RE-001

GMM の分布選択に基づく アンカーモデルのクラスタリングによる話者認識

Speaker Recognition Using Anchor Model Clustering Based on Selection of Gaussian Mixtures

細川光政† 西田昌史† 山本誠一†
Mitsumasa Hosokawa, Masafumi Nishida, Seiichi Yamamoto

1．はじめに

近年，セキュリティのための生体認証としての話者認識，会議や討論などの複数話者の音声を対象としたデジタルアーカイブや情報検索などにおいて話者認識技術を応用した話者分類に関する研究がさかんに行われている[1].

従来の話者認識の手法としては，登録話者の音声データから抽出した特徴を統計的にモデル化する Gaussian Mixture Model (GMM)がよく用いられてきた[2][3]．この GMM による手法では多くの学習データが得られれば高い認識精度が得られるが，学習データ量が少ない場合には認識精度が劣化してしまう．それに対して，登録話者のモデルを仮定せずに登録話者以外の多くの話者モデルを用いることで，少量の音声データで認識を行うアンカーモデルという手法が提案されている．このアンカーモデルに基づいた手法は，会議や討論などの音声データベースを対象とした話者インデキシング[4][5]や話者照合[6]による手法に用いられており，アンカーモデルによる話者空間を判別分析などで構成する手法[7]なども提案されている．また，話者ごとに音素モデルを学習することで，これらをアンカーモデルとして話者識別を行う手法が提案されている[8].

従来のアンカーモデルによる手法では，アンカーモデルを無作為に選択しており，多くの話者モデルを用意することで高い認識精度を実現している．そのため選択された中には音響的に類似したモデルも含まれており，モデル数の増加に伴い計算量が増加する．そこで，cross likelihood ratio(CLR)を用いたアンカーモデルのクラスタリング手法が提案されている[9]．しかし，CLR は GMM 間の尤度比に基づく距離尺度で，尤度を求める際に音声データを必要とし多くの計算量がかかるといった問題点がある．

それに対し，Universal Background Model (UBM)を初期モデルとした Maximum a posteriori (MAP)推定により学習した GMM をアンカーモデルとして用い，GMM 間の Kullback–Leibler (KL)距離に基づいたアンカーモデルの階層的クラスタリング手法を提案し，認識精度を維持したままアンカーモデル数を削減できることを明らかにした[10]．本手法では，音声データを用いずに GMM のみを用いてクラスタリングを行うことができる．しかし，クラスタリングする際の GMM 間の KL 距離ならびにクラスタリング後の GMM を用いたアンカーモデルによる認識において，全混合分布間の距離ならびに尤度計算を行っていたため、処理コストがかかっていた．

本研究では UBM を初期モデルとして MAP 推定によりアンカーモデルを学習する際に得られる事後確率に着目し，事後確率が大きい上位の分布のみを選択してクラスタリングならびにアンカーモデルによる認識を行う手法を提案する．GMM の事後確率が大きい分布のみを用いた手法は，言語識別などの分野で用いられている [11]．事後確率が大きい分布はその話者の特徴を顕著に表していると考えられるので，それらの分布にしぼることで認識精度を向上させることができ，さらにクラスタリングならびに認識時の処理を高速化することができると考えられる．本手法の有効性を示すために，従来よく用いられている Bayesian Information Criterion (BIC)に基づく話者クラスタリング手法[12]との比較実験を行う．なお、本研究は発話内容に依存しないテキスト独立型の話者識別を行う．

2．アンカーモデルによる話者認識

2.1 Universal Background Model を用いたモデル学習

アンカーモデルによる話者認識では，認識対象以外の多くの話者の音声データを集め，話者ごとに GMM を学習する．本研究では，多数話者の音声データから学習した UBM を初期モデルとして，各アンカーモデルの学習データにより MAP 推定を行うことで話者モデルである GMM を学習する．

$$\Pr(i \mid x_t) = \frac{w_i p_i(x_t)}{\sum_{j=1}^{M} w_j p_j(x_t)} \tag{1}$$

$$n_i = \sum_{t=1}^{T} \Pr(i \mid x_t) \tag{2}$$

$$E_i(x) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i \mid x_t) x_t \tag{3}$$

