音声と発話に伴う口唇の動き特徴を用いた話者認識に関する検討

佐藤　慶幸†　　西田　眞†　　西　健治‡
†秋田大学　　‡株式会社アルファシステムズ

1. はじめに

近年，バイオメトリクスに関する研究が盛んに行われている．中でも画像データと音声データを組み合わせた認証は，非接触でユーザへの負担が軽いことから有用な個人識別手法として期待されている．これまでに報告されている事例[1]では，登録者本人の入力データと教師データから統合尤度を算出し話者認識を行っている（以下，従来法とする）．統合尤度を用いた場合においても，登録者本人の特徴に加え，他の登録者の特徴との関連を考慮して話者認識を行うことにより良好な結果の得られることが予想される．

そこで本研究では，他の登録者の特徴との関連を考慮した話者認識手法を提案し，その有用性について検討を加えた．

2. 話者認識法

2.1 使用データ

本研究では，通常の室内環境下（蛍光灯による照明）かつ雑音の少ない状況下で CCD ビデオカメラ（SONY 製：DCR-TRV900）により動画データを取得した．被験者は 10 名，発話内容は「あきたたろう」，発話回数は 15 回である．取得動画データを 3 グループに分け，入力データ，教師データおよび参照データとして用いた．なお，参照データとは統合尤度算出時（詳細については後述）に特徴量補正として用いるデータである．

取得した動画データを毎秒 30 フレームの静止画像に変換し，口唇画像データとして解析に用いた．一方，動画データから 32kHz，16bit の音声データを取得し，16kHz にダウンサンプリングしたデータを音声データとして解析に用いた．

2.2 特徴量算出

口唇画像データからは，フレーム毎に口唇の横幅 diX および縦幅 diY を取得し，これらを要素とするベクトルの大きさ $|\overrightarrow{diZ}|$ をパラメータとして用いた（図 1 参照）．また，音声データに

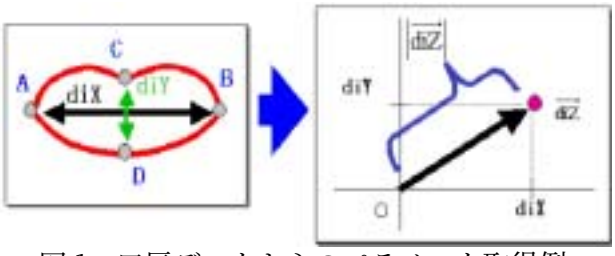

図 1　口唇データからのパラメータ取得例

フーリエ変換およびケプストラム分析を行い，音声信号の特徴パラメータを取得した[2]．さらに各パラメータに対し DP マッチング[3]を施して口唇画像データおよび音声データの特徴量を算出した．

2.3 統合尤度算出法

従来法では，特徴 i の登録者 A に対する特徴量 $\theta_i(A)$ の補正 $d_i(A)$ を(1)式により求めている．

$$d_i(A)=\frac{\theta_i(A)-\mu_i(A)}{\sigma_i(A)} \quad \cdots\cdots(1)$$

ここで，$\mu_i(A)$ は登録者 A の教師データと参照データによる特徴量の平均であり，$\sigma_i(A)$ は標準偏差である．(1)式で求めた各特徴量 $d_i(A)$ に重み付けをして統合尤度を求め，話者認識を行っている．このとき，従来法では入力データと登録データの一致度を表す補正値 $d_i(A)$ を被験者ごとに線形結合し統合尤度を算出しているため，他の登録者の特徴量は考慮されていない．

そこで本研究では，他の登録者の特徴量も考慮した統合尤度の算出法を提案する(以下，提案手法とする)．提案手法では，(2)式を用いて特徴 i における登録者 B との特徴量（以下，スコアとする）$S_i(B)$ を求め，各スコア $S_i(B)$ に重み付けをし，登録者 B に対する統合尤度を求めた．

$$S_i(B)=\sum_{m=1}^{M}\left(\frac{\theta_i(mI)-\mu_i(mB)}{\sigma_i(mB)}\right)^2 \quad \cdots\cdots(2)$$

ここで，M は登録者数，$\theta_i(mI)$ は登録者 m の教師データと入力データ I による特徴量，$\mu_i(mB)$ は登録者 m の教師データと登録者 B の参照データによる特徴量の平均，$\sigma_i(mB)$ は標準偏差である．

3. シミュレーション

3.1 シミュレーション条件

A study on speaker recognition using speech sound and movement features of the lip in speech.
†Yoshiyuki Sato , Makoto Nishida (Akita University)
‡Kenji Nishi (Alpha Systems Inc.)

