社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－ＳＬＰ－69 (32)
2007／12／20

大規模話者骨導音声データベースを用いたテキスト独立型話者照合実験

喜多 雅彦† 黒岩 眞吾†† 柘植 覚† 蒔苗 久則††† 長内 隆†††
鎌田 敏明††† 谷本 益巳††† 土屋 誠司† 福見 稔† 任 福継†

† 徳島大学大学院
†† 千葉大学大学院
††† 科学警察研究所

E-mail: †{kita__m,tsuge,tsuchiya,fukumi,ren}@is.tokushima-u.ac.jp, ††kuroiwa@faculty.chiba-u.jp,
†††{makinae,osanai,kamada,tanimoto}@nrips.go.jp

あらまし 本稿では，科学警察研究所によって構築された大規模話者骨導音声データベースを用いた話者照合実験を行った結果を報告する．実験には，664 名 (男性 336 名，女性 328 名) のコンデンサマイクで収録された音声 (気導音)，骨導マイクで収録された音声 (骨導音) を用いた．実験では，以前我々が提案した複数話者モデルの順位情報を用いた話者照合手法を評価した．また，話者モデルとして GMM とベクトル量子化 (VQ) セントロイドの比較，発声時期の違いによる照合精度の比較を行った．実験結果より，提案手法は従来の T-Norm を用いた話者照合手法より高い照合精度を示すことが観測された．さらに，話者モデルの違いによる照合精度の比較結果より，気導音では VQ セントロイドを用いた方が照合精度が高く，骨導音では GMM を用いた方が高いことが観測された．また，骨導音による照合精度は気導音より低く，さらに骨導音は時期差が生じた場合，照合精度低下が著しいことが観測された．

キーワード 生体認証，話者照合，骨導音

Text-independent speaker verification experiment using a large-scale bone-conducted speech database

Masahiko KITA†, Shingo KUROIWA††, Satoru TSUGE†, Hisanori MAKINAE†††,
Takashi OSANAI†††, Toshiaki KAMADA†††, Masumi TANIMOTO†††, Seiji TSUCHIYA†,
Minoru FUKUMI†, and Fuji REN†

† The University of Tokushima
†† Chiba University
††† National Research Institute of Police Science

E-mail: †{kita__m,tsuge,tsuchiya,fukumi,ren}@is.tokushima-u.ac.jp, ††kuroiwa@faculty.chiba-u.jp,
†††{makinae,osanai,kamada,tanimoto}@nrips.go.jp

Abstract In this paper, we conducted a speaker verification experiment using large-scale speech database maintained by National Research Institute of Police Science, Japan. In this exepriment, we used speech data of 664 people collected by a capacitor microphone and a bone-conducted microphone. From experimental results, we confirmed that our proposed method that uses rank information obtained by multiple speaker model in previous work improved verification performance than a conventional method using T-norm score. In addition, we compared the speaker model based on GMMs and that based on VQ centroids. From this comparison, we can see that the speaker model based on VQ centroids is higher performance than that based on GMMs under the condition of the capacitor microphone speech. However, VQ centroids degraded the performance of that based on GMMs under the condition of the bone-conducted speech. Moreover, the performances of the bone-conducted speech significant degraded performance if there were difference of the speaking session between the registration and the testing.

Key words bioinformatics, speaker verification, bone-conduction speech

1. はじめに

近年，虹彩や指紋，顔，静脈，音声などの生体情報を用いた生体認証技術が注目されている．これら生体認証技術の中でも音声による個人認証 (話者認識) は携帯電話などに備え付けられたマイクによる入力も可能であるため，特別なハードウェアを必要とせず，遠隔地からの個人認証を行うことも可能である．さらに，普段から使用しているハードウェアであるため，ユーザの心理的負担が少ないなどの利点が多くある．そこで，我々は音声による生体認証に着目し，話者認識の研究を進めている．

現在，話者認識の研究を行うための音声データベースとして，YOHO，TIMIT，NIST などがあり [1]，これらのデータベースを用いて様々な研究が行われている [2], [3]．また，日本語の音声データベースも複数存在している．しかし，多くの日本語音声データベースは 1 話者が 1 時期にしか発声をしていないため，話者認識の問題として挙げられる発声時期が異なる場合の性能劣化に関する研究に，これらのデータベースは利用できない．1 話者が複数の時期に発声している音声データベースとして文献 [4] が挙げられるが，収録されている話者数が少人数であるため，大規模な実験を行うには不向きである．