ここで，x_tは各アンカーモデルの学習データ，Tは各アンカーモデルの学習データの総フレーム数，MはUBMの混合分布数，w_iはUBMの各混合分布の重みを表す．以上で求めた結果をもとに，UBMの各混合分布の重みw，平均μ，分散σ^2を以下の式により適応する．

$$\hat{w}_i = [\alpha_i n_i / T + (1-\alpha_i) w_i]\gamma \tag{4}$$

$$\hat{\mu}_i = \alpha_i E_i(x) + (1-\alpha_i)\mu \tag{5}$$

$$\hat{\sigma}_i^{\,2} = \alpha_i E_i(x^2) + (1-\alpha_i)(\sigma_i^{\,2} + \mu_i^{\,2}) - \hat{\mu}_i^{\,2} \tag{6}$$

† 同志社大学
Doshisha University

ここで，γは混合分布の重みの総和を制御する係数を表し，適応データの割合を制御する係数は，$\alpha_i = n_i/(n_i + r)$により求める．

2.2 アンカーモデルによる認識

アンカーモデルによる認識では,認識対象以外の多くの話者の音声データを集め，話者ごとに UBM を初期モデルとした MAP 推定により GMM を学習する．

アンカーモデルに基づいた手法では，j番目の発話の話者ベクトルVは式(7)のように求められる．ここでx_jはj番目の発話の入力特徴時系列全体を表し，$P(x_j|A_u)$はアンカーモデルA_uのGMMに対するx_jの対数尤度を表す．Uはアンカーモデルの総数である．x_jを発声する識別対象話者はアンカーモデルとして利用されているU人の話者には含まれない．

入力された発話と認識対象以外の各話者の尤度を求め，この尤度を要素とする話者ベクトルV_jを求め，登録話者のベクトルと入力話者のベクトル間のユークリッド距離を求め，距離が最短となる話者ベクトルをもつ話者が入力音声の話者であると識別する．

本研究では，尤度を求める際に GMM 全ての分布を使用せずに MAP 推定を行った際に式(2)により得られる事後確率の高い上位の分布のみを選択する．また，話者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．

$$V_j = \begin{bmatrix} \dfrac{P(x_j \mid A_1) - \mu_j}{\sigma_j} \\ \dfrac{P(x_j \mid A_2) - \mu_j}{\sigma_j} \\ \vdots \\ \dfrac{P(x_j \mid A_U) - \mu_j}{\sigma_j} \end{bmatrix} \tag{7}$$

$$\mu_j = \frac{1}{U}\sum_{u=1}^{U} P(x_j \mid A_u) \tag{8}$$

$$\sigma_j = \sqrt{\frac{1}{U}\sum_{u=1}^{U}(P(x_j \mid A_u) - \mu_j)^2} \tag{9}$$

図 1 に 3 次元での話者ベクトル空間の概念図を示す．それぞれの軸は，認識対象以外の話者であるアンカーモデルを表している．

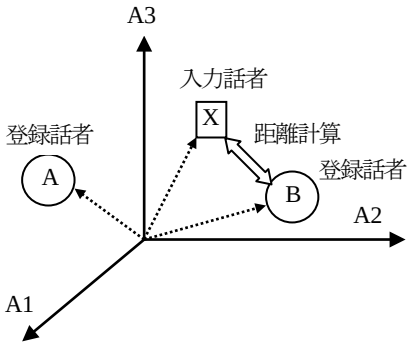

図1　アンカーモデルによる認識

GMM に基づく従来の話者認識手法では，識別対象話者の話者モデルを作成する必要があり，学習用の発話が複数文必要であった．それに対してアンカーモデルによる認識手法では，識別対象話者のためにモデルを学習する必要がなく，話者ベクトルの生成には 1 発話程度あればよい．

しかしながら，認識対象以外の不特定多数の話者の音声データからアンカーモデルを作成する必要があり，モデル数が多いほど処理時間がかかってしまうという問題がある．また，従来アンカーモデルは実験的に選択されており，登録話者を識別するにあたりどのような話者をアンカーモデルとして用意すべきかが重要である．

3. BIC によるアンカーモデルのクラスタリング

BIC に基づくアンカーモデルのクラスタリング手法について述べる．BIC は，ベイズ推定に基づいてモデル選択を行う基準として用いられている．各話者のデータに対して単一ガウス分布を仮定し，その分散比に基づいてクラスタリングを行う．この手法では，2 つの話者が似た特徴を持つと仮定した場合と，異なる特徴を持つと仮定した場合の BIC 値の差分に基づいて判定する．