本研究では，口唇画像データと音声データを用い，それらを単独で用いた場合と組み合わせた場合について，下記条件①，②によりそれぞれシミュレーションを行った．

① 被験者全員を登録者とした場合
② 被験者を半分ずつ登録者と未登録者に分けた場合

統合尤度を算出する際の各データに対する重み係数は，0.1：0.9～0.9：0.1 の範囲において 0.1 刻みで変化させてシミュレーションを行った．

3.2 シミュレーション結果の評価方法

シミュレーション結果の評価は以下の基準により行った．

① 正答：登録者本人の入力データを正しく推定し，未登録者を棄却する
② 誤識別：登録者本人の入力データを他の登録者として推定する
③ 誤棄却：登録者本人の入力データを未登録者として推定する
④ 誤受領：未登録者の入力データを登録者として推定する

また，正答率は次式を用いて算出した．

$$(\text{正答率}) = \frac{(\text{正答数})}{(\text{全入力数})} \qquad \cdots\cdots(3)$$

4. 実験結果および検討

4.1 被験者を全員登録者とした場合

シミュレーション結果の一例を表 1 に示す．本研究で使用したデータでは，従来法・提案手法ともに，音声データの重みを大きく（0.6～0.9）することにより誤識別は認められず，100%の正答率が得られた．

4.2 被験者を登録者と未登録者に分けた場合

シミュレーションにより得られた正答率の一例を表 2 に，誤棄却数および誤受領数の一例を表 3 にそれぞれ示す．従来法では最大で正答率 98.3%（300 例中 295 例正答），提案手法では最大で正答率 99.3%(300 例中 298 例正答)の結果が得られた．各条件における誤棄却数および誤受領数（表 3 参照）に注目すると，従来法では音声データのみを用いた場合と比較し，誤受領数は減少しているものの誤棄却数は増加していることがわかる．これに対し，提案手法では誤棄却数を増加させることなく誤受領数を減少させることが可能である．このことは本研究で提案した手法が有用であることを示唆するものである．

5. まとめ

本研究で得られた結果を以下にまとめる．

① 被験者全員を登録者とした場合，提案手法は従来法と同等の結果が得られた．
② 被験者を登録者と未登録者に分けた場合，提案手法は誤棄却数を増加させることなく誤受領数を減少させることが可能である．

最後に，本研究においてデータ取得にご協力いただいた高橋正人氏に謝意を表します．

表 1　正答率の一例(全員登録者)

	正答率(%)
口唇画像	76.0
音声	100.0
従来法	
口唇:音声	
0.1:0.9	100.0
0.2:0.8	100.0
0.3:0.7	100.0
0.4:0.6	100.0
提案手法	
0.1:0.9	100.0
0.2:0.8	100.0
0.3:0.7	100.0
0.4:0.6	100.0

表 2　正答率の一例(登録者と未登録者)

	正答率(%)
口唇画像	72.7
音声	97.7
従来法	
口唇:音声	
0.1:0.9	98.0
0.2:0.8	98.3
0.3:0.7	98.0
0.4:0.6	95.7
提案手法	
0.1:0.9	96.7
0.2:0.8	99.0
0.3:0.7	99.3
0.4:0.6	98.7

表 3　誤棄却数と誤受領数（登録者と未登録者）

	誤棄却数	誤受領数
口唇画像	36	29
音声	1	6
従来法		
口唇:音声		
0.1:0.9	4	2
0.2:0.8	4	1
0.3:0.7	5	1
0.4:0.6	9	4
提案手法		
0.1:0.9	1	9
0.2:0.8	1	2
0.3:0.7	1	1
0.4:0.6	2	2

参考文献

[1] 前田茂則他：「顔画像特徴，歩行画像特徴および音声特徴の統合による個人識別」 電子情報通信学会論文誌 D-Ⅱ，Vol.j79-D-Ⅱ，No.4，pp.600-607 (1996)
[2] 鹿野清宏他：「音声認識システム」，オーム社出版局 (2001)
[3] 谷萩隆嗣：「音声と画像のデジタル信号処理」，コロナ社 (1996)