近年，科学警察研究所 (以下，科警研) によって大規模話者骨導音声データベースが構築された [5]．このデータベースでは，1 話者が 2 時期で発声しており，収録されている話者数も 664 名 (2 時期目まで収録されている話者数は 632 名) であることから大規模な実験を行うことが可能である．さらに，通常のコンデンサマイクだけでなく，骨導マイクによる音声データも収録されているため，気導音と骨導音との認識性能を比較，現状の手法が骨導音に適用することができるかの検証も可能である．

一般的に，話者照合を行う場合，申告話者のモデルとコホートモデルやガーベッジモデルとの類似度 (尤度やベクトル量子化 (Vector Quantization : VQ) 歪み) を比較し，閾値処理を行うことにより，入力音声が申告話者のものであるか判定し受理・棄却を行っている．しかし，この場合，発声長や発声内容，発声環境などで最適な閾値が異なり，最適な閾値を決定することが困難となる [6]．そこで我々は，複数の話者モデル内における申告話者モデルの順位に着目した，複数話者モデルの順位情報を用いた話者照合手法を提案した [7]．この手法は，複数話者モデル内での申告話者モデルの順位のみを用いて話者照合を行うため，尤度比等のダイナミックレンジを事前に予測できない値に閾値を設ける手法に比べ，頑健な話者照合が可能である．

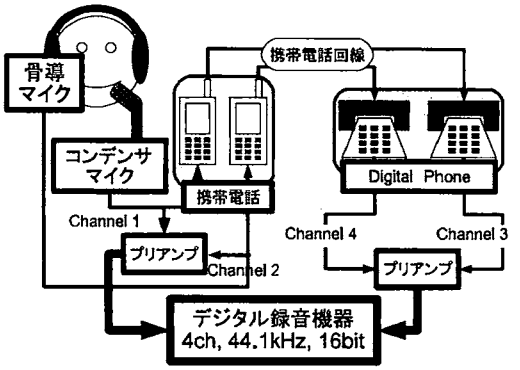

図 1　データベースの収録環境図

本稿では，科警研が構築したデータベースを用いたテキスト独立型話者照合実験結果を報告する．実験には，男性 336 名，女性 328 名 (時期差がある場合は男性 313 名，女性 319 名) の音声データを用い，提案手法と従来手法である T-Norm [8] による話者照合精度の比較を行った．また，GMM による方法と VQ 歪みによる話者照合性能の比較，GMM の混合数，VQ セントロイド数などの比較を気導音，骨導音において行った．

2. 大規模話者骨導音声データベース

本節では，科警研により構築された大規模話者骨導音声データベース [5] の紹介を行う．図 1 に，データベースの収録環境を示す．図に示す通り，このデータベースには，

- コンデンサマイク (Channel 1)
- 骨導マイク (Channel 2)
- コンデンサマイク収録時の音声を携帯電話網を介して収録 (Channel 3)
- 骨導マイク収録時の音声を携帯電話網を介して収録 (Channel 4)

