論文 / 著書情報
Article / Book Information

論題(和文)	音声と顔画像を用いたマルチモーダル話者照合
Title(English)	
著者(和文)	広瀬智治, 岩野 公司, 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. 3-3-2, pp. 107-108
Citation(English)	, Vol. , No. 3-3-2, pp. 107-108
発行日 / Pub. date	2003, 3

音声と顔画像を用いたマルチモーダル話者照合 *

◎広瀬 智治　岩野 公司　古井 貞熙 (東工大)

1　はじめに

近年，携帯端末の急速な普及により，携帯端末の認証機能を使ったオンラインショッピングや電子決済サービスの導入が進められている．しかし，その端末を使用しているのが所有者本人であるかという認証は現在のところ暗証番号によるものであり，忘却，盗難，解読の恐れがある．これに対して，バイオメトリクスを用いた個人認証では，盗難，偽造，忘却，紛失の恐れが小さく，常に人間に附随するものなので「なりすまし」などを防ぐことが可能である．

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．その中でも音声や顔を用いた個人認証は，人間が普段用いている個人認証法であり，馴染みやすく心理的抵抗感が少ない．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る認証性能の劣化，顔画像は 1) 髪型などの時期差変動，2) 照明条件の影響による認証性能の低下があり，1 つのバイオメトリクスのみを用いた場合では実用的な性能を得ることが困難である．そこで，複数のバイオメトリクスを併せて利用して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．マルチモーダル・バイオメトリクス話者照合の例としては，音声と顔動画像を用いたものなどがあるが [1]，研究例は未だ少ない．そこで，本研究では，この音声と顔動画像の 2 つのバイオメトリクスを組み合わせた話者照合手法について検討を行う．本稿では，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　音声と顔動画像を用いたマルチモーダル話者照合

2.1　音声と顔動画像情報の融合

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，顔画像特徴量を x_f とし，S^c_s は申告話者の音声，S^c_f は申告話者の顔動画像情報を表す．

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

ここで，音声・顔画像特徴量の生起確率 $p(x_s), p(x_f)$ 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 $p(x_s|S^g_s), p(x_f|S^g_f)$ を用いて表すと，

$$p(S^c|x) = \frac{p(x_s|S^c_s)p(S^c_s)}{p(x_s|S^g_s)p(S^g_s)}\cdot\frac{p(x_f|S^c_f)p(S^c_f)}{p(x_f|S^g_f)p(S^g_f)} \quad (2)$$

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

$$p(S^c|x) \propto \frac{p(x_s|S^c_s)}{p(x_s|S^g_s)}\cdot\frac{p(x_f|S^c_f)}{p(x_f|S^g_f)} \quad (3)$$

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・顔動画像それぞれの照合スコアを，

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

と定義し，最終的な融合スコア p_{sf} を，

$$p_{sf} = \lambda_s p_s + \lambda_f p_f\ \ (\lambda_s + \lambda_f = 1) \quad (5)$$

と定義する．λ_s, λ_f は音声・顔動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・顔動画像データ

音声・顔動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．この際，連続数字読み上げ時の顔正面の動画像を撮影し（15Hz プログレッシブ，24bit カラー，解像度 720×480），動画像から静止画像列を抽出して使用した．撮影時の照明条件はほぼ一定で，なるべく頭の位置がずれないように後頭部を壁につけている．

