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

2006－SLP－64 (17)
2006／12／21

音素モデルを用いた話者ベクトルに基づく話者識別の検討

赤津 達也† 加藤 正治†† 小坂 哲夫†† 好田 正紀††

† 山形大学工学部
〒992-8510 山形県米沢市城南 4-3-16
E-mail: †tatsuya17akatsu@yahoo.co.jp. ††{katoh,tkosaka,kohda}@yz.yamagata-u.ac.jp

あらまし 本稿では, 音素モデルを用いた話者ベクトルに基づくテキスト独立型話者識別について述べる. 本研究の目的は. 音声の音素コンテキスト情報を使用することによって, 話者識別の性能を改善させることである. 本話者識別システムはアンカーモデルに基づいている. このシステムでは, 識別対象話者の発声とアンカーモデル間の尤度からなる話者ベクトルによって, 各々の話者が話者空間に配置される. 話者識別の性能の改善のために, アンカーモデルとしてガウス混合モデル (GMM) の代わりに音素モデルが使用される. 音素 HMM の対数尤度の計算には, 音素認識装置を使用する. また, アンカーモデルのパラメータ数についても検討を行う. 本手法は 30 名の日本語話者識別タスクで評価を行った. この結果, GMM ベースのシステムと比較し, 72.1%の相対的改善を得られた.

キーワード 話者認識, 話者識別, 隠れマルコフモデル (HMM), ガウス混合モデル (GMM), 音素クラス

An investigation on the speaker vector-based speaker identification with phonetic modeling

Tatsuya AKATSU†, Masaharu KATOH††, Tetsuo KOSAKA††, and Masaki KOHDA††

† Faculty of Engineering, Yamagata University
Jounan 4-3-16, Yonezawa-city, Yamagata, 992-8510 Japan
E-mail: †tatsuya17akatsu@yahoo.co.jp, ††{katoh,tkosaka,kohda}@yz.yamagata-u.ac.jp

Abstract This paper presents a phonetic based approach for speaker identification performed in text-independent mode. The aim of this work is to improve identification performance by using information about the phonetic content of the speech. The identification systems is based on the technique of anchor models. In this system, the location of each speaker is represented by the speaker vector which consists of the set of the likelihood between a target utterance and the anchor models. In order to improve the identification performance, phonetic modeling is used instead of Gaussian Mixture Model (GMM) scheme as anchor models. This approach utilizes a phonetic speech recognizer to calculate the log-likelihood with phonetic HMMs. We also investigate the number of parameters of anchor models. The proposed method was evaluated on Japanese speaker identification task with 30 speakers. It showed that the proposed method achieved 72.1% relative improvement over the GMM-based system.

Key words Speaker recognition, speaker identification, hidden Markov model(HMM), Gaussian mixture model(GMM), phonetic class

1. はじめに

本稿ではアンカーモデルに基づいた, 音素モデルを使用する話者識別について述べる. アンカーモデルシステムはすでに話者インデキシングのために [1] で提案されている. また, 話者識別 [2] および話者照合 [3] で既に使用されている. アンカーモデルに基づく話者識別の基本的な考え方は, 多数の参照話者のモデル (アンカーモデル) を用いて, 入力話者と多数話者の相対的位置関係を識別に用いるということである. この方法において, 各々の話者の位置は話者ベクトルによって表される. 話者ベクトルは, 識別対象話者の発声と多数のアンカーモデル間の尤度から求められる. それは識別対象話者の発声の, 話者空間への写像と考えることが出来る. 本手法のひとつの利点は, アンカーモデルのセットが識別対象話者のモデルを含まないため, 新しい識別対象話者のためにモデルを学習する必要がないという点である. これにより, ユーザーがモデルを学習するために

繰り返し発声を行う手間を省くことができる．しかしこの方法では，話者識別の性能が不十分という問題がある．例えば[2]では，アンカーモデルに16混合のGMMを利用し，次元数が最高で500，識別対象話者数50名よる話者識別タスクで，識別率76.6%と報告されている．

