像を図 2 に示す.

2.4 音声・耳介画像のモデル化

音声特徴量は数字 HMM でモデル化を行う．数字列を w としたとき (3) 式右辺の音声による照合スコアは以下のように表される．

$$\frac{p(x_s|S_s^c)}{p(x_s|S_s^g)} = \frac{\Sigma_w p(x_s|S_s^c,w)p(w)}{\Sigma_w p(x_s|S_s^g,w)p(w)} \approx \frac{\max_w p(x_s|S_s^c,w)}{\max_w p(x_s|S_s^g,w)} \quad (6)$$

したがって，特定話者・不特定話者モデルを用いて連続数字認識を行い，その認識結果の尤度を照合スコアの計算に利用することになる．今回は，認識結果の数字列情報は利用しないため，テキスト独立型の条件で話者照合を行っている．

耳介画像特徴量は，混合正規分布でモデル化する．学習データ中には，一話者あたり 3 時期分の 3 枚の画像データしか存在していないが，うなずきによる耳画像の回転による揺れを確率分布で表すため，円形画像を -30 ～ 30 度まで一度ずつ回転させて画像を作成し，それぞれの画像特徴量を求めて，全てをモデル学習に利用する．なお，評価データには，このような回転処理は行わない．

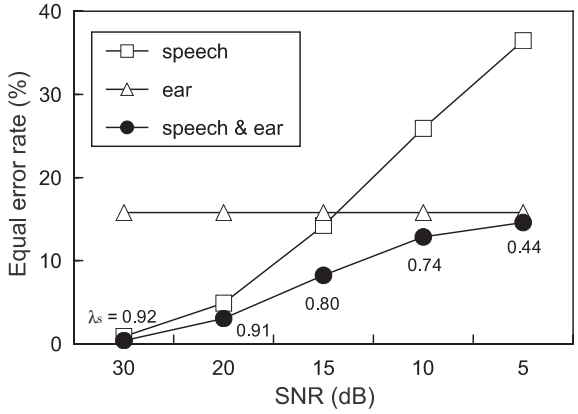

図 3. 照合実験結果

3 照合実験結果

話者照合実験の結果を図 3 に示す．図は各 SN 比条件における，音声のみを利用した手法（$\lambda_s = 1.0$），耳介画像のみを利用した手法（$\lambda_s = 0$），提案した融合法，による話者照合の等誤り率（equal error rate）を表している．融合法の重み係数 λ_s, λ_e は，各 SN 比における実験ごとに最適値を設定した．音声・画像モデルの混合数は，SN 比が 30dB において最も高い性能が得られた条件を，全ての実験において用いており，音声モデルの混合数は特定話者・不特定話者音声モデルともに 4，画像モデルの混合数は特定話者モデルで 1，不特定話者モデルで 16 とした．

音声のみを利用した手法では，SN 比が小さくなるにつれて，照合性能が大きく劣化していることがわかる．また，全ての SN 比条件で融合法の性能が，音声・耳介画像を単独で用いる場合より上回っていることが確認できる．特に SN 比 15dB の時に照合性能の改善が大きく，音声のみの誤り率からは 42.2%，耳介画像のみの誤り率からは 47.9%，相対的に誤り率が削減された．融合法の結果に付随している数字は，最適な音声スコアの重み係数 λ_s を表している．SN 比が小さく，音声特徴量の信頼性が乏しいほど，その値が小さくなり，より耳介画像情報を利用していることがわかる．なお，耳介画像における等誤り率は 15.8% であった.

4 まとめ

本稿では，音声と耳介画像情報を用いたマルチモーダル話者照合を提案し，時期差のある音声・画像データを用いて性能を評価し，提案手法の雑音環境下における有効性を確認した．今後の課題としては，耳介画像のみによる個人照合の性能改善，時期差データに対する頑健性の証明などがあげられる．

謝辞 本研究は NTT ドコモ株式会社の援助を受けて行われました．ここに謝意を表します.

参考文献

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

[2] M. Burge and W. Burger, “Ear biometrics,” in *Biometrics: Personal Identification in Networked society*, A. Jain, R. Bolle and S. Pankanti, Eds. pp. 273–285, Kluwer Academic, Boston, MA (1999).

[3] 田代 訓章，篠原 克幸，阿部 雅也，岡村 勉，“耳介の構成要素の輪郭および重ね合わせによる個人認証,” 映像メディア学会技術報告，vol.25, no.22, pp.7–13 (2001-3).