2 つの話者をマージしたときの共分散行列をΣ_0，1 人目の話者の共分散行列をΣ_1，2 人目の話者の共分散行列をΣ_2，各話者のフレーム数をN_i，特徴ベクトルの次元数をdとするとBIC値の差分は式(10)により求まる．αは，重み係数であり，実験的に決める必要がある．

$$\Delta BIC = \frac{N_1 + N_2}{2}\log|\Sigma_0| - \frac{N_1}{2}\log|\Sigma_1| - \frac{N_2}{2}\log|\Sigma_2| - \alpha\frac{1}{2}\left(d + \frac{d(d+1)}{2}\right)\log(N_1 + N_2) \tag{10}$$

ΔBIC 値が負であれば 2 つの話者をマージする．BIC 値が最も大きい話者間から順次マージし,全ての話者間で BIC 値が正になれば，どの話者もマージすべきでないとしてクラスタリングを終了する．以上で得られたクラスタごとに，UBM を初期モデルとした MAP 推定により GMM を再学習してアンカーモデルとする．こうして得られたアンカーモデルをもとに，MAP 推定を行う際に式(2)により得られる事後確率の高い分布のみを選択して尤度計算を行い認識を行う．

4. KL 距離に基づくアンカーモデルの階層的クラスタリング

本手法では，アンカーモデルをクラスタリングするにあたり，GMM 間の KL 距離を用いた．なお，GMM は UBM を初期モデルとした MAP 推定により学習した．一般的に，KL 距離は単一ガウス分布間の距離尺度であるので，本研究では式(11)のように混合分布間の距離尺度に拡張して用いた[13]．また，MAP 推定を行う際に式(2)により得られる事後確率の上位分布のみ選択し，分布間の距離を求める．

$$d(t,s)=\sum_{p=1}^{M} w_p \min_q KL(p,q)$$

$$KL(p,q)=\sum_{i=1}^{d}\{\frac{{\sigma_{pi}}^2-{\sigma_{qi}}^2+(\mu_{qi}-\mu_{pi})^2}{{\sigma_{qi}}^2} + \frac{{\sigma_{qi}}^2-{\sigma_{pi}}^2+(\mu_{qi}-\mu_{pi})^2}{{\sigma_{pi}}^2}\} \quad (11)$$

ここで，pは話者tのモデルの分布番号，qは話者sのモデルの分布番号，M は話者モデルの混合分布数，w_pは混合分布の重み，d は特徴ベクトルの次元数を示している．また，μ，σは混合分布の平均ベクトル，共分散行列の要素を表している．

GMM 間の KL 距離が閾値よりも小さい話者をマージし，それぞれをクラスタとする．そして，クラスタ毎に UBM を初期モデルとした MAP 推定により GMM を再学習してアンカーモデルとする．

クラスタリングの処理の流れを以下に示す．

(1) アンカーモデルの GMM 間の KL 距離を全てのモデル間で計算する．ここで，計算対象となるのは MAP 推定の式(2)により得られる事後確率が上位の分布のみである．
(2) KL 距離が最小となるモデル同士をマージし新たなクラスタとする．ここで，マージされた GMM は再学習しない．
(3) (2) でマージしたモデル以外で KL 距離が最小となる話者を距離が閾値よりも小さければマージする．全てのモデル同士の KL 距離が閾値より大きくなるまで (2)，(3) を繰り返す．
(4) (3) までの処理で得られたクラスタと単独モデルの KL 距離が最小となるクラスタを探す．ここで，クラスタと単独モデルとの距離は，クラスタ内の各 GMM との KL 距離の平均距離により求める．この距離が閾値より大きくなるまで処理を繰り返す．
(5) クラスタ同士の KL 距離を比較し，距離が最小となるクラスタ同士をマージする．ここで，クラスタ間の距離はクラスタ内の各 GMM 間の KL 距離の平均距離により求める．この距離が閾値より大きくなるまで処理を繰り返す．
(6) 以上より得られたクラスタごとに UBM-MAP により GMM を再学習し，これらをアンカーモデルとする．認識を行う際には，MAP 推定の式(2)により得られる事後確率の高い上位の分布のみを選択して尤度計算を行う．