本研究の目標は，アンカーモデルとしてGMMのかわりに音素モデルを使用することによって，本識別法のパフォーマンスを向上させることである．GMMベースの音響モデルは，各話者において全音素イベントをカバーしている．それは話者間の音響特徴の全体的な違いを表現する．しかし，音素個々の発音の違いを表現することはできない．したがって，本研究では音素特徴における詳細な違いを見つけ，それを話者識別のための情報として使用することを試みる．音素特徴を発見するために，アンカーモデルとして話者依存音素HMMのセットを使用する．識別対象話者の発声とアンカーモデル間の尤度計算は，音素対文法のHMMベース音素認識装置により行われる．

本提案手法を評価するために，音素ベースシステムとGMMベースシステムをアンカーモデルフレームワークで比較する．また，アンカーモデルのパラメータ数も検討する．アンカーモデルのガウス分布数および話者空間の次元数の両方について，検討を行う．この目的のために，2000名以上の話者が含まれている大規模なスピーチコーパスを，アンカーモデルの学習のために使用した[4]．

以下では，関連研究を紹介する．近年，音素ベースの話者識別手法がいくつか提案されている．Hebertらは，音素クラスの木構造に基づいた話者照合法を提案した[5]．この論文では，音素クラスに基づいたシステムは，従来法であるGMMアプローチよりも優れた性能が得られたことを示している．Parkらは，各話者について音素クラスGMMを使用する話者識別手法を提案した[6]．識別対象話者のためにモデルが必要だという点で，この2つの方法は本提案手法とは異なる．Andrewsらは，音響特徴ベクトルに基づいた方法のかわりに，音素列にのみ基づいた話者認識システムを開発した[7]．この方法において，評価話者モデルは音響モデルではなくn-phone頻度数を使用して生成される点が本手法とは異なる．

本稿は以下のように構成される．2章では話者識別の方法について記述する．3章ではタスクおよびデータセットについて記述する．また，実験条件についても示す．4章では実験結果および考察を記述する．最後に，5章では結論および検討課題を示す．

2. 音素モデルを用いた話者識別

2.1 アンカーモデルを用いた話者空間の構成

図1に話者ベクトル空間の3次元での概念図を示す．アンカーモデルに基づいたシステムでは，識別対象話者の音声はアンカーモデルの尤度によるベクトルによって特徴付けられる．入力音声は以下のベクトルによって表される．

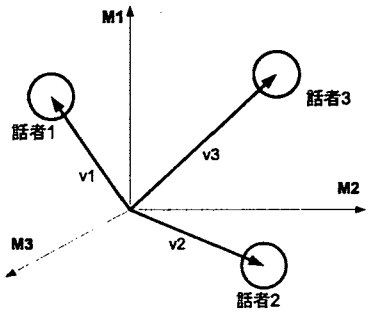

図1 ベクトル空間の概念図
Fig. 1 A conception diagram of vector space

$$V_j = \begin{bmatrix} \frac{p(s_j|M_1)-\mu_j}{\sigma_j} \\ \frac{p(s_j|M_2)-\mu_j}{\sigma_j} \\ \vdots \\ \frac{p(s_j|M_N)-\mu_j}{\sigma_j} \end{bmatrix} \quad (1)$$

μ_j ： $\{p(s_j|M_i)\}$ の平均値

σ_j ： $\{p(s_j|M_i)\}$ の標準偏差

ここで $p(s_j|M_n)$ はアンカーモデル M_n における入力特徴 s_j の対数尤度である．ベクトルは発話間のスコア変動を抑えるために平均0，分散1に正規化される．識別対象話者 j はアンカーモデルとして利用されている N 人の話者には含まれない．本手法では，入力音声から話者ベクトルを生成し，識別対象話者の登録音声の話者ベクトルとのユークリッド距離を計算して，距離が最短のものを入力音声の話者であると識別する．従来のGMMに基づく一般的手法では，識別対象話者の話者モデルを作成する必要があり，学習用の発声が10文程度は必要であった．提案手法では識別対象話者のためにモデルを学習する必要がないので，登録ベクトル生成には1発声程度あれば良い．