以上の 4 チャネルで収録された音声データが含まれている．収録された 4 チャネルの音声データは 20ms 以内に時間同期がとられ，データベースに格納され

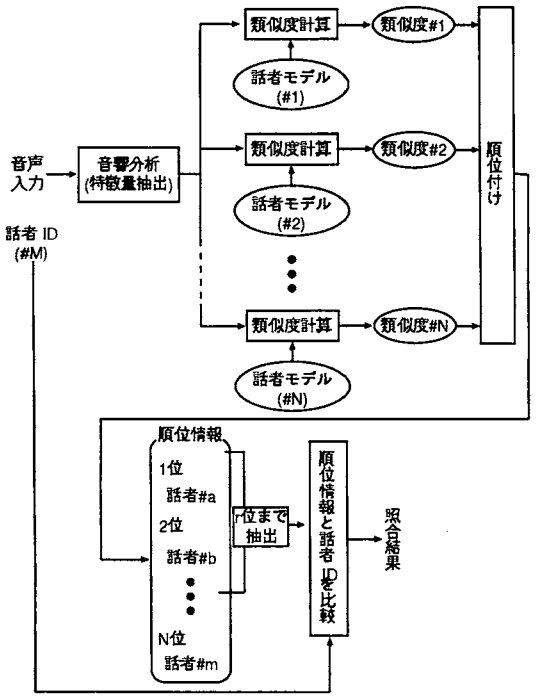

図 2　提案手法の流れ

ている．

このデータベースに収録されている話者数は男性 336 名，女性 328 名の計 664 名であり，各話者が 2 時期にわたって発声を行っている．ただし，このうち 32 名 (男性 24 名，女性 8 名) は 1 時期目のみの発声となっている．各話者は各時期において，ATR 音素バランス文などの発声セットを 2 回ずつ発声している．そのため，このデータベースには話者ごとに同じ発声内容の音声が 4 回収録されている．これらの音声データはサンプリング周波数 44.1kHz，量子化ビット数 16bit で収録されている．話者と収録時期，発声内容等に関するデータベースの詳細については文献 [5] を参照してほしい．

3. 複数話者による順位情報を用いた話者照合手法

本節では，我々が提案した順位情報を用いた話者照合手法を説明する．図 2 に提案手法の流れを示し，以下で提案手法の詳細を示す．

3.1 各話者モデルとの類似度計算

登録話者を照合する話者モデルは登録音声を用い，順位情報を得るための本人以外の話者モデルは，他者の登録音声および既存のデータベースを用い学習する．このモデルとして，確率モデルである GMM やノンパラメトリック手法である VQ セントロイドが考えられる．これらのモデルを用い，入力音声 $\boldsymbol{x}$ に対し，学習した全ての話者モデルの距離 (類似度) を計算し，類似度集合 ($\mathrm{Q}_{\boldsymbol{x}}$) を求める．

$$\mathrm{Q}_{\boldsymbol{x}} = \{\mathrm{P}(\boldsymbol{x}|\lambda_1), \mathrm{P}(\boldsymbol{x}|\lambda_2), \cdots, \mathrm{P}(\boldsymbol{x}|\lambda_n)\} \qquad (1)$$

ここで，λ_i は話者 ID が i の話者モデルを示し，$\mathrm{P}(\boldsymbol{x}|\lambda_i)$ は話者モデル λ_i の $\boldsymbol{x}$ に対する類似度 (話者モデルが GMM の場合，対数尤度，話者モデルが VQ セントロイドの場合，VQ 歪み) を示す．

3.2 順位情報を用いた話者照合

前節で得られた式 (1) で計算される類似度情報および各話者モデルの順位を返り値として示す関数 (RANK()) を用い，照合する話者 ($target$) の複数の話者モデル内の順位情報 (R_{target}) を得る．

$$R_{target} = RANK(Q_x, \lambda_{target}) \qquad (2)$$

求められた複数話者モデル内の申請者モデルの順位情報を用い，以下の通り

$$R_{target} \leq R_M \Longrightarrow accept \qquad (3)$$

$$R_{target} > R_M \Longrightarrow reject \qquad (4)$$

決められた順位 (R_M) 以内なら受理，決められた順位より低い場合は棄却する．本提案手法では，申請者話者モデルとユニバーサルモデルやコホートモデルとの尤度による比較を行い，その尤度比を閾値により判別し，受理，棄却を決定しない．そのため，発話長，発話内容などにより変動する尤度を直接的には用いないため，頑健な話者照合が行えると考えられる．

4. テキスト独立型話者照合実験

2. で紹介した大規模話者骨導音声データベースを用いて，テキスト独立型の話者照合実験を行った．データベースに収録されている音声データの標本化周波数は 44.1kHz であるが，本実験では，8kHz にダウンサンプリングした後に使用した．

4.1 実験条件

音声データは 664 名 (男性 336 名，女性 328 名) が発声した ATR 音素バランス文 (気導音 : Channel 1,

表 1 音響分析条件

サンプリング周波数	8kHz
フレーム周期	10ms
フレーム長	25ms
窓タイプ	ハミング窓
メルフィルタバンク数	24

表 2 テストセット区分

	発声時期	本人発声数	詐称者発声数
1_1	1 時期目，1 回目	3,320	332,000
1_2	1 時期目，2 回目	3,320	332,000
2_1	2 時期目，1 回目	3,159	315,900
2_2	2 時期目，2 回目	3,160	316,000