5. 評価実験

5.1 実験条件

本研究では，NTT の話者認識用データベースを用いて話者認識実験を行った．話者 30 名（男性 21 名・女性 9 名）が約 1 年間の 7 時期（1990 年 8 月・9 月・12 月，1991 年 3 月・6 月・9 月，1992 年 3 月）に発声した各時期 10 文章データで，各文章における 3 種類の発声速度（普通，遅い，速い）の計 30 発話である．

UBM ならびにアンカーモデルの学習データには，認識対象のデータと異なる「日本語話し言葉コーパス」（CSJ）に含まれる講演音声を用いた．1 人あたり 300ms 以上の無音区間を基準に発話を分割し無音区間を除いた約 60 秒の発話で，600 名（男性 300 名，女性 300 名）の話者のデータを UBM の学習に，それとは異なる 500 名の話者をアンカーモデルの学習に用いた．UBM の混合分布数は 256 とした．

アンカーモデルによる認識では，学習データとして最初の時期 90 年 8 月の普通の速さ 1 発話を用いて行い，認識では全 7 時期の学習とは異なる 5 文の 3 速度の 15 文章で，話者ごとに合計 105 発話を用いた．本実験で用いた音声データは，フレーム長 25ms，フレーム周期 10ms で音響分析を行い，12 次 MFCC の特徴量を求めている．

5.2 実験結果と考察

GMM の全ての分布を使用する通常のアンカーモデルによる認識結果を表 1 に，MAP 推定を行う際に得られる事後確率の上位分布を選択した際の分布数を変えたときの結果を図 2 に示す．アンカーモデル数は全て 500 である．

表 1　通常のアンカーモデルによる認識結果

アンカーモデル数 500	認識率（%）
分布数 256	80.1

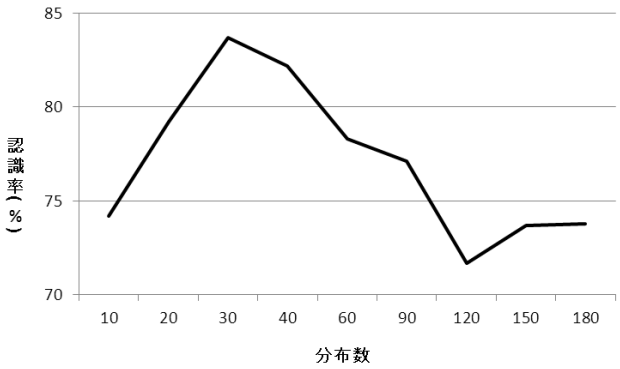

図 2　分布数の選択によるアンカーモデルの認識結果

全ての分布を使用した認識率は 80.1%，分布数が 10 個のとき 74.2%，20 個のとき 79.2%，30 個のとき 83.7%，40 個のとき 82.2%，60 個のとき 78.3%，90 個のとき 77.1%，120 個のとき 71.7%，150 個のとき 73.7%，180 個のとき 73.8%となり，全ての分布を使用して認識した結果よりも事後確率が上位 30 個の分布を選択したときに最も高い認識精度となった．以後の実験においては，事後確率が上位 30 個の分布のみを用いて行う．

次に BIC と KL 距離に基づいてアンカーモデルのクラスタリングを行い，得られたアンカーモデルにより認識を行った．BIC に基づくクラスタリングを行った結果のモデル数と認識率を表 2 に，KL 距離に基づくクラスタリングを行った結果のモデル数と認識率を表 3 に示す．各モデル数は，BIC の α の値と KL 距離の閾値を変化させて得られた結果である．閾値は BIC においてモデル数 250 のとき 0.5，169 のとき 2.3，140 のとき 2.45，KL 距離においてモデル数 252 のとき 0.06，165 のとき 0.07，134 のとき 0.075 である．

BIC に基づくクラスタリングの結果と KL 距離に基づくクラスタリングによる結果を比較すると，ほぼ同数のモデル数のときに KL 距離の方が高い認識精度を得ることができた．このことから KL 距離に基づくクラスタリングが有効であることが明らかになった．BIC による手法では，単一分布にてモデルを表現しクラスタリングを行うが，KL 距離による手法では混合分布で表現されるためより特徴を細かくとらえることができ，精度が向上していると考えられる．また，事後確率を基に分布を選択することでクラスタリング時の計算量を削減することができた．