2.2 アンカーモデルの音素表現

以前の研究においては[1][2]，GMMがアンカーモデルとして使用されてきたが，従来のGMMベースの方法では音素コンテキストの情報を無視している．しかし，いくつかの音素クラスは高い話者性を有していると考えられる．そこで本研究では性能向上を目的として，音素HMMを使用する．本手法では式(1)で現れる尤度 $p(s_j|M_n)$ を計算するために，音素認識装置を用いる．話者 n における対数尤度は，話者依存音素HMMの認識装置によって得られる．本研究では音素の種類数は35とした．認識装置は未知発話をデコードするので，識別システムはテキスト独立型として実行される．認識では，音素対文法によってデコードを行い，このときの音響尤度を話者ベクトルの計算に利用する．ただしビームから落ちた場合には，最小尤度と置き換える．図2は1000次元で構成された話者ベクトルの一例を示している．x軸はある話者の26番目の発話の話者ベクトルの値を表し，y軸は同じ話者の27番目の発話の値を表す．それぞれの点は1000のアンカーモデルに対する個々の値を表している．同じ話者によって与えられた両発話は，発話内容が異なる．

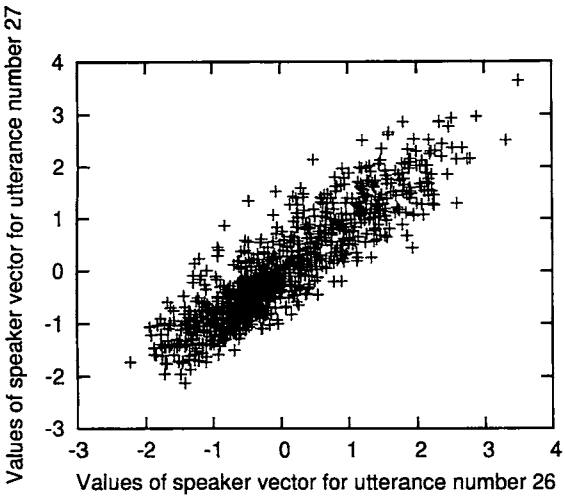

図 2 話者ベクトルの値の例

Fig.2 Example of values of speaker vector

内容は同じではないが, 2 つの値のセットは高い相関を示す. これは, 発話内容が異なっても, 発声話者が同じなら, 話者空間上で近い位置に写像されることを意味する. またテキスト独立型話者識別が本手法で実行可能であることを示唆している.

2.3 音素構造 GMM

GMM ベースシステムは音素全体を利用して音響モデルの生成を行っており, 各音素の情報を無視している. しかし音素 HMM のように学習を音素ごとに行えば, 識別性能が改善する可能性が考えられる [6]. 音素情報を利用した研究として, 例えば [9] では, Faltlhauser らによって音素構造 GMM が提案されている. これは各話者に対する個々の音素クラスの GMM を学習し, それらをひとつのモデルへ混合重みを与え, 組み合わせることによって作成される. 音素構造 GMM の概念図を図 3 に示す. また [10] では, Sauma によって音素クラス GMM の研究が行われている. この技術は音素構造 GMM と似ているが, 一つの話者 GMM へは組み合わせずに, 個々の音素クラスが保持されている点が異なる. 本研究では従来の GMM および HMM に加えて音素構造 GMM についても検討を行い, 性能を比較する.

3. 実験条件

評価には, スピーチコーパスとして ATR SDB-I を使用する [4]. このコーパスは多数の話者読み上げ音声と対話音声からなり, 話者による音響的変動をカバーしている. アンカーモデルを表現するために, 744 名の男性と 1288 名の女性からなる計 2032 名の話者によって発声される, 音素バランス音声データを使用する. このため話者空間の最大次元数は 2032 となる. 発話の総数は 51131 である. 個々の話者の発声量が少ないため, アンカーモデルの学習には, ML 推定の代わりに MAP 推定を使用する. 評価データセットは 30 名の話者からなり, それぞれ 25 の発声データを持っている. 評価セットの平均発話長は約 5.5 秒である.