骨導音：Channel 2) を用いた．話者モデル作成データ (登録用音声データ) として，気導音，骨導音ともに各話者が 1 時期目の 1 回目に発声した音素バランス文 5 発声を用いた．評価用データは，各話者 1 時期目に発声した 10 文 (各回 5 文)，2 時期目に発声した 10 文 (各回 5 文) を用いた．また，評価用データには登録用データとは異なる発声内容のものを用いている．それぞれのテストセットでの全発声数は表 2 に示す通りである．ただし，2 時期目の 1 回目発声のテストセットは分析ミスのため 1 発声少なくなっている．

特徴パラメータには表 1 に示す音響分析条件により分析し，1 発声毎に CMS を行った MFCC (Mel Frequency Cepstrum Coefficient) 12 次元，その一次回帰係数 12 次元，対数パワーの一次回帰係数の合計 25 次元を用いた．この特徴パラメータに対し，無声部分や頼りない発声部分の影響を避けるためにパワーの閾値を決め，フレーム毎にパワーによるデータの取捨選択を行った．実験には以下のパワー閾値を用いた．

$$PowerThreshold = (P_{95\%tile} - P_{10\%tile}) \times 0.2 + P_{10\%tile} \quad (5)$$

ここで，$P_{95\%tile}, P_{10\%tile}$ は，特徴量の各フレームのパワーをソートして，パワーの小さい方を 0% としたときの 95%, 10% 地点の値である．

話者モデルとして，GMM，VQ セントロイドを用いた．GMM の混合数と VQ セントロイド数は各実験で変更している．そのため混合数などについては実験の各節で示す．

話者照合性能は，DET(Detection Error Tradeoff) 曲線を用いて比較した．DET 曲線は，横軸に詐称者受理率を，縦軸に本人棄却率を取り，照合アルゴリズムや認証システムの精度を曲線で表したものである．

詐称者用データには，本人データの音声を本人以外の話者に用いることで詐称者とした．この際，実験時間の短縮のためデータベース全員ではなくそれぞれ男女 100 名に用いることとした．また，詐称者のデータを入力する際には詐称者本人のコードブックは登録話者モデルから除いた．

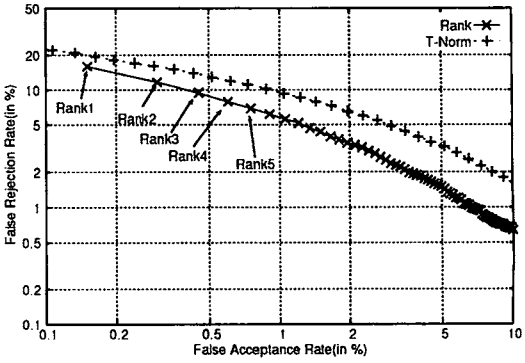

図 3 提案手法と T-Norm の結果の DET 曲線

上記の条件で，以下の比較を行った．

- 提案手法と T-Norm との比較 (4.2)
- 気導音と骨導音および話者モデルの比較 (4.3)
- テストセットの比較 (4.4)

4.2 T-Norm との比較

本節では提案手法の有効性を検証するため，従来手法の T-Norm との比較実験を行った．比較は VQ セントロイド (VQ セントロイド数 128) を話者モデルとして用い，気導音においてテストセットに分割せず行った．

また，T-Norm による正規化に用いるコホート話者は，登録用データのうち各話者 1 発声について VQ 歪み距離を求め，距離の近い 300 人とした．

比較実験結果を図 3 に示す．図中の × は提案手法の DET 曲線を示し，＋は T-Norm の DET 曲線を示す．また，図中の RankN は N 番目までの順位話者を正解とする受理順位，すなわち式 (3),(4) での R_M