表 2　BIC に基づくクラスタリングの結果

モデル数	認識率（%）
250	78.5
169	77.7
140	76.1

表 3　KL 距離に基づくクラスタリングの結果

モデル数	認識率（%）
252	80.4
165	80.1
134	77.6

また，クラスタリングを行わずにアンカーモデルに用いる話者モデル数を変えたときの結果を表 4 に示す．この結果も分布数を事後確率の上位 30 個選択した認識結果である．

表 4 アンカーモデル数の違いによる認識結果

モデル数	認識率（%）
250	78.2
160	76.2
130	75.9

表 2 と表 4 の結果から，提案手法によるクラスタリングは，クラスタリングを行わなかったときに比べても高い認識精度を得ることができた．

6. おわりに

本研究では，UBM を初期モデルとした MAP 推定により学習した GMM をアンカーモデルとして用い，MAP 推定によって得られる事後確率の上位分布のみを用いて KL 距離によるクラスタリングならびに認識を行う手法を提案した．本手法を従来の BIC に基づく階層的クラスタリング手法との比較実験を行った結果，ほぼ同じクラスタ数のときの認識精度を比較した場合に提案手法のほうが高い認識精度が得られた．また，クラスタリングを行わない場合に比べても高い認識精度が得られた．したがって，提案手法によりアンカーモデル数ならびに GMM の分布数を削減することの有効性を示すことができた．

今後は，提案手法において処理効率や認識精度の観点で詳細な分析を行う予定である．また，認識対象話者の識別に有効なアンカーモデルの構成方法についてさらに検討を行い，より多くのデータを対象に評価実験を行っていく予定である．

参考文献

[1] S. E. Tranter and D. A. Reynolds, “An Overview of Automatic Speaker Diarization Systems ”, IEEE Transactions on Audio, Speech, and Language Processing, Vol.14, No.5, pp.1557-1565, 2006.

[2] D.A.Reynolds，T.F.Quatieri，and R. B. Dunn,"Speaker verification using adapted Gaussian mixture models," Digit. Signal Process，vol.10, pp.19-41，2000.

[3] S. Nakagawa，W. Zhang, and M. Takahashi，“Text-Independent/Text-Prompted Speaker Recognition by Combining Speaker-Specific GMM with Speaker Adapted Syllable-Based HMM”，IEICE TRANS.INF.&SYST，vol.E89-D, No.3, pp.1058-165, 2006.

[4] D. Sturim，D. Reynolds，E. Singer，and J. Campbell，“Speaker indexing in large audio databases using anchor models”, Proc. ICASSP，Vol.1，pp.429-432，2001.

[5] 秋田祐哉，河原達也，“多数話者モデルを用いた討論音声の教師なし話者インデキシング”，電子情報通信学会論文誌，Vol.J87-D-Ⅱ No.2，pp.495-503，2004.

[6] Y. Yang, M. Yang, Z. Wu, "A Rank based Metric of Anchor Models for Speaker Verification", Proc. ICME, pp.1097-1100, 2006.

[7] Yassine Mami，Delphine Charlet，“Speaker recognition by location in the space of reference speakers”, Speech Communication 48，pp.127-141，2006.

[8] 小坂哲夫，赤津達也，加藤正治，好田正紀，“音素モデルを用いた話者ベクトルに基づく話者識別”，電子情報通信学会論文誌，Vol.J90-D No.12，pp.3201-3209，2007.

[9] Y. Mami，D. Charlet，“Speaker identification by anchor models with PCA/LDA post-processing”, Proc. ICASSP，pp.180-183，2003.

[10] 細川 光政，西田 昌史，山本 誠一，“GMM 間の KL 距離に基づく Anchor Model のクラスタリングによる話者認識”，情報処理学会第 73 回全国大会，6P-7，pp.2_121-2_122，2011.

[11] E.Wong, J.pelecanos, S. Myers and S. Sridharan, “Language identification using efficient Gaussian mixture model analysis”, Proc. SST, pp.78-83, 2000.

[12] S.Chen and P. Gopalakrishnan, “Speaker, environment and channel change detection and clustering via the Bayesian information criterion”, Proc. DARPA Broadcast News Transcription and Understanding Workshop, pp.127-132, 1998.

[13] 西田昌史，堀内靖雄，市川熹，河原達也，”統計的モデル選択に基づくクラスタリングを用いた話者適応”，日本音響学会講演論文集，2-11-5，pp.109-110，2004.