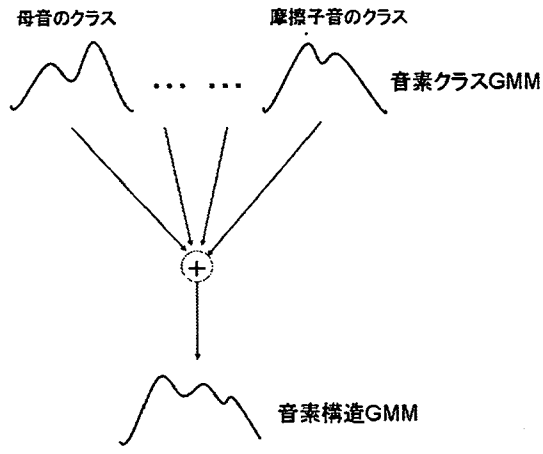

図 3 音素構造 GMM の概念図

Fig.3 A conception diagram of phonetically structured GMM

表 1 分析条件

Table 1 Analysis Conditions

フロントエンド	ETSI Advanced front-end (AFE-WI008) Blind Equalization なし
標本化	16kHz
量子化	16bit
周期	10msec
フレーム長	25msec
分析窓	ハミング窓
高域強調	$1-0.9z^{-1}$
分析	MFCC(1-12 次), 対数パワー +Δ+ΔΔ (計 39 次元)

表 1 に分析条件を示す. 本研究では将来的に雑音状況下での話者識別法の開発を予定しているので, 分析には雑音に頑健なアルゴリズムが使用されている ETSI advanced front-end (AFE-WI008) を使用する [8]. このフロントエンドは雑音対策として, 加算性雑音には雑音除去を, 乗算性雑音には blind equalization を用いている. 予備実験の結果から, blind equalizaton が話者識別の性能に悪影響を与えることが示されたため, 本研究では blind equalizaton 処理は省略する.

識別性能の評価に当たっては, 登録発声内容の違いによる識別性能の変動を平均化するために, 以下の方法を採用する.

- 各評価話者の 25 発声のうち 24 発声を評価用, 残り 1 発声を登録用として使用する
- 25 の異なる登録について実験し, その平均を識別率とする

4. 実験結果および考察

まず, 話者空間における次元数の影響について検討を行った. GMM および HMM の混合数についてもまた, 検討を行った. 例えば [2] では, 話者空間は 16 混合 GMM, 500 次元からなっている. しかし話者空間におけるパラメータ数の詳細な研究はこれまで行われておらず, このパラメータ数では十分かどうか

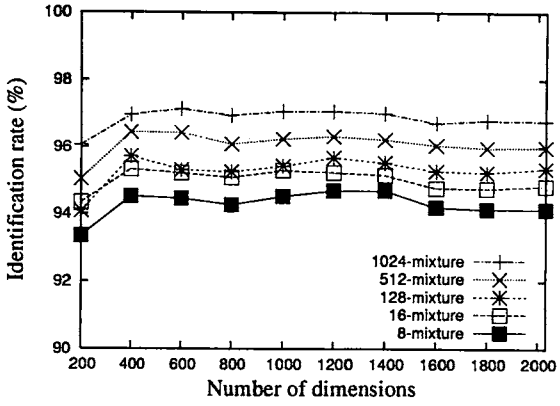

図 4 アンカーモデルとして GMM を使用したときの話者識別率（評価話者数 8 名）

Fig. 4 Speaker identification rate by using GMMs as anchor models. (#test speakers = 8)

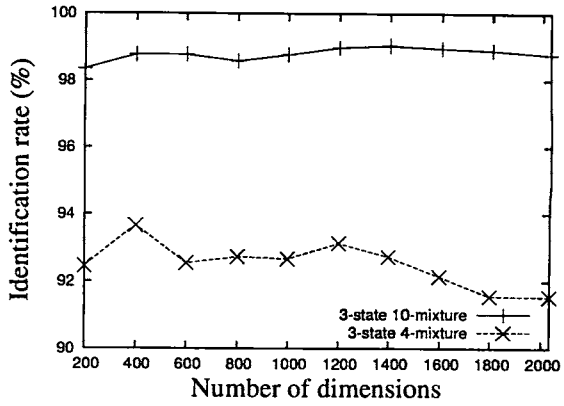

図 5 アンカーモデルとして HMM を使用したときの話者識別率（評価話者数 8 名）

Fig. 5 Speaker identification rate by using HMMs as anchor models. (#test speakers = 8)