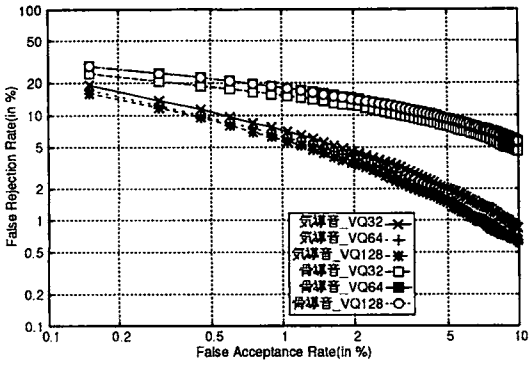

図 4 話者モデルとして VQ セントロイドを用いた場合の気導音，骨導音の照合精度比較

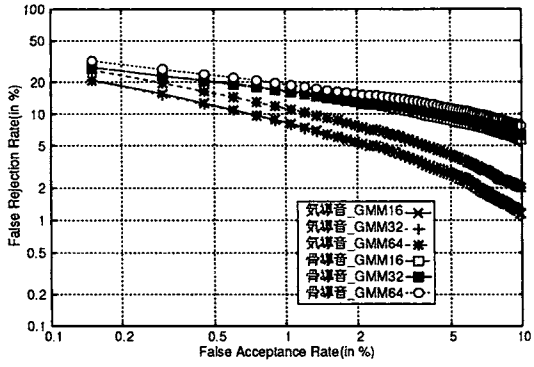

図 5 話者モデルとして GMM を用いた場合の気導音，骨導音の照合精度比較

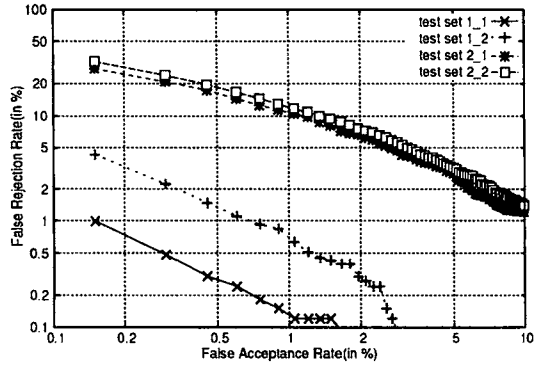

図 6 気導音でのテストセット区分毎の DET 曲線 (VQ)

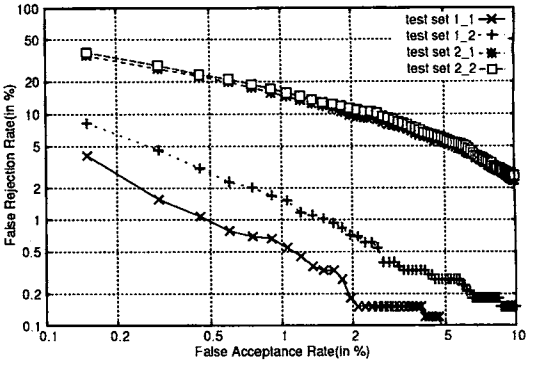

図 7 気導音でのテストセット区分毎の DET 曲線 (GMM)

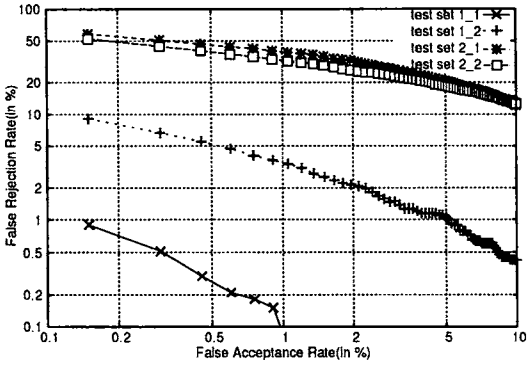

図 8 骨導音でのテストセット区分毎の DET 曲線 (VQ)

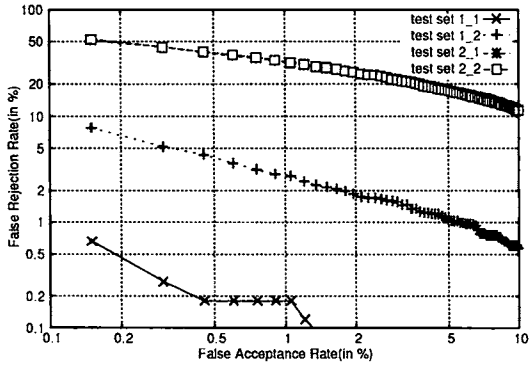

図 9 骨導音でのテストセット区分毎の DET 曲線 (GMM)

を示す．この結果より，提案手法は従来の T-Norm より高い話者照合精度を示すことがわかる．この結果より，以下の比較実験は提案手法のみで行う．

4.3 気導音と骨導音および話者モデルの比較

本節では気導音と骨導音の違いによる話者照合精度の比較および，話者モデルとして用いる VQ セントロイドと GMM の比較を行った．比較はテストセットに分割せず行った．比較実験結果を図 4, 5 に示す．