学習データは $1 \sim 3$ 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．

動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行った．

(1) 解像度 720×480，24bit カラー画像について両眼と顎の位置を元に，大きさ・傾きの正規化を行い顔領域を切り出す．位置と傾きの設定は手動で行った．

(2) 解像度の縮小とガウシアンフィルタによるフィルタリングを行って，解像度 40×40，8bit グレイスケール画像を生成する（図 1）．

∗ Multi-modal speaker verification using speech and face images

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

図 1. 顔画像の例

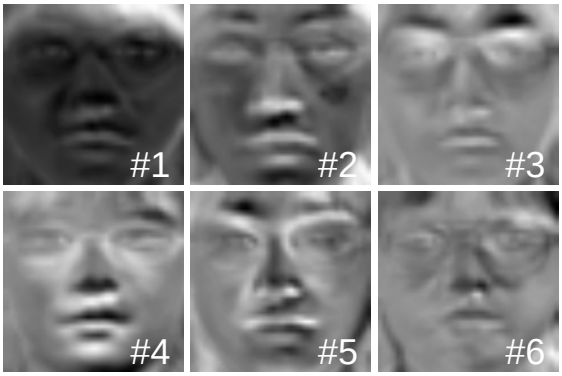

図 2. 固有顔画像の例

2.3 音声・顔画像特徴量

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．また発声ごとに CMS を行った．フレーム間隔は 10ms，特徴量抽出時のフレーム長は 25ms である．

画像に関しては，固有顔の手法 [2] に基づいて特徴量を抽出した．まず，学習データに含まれる話者の 1 時期目の画像のうち 1 枚ずつを用いて主成分分析を行い，固有顔（eigen face）を作成しておく．この時，各固有顔（固有ベクトル）は固有顔空間を張る軸になる．動画像を固有顔空間に射影した際の主成分得点の時系列を画像特徴量とした．本実験で使用する固有顔空間の次元数は 18 次元とした．第 1 ∼ 6 固有顔画像を図 2 に示す．また，画像特徴量を時間軸方向に補間し，音声と同じフレーム間隔とした．

2.4 音声・顔動画像のモデル化

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．

画像特徴量は GMM でモデル化し，頭部や口の動きによる顔画像の揺れを確率分布で表している．

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（$\lambda_s = 1.0$），顔動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）

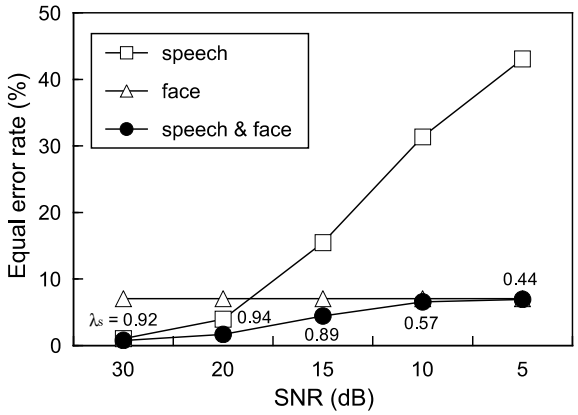

図 3. 照合実験結果

を表している．融合法の重み係数 λ_s, λ_f は，各 SN 比における実験ごとの最適値を設定した．音声・画像モデルの混合数は，実験的に最も高い性能が得られたものを用いている．音声モデルの混合数は SN 比 30dB における最適値を用い，特定話者モデルで 8，不特定話者音声モデルで 64，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・顔動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た．SN 比 15dB においては，音声のみの誤り率から 68.7%，顔動画像のみの誤り率から 31.2%，SN 比 20dB においては，音声のみの誤り率から 58.0%，顔動画像のみの誤り率から 76.5%，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，より顔動画像情報を利用していることがわかる．なお，顔動画像における等誤り率は 7.0% であった．

4 まとめ

本稿では，音声と顔動画像情報を用いたマルチモーダル話者照合について検討した．時期差のある音声・画像データを用いてその性能を評価し，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，顔動画像のみによる個人照合性能の改善として，時系列変化を考慮したモデリングの検討，特徴量の改善，また，重み係数や照合判定閾値の事前決定法についても検討する必要がある．

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] S. Ben-Yacoub, J. Luettin, K. Jonsson, J. Matas, and J. Kittler, “Audio-Visual Person Verification,” Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol.1, pp.580-585 (1999-6).

[2] M. Turk and A. Pentland, “Face Recognition Using Eigenfaces,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.586-591 (1991-6).