不明である．

本稿では，評価話者 30 名のうちの 8 名を用いて，最適なパラメータ数の検討を行った．さまざまな混合数の 2032 個の GMM モデルを学習し，アンカーモデルとして使用した．同様に HMM モデルの 2032 個のセットもまた学習した．各話者における monophone HMM の数は 35 である．比較として，2 タイプの HMM（3 状態 4 混合，3 状態 10 混合）についても検討を行う．

図 4 はアンカーモデルとして GMM を使用したときの話者識別率を示している．さまざまな混合数の GMM の比較を行った．この図において，x 軸は話者空間の次元数を表す．この実験では，アンカーモデルは 2032 のモデルからランダムに選択されている．次元数削減は本手法において重要な問題であるが，本稿では検討を行っていない．種々の削減方法が [2] で提案されている．実験結果より，性能は数百次元で飽和しているのが分かる．対照的に，混合数増加によって識別性能は改善している．最良の性能の 97.13%は 1024 混合 GMM の 600 次元で得られている．図 5 はアンカーモデルとして monophone HMM を使用したときの結果を示す．次元数においては，HMM の性能は GMM と同様の傾向を示す．また実験結果から，HMM の混合数増加は効果的であることが示された．また GMM と比較すると，HMM の性能のほうが良い結果となった．最も良い識別率の 99.04%は 3 状態 10 混合 HMM の 1400 次元で得られている．

以上の結果を踏まえ，HMM のさらに進んだ詳細な検討を評価話者 30 名で実施する．図 6 は評価話者 30 名での混合数と識別率の関係を示す．この実験では，1 状態および 3 状態の HMM を使用した．話者空間の次元数は 1000 で固定した．1 状態と 3 状態を比較してみると，8 混合以降では 3 状態のほうが良い性能が得られている．また，混合数 8 付近で顕著な改善が現れている．要因としては，これらの混合数付近で有音と無音の境界の精度が改善したためであると考えられる．

HMM と GMM 間の性能比較を図 7 に示す．また，音素構造

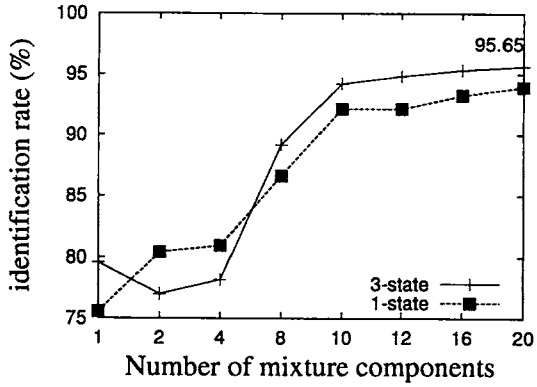

図 6 HMM のさまざまな混合数での結果（状態数 1 または 3，次元数 1000，評価話者数 30 名）

Fig. 6 Results with various number of mixture components of HMMs. (#state = 1 or 3, #dimentions = 1000, #test speakers = 30)

GMM の性能についても比較として掲載する．音素構造 GMM は，pdf の総数を考慮して．3 状態 10 混合 HMM に含まれるガウス分布を使用して作成する．以下に音素構造 GMM の作成法を記述する．

- 3 状態 10 混合 HMM の学習モデルから，無音モデルを除く 34 音素についての学習結果を取り出す（3 状態× 10 混合 × 34 音素 = pdf 数：1020）
- 重みのみを再学習し，得られた値を用いて音素構造 GMM とする

次元数は 1000 で，アンカーモデルの 4 タイプ（3s20mHMM：3 状態 20 混合 HMM，3s10mHMM：3 状態 10 混合 HMM，phoneGMM：1020 混合音素構造 GMM，1024mGMM：1024 混合 GMM）について比較を行っている．各話者の無音モデルを除いた pdf の総数は，3s10mHMM で 1020（=3 状態× 10 混合