図より，気導音は骨導音よりも高い話者照合精度を示すことがわかる．この原因として，データベース内の骨導音には，収録の際に生じたと考えられる

雑音が多く見られたことが考えられる．骨導音の照合精度が低い原因の解明は今後の課題である．また，結果より本実験条件ではVQセントロイド数が気導音では128，骨導音では32が，GMMでは気導音では32，骨導音では16が高い照合精度を示すことがわかる．この結果より，以下の実験では気導音ではVQセントロイド数128，GMMの混合数32を用い，骨導音ではVQセントロイド数32，GMMの混合数16を用いて各テストセットごとの誤り傾向を分析する．

4.4 テストセットの比較

本節では，前節の実験より照合精度が高かったVQセントロイド数(気導音：128，骨導音：32)とGMMの混合数(気導音：32，骨導音：16)を用いて各テストセットごとの照合精度の比較を行った．図6～9に，各テストセット区分における気導音と骨導音および，VQ歪みとGMMでのそれぞれの結果を示す．

図より，発声時期が異なる場合に発声時期が同じ場合での同じ順位と比べ，大きく照合誤り率が増加してしまっていることがわかる．特に骨導音での結果は気導音に比べて非常に時期差の影響を受けてしまっていることがわかる．また，発声時期が同じであっても1回目の発声と2回目の発声では2回目の照合誤り率の方が大きい．特に骨導音ではこの傾向が顕著に表れている．この原因としては，2回目の発声が1回目の発声セットを収録した後に収録されたことによる疲れの影響や，1回目と2回目の収録時のマイク位置が変化したことなどが考えられる．

5. ま と め

本稿では，科警研によって構築された大規模話者骨導音声データベースを用いたテキスト独立型話者照合実験結果を報告した．実験には，男性336名，女性328名のコンデンサマイク収録の音声データ(気導音)および骨導マイク収録の音声データ(骨導音)を用い，我々が提案している，順位情報を用いた話者照合手法の検証を行った．また，話者モデルとしてGMMを用いた場合とVQセントロイドを用いた場合の比較，発声の時期の違いなどの比較も行った．この際，音声データは8kHzにダウンサンプリングしている．実験結果より，提案手法は従来手法のT-Normよりも高い照合精度を示すことがわかった．また，骨導マイクでの結果はコンデンサマイクでの結果よりも照合誤り率が増加し，時期差が生じた場合も照合精度が劣化することも確認できた．今後は骨導マイクの特徴をさらに調査し，骨導マイクでの有効な話者照合手法について検討していく．

謝辞 本研究の一部は文部科学省科学研究費，若手研究(B)19700172，基盤研究(B)17300065，基盤研究(B)19300029，萌芽研究 17656128の補助を受けた．

文　献

[1] J. Godfrey, D. Graff, and A. Martin, "Public databases forspeaker recognition and verification", Proc. of ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 39-42, 1994.

[2] Y. Liu, et.al. "The role of dynamic features in text-dependent and -independent speaker verification", Proc. ICASSP, pp.669-672, 2006.

[3] A. Stolcke, et.al. "MLLR transforms as features in speakerrecognition", Proc. of Interspeech, pp. 2425-2428, 2005.

[4] T. Matsui and K. Aikawa, "Robust model for speaker verification against session-dependent utterance variation", Proc.ICASSP, pp. 117-120, 1998.

[5] 蒔苗久則，長内隆，鎌田敏明，谷本益巳，"大規模話者骨導音声データベースの構築と予備的な解析"，信学技報，Vol.107, No.165, PP.97-102, SP2007.

[6] 松井知子，西谷隆，古井貞熙，"話者照合におけるモデルとしきい値の更新法"，電子情報通信学会論文誌，J81-D-II, No.2, pp.268-276, 1998.

[7] 喜多雅彦，柘植覚，黒岩眞吾，任福継，"多数の話者モデル内での順位情報を用いた話者照合"，人工知能学会研究会資料，SIG-SLUD-A701, pp.7-12, 2007.

[8] R.Auckenthale, M.Carey, and H.L.Thomas, "Score normalization for text-independent speaker verification systems", Digital Speech Processing, Vol.10, pp.42-54, 2000.