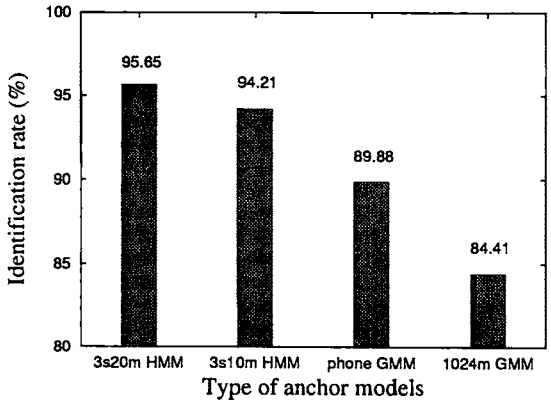

図 7 HMM と GMM の性能比較 (評価話者数 30 名)

Fig. 7 Performance comparison of HMM vs. GMM.(#test speakers = 30)

× 34 音素), 1024mGMM で 1024 である. 音素構造 GMM についても 3s10mHMM と同数であり, pdf の総数はこの 3 タイプのシステムにおいてほぼ同等である. 比較すると, HMM ベースシステムのほうが GMM ベースシステム以上の改善を示している. 計算量を考慮し, 2048mGMM の場合については行っていない. しかし, 3s10mHMM と 1024mGMM 間のパフォーマンスの違いが顕著であるので, 2048mGMM よりも, 同等の pdf 数の HMM ベースシステムのほうが性能が良好であることが予想される. また, 音素構造 GMM の性能は通常の GMM システムと比べ改善されている. しかし, HMM システムほどの改善は得られず, GMM よりも HMM を用いたほうが優れた性能を得ることができる. 最終的に, 95.65%の識別率が 3s20mHMM システムの 30 名の話者識別タスクで得られた. GMM ベースシステムと比べると, 3s20mHMM は 72.1%, 3s10mHMM は 62.9%の相対的改善となる.

5. 結　論

本稿では音素モデルを用いた話者ベクトルに基づく話者識別手法について検討を行った. 本手法は, 新たな識別対象話者のモデルのための学習を必要としない, テキスト独立型手法である. 識別性能改善のために, アンカーモデルとしてガウス混合モデル (GMM) の代わりに音素 HMM を使用した. 本提案手法は日本語話者識別タスクで評価を行った. この結果, GMM ベースシステムの性能と比較して, 顕著な改善が得られた. 95.65%の識別率が 3 状態 20 混合 HMM の 30 名の話者識別タスクから得られている.

本稿では, 全音素クラスが識別に使用された. しかし, ある音素クラスが他のものより高い話者特徴を有している可能性がある. そこで今後は, 話者識別における音素クラスの影響を検討し, より正確な話者識別システムの開発を行う. また, 計算量削減のために話者空間の次元数削減についても検討を行う予定である.

文　献

[1] D.Sturim, D.Reynolds, E.Singer, and J.Campbell, "Speaker indexing in large audio databases using anchor models," in *ICASSP01*, 2001, vol.1, pp.429-432.

[2] Yassine Mami and Delphine Charlet, "Speaker identification by anchor models with pca/lda post-processing," in *ICASSP03*, 2003, vol.1, pp.180-183.

[3] Delphine Charlet Mikael Collet, Yassine Mami and Frederic Bimbot, "Probabilistic anchor models approach for speaker verification," in *INTERSPEECH05*, 2005, pp.2005-2008.

[4] A.Nakamura et al., "Japanese speech databases for robust speech recognition," in *ICSLP96*, 1996, pp.2199-2202.

[5] M.Hebert and L.P.Heck, "Phonetic class-based speaker verification," in *EUROSPEECH03*, 2003, pp.1665-1668.

[6] A.Park and T.J.Hazen, "Asr dependent techniques for speaker identification," in *ICSLP02*, 2002, pp.1337-1340.

[7] M.A.Kohler W.D.Andrews and J.P.Campbell, "Phonetic speaker recognition," in *EUROSPEECH01*, 2001, pp.149-153.

[8] ETSI ES 202 050 V1.1.1, "Stq; distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," *ETSI standard*, 2002.

[9] R.Faltlhauser and G.Ruske, "Improving Speaker Recognition Performance Using Phonetically Structured Gaussian Mixture Models," in *Proc.Eurospeech*, Aalborg, Sept.2001, pp.751-754.

[10] S.Sarma and V.Zue, "Segment-based Speaker Verification System Using SUMMIT," in *Proc.Eurospeech*, Rhodes, Sept.1997, pp